
Amazon EC2 Auto Scaling

Guia do usuário



Amazon EC2 Auto Scaling: Guia do usuário

Copyright © 2023 Amazon Web Services, Inc. and/or its affiliates. All rights reserved.

As marcas comerciais e imagens de marcas 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 desprestigue a Amazon. Todas as outras marcas comerciais que não pertencem à Amazon pertencem a seus respectivos proprietários, que podem ou não ser afiliados, conectados ou patrocinados pela Amazon.

Table of Contents

O que é o Amazon EC2 Auto Scaling	1
Componentes do Auto Scaling	1
Preços do Amazon EC2 Auto Scaling	2
Conceitos básicos	2
Trabalhar com grupos do Auto Scaling	2
Benefícios do Auto Scaling	3
Exemplo: atender a demanda variável	3
Exemplo: arquitetura de aplicação Web	5
Exemplo: distribuir instâncias entre zonas de disponibilidade	6
Ciclo de vida da instância	8
Escalonamento horizontal	8
Instâncias em serviço	9
Reducir a escala na horizontal	9
Anexar uma instância	10
Desvincular uma instância	10
Ganchos do ciclo de vida	10
Entrar e sair de espera	11
Cotas	11
Taxes de terminação do EC2	12
Outros produtos da	13
Configurar	14
Preparação para usar o Amazon EC2	14
Preparar-se para usar a AWS CLI	14
Conceitos básicos	15
Resumo da demonstração	15
Preparar para a demonstração	16
Etapa 1: Criar um modelo de execução	16
Etapa 2: Criar um grupo do Auto Scaling com uma única instância	18
Etapa 3: Verificar seu grupo do Auto Scaling	19
Etapa 4: Terminar uma instância no seu grupo do Auto Scaling	19
Etapa 5: Próximas etapas	20
Etapa 6: Limpar	20
Modelos de execução	22
Permissões	22
Criar um modelo de execução para um grupo do Auto Scaling	23
Criar seu modelo de execução (console)	23
Criar um modelo de execução com base em uma instância existente (console)	31
Informações adicionais	31
Limitações	31
Migre para lançar modelos	32
Etapa 1: Encontre grupos de Auto Scaling que usam configurações de inicialização	32
Etapa 2: Copiar uma configuração de inicialização para um modelo de inicialização	34
Etapa 3: atualizar um grupo de Auto Scaling para usar um modelo de lançamento	35
Etapa 4: substitua suas instâncias	35
Informações adicionais	36
Request Spot Instances	36
Exemplos da AWS CLI para trabalhar com modelos de execução	37
Exemplo de uso	38
Criar um modelo de execução básico	38
Especificar etiquetas que marcam instâncias ao iniciar	39
Especificar uma função do IAM a ser transmitida às instâncias	39
Atribuir um endereço IP público	39
Especificar um script de dados do usuário que configura instâncias ao iniciar	40
Especificar um mapeamento de dispositivos de blocos	40

Especificar hosts dedicados para trazer licenças de software de fornecedores externos	40
Especificar uma interface de rede existente	40
Criar várias interfaces de rede	41
Gerenciar modelos de execução	41
Atualizar um grupo do Auto Scaling para usar um modelo de execução	43
Usar parâmetros do Systems Manager em vez de IDs de AMI	44
Limitações	48
Configurações de execução	49
Criar uma configuração de execução	49
Criar uma configuração de execução (console)	50
Criar uma configuração de execução (AWS CLI)	51
Configurar IMDS	51
Criar uma configuração de execução usando uma instância do EC2	53
Criar uma configuração de execução usando uma instância do EC2	54
Criar uma configuração de execução a partir de uma instância e substituir os dispositivos de blocos (AWS CLI)	55
Criar uma configuração de execução e substituir o tipo de instância (AWS CLI)	56
Alterar uma configuração de execução	57
Configurar locação da instância	58
Grupos do Auto Scaling	61
Crie grupos de Auto Scaling usando modelos de lançamento	62
Criar um grupo usando um modelo de execução	62
Criar um grupo usando o assistente de execução do EC2	64
Usar vários tipos de instâncias e opções de compra	67
Criar um grupo usando seleção de tipo de instância baseada em atributos	92
Crie grupos de Auto Scaling usando configurações de lançamento	99
Criar um grupo usando uma configuração de execução	100
Criar um grupo usando uma instância do EC2	102
Atualizar um grupo de Auto Scaling	106
Atualizar instâncias do Auto Scaling	107
Substituir instâncias	108
Substituir instâncias com base em uma atualização de instância	108
Substituir instâncias com base na vida útil máxima da instância	135
Marcar grupos e instâncias	138
Restrições de nomeação e uso de tags	139
Ciclo de vida de marcação de instâncias do EC2	139
Marcar seus grupos do Auto Scaling	140
Excluir tags	142
Etiquetas para segurança	142
Controlar o acesso usando etiquetas	143
Usar etiquetas para filtrar grupos do Auto Scaling	144
Excluir infraestrutura do Auto Scaling	146
Excluir seu grupo do Auto Scaling	146
(Opcional) Excluir a configuração de execução	147
(Opcional) Excluir o modelo de execução	147
(Opcional) Excluir o平衡ador de carga e grupos de destino	148
(Opcional) Excluir CloudWatchalarms	149
AWSExemplos de SDK para trabalhar com grupos de Auto Scaling	149
Criar um grupo do Auto Scaling	149
Atualizar um grupo de Auto Scaling	155
Excluir um grupo de Auto Scaling	159
Recursos relacionados	163
Escalar o grupo	165
Opcões de escalabilidade	165
Definir limites de capacidade	166
Manter um número fixo de instâncias	168
Escalabilidade manual	168

Alterar o tamanho do grupo do Auto Scaling (console)	168
Alterar o tamanho do grupo do Auto Scaling (AWS CLI)	169
Anexar instâncias do EC2 a seu grupo do Auto Scaling	171
Desvincular instâncias do EC2 do seu grupo do Auto Scaling	174
Escalabilidade dinâmica	178
Como funcionam as políticas de escalabilidade dinâmica	178
Várias políticas de escalabilidade dinâmica	179
Políticas de escalabilidade de rastreamento de destino	180
Políticas de escalabilidade simples e em etapas	190
Definir valores de aquecimento ou desaquecimento padrão	200
Escalabilidade baseada no Amazon SQS	210
Verificar uma ação de escalabilidade	215
Desabilitar uma política de escalabilidade	216
Excluir uma política de escalabilidade	218
Exemplos da AWS CLI para políticas de escalabilidade	220
Escalabilidade preditiva	222
Como a escalabilidade preditiva funciona	223
Práticas recomendadas	223
Criar uma política de escalabilidade preditiva (console)	224
Criar uma política de escalabilidade preditiva (AWS CLI)	227
Limitações	229
Supported Regions (Regiões compatíveis)	229
Avaliar as políticas de escalabilidade preditiva	230
Substituir a previsão	236
Usar métricas personalizadas	239
Escalabilidade programada	247
Considerações	247
Programações recorrentes	248
Criar e gerenciar ações programadas (console)	248
Criar e gerenciar ações programadas (AWS CLI)	250
Limitações	252
Ganchos do ciclo de vida	252
Disponibilidade de ganchos do ciclo de vida	253
Considerações e limitações	254
Recursos relacionados	255
Como os ganchos do ciclo de vida funcionam	255
Preparar para adicionar um gancho de ciclo de vida	257
Recuperar o estado de destino do ciclo de vida	262
Adicionar ganchos do ciclo de vida	263
Concluir uma ação do ciclo de vida	266
Tutorial: configurar dados do usuário para recuperar o estado de destino do ciclo de vida por meio de metadados de instância	267
Tutorial: Configurar um gancho do ciclo de vida que invoca uma função do Lambda	273
Grupos de alta atividade	279
Conceitos principais	280
Pré-requisitos	282
Criar um grupo de alta atividade	283
Atualizar um grupo de alta atividade	283
Excluir um grupo de alta atividade	284
Limitações	284
Usar ganchos de ciclo de vida	284
Visualizar status da verificação de integridade	287
Exemplos da AWS CLI para trabalhar com grupos de alta atividade	289
Controlar o término de instâncias	292
Cenários de término	292
Trabalhar com políticas de término	295
Criar uma política de término personalizada com o Lambda	298

Usar proteção de redução na escala na horizontal de instâncias	302
Design para o encerramento de instâncias sem problemas	305
Remover instâncias temporariamente	308
Como o estado de espera funciona	308
Considerações	309
Status de integridade de uma instância em um estado de espera	309
Remoção temporária de uma instância (console)	309
Remover uma instância temporariamente (AWS CLI)	310
Suspender-retomar processos	312
Tipos de processos	313
Considerações	313
Suspender e retomar processos (console)	316
Suspender e retomar processos (AWS CLI)	316
Monitor	318
Verificar a integridade da instância	319
Tipo de verificação de integridade	320
Verificações de integridade do Amazon EC2	320
Verificações de integridade do Elastic Load Balancing	321
Verificações de integridade do VPC Lattice	322
Tarefas personalizadas de detecção de integridade	322
Substituição de instância não íntegra	323
Como o Amazon EC2 Auto Scaling minimiza o tempo de inatividade	324
Considerações sobre a verificação de integridade	324
Informações adicionais	325
Período de carência da verificação de integridade	325
Monitoramento com o AWS Health Dashboard	327
MonitorCloudWatchmétricas	328
Visualizar grafos de monitoramento no console do Amazon EC2 Auto Scaling	328
CloudWatchmétricas para o Amazon EC2 Auto Scaling	332
Configurar monitoramento para instâncias do Auto Scaling	337
Registrar chamadas de API com o AWS CloudTrail	339
Informações do Amazon EC2 Auto Scaling emCloudTrail	339
Noções básicas sobre entradas do arquivo de log do Amazon EC2 Auto Scaling	340
Recursos relacionados	341
Monitorar com notificações do Amazon SNS	341
Notificações do SNS	342
Configurar notificações do Amazon SNS para o Amazon EC2 Auto Scaling	343
Política de chaves para um tópico criptografado do Amazon SNS	345
Trabalhar com outros serviços	346
Rebalanceamento de capacidade	346
Visão geral	347
Comportamento de reequilíbrio de capacidade	347
Considerações	348
Habilitar o rebalanceamento de capacidade (console)	349
Habilitar o rebalanceamento de capacidade (AWS CLI)	350
Recursos relacionados	353
Limitações	353
Reservas de capacidade	354
Etapa 1: Criar as reservas de capacidade	354
Etapa 2: Criar um grupo de reserva de capacidade	356
Etapa 3: criar um modelo de lançamento	357
Etapa 4: criar um grupo de Auto Scaling	358
Recursos relacionados	360
AWS CloudShell	360
AWS CloudFormation	360
Amazon EC2 Auto Scaling e modelos AWS CloudFormation	361
Saiba mais sobre o AWS CloudFormation	361

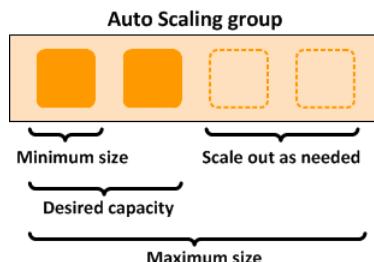
Migrar configurações de execução	361
Compute Optimizer	367
Limitações	367
Descobertas	367
Exibir recomendações	368
Considerações para avaliação das recomendações	368
Elastic Load Balancing	369
Tipos de Elastic Load Balancing	370
Pré-requisitos	371
Anexar um平衡ador de carga	372
Configurar um平衡ador de carga do console do Amazon EC2 Auto Scaling	374
Verificar o status do anexo	375
Adicionar verificações de integridade do Elastic Load Balancing	376
Adicionar zonas de disponibilidade	377
Exemplos da AWS CLI para trabalhar com Elastic Load Balancing	379
Tutorial: Configurar uma aplicação escalonada e com balanceamento de carga	385
Lattice de VPC	393
Preparar para anexar um grupo de destino	394
Anexar um grupo-alvo do VPC Lattice	396
Verificar o status do anexo	399
EventBridge	400
Referência de eventos do Amazon EC2 Auto Scaling	401
Exemplos de eventos e padrões de piscinas aquecidas	407
Crie EventBridge regras	411
Amazon VPC	414
EC2-Classic	415
VPC padrão	416
VPC não padrão	416
Considerações sobre a escolha de sub-redes da VPC	417
Endereçamento IP em uma VPC	417
Interfaces de rede em uma VPC	417
Locação de localização de instância	418
AWS Outposts	418
Mais recursos para saber mais sobre VPCs	418
Segurança	419
Proteção de dados	419
Usar AWS KMS keys para criptografar volumes do Amazon EBS	420
Gerenciamento de identidade e acesso	421
Controle de acesso	421
Como o Amazon EC2 Auto Scaling funciona com o IAM	421
Permissões de API	428
Políticas gerenciadas	429
Funções vinculadas ao serviço	432
Exemplos de políticas baseadas em identidade	437
Prevenção do problema do substituto confuso entre serviços	442
Suporte a modelo de execução	443
Funções do IAM para aplicações que são executadas em instâncias do Amazon EC2	448
Política de chaves do AWS KMS para uso com volumes criptografados	450
Validação de conformidade	455
Conformidade do PCI DSS	455
Resiliência	456
Segurança da infraestrutura	457
Usar endpoints da VPC para conectividade privada	457
Criar um VPC endpoint de interface	458
Criar uma política de endpoint da VPC	458
Solução de problemas	459
Recuperar uma mensagem de erro	459

Recursos adicionais para solução de problemas	460
Falha ao iniciar instância	461
A configuração solicitada não é suportada atualmente.	462
O grupo de segurança <nome do grupo de segurança> não existe. Falha ao ativar a instância EC2.	462
O par de chaves <par de chaves associado à sua instância do EC2> não existe. Falha ao ativar a instância EC2.	463
A Zona de disponibilidade solicitada não é mais suportada. Tente sua solicitação novamente....	463
O tipo de instância solicitado (<tipo de instância>) não tem suporte na Zona de disponibilidade solicitada (<Zona de disponibilidade da instância>)....	463
Seu preço de solicitação spot de 0,015 é inferior ao preço mínimo de atendimento de solicitação spot exigido de 0,0735....	464
Nome de dispositivo inválido <nome do dispositivo> / Carregamento do nome de dispositivo inválido. Falha ao ativar a instância EC2.	464
O valor (<nome associado ao dispositivo de armazenamento de instâncias>) do parâmetro virtualName é inválido....	464
Mapeamentos de dispositivos de blocos do EBS não suportados para AMIs de armazenamento de instância.	465
Os placement groups não podem ser usados com instâncias do tipo 'm1.large'. Falha ao ativar a instância EC2.	465
de de de de InternalError: Erro do cliente na inicialização....	465
No momento, não temos capacidade de <tipo de instância> suficiente para tipo de instância na zona de disponibilidade solicitada. Falha ao ativar a instância EC2.	466
Não há capacidade spot disponível que corresponda à sua solicitação. Falha ao ativar a instância EC2.	467
<número de instâncias> instância(s) já estão em execução. Falha ao ativar a instância EC2.	467
Problemas de AMI	467
O ID da AMI <ID da sua AMI> não existe. Falha ao ativar a instância EC2.	468
A AMI <ID da AMI> está pendente e não pode ser executada. Falha ao ativar a instância EC2.	468
O valor do (<ID da ami>) para o parâmetro virtualName é inválido.	468
A arquitetura do tipo de instância solicitado (i386) não corresponde à arquitetura no manifesto da ami-6622f00f (x86_64). Falha ao ativar a instância EC2.	469
Problemas do balanceador de carga	469
Um ou mais grupos de destino não encontrados. Falha na validação da configuração do balanceador de carga.	470
Não é possível encontrar o Load Balancer <seu load balancer>. Falha na validação da configuração do balanceador de carga.	470
Não há nenhum balanceador de carga ATIVO chamado <nome do balanceador de carga>. Falha ao atualizar a configuração do balanceador de carga.	470
A instância do EC2 <ID da instância> não está na VPC. Falha ao atualizar a configuração do balanceador de carga.	471
A instância do EC2 <ID da instância> está na VPC. Falha ao atualizar a configuração do balanceador de carga.	471
Problemas em modelos de execução	471
Você deve usar um modelo de inicialização totalmente formado válido (valor inválido)	471
Você não está autorizado a usar o modelo de execução (permissões insuficientes)	472
Verificações de integridade	473
Uma instância foi retirada de serviço em resposta a uma falha de verificação de status de instância do EC2	473
Uma instância foi retirada de serviço em resposta a uma reinicialização programada do EC2	474
Uma instância foi retirada de serviço em resposta a uma verificação de integridade do EC2 que indicou que ela tinha sido terminada ou interrompida	474
Uma instância foi retirada de serviço em resposta a uma falha na verificação de integridade do sistema ELB	475
Informações relacionadas	477
Histórico do documento	479
	di

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 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 uma instância, uma capacidade desejada de duas instâncias e um tamanho máximo de quatro 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.



Para obter mais informações sobre os benefícios do Amazon EC2 Auto Scaling consulte [Benefícios do Amazon EC2 Auto Scaling \(p. 3\)](#).

Para configurar a autoescalabilidade para recursos escaláveis para serviços da Amazon Web Services além do Amazon EC2, consulte o [Manual do usuário do Application Auto Scaling](#).

Componentes do Auto Scaling

A tabela a seguir descreve os principais componentes do Amazon EC2 Auto Scaling.

	<p>Grupos</p> <p>Suas instâncias do EC2 são organizadas em groups para que possam ser tratadas como uma unidade lógica para fins de escalabilidade e gerenciamento. Ao criar um grupo, você pode especificar o número mínimo, máximo e desejado de instâncias do EC2. Para obter mais informações, consulte Grupos do Auto Scaling (p. 61).</p>
	<p>Modelos de configuração</p> <p>Seu grupo usa um modelo de execução ou uma configuração de execução (não recomendada, oferece menos recursos), como um</p>

	<p>modelo de configuração para suas instâncias do EC2. Você pode especificar 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 ter mais informações, consulte Modelos de execução (p. 22) e Configurações de execução (p. 49).</p>
	<p>Opções de escalabilidade</p> <p>O Amazon EC2 Auto Scaling fornece várias formas de escalar seus grupos do Auto Scaling. Por exemplo, você pode configurar um grupo para escalar com base na ocorrência de condições especificadas (escalabilidade dinâmica) ou em uma programação. Para obter mais informações, consulte Opções de escalabilidade (p. 165).</p>

Preços do Amazon EC2 Auto Scaling

Como não há tarifas adicionais para o Amazon EC2 Auto Scaling, é fácil testá-lo e ver como ele pode beneficiar sua arquitetura da AWS. Você paga apenas pelo AWS Recursos (por exemplo, instâncias do EC2, volumes do EBS e CloudWatch Alarms) que você usa.

Conceitos básicos

Para começar, conclua o tutorial [Conceitos básicos do Amazon EC2 Auto Scaling \(p. 15\)](#) para criar um grupo do Auto Scaling e ver como ele responde quando uma instância desse grupo é encerrada.

Para ver tutoriais adicionais que se concentram em casos de uso específicos, consulte os seguintes tópicos:

- [Tutorial: Configurar uma aplicação escalonada e com平衡amento de carga \(p. 385\)](#). Este tutorial mostra como configurar seu grupo de Auto Scaling para receber tráfego de um balanceador de carga do Elastic Load Balancing.
- [Tutorial: Configurar um gancho do ciclo de vida que invoca uma função do Lambda \(p. 273\)](#). Este tutorial mostra como usar a Amazon EventBridge para criar regras que invocam funções do Lambda com base em eventos que acontecem com as instâncias em seu grupo de Auto Scaling.
- [Tutorial: configurar dados do usuário para recuperar o estado de destino do ciclo de vida por meio de metadados de instância \(p. 267\)](#). Este tutorial mostra como usar o Instance Metadata Service (IMDS) para invocar uma ação de dentro da própria instância.

Trabalhar com grupos do Auto Scaling

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

- AWS Management Console: fornece uma interface da Web que você pode usar para acessar os grupos do Auto Scaling. Se você cadastrou uma Conta da AWS, poderá acessar seus grupos do Auto Scaling fazendo login no AWS Management Console, usando a caixa de pesquisa na barra de navegação para procurar grupos do Auto Scaling e escolhendo Auto Scaling groups (grupos do 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 [Preparar-se para usar a](#)

[AWS CLI \(p. 14\)](#). 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 escreve no PowerShellmeio ambiente. Para começar a usar, consulte o [Guia do usuário do AWS Tools for Windows PowerShell](#). Para obter mais informações, consulte [Referência de Cmdlets do AWS Tools for PowerShell](#).
- AWS SDKs: fornecem operações de API específicas da linguagem e cuidam de muitos dos detalhes da conexão, como cálculo de assinaturas, tratamento de novas tentativas de solicitação e tratamento de 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](#).
- AWS CloudFormation— Suporta a criação de grupos de Auto Scaling usandoCloudFormationmodelos. Para obter mais informações, consulte [Criar um grupo do Auto Scaling com AWS CloudFormation \(p. 360\)](#).

Para se conectar a um AWS service (Serviço da AWS) de forma programática, use um endpoint. Para obter informações sobre endpoints para chamadas para o Amazon EC2 Auto Scaling, consulte[Endpoints e cotas do Amazon EC2 Auto Scaling](#)naReferência geral da AWS.

Benefícios do Amazon EC2 Auto Scaling

A adição do Amazon EC2 Auto Scaling à arquitetura da sua aplicação é uma maneira de maximizar os benefícios da Nuvem AWS. 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.

Índice

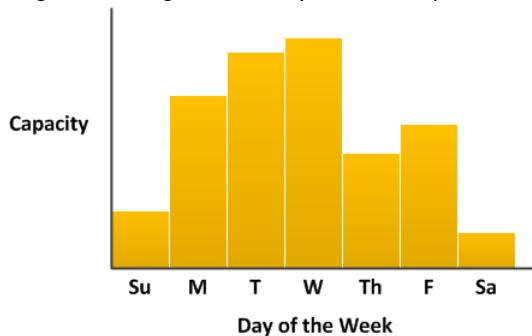
- [Exemplo: atender a demanda variável \(p. 3\)](#)
- [Exemplo: arquitetura de aplicação Web \(p. 5\)](#)
- [Exemplo: distribuir instâncias entre zonas de disponibilidade \(p. 6\)](#)
 - [Distribuição de instâncias \(p. 7\)](#)
 - [Atividades de rebalanceamento \(p. 7\)](#)

Exemplo: atender a demanda variável

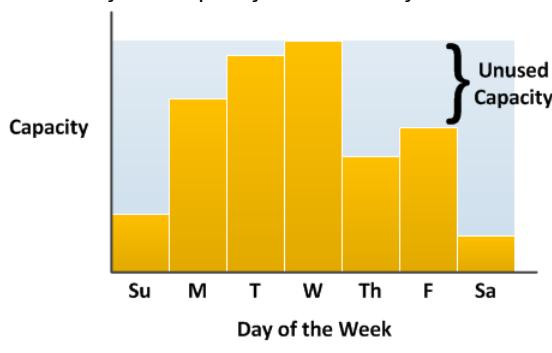
Para demonstrar alguns dos benefícios do Amazon EC2 Auto Scaling, considere uma aplicação Web básica em execução na AWS. Essa aplicação permite que os funcionários pesquisem salas de conferência que podem usar para reuniões. Durante o início e o fim da semana, o uso dessa aplicação é mínimo.

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.

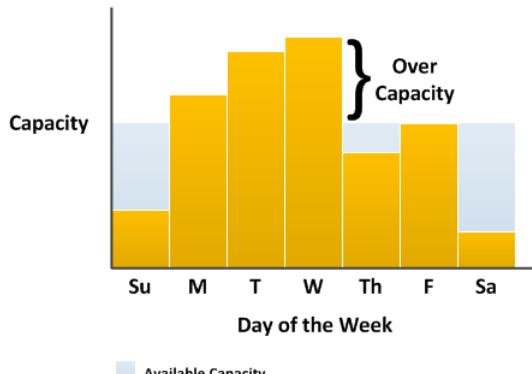


Tradicionalmente, há duas maneiras de planejar essas alterações na capacidade. A primeira opção é adicionar servidores suficientes para que a aplicação sempre tenha capacidade suficiente para atender à demanda. A desvantagem dessa opção, no entanto, é que há dias em que a aplicação não precisa de toda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



Available Capacity

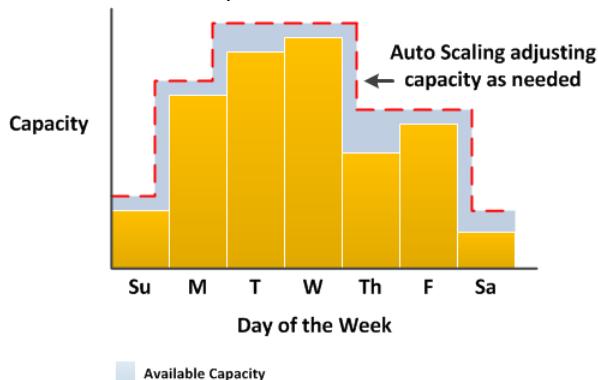
A segunda opção é ter capacidade suficiente para lidar com a demanda média na aplicação. Essa opção é mais barata, porque você não está comprando equipamento que usará apenas ocasionalmente. No entanto, você corre o risco de criar uma experiência do cliente insatisfatória quando a demanda na aplicação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



Available Capacity

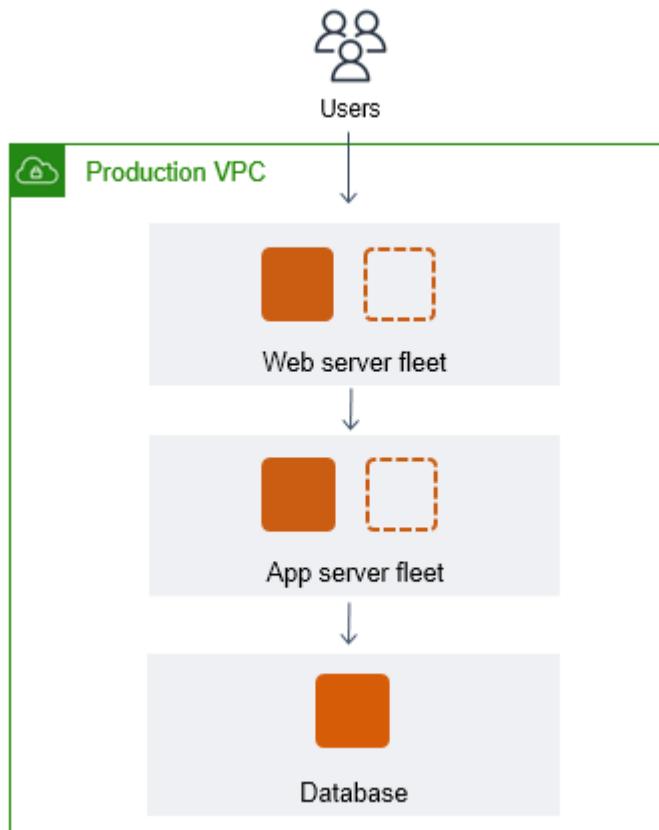
Ao adicionar o Amazon EC2 Auto Scaling a essa aplicação, você passa a ter uma terceira opção disponível. Você pode adicionar novas instâncias à aplicação somente quando necessário e encerrá-las quando não forem mais necessárias. Como o Amazon EC2 Auto Scaling usa instâncias do EC2, você só

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 uma AmazonCloudWatchalarme) 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平衡ador de carga em sua arquitetura. Para obter mais informações, consulte [Elastic Load Balancing \(p. 369\)](#).

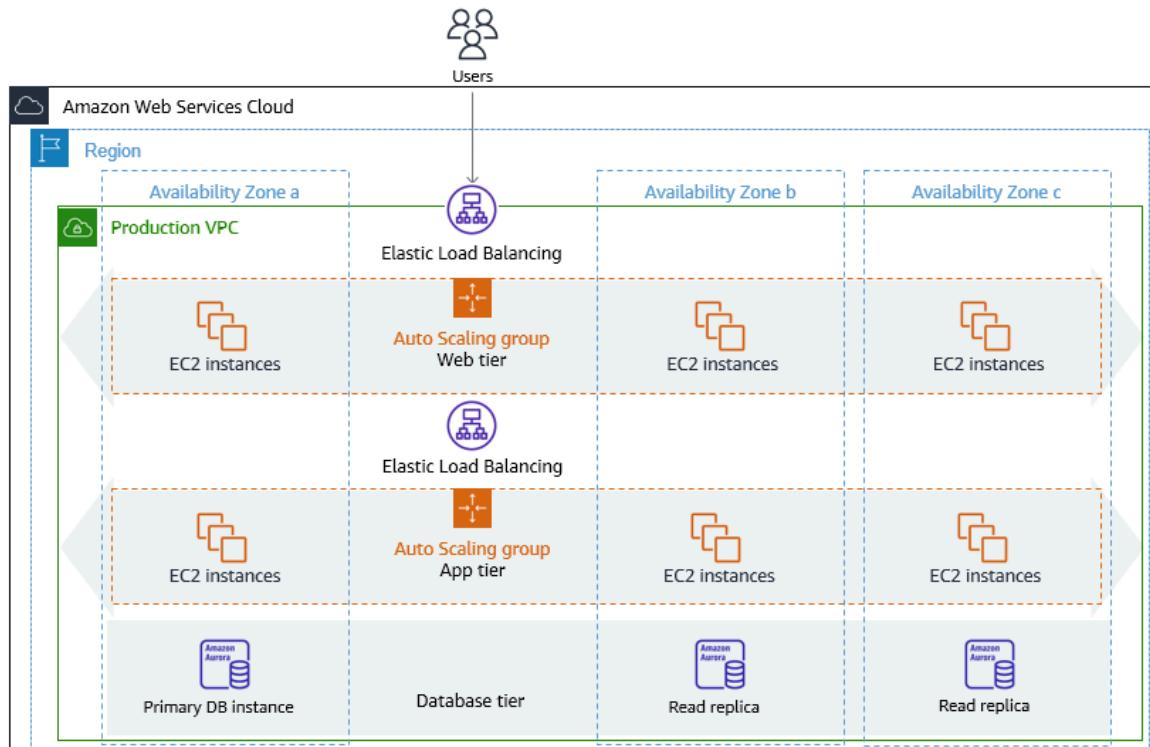
Exemplo: distribuir instâncias entre zonas de disponibilidade

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 \(p. 456\)](#).

Uma zona de disponibilidade é identificada pelo código da Região da AWS seguido por um identificador alfabético (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, você deve 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 selecionará aleatoriamente uma sub-rede dessa 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 desequilibrada. 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. Pode fazer isso por uma margem de 10% ou por 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. Isso pode acontecer devido a um rezoneamento solicitado pelo usuário ou para compensar problemas de disponibilidade de zona. A extensão dura somente o tempo necessário para rebalancear o grupo (em geral, alguns minutos).

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 [Usar o rebalanceamento de capacidade para lidar com interrupções de spot do Amazon EC2 \(p. 346\)](#).

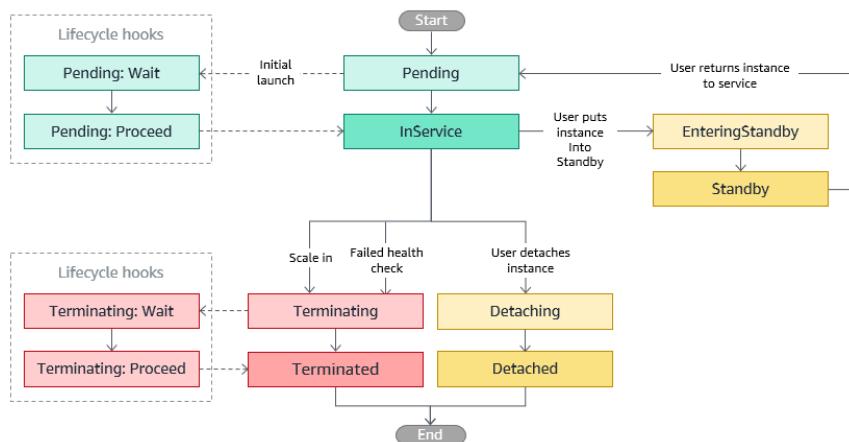
Ciclo de vida das instâncias do Amazon EC2 Auto Scaling

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:

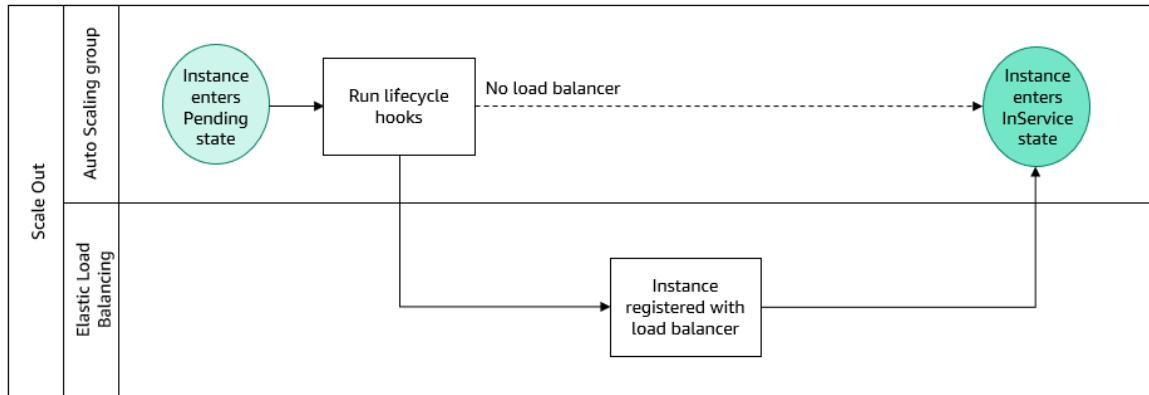
- Você aumenta o tamanho do grupo manualmente. Para obter mais informações, consulte [Escalabilidade manual para o Amazon EC2 Auto Scaling \(p. 168\)](#).
- Você cria uma política de escalabilidade para aumentar automaticamente o tamanho do grupo com base em um aumento especificado na demanda. Para obter mais informações, consulte [Escalabilidade dinâmica para o Amazon EC2 Auto Scaling \(p. 178\)](#).
- 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 \(p. 247\)](#).

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 [Ganchos do ciclo de vida \(p. 10\)](#).

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`.

Veja a seguir um resumo do fluxo de trabalho para registrar uma instância com um balanceador de carga para um evento de aumento da escala na horizontal.



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 \(p. 292\)](#).
- Você coloca a instância em um estado Standby. Para obter mais informações, consulte [Entrar e sair de espera \(p. 11\)](#).
- Você desvincula a instância do grupo do Auto Scaling. Para obter mais informações, consulte [Desvincular uma instância \(p. 10\)](#).
- 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 do Auto Scaling \(p. 319\)](#).

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:

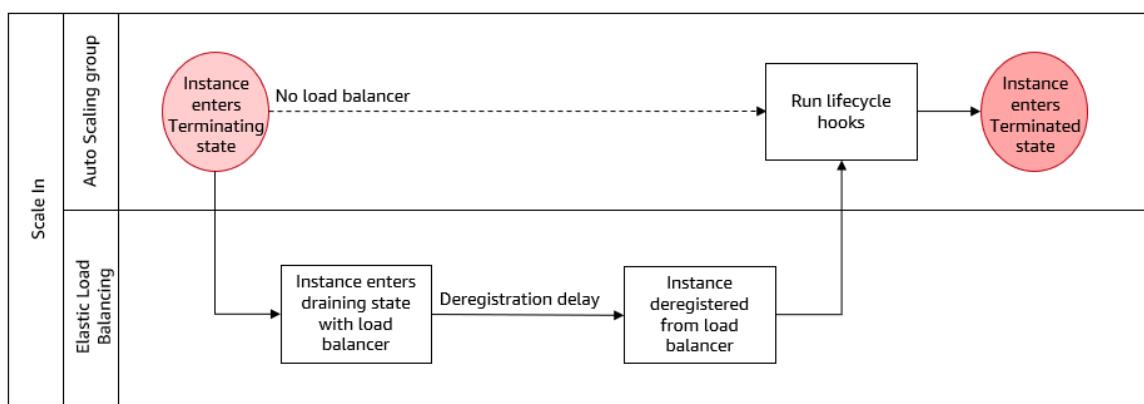
- Você reduz o tamanho do grupo manualmente. Para obter mais informações, consulte [Escalabilidade manual para o Amazon EC2 Auto Scaling \(p. 168\)](#).
- Você cria uma política de escalabilidade para reduzir automaticamente o tamanho do grupo com base em uma redução especificada na demanda. Para obter mais informações, consulte [Escalabilidade dinâmica para o Amazon EC2 Auto Scaling \(p. 178\)](#).
- 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 \(p. 247\)](#).

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 no processo de terminação do grupo do Auto Scaling e de encerramento entram no estado Terminating e não podem ser recolocadas em serviço. Se adicionar um gancho do ciclo de vida a seu grupo do Auto Scaling, você poderá executar uma ação personalizada aqui. Finalmente, as instâncias são completamente encerradas e entram no estado Terminated.

Se o grupo do Auto Scaling estiver configurado para receber tráfego de um平衡ador de carga do Elastic Load Balancing, o Amazon EC2 Auto Scaling cancelará automaticamente o registro da instância em encerramento do balanceador de carga antes de executar ganchos do ciclo de vida. 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.

Veja a seguir um resumo do fluxo de trabalho para cancelar o registro de uma instância com um balanceador de carga para um evento de redução da escala na horizontal.



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 [Anexar instâncias do EC2 a seu grupo do Auto Scaling \(p. 171\)](#).

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 EC2 do seu grupo do Auto Scaling \(p. 174\)](#).

Ganchos 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 \(p. 252\)](#).

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 temporariamente instâncias do grupo do Auto Scaling \(p. 308\)](#).

Cotas do Amazon EC2 Auto Scaling

Sua Conta da AWS tem cotas padrão, anteriormente chamadas de limites, para cada serviço da AWS. A menos que especificado de outra forma, cada cota é específica da região. Você pode solicitar aumentos para algumas cotas e outras cotas 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 [Requesting a Quota Increase](#) (Solicitação de 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.

Todas as solicitações são enviadas ao AWS Support. Você pode acompanhar seu caso de solicitação no console do AWS Support.

Recursos do Amazon EC2 Auto Scaling

Sua Conta da AWS tem as seguintes cotas relacionadas ao número de grupos do Auto Scaling e configurações de execução 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

Sua 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
Ganchos 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-alvo do Elastic Load Balancing por grupo de Auto Scaling	50
Grupos-alvo do VPC Lattice por grupo de 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.

Operação	Máximo de membros da matriz
<u>AttachInstances</u>	20 IDs de instância
<u>AttachLoadBalancers</u>	10平衡adores de cargas
<u>AttachLoadBalancerTargetGroups</u>	10 grupos de destino
<u>BatchDeleteScheduledAction</u>	50 ações programadas
<u>BatchPutScheduledUpdateGroupAction</u>	50 ações programadas
<u>DetachInstances</u>	20 IDs de instância
<u>DetachLoadBalancers</u>	10 balanceadores de cargas
<u>DetachLoadBalancerTargetGroups</u>	10 grupos de destino
<u>EnterStandby</u>	20 IDs de instância
<u>ExitStandby</u>	20 IDs de instância
<u>SetInstanceProtection</u>	50 IDs de instância

Taxas de terminação do EC2

O Amazon EC2 Auto Scaling determina dinamicamente o número de operações de encerramento de instâncias do EC2 que ele pode realizar no momento em que seu grupo de Auto Scaling se expande. Isso significa que você pode ver variações no número de instâncias encerradas por vez em todos os grupos do

Auto Scaling. Essas variações são causadas por considerações externas, como se o Amazon EC2 Auto Scaling deve cancelar o registro de instâncias com um平衡ador de carga.

Outros produtos da

As cotas para outros serviços, como o Amazon EC2, podem afetar seus grupos do Auto Scaling. Para cotas para outros AWS serviços, consulte [Endpoints e cotas de serviço](#) na Referência geral da Amazon Web Services. Para obter cotas para modelos de lançamento, consulte [Restrições do modelo de lançamento](#) na Guia do usuário do Amazon EC2 para instâncias Linux.

Configurar o Amazon EC2 Auto Scaling

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

Tarefas

- [Preparação para usar o Amazon EC2 \(p. 14\)](#)
- [Preparar-se para usar a AWS CLI \(p. 14\)](#)

Preparação para usar o Amazon EC2

Se você não tiver usado o Amazon EC2 anteriormente, execute as tarefas descritas na documentação do Amazon EC2. Para obter mais informações, consulte [Configuração com o Amazon EC2](#) no Manual do usuário do Amazon EC2 para instâncias do Linux ou [Configuração com o Amazon EC2](#) no Manual do usuário do Amazon EC2 para instâncias do Windows.

Preparar-se para usar a AWS CLI

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

Para usar a AWS Command Line Interface (AWS CLI), baixe, instale e configure a versão 1 ou 2 da 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 versão 2 da AWS CLI, consulte [Instalar ou atualizar a versão mais recente da AWS CLI](#) no Guia do usuário da versão 2 da AWS CLI.

O AWS CloudShell permite pular a instalação da AWS CLI em seu ambiente de desenvolvimento e usar o AWS Management Console em seu lugar. Além de evitar a instalação, não é necessário configurar credenciais nem especificar uma região. Sua sessão do AWS Management Console fornece esse contexto para a AWS CLI. O AWS CloudShell pode ser usado em Regiões da AWS compatíveis. Para obter mais informações, consulte [Crie um grupo do Auto Scaling na linha de comando usando o AWS CloudShell. \(p. 360\)](#).

Para obter mais informações, consulte [escalabilidade automática](#) na Referência de comandos da AWS CLI.

Conceitos básicos do Amazon EC2 Auto Scaling

Ao usar o Amazon EC2 Auto Scaling, você deve usar determinados blocos de construção para começar. Este tutorial orienta você durante o processo de configuração de elementos essenciais para criar uma infraestrutura básica para o Amazon EC2 Auto Scaling.

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 imagens de máquina da Amazon (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 configurar um servidor.

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

Note

Para ver um vídeo de introdução, consulte [AWSre:Invent 2018: gerenciamento de capacidade facilitado com o Amazon EC2 Auto Scaling on YouTube](#)

Tarefas

- [Resumo da demonstração \(p. 15\)](#)
- [Preparar para a demonstração \(p. 16\)](#)
- [Etapa 1: Criar um modelo de execução \(p. 16\)](#)
- [Etapa 2: Criar um grupo do Auto Scaling com uma única instância \(p. 18\)](#)
- [Etapa 3: Verificar seu grupo do Auto Scaling \(p. 19\)](#)
- [Etapa 4: Terminar uma instância no seu grupo do Auto Scaling \(p. 19\)](#)
- [Etapa 5: Próximas etapas \(p. 20\)](#)
- [Etapa 6: Limpar \(p. 20\)](#)

Resumo da demonstração

Nesta explicação, você:

- Crie um modelo de configuração que defina suas instâncias do EC2. Pode escolher as instruções do modelo de execução ou da configuração de execução. Embora seja possível usar uma configuração de execução, recomendamos um modelo de execução para que você possa usar os recursos mais recentes do Amazon EC2 e do Amazon EC2 Auto Scaling.
- Cria um grupo do Auto Scaling com uma única instância nele.
- Termina a instância e verifica 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.

Se tiver criado sua Conta da AWS há menos de 12 meses e ainda não tiver excedido os benefícios do [nível gratuito](#) para o Amazon EC2, não haverá cobrança para concluir este tutorial, pois ajudamos você a selecionar um tipo de instância que esteja dentro dos benefícios do nível gratuito. Caso contrário, ao seguir este tutorial, você incorrerá em taxas de uso padrão do Amazon EC2 a partir do momento em que a instância for executada até você excluir o grupo do Auto Scaling (que é a tarefa final deste tutorial) e o status da instância for alterado para terminated.

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. Para obter mais informações, consulte [Configuração do Amazon EC2](#) no Manual do usuário do Amazon EC2 para instâncias do Linux.

Se for novo no Amazon EC2 Auto Scaling e quiser começar a usar o serviço, você pode usar a VPC default (padrão) para sua 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 ver suas VPCs na página [Your VPCs](#) (Suas VPCs) do console do Amazon Virtual Private Cloud (Amazon VPC).

Etapa 1: Criar um modelo de execução

Nesta etapa, você faz login no console do Amazon EC2 com suas credenciais de Conta da AWS e cria um modelo de execução que especifique o tipo de instância do EC2 que o Amazon EC2 Auto Scaling criará 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.

Note

Para usar uma configuração de execução em vez disso, consulte [Create a launch configuration](#).

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 superior, selecione uma 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 Create launch template (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.

Note

Se sua conta tiver menos de 12 meses de vida, você poderá usar uma instância **t2.micro** gratuitamente em determinados limites de uso. Para obter mais informações, consulte o [nível gratuito da AWS](#).

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), Security groups (Grupos de segurança), escolha um grupo de segurança na mesma VPC que pretende 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.
10. É possível deixar Advanced network configuration (Configuração de rede avançada) vazio. Deixar essa definição vazia cria uma interface de rede primária com endereços IP que selecionamos para sua instância, com base na sub-rede em que a interface de rede está estabelecida. Se, em vez disso, você optar por configurar uma interface de rede, o grupo de segurança deverá fazer parte dela.
11. Escolha Create launch template (Criar modelo de execução).
12. Na página de confirmação, escolha Create Auto Scaling group (Criar grupo do Auto Scaling).

Se você não estiver usando modelos de execução no momento e preferir não criar um agora, será possível criar uma configuração de execução.

Uma configuração de execução é semelhante a um modelo de execução, uma vez que especifica o tipo de instância do EC2 que o Amazon EC2 Auto Scaling cria para você. Crie uma configuração de execução incluindo 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 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 você deseja visualizar a página de configurações de inicialização.
2. Na barra de navegação superior, selecione uma Região da AWS. A configuração de execução e os recursos do grupo do Auto Scaling que você cria são vinculados à região que você especifica.
3. Selecione Create launch configuration (Criar uma configuração de execução) e insira **my-first-launch-configuration** no campo Name (Nome).
4. Em Amazon machine image (AMI) (Imagen 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](#).
- 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). Observe que essas AMIs estão marcadas como “Free tier eligible” (Qualificáveis para nível gratuito).
5. Em Instance type (Tipo de instância), selecione uma configuração de hardware para sua instância.

Note

Se sua conta tiver menos de 12 meses de vida, você poderá usar uma instância `t2.micro` gratuitamente em determinados limites de uso. Para obter mais informações, consulte o [nível gratuito da AWS](#).

6. Em Additional configuration (Configurações adicionais), para Advanced details (Detalhes avançados), IP address type (Tipo de endereço IP), faça uma seleção. Para fornecer conectividade com a Internet a instâncias em uma VPC, escolha uma opção que atribua um endereço IP público. Se uma instância for executada em uma VPC padrão, o padrão é atribuir um endereço IP público. Se você quiser fornecer conectividade com a Internet às suas instâncias, mas não tem certeza se

tem uma VPC padrão, escolha Assign a public IP address to every instance (Atribuir um endereço IP público a cada instância).

7. Para Security groups (Grupos de segurança), escolha um grupo de segurança existente. 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.
8. Em Key pair (login) (Par de chaves - login), escolha uma opção em Key pair options (Opções de par de chaves) conforme orientado. A conexão a uma instância não está incluída como parte deste tutorial. Portanto, você pode selecionar Proceed without a key pair (Continuar sem um par de chaves), a menos que você pretenda conectar-se à instância usando SSH.
9. 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 com uma única instância

Agora use o Amazon EC2 Auto Scaling para criar um grupo do Auto Scaling e adicionar o modelo de execução ou a configuração de execução ao grupo. Inclua também informações como as sub-redes da VPC para as instâncias.

Use o procedimento a seguir para continuar de onde parou depois que criar um modelo de execução ou uma configuração de execução.

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 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).

3. Na seção Network (Rede), mantenha a VPC definida como a VPC padrão para a Região da AWS escolhida ou selecione sua própria 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 \(p. 417\)](#).
5. [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 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 [Auto Scaling grupos](#) 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 PreInService. 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 Health status (Status de integridade) mostra o resultado da verificação de integridade da instância do EC2 em 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 [Auto Scaling grupos](#) 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).

O desaquecimento padrão para o grupo do Auto Scaling é de 300 segundos (5 minutos), de forma que demora 5 minutos até que você veja a ação de escalabilidade. 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.

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 para a escalabilidade automática recém-criada. 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 [Conectar-se à instância do Linux usando o Session Manager](#) e [Conectar-se à instância do Linux usando o SSH](#) no Guia do usuário do Amazon EC2 para instâncias do Linux.
- 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 [Monitorar com notificações do Amazon SNS \(p. 341\)](#).
- Escalar manualmente seu grupo do Auto Scaling para testar a notificação do SNS. Para obter mais informações, consulte [Escalabilidade manual \(p. 168\)](#).

Você também pode começar a se familiarizar com os conceitos de escalonamento lendo sobre [Políticas de escalabilidade de rastreamento de destino \(p. 180\)](#). Se a carga do seu aplicativo 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 capacidade no grupo do Auto Scaling \(p. 166\)](#).

Se você planeja anexar um平衡ador de carga ao grupo do Auto Scaling, você pode aprender como criar um rapidamente usando o console do Amazon EC2 Auto Scaling. Para obter mais informações, consulte [Configurar um Application Load Balancer ou Network Load Balancer pelo console do Amazon EC2 Auto Scaling \(p. 374\)](#).

Etapa 6: Limpar

Você pode excluir sua infraestrutura de escalabilidade ou excluir apenas o grupo do Auto Scaling e manter o modelo de execução ou configuração 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 [Auto Scaling grupos](#) do Amazon EC2.
2. Marque a caixa de seleção ao lado do seu grupo do Auto Scaling (my-first-asg).
3. Escolha Delete (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 Desired (Desejado), Min (Mínimo) e Max (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.

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 a configuração de execução (my-first-launch-configuration).
3. Escolha Ações, Excluir configuração de execução.
4. Quando a confirmação for solicitada, escolha Delete (Excluir).

Modelos de execução

Um modelo de execução é semelhante a uma [configuração de execução \(p. 49\)](#), 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. Você deve 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 \(p. 67\)](#).

Com modelos de execução, você também pode usar recursos mais recentes do Amazon EC2. Isso inclui, parâmetros do Systems Manager (ID de AMI), a atual geração de volumes de IOPS provisionadas do EBS (io2), a marcação de volume do EBS, [instâncias T2 ilimitadas](#), Elastic Inference e [hosts dedicados](#), entre outros. Os hosts dedicados são servidores físicos com a capacidade das instâncias do EC2 exclusivos para seu uso. Enquanto as [Instâncias dedicadas](#) do Amazon EC2 também são executadas em hardware dedicado, a vantagem de usar hosts dedicados sobre instâncias dedicadas é que você pode trazer licenças de software qualificadas de fornecedores externos e usá-las em instâncias do EC2.

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.

Índice

- [Permissões \(p. 22\)](#)
- [Criar um modelo de execução para um grupo do Auto Scaling \(p. 23\)](#)
- [Migre para lançar modelos \(p. 32\)](#)
- [Solicitar instâncias spot para aplicações flexíveis e com tolerância a falhas \(p. 36\)](#)
- [Exemplos para criação e gerenciamento de modelos de execução com a AWS Command Line Interface \(AWS CLI\) \(p. 37\)](#)
- [Usar parâmetros do AWS Systems Manager em vez de IDs de AMI em modelos de execução \(p. 44\)](#)

Permissões

Os procedimentos nesta seção pressupõem que você já tenha as permissões necessárias para usar modelos de execução. Com permissões vigentes, você pode criar e gerenciar modelos de execução. Você

também pode criar e atualizar grupos do Auto Scaling e especificar um modelo de execução em vez de uma configuração de execução.

Ao atualizar ou criar um grupo do Auto Scaling e especificar um modelo de execução, suas permissões de `ec2:RunInstances` são verificadas. Se você não tiver permissões suficientes, receberá um erro informando que não está autorizado a usar o modelo de execução.

Algumas funcionalidades adicionais na solicitação exigem permissões adicionais, como a capacidade de passar uma função do IAM para instâncias provisionadas ou adicionar tags a instâncias e volumes provisionados.

Para obter informações sobre como um administrador concede permissões, consulte [Suporte a modelo de execução \(p. 443\)](#).

Criar um modelo de execução para um grupo do Auto Scaling

Antes de criar um grupo do Auto Scaling usando um modelo de execução, você deve criar um modelo de execução com os parâmetros necessários para executar uma instância do EC2. Esses parâmetros incluem o ID da Imagem de máquina da Amazon (AMI) e um tipo de instância.

Um modelo de execução fornece funcionalidade completa para o Amazon EC2 Auto Scaling e também recursos mais recentes do Amazon EC2, como a geração atual de volumes de IOPS provisionadas do EBS Amazon (io2), marcação de volume do EBS, instâncias T2 ilimitadas, Elastic Inference e hosts dedicados.

Siga os procedimentos abaixo para criar novos modelos de execução.

Índice

- [Criar seu modelo de execução \(console\) \(p. 23\)](#)
 - [Altere as configurações da interface de rede padrão \(p. 25\)](#)
 - [Modificar a configuração do armazenamento \(p. 27\)](#)
 - [Definir configurações avançadas para seu modelo de execução \(p. 29\)](#)
- [Criar um modelo de execução com base em uma instância existente \(console\) \(p. 31\)](#)
- [Informações adicionais \(p. 31\)](#)
- [Limitações \(p. 31\)](#)

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 executar instâncias com uma AMI AWS Graviton ou Graviton2 baseada em Arm, você deve especificar um tipo de instância compatível com Arm.

Criar seu modelo de execução (console)

As etapas a seguir descrevem como configurar seu modelo 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 Create launch template (Criar modelo de execução). Insira um nome e forneça uma descrição para a versão inicial do modelo de execução.
4. (Opcional) Abaixo Orientação de Auto Scaling, marque a caixa de seleção para que o Amazon EC2 forneça orientação para ajudar a criar um modelo para usar 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 aplicativos e sistemas operacionais (Amazon Machine Image): (Obrigatório) escolha o ID da AMI para suas instâncias. Você pode pesquisar todas as AMIs disponíveis ou selecionar uma AMI na lista Recents (Recentes) ou Quick Start (Início rápido). Caso não veja a AMI de que precisa, escolha Browser more AMIs (Pesquisar mais AMIs) para navegar pelo catálogo completo de AMIs.

Para escolher uma AMI personalizada, primeiro você deve criar uma AMI desde uma instância personalizada. Para mais informações, consulte [Create an AMI](#) (Criar uma AMI) no Guia do usuário do Amazon EC2 para instâncias do Linux.

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

Como alternativa, para iniciar um grupo do Auto Scaling com vários tipos de instâncias, escolha Advanced (Avançado), Specify instance type attributes (Especificando atributos de tipo de instância) e especifique as seguintes opções:

- Number of vCPUs (Número de vCPUs): insira o número mínimo e máximo de vCPUs. 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.
- Expand 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 [InstanceRequirementsRequest](#) na Referência da API Amazon EC2.
- Tipos de instância resultantes: é possível 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.

Para obter mais informações, consulte [Criar um grupo do Auto Scaling usando seleção de tipo de instância baseada em atributos \(p. 92\)](#).

- 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 e instância do Linux](#) no Guia do usuário do Amazon EC2 para instâncias do Linux.

- 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 para instâncias do Linux.

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 executará 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 sua VPCs](#) no Guia do usuário da Amazon VPC.

- e. Faça um dos seguintes procedimentos:
 - Altere as configurações da interface de rede padrão. Por exemplo, você pode habilitar ou desabilitar o recurso de endereçamento IPv4 público, que substitui a configuração de atribuição automática de endereços IPv4 públicos na sub-rede. Para obter mais informações, consulte [Alterar as configurações da interface de rede padrão \(p. 25\)](#).
 - Ignore essa etapa para manter as configurações da interface de rede padrão.
 - f. Faça um dos seguintes procedimentos:
 - Modificar a configuração do armazenamento. Para obter mais informações, consulte [Modificar a configuração do armazenamento \(p. 27\)](#).
 - 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). Para obter mais informações, consulte [Definir configurações avançadas para seu modelo de execução \(p. 29\)](#).
 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.

Altere as configurações da interface de rede padrão

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

Considerações e limitações

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

- Você deverá configurar o grupo de segurança como parte da interface de rede, e não na seção Security Groups (Grupos de segurança) do modelo. Não é possível especificar grupos de segurança nos dois locais.
- Não é possível atribuir endereços IP privados adicionais, conhecidos como endereços IP privados secundários, a uma interface de rede.
- Se você especificar um ID de interface de rede existente, poderá executar apenas uma instância. Para tanto, você deve usar a AWS CLI ou um SDK para criar o grupo do Auto Scaling. Ao criar o grupo, você deve 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 atribuir automaticamente um endereço IPv4 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. Para obter mais informações, consulte [Fornecer conectividade de rede para suas instâncias do Auto Scaling usando a Amazon VPC \(p. 414\)](#).
- 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. Auto-assign public IP (Atribuir IP público automaticamente): altere se sua interface de rede com um índice de dispositivo de 0 recebe um endereço IPv4 público. Por padrão, as instâncias em uma sub-rede padrão recebem um endereço IPv4 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](#) no Guia do usuário do Amazon EC2 para instâncias do Linux.
- 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 (HPC), 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 para instâncias do Linux.
- 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 para instâncias do Linux.

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

Modificar a configuração do armazenamento

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 ajudá-lo 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 obter mais informações, consulte [Armazenamento de instância do Amazon EC2](#) e [Volumes do Amazon EBS](#) no Manual do usuário do Amazon EC2 para instâncias do Linux.
 - 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 EC2 para instâncias do Linux.
 - e. Volume type (Tipo de volume): para volumes do EBS, escolha o tipo de volume. Para ter mais informações, consulte [Tipos de volumes do Amazon EBS](#) no Guia do usuário do Amazon EC2 para instâncias do Linux.
 - f. IOPS: se você tiver selecionado um SSD de IOPS provisionadas (io1 e io2) ou um tipo de volume de SSD de uso geral (gp3), poderá inserir o número de operações de E/S por segundo (IOPS) com o qual o volume seja compatível. 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 a criptografia no [Amazon EBS](#) e [Criptografia por padrão](#) no Manual do usuário do Amazon EC2 para instâncias do Linux.

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 AWS KMS key especificado.

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	
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 escolheu Yes (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).

Definir configurações avançadas para seu modelo de execução

Você pode definir quaisquer recursos adicionais que suas instâncias do Auto Scaling precisem. Por exemplo, você pode escolher uma função 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.

As etapas a seguir discutem as configurações mais úteis para se observar com atenção. Para obter mais informações sobre qualquer uma das configurações em Advanced details (Detalhes avançados), consulte [Criação de um modelo de execução](#) no Manual do usuário do Amazon EC2 para instâncias do Linux.

Para definir configurações avançadas

1. Em Advanced details (Detalhes avançados), expanda a seção para visualizar os campos.
2. Em Purchase 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, e escolher Customize (Personalizar) para alterar as configurações padrão da instância spot. Para um grupo do Auto Scaling, você deve 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 \(p. 36\)](#).

Note

O Amazon EC2 Auto Scaling permite substituir o tipo de instância no modelo de execução para criar um grupo do Auto Scaling que use vários tipos de instância e execute instâncias spot e sob demanda. Para isso, você deve deixar Purchasing option (Opção de compra) não especificado no modelo de execução.

Se tentar criar um grupo de instâncias mistas usando um modelo de execução com Purchasing option (Opção de compra) especificado, você receberá o erro a seguir.

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.

Para obter informações sobre a criação de grupos de instâncias mistas, consulte [Grupos do Auto Scaling com vários tipos de instâncias e opções de compra \(p. 67\)](#).

3. Em IAM instance profile (Perfil da instância do IAM), é possível especificar um perfil de instância do AWS Identity and Access Management (IAM) para associar às instâncias. Ao escolher um perfil da instância, você associa a função do IAM correspondente às instâncias do EC2. Para obter mais informações, consulte [Funções do IAM para aplicações que são executadas em instâncias do Amazon EC2 \(p. 448\)](#).
4. Em Termination protection (Proteção contra término), escolha se deseja proteger as instâncias contra término accidental. Quando você habilita a proteção contra término, ela fornece proteção adicional contra término, mas não protege contra o término iniciado pelo Amazon EC2 Auto Scaling. Para controlar se um grupo do Auto Scaling pode terminar uma instância específica, use [Usar proteção de redução na escala na horizontal de instâncias \(p. 302\)](#).
5. Para DetalhadoCloudWatchmonitoramento, escolha se deseja permitir que as instâncias publiquem dados métricos em intervalos de 1 minuto na AmazonCloudWatch. Aplicam-se cobranças adicionais. Para obter mais informações, consulte [Configurar monitoramento para instâncias do Auto Scaling \(p. 337\)](#).
6. Para Elastic inference (Inferência elástica) escolha uma aceleradora de inferência elástica a ser anexada à instância de CPU do EC2. Aplicam-se cobranças adicionais. Para obter mais informações, consulte [Trabalhando com o Amazon Elastic Inference](#) no Guia do desenvolvedor do Amazon Elastic Inference.
7. Em T2/T3 Unlimited (T2/T3 ilimitado), escolha se habilita as aplicações a terem intermitência acima da linha de base pelo tempo que for necessário. Este campo é válido somente para instâncias T2, T3 e T3a. Podem se aplicar cobranças adicionais. Para obter mais informações, consulte [Usar um grupo do](#)

[Auto Scaling para iniciar uma instância expansível como ilimitada](#) no Manual do usuário do Amazon EC2 para instâncias do Linux.

8. Em Placement group name (Nome do grupo de posicionamento), você pode especificar um grupo de posicionamento no qual executar as instâncias. Nem todos os tipos de instância podem ser executados em um grupo de posicionamento. Se você configurar um grupo do Auto Scaling usando um comando da CLI que especifica um grupo de posicionamento diferente, o grupo de posicionamento para o grupo do Auto Scaling terá precedência.
9. Em Capacity Reservation (Reserva de capacidade), você pode especificar se deseja iniciar instâncias em capacidade compartilhada, qualquer reserva de capacidade open, uma reserva de capacidade específica ou um grupo de reserva de capacidade. Para obter mais informações, consulte [Use reservas de capacidade sob demanda para reservar capacidade em zonas de disponibilidade específicas \(p. 354\)](#).
10. Em Tenancy (Locação), você pode escolher iniciar suas instâncias em hardware compartilhado (Shared), em hardware dedicado (Dedicated) ou, ao usar um grupo de recursos de host, em hosts dedicados (Dedicated host). Podem se aplicar cobranças adicionais.

Se você escolher Dedicated Hosts (Hosts dedicados), preencha as informações a seguir:

- Em Tenancy host resource group (Grupo de recursos de host de locação), você pode especificar um grupo de recursos de host para uma AMI de BYOL a ser usada em hosts dedicados. Não é necessário já ter alocado hosts dedicados em sua conta antes de usar esse recurso. De qualquer forma, suas instâncias serão executadas automaticamente em hosts dedicados. Observe que uma AMI baseada em uma associação de configuração de licença pode ser mapeada para apenas um grupo de recursos de host por vez. Para obter mais informações, consulte [Grupos de recursos de host](#) no Manual do usuário do AWS License Manager.
11. Em License configurations (Configurações de licença), especifique a configuração de licença a ser usada. Você pode iniciar instâncias com relação à configuração de licença especificada para rastrear o uso da licença. Para obter mais informações, consulte [Criar uma configuração de licença](#) no Manual do usuário do License Manager.
 12. Para configurar opções de metadados de instância para todas as instâncias associadas a esta versão do modelo de execução, 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. Só é possível especificar a condição para exigir IMDSv2 quando o endpoint HTTP estiver habilitado.
 - b. Em Metadata version (Versão dos metadados), você pode escolher 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 IMDSv1 e IMDSv2.
 - 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.

Para obter mais informações, consulte [Configuração do serviço de metadados de instância](#) no Manual do usuário do Amazon EC2 para instâncias do Linux.

13. Em User data (Dados do usuário), você pode adicionar scripts de shell e diretivas de cloud-init para personalizar uma instância na inicialização. Para obter mais informações, consulte [Executar comandos na instância do Linux na inicialização](#) no Guia do usuário do Amazon EC2 para instâncias do Linux.

Note

Se você executar scripts na inicialização, será necessário mais tempo para que uma instância esteja pronta para uso. No entanto, você pode conceder um tempo extra para que os scripts

sejam concluídos antes que a instância entre no estado InService adicionando um gancho do ciclo de vida ao grupo do Auto Scaling. Para obter mais informações, consulte [Ganchos do ciclo de vida do Amazon EC2 Auto Scaling \(p. 252\)](#).

14. Escolha Create launch template (Criar modelo de execução).
15. Para criar um grupo do Auto Scaling, escolha Create Auto Scaling group (Criar grupo do Auto Scaling) na página de confirmação.

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 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.

Informações adicionais

Para obter mais informações sobre como criar modelos de execução, consulte:

- Seção [Execução de uma instância a partir de um modelo de execução](#) do Manual do usuário do Amazon EC2 para instâncias do Linux
- Seção [Trechos de modelos do Auto Scaling](#) do Guia do usuário do AWS CloudFormation
- [AWS::EC2::LaunchTemplate](#) seção do AWS CloudFormation Guia do usuário

Para os procedimentos de criação de um grupo do Auto Scaling com um modelo de execução, veja os tópicos a seguir:

- [Criar um grupo do Auto Scaling usando um modelo de execução \(p. 62\)](#)
- [Grupos do Auto Scaling com vários tipos de instâncias e opções de compra \(p. 67\)](#)
- [Criar um grupo do Auto Scaling usando seleção de tipo de instância baseada em atributos \(p. 92\)](#)

Limitações

- O Amazon EC2 permite que você configure uma sub-rede em um modelo de execução. No entanto, as configurações de sub-rede do grupo do Auto Scaling têm precedência sobre as configurações de sub-rede do modelo de execução.
- Como as configurações de sub-rede no modelo de execução são ignoradas em favor do que é especificado no grupo do Auto Scaling, todas as interfaces de rede criadas para uma determinada

instância serão conectadas à mesma sub-rede que a instância. Para outras limitações em interfaces de rede definidas pelo usuário, consulte [Altere as configurações da interface de rede padrão](#) (p. 25).

- Um modelo de execução permite que você defina configurações adicionais no grupo do Auto Scaling para executar vários tipos de instâncias e combinar opções de compra spot e sob demanda, conforme descrito em [Grupos do Auto Scaling com vários tipos de instâncias e opções de compra \(p. 67\)](#). Não há compatibilidade para a execução de instâncias com essa combinação se você especificar uma solicitação de instância spot no modelo de execução.
 - O suporte para hosts dedicados (locação de host) só estará disponível 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.

Migre para lançar modelos

A partir de 2023, as configurações de lançamento não oferecem suporte a nenhum novo tipo de instância do Amazon EC2 lançado após 31 de dezembro de 2022. Para obter mais informações, consulte [Configurações de execução](#) (p. 49).

Para migrar de uma configuração de inicialização para um modelo de execução, consulte as etapas a seguir.

Important

Antes de continuar, confirme se você tem as permissões necessárias para trabalhar com modelos de lançamento. Para obter mais informações, consulte [Suporte a modelo de execução](#) (p. 443).

Etapa 1: Encontre grupos de Auto Scaling que usam configurações de inicialização

Para identificar se você tem grupos de Auto Scaling que ainda estão usando configurações de inicialização, execute o seguinte [describe-auto-scaling-groups](#) comando usando o AWS CLI. Substituir **REGIÃO** com o seu Reqião da AWS.

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

A seguir está 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"
},
...
additional groups
]
```

Como alternativa, para remover tudo, exceto os nomes dos grupos do Auto Scaling com os nomes de suas respectivas configurações de inicialização 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 da saída da AWS CLI](#) no Guia do usuário da AWS Command Line Interface.

Etapa 2: Copiar uma configuração de inicialização para um modelo de inicialização

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

A cópia de várias configurações de inicialização resulta em modelos de execução com nomes idênticos. Para alterar o nome dado a um modelo de execução durante o processo de cópia, você deve copiar as configurações de inicialização uma por 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 Auto Scaling, escolha Launch Configurations (Configurações de execução).
3. 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.
4. 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.
5. (Opcional) Selecione Crie um grupo de Auto Scaling usando o novo modelo.

Você pode pular essa etapa para concluir a cópia da configuração de inicialização. Você não precisa criar um novo grupo de Auto Scaling.

6. 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 de Auto Scaling para usar um modelo de lançamento

Depois de criar um modelo de lançamento, você está pronto para adicioná-lo ao seu grupo de Auto Scaling.

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

1. Abra o console do Amazon EC2 em <https://console.aws.amazon.com/ec2/> e escolha grupos do Auto Scaling 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. Escolha Atualizar.

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

O seguinte `update-auto-scaling-group` comando atualiza o grupo de Auto Scaling especificado para usar a versão inicial do modelo de lançamento 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 mais exemplos de uso de comandos da CLI para atualizar um grupo de Auto Scaling para usar um modelo de inicialização, consulte [Atualizar um grupo do Auto Scaling para usar um modelo de execução \(p. 43\)](#).

Etapa 4: substitua suas instâncias

Depois de 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 permitir que o escalonamento automático substitua gradualmente as instâncias existentes por novas instâncias com base nas do grupo [políticas de rescisão \(p. 292\)](#), ou você pode encerrá-los. O encerramento manual força seu grupo de Auto Scaling a lançar novas instâncias para manter a capacidade desejada pelo grupo. Para obter mais informações, consulte [Terminar uma instância](#), no Guia do usuário do Amazon EC2 para instâncias do Linux.

Como alternativa, você pode iniciar uma atualização de instância para substituir as instâncias em seu grupo de Auto Scaling, em vez de substituir manualmente as instâncias algumas por vez. Para obter mais informações, consulte [Substituir instâncias do Auto Scaling com base em uma atualização de instância \(p. 108\)](#). Se o grupo for grande, uma atualização da instância pode ser particularmente útil.

Informações adicionais

Para mais informações sobre como migrar para modelos de execução, consulte [Amazon EC2 Auto Scaling will no longer add support for new EC2 features to Launch Configurations](#) (O Amazon EC2 Auto Scaling não oferecerá mais suporte para novos recursos do EC2 para configurações de execução) no Blog AWS Compute.

Para um tópico que explica como migrar AWS CloudFormation pilhas de configurações de lançamento a modelos de lançamento, consulte [Migrar AWS CloudFormation pilhas de configurações de execução \(p. 361\)](#).

Para obter instruções que mostram como criar um novo modelo de lançamento para o Amazon EC2 Auto Scaling a partir do console, consulte [Criar um modelo de execução para um grupo do Auto Scaling \(p. 23\)](#).

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 [Definir configurações avançadas para seu modelo de execução \(p. 29\)](#).

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 executar 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 \(p. 67\)](#).

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 \(p. 49\)](#). 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 [Spot Instances](#) (Instâncias spot) e [Pricing and savings](#) (Custos e economias) no Guia do usuário do Amazon EC2 para instâncias Linux.
- 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.

Exemplos para criação e gerenciamento de modelos de execução com a AWS Command Line Interface (AWS CLI)

Você pode criar e gerenciar modelos de execução usando o AWS Management Console, AWS CLI ou SDKs. Esta seção mostra exemplos de criação e gerenciamento de modelos de execução para o Amazon EC2 Auto Scaling na AWS CLI.

Índice

- [Exemplo de uso \(p. 38\)](#)
- [Criar um modelo de execução básico \(p. 38\)](#)
- [Especificar etiquetas que marcam instâncias ao iniciar \(p. 39\)](#)
- [Especificar uma função do IAM a ser transmitida às instâncias \(p. 39\)](#)
- [Atribuir um endereço IP público \(p. 39\)](#)
- [Especificar um script de dados do usuário que configura instâncias ao iniciar \(p. 40\)](#)
- [Especificar um mapeamento de dispositivos de blocos \(p. 40\)](#)
- [Especificar hosts dedicados para trazer licenças de software de fornecedores externos \(p. 40\)](#)
- [Especificar uma interface de rede existente \(p. 40\)](#)
- [Criar várias interfaces de rede \(p. 41\)](#)
- [Gerenciar modelos de execução \(p. 41\)](#)

- [Atualizar um grupo do Auto Scaling para usar um modelo de execução \(p. 43\)](#)

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-template](#) comando da seguinte forma, com essas 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 IPv4 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.micro"}]
```

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 uma função do IAM a ser transmitida às instâncias

O exemplo a seguir especifica o nome do perfil da instância associada à função do IAM a ser passada às instâncias na execução. Para obter mais informações, consulte [Funções do IAM para aplicações que são executadas em instâncias do Amazon EC2 \(p. 448\)](#).

```
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 seguinte `create-launch-template` exemplo configura o modelo de execução para atribuir endereços públicos às instâncias lançadas 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", "InstanceType": "t2.micro"}'
```

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 comando [create-launch-template](#) requer dados de 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 seguinte exemplo cria um modelo de lançamento com um mapeamento de dispositivos de blocos: um volume 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-04d5cc9b88example", "In-
```

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 nas instâncias do EC2 usando o comando [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 seguinte exemplo 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.micro"}'
```

Criar várias interfaces de rede

O seguinte [create-launch-template](#) exemplo 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", "InstanceType": "t2.micro"}]'
```

Se você usa um tipo de instância que suporta várias placas de rede e Elastic Fabric Adapters (EFAs), você pode adicionar uma interface secundária a uma placa de rede secundária e habilitar o EFA usando o seguinte [create-launch-template](#) comando. Para obter mais informações, consulte [Adicionando um EFA a um modelo de execução](#) no Manual do usuário do Amazon EC2 para instâncias do Linux.

```
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-09d95fab7fxexa", "InstanceType": "t2.micro"}]'
```

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 usam 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. Como exemplo, consulte [Como posso fazer minha interface de rede secundária funcionar na minha instância do Ubuntu EC2?](#) na Central de conhecimento da AWS.

Gerenciar modelos de execução

A AWS CLI inclui vários outros comandos que ajudam você a gerenciar seus modelos de execução.

Índice

- [Listar e descrever modelos de execução \(p. 41\)](#)
- [Criar uma versão de modelo de execução \(p. 43\)](#)
- [Excluir uma versão de modelo de execução \(p. 43\)](#)
- [Excluir um modelo de execução \(p. 43\)](#)

Listar e descrever modelos de execução

Você pode usar dois AWS CLI comandos para obter informações sobre seus modelos de lançamento: [describe-launch-templates](#) e [describe-launch-template-versions](#).

O [describe-launch-templates](#) comando permite que você obtenha uma lista de qualquer um dos modelos de inicialização 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
```

Esta é uma resposta de exemplo.

```
{  
    "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 seguinte [describe-launch-template-versions](#) comando fornece informações que descrevem as versões do modelo de lançamento especificado.

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

Esta é uma resposta de exemplo.

```
{  
    "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"  
                        ],  
                        "SubnetId": "subnet-00000000000000000000000000000000"  
                    }  
                ]  
            }  
        }  
    ]  
}
```

```
        "AssociatePublicIpAddress": true
    }
],
"DefaultVersion": true,
"CreateTime": "2020-02-28T19:52:27.000Z"
]
}
```

Criar uma versão de modelo de execução

O seguinte [create-launch-template-version](#) comando cria uma nova versão do modelo de execução 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-template](#) comando.

Excluir uma versão de modelo de execução

O seguinte [delete-launch-template-versions](#) comando exclui a versão do modelo de lançamento especificada.

```
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 seguinte [delete-launch-template](#) comando. 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-group](#) comando para adicionar um modelo de lançamento a um grupo existente de Auto Scaling.

Note

Se você alternar seu grupo do Auto Scaling do uso uma configuração de execução, certifique-se de que suas permissões estejam atualizadas. Para usar um modelo de execução, você precisa de [permissões](#) específicas.

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

O seguinte [update-auto-scaling-group](#) comando atualiza o grupo de Auto Scaling especificado para usar a versão mais recente do modelo de inicialização 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 seguinte comando atualiza o grupo especificado do Auto Scaling para usar uma versão específica do modelo de lançamento especificado.

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

Usar parâmetros do AWS Systems Manager em vez de IDs de AMI em modelos de execução

O Amazon EC2 Auto Scaling oferece suporte ao uso de parâmetros do AWS Systems Manager que referenciam IDs de imagem de máquina da Amazon (AMI) em modelos de execução. Com os parâmetros do Systems Manager, é possível atualizar grupos do Auto Scaling para usar novos IDs de AMI sem precisar criar novos modelos de execução ou novas versões dos modelos de execução sempre que um ID de AMI for alterado. Esses IDs podem ser alterados regularmente, como quando uma AMI recebe as atualizações de sistema operacional ou de software mais recentes.

Você pode criar, atualizar ou excluir parâmetros do Systems Manager usando o Parameter Store, um recurso do 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.

Important

Para especificar parâmetros do Systems Manager em modelos de execução, é necessário ter permissão para usar a ação `ssm:GetParameters`. Também é necessário ter permissão para usar a ação `ssm:GetParameters` para usar um modelo de execução que especifique um parâmetro do Systems Manager. Isso permite 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.

Para mais informações, consulte os seguintes recursos :

- Para criar um parâmetro no console, consulte [Crie um parâmetro do Systems Manager \(console\)](#) no Guia do usuário do AWS Systems Manager.
- Para criar um parâmetro com a AWS CLI, consulte [Crie um parâmetro do Systems Manager \(AWS CLI\)](#) no Guia do usuário do AWS Systems Manager.
- Para criar versões e rótulos de parâmetros, consulte [Como trabalhar com versões de parâmetros e Trabalhar com rótulos de parâmetros](#) no Guia do usuário do AWS Systems Manager.
- Para obter informações sobre verificar se o valor do parâmetro inserido é um ID de AMI válido, consulte [Suporte a parâmetros nativos para IDs de imagem de máquina da Amazon](#) no Guia do usuário do AWS Systems Manager.
- Para obter mais informações sobre como usar parâmetros que referenciam IDs de AMI em modelos de execução, consulte [Usar um parâmetro do Systems Manager em vez de um ID de AMI](#) no Guia do usuário do Amazon EC2 para instâncias do Linux.
- Você também pode referenciar parâmetros públicos do AWS Systems Manager para AMIs públicas mantidas pela AWS em seus modelos de execução. Para saber como referenciar esses parâmetros

públicos, consulte [Encontre a AMI mais recente do Amazon Linux usando Systems Manager no Guia do usuário do Amazon EC2 para instâncias do Linux](#) e [Encontre a AMI mais recente do Amazon Linux usando Systems Manager](#) no Guia do usuário do Amazon EC2 para instâncias do Windows.

Console

Para criar um modelo de execução usando um parâmetro do AWS Systems Manager

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 Application and OS Images (Amazon Machine Image) (Imagens de aplicações e sistemas operacionais [imagem de máquina da Amazon]), escolha Browse more AMIs (Procurar mais AMIs).
5. Escolha o botão de seta à direita da barra de pesquisa e escolha Especificar valor personalizado/parâmetro do Systems Manager.
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:
 - **resolve:ssm:parameter-name**
 - **resolve:ssm:parameter-name:version-number**
 - **resolve:ssm:parameter-name:label**
 - **resolve:ssm:public-parameter**
 - b. Escolha Save (Salvar).
7. Especifique outros parâmetros do modelo de execução, se necessário, e escolha Criar modelo de execução.

Para obter mais informações sobre como criar um modelo de execução no console, consulte [Criar um modelo de execução para um grupo do Auto Scaling \(p. 23\)](#).

AWS CLI

Exemplo de modelo de execução que especifica um parâmetro de propriedade do cliente

1. 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 para uma versão específica do parâmetro `golden-ami`.

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

A seguir está 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
  }
}
```

2. Para confirmar que o modelo de execução recebe seu ID de AMI preferido, use o [describe-launch-template-versions](#) comando. O comando usa a opção `--resolve-alias` para resolver o parâmetro com o ID de AMI real.

```
aws ec2 describe-launch-template-versions --launch-template-name my-template-for-
auto-scaling \
--versions $Default --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-04d5cc9b88example`.

```
{
  "LaunchTemplateVersions": [
    {
      "LaunchTemplateId": "lt-068f72b724example",
      "LaunchTemplateName": "my-template-for-auto-scaling",
      "VersionNumber": 1,
      "CreateTime": "2022-12-27T17:11:21.000Z",
      "CreatedBy": "arn:aws:iam::123456789012:user/Bob",
      "DefaultVersion": true,
      "LaunchTemplateData": [
        {
          "ImageId": "ami-04d5cc9b88example",
          "InstanceType": "t2.micro"
        }
      ]
    }
]
```

}

Exemplo de modelo de execução que especifica um parâmetro público de propriedade da AWS

1. 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 fornecido pela AWS para iniciar instâncias usando a AMI do Amazon Linux 2 mais recente na Região da AWS 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"
}
```

Esta é uma resposta de exemplo.

```
{
  "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
  }
}
```

2. Para confirmar se o modelo de execução recebe a ID de AMI correta, use o [describe-launch-template-versions](#) comando. O comando usa a opção `--resolve-alias` para resolver o parâmetro com o ID de AMI real.

```
aws ec2 describe-launch-template-versions --launch-template-name my-template-for-
auto-scaling \
--versions $Default --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",  
        }  
    }  
}
```

Limitações

- Atualmente, não é possível usar um modelo de execução que especifique um parâmetro do Systems Manager em vez de um ID de AMI em um grupo do Auto Scaling com uma [política de instâncias mistas \(p. 67\)](#).
- Os parâmetros do Systems Manager Parameter Store que você criar devem existir na mesma conta em que as instâncias são executadas.
- O Amazon EC2 Auto Scaling é compatível apenas com a especificação de IDs de AMI como parâmetros.
- Embora seu grupo do Auto Scaling tenha um modelo de execução que refcrcie um parâmetro em vez de um ID de AMI, não é possível iniciar uma atualização de instância que especifique uma configuração desejada, ou use a opção de ignorar correspondência para substituir as instâncias no grupo.
- 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 [Preço do AWS Systems Manager](#).

Configurações de execução

Important

As configurações de inicialização não adicionam mais suporte para novos tipos de instância do Amazon EC2 lançados após 31 de dezembro de 2022. Além disso, todas as novas contas criadas em ou após 1º de junho de 2023 não terão a opção de criar novas configurações de inicialização por meio do console. No entanto, API, CLI e CloudFormation acesso estarão disponíveis para novas contas criadas entre 1º de junho de 2023 e 31 de dezembro de 2023 para oferecer suporte a clientes com casos de uso de automação. Novas contas criadas em ou após 1º de janeiro de 2024 não poderão criar novas configurações de execução usando o console, API, CLI e CloudFormation. Para obter informações sobre como migrar seus grupos do Auto Scaling para lançar modelos, consulte. [Migre para lançar modelos \(p. 32\)](#)

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, você deverá criar uma configuração de execução e, em seguida, atualizar seu grupo do Auto Scaling com ela.

Índice

- [Criar uma configuração de execução \(p. 49\)](#)
- [Criar uma configuração de execução usando uma instância do EC2 \(p. 53\)](#)
- [Alterar a configuração de execução de um grupo do Auto Scaling \(p. 57\)](#)
- [Configurar a locação de instância com uma configuração de execução \(p. 58\)](#)

Criar uma configuração de execução

Important

As configurações de inicialização não adicionam mais suporte para novos tipos de instância do Amazon EC2 lançados após 31 de dezembro de 2022. Além disso, todas as novas contas criadas em ou após 1º de junho de 2023 não terão a opção de criar novas configurações de inicialização por meio do console. No entanto, API, CLI e CloudFormation acesso estarão disponíveis para novas contas criadas entre 1º de junho de 2023 e 31 de dezembro de 2023 para oferecer suporte a clientes com casos de uso de automação. Novas contas criadas em ou após 1º de janeiro de 2024 não poderão criar novas configurações de execução usando o console, API, CLI e CloudFormation. Para obter informações sobre como migrar seus grupos do Auto Scaling para lançar modelos, consulte. [Migre para lançar modelos \(p. 32\)](#)

Ao criar uma configuração de execução, você deve especificar informações sobre as instâncias do EC2 a serem executadas. Inclua o ID da imagem de máquina da Amazon (AMI), o tipo de instância, um par de chaves, grupos de segurança e um mapeamento de dispositivos de blocos. Como alternativa, você pode criar uma configuração de execução usando atributos de uma instância do EC2 em ativação. Para obter mais informações, consulte [Criar uma configuração de execução usando uma instância do EC2 \(p. 53\)](#).

Depois de criar uma configuração de execução, você pode criar um grupo do Auto Scaling. Para obter mais informações, consulte [Criar um grupo do Auto Scaling usando uma configuração de execução \(p. 100\)](#).

Um grupo do Auto Scaling é associado a uma configuração de execução de cada vez, e você não pode modificar uma configuração de execução depois de criá-la. Portanto, se você quiser alterar a configuração de execução para um grupo do Auto Scaling existente, deverá atualizá-lo com a nova configuração de execução. Para obter mais informações, consulte [Alterar a configuração de execução de um grupo do Auto Scaling \(p. 57\)](#).

Índice

- [Criar uma configuração de execução \(console\) \(p. 50\)](#)
- [Criar uma configuração de execução \(AWS CLI\) \(p. 51\)](#)
- [Configurar as opções de metadados da instância \(p. 51\)](#)

Criar uma configuração de execução (console)

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 da.
3. No painel de navegação esquerdo, em Grupos do Auto Scaling, escolha Grupos do Auto Scaling.
4. Escolha Configurações do execução na parte superior da página. Quando a confirmação for solicitada, escolha Exibir configurações de inicialização para confirmar que você deseja visualizar a página de configurações de inicialização.
5. Selecione [Create launch configuration \(Criar uma configuração de execução\)](#), e insira um nome para sua configuração de execução.
6. Em Amazon machine image (AMI) (Imagen 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 \(p. 36\)](#).

- b. (Opcional) Em IAM instance profile (Perfil de instância do IAM) selecione uma função a ser associada às instâncias. Para obter mais informações, consulte [Funções do IAM para aplicações que são executadas em instâncias do Amazon EC2 \(p. 448\)](#).
 - c. (Opcional) Em Monitoring (Monitoramento), escolha se as instâncias devem publicar dados métricos em intervalos de 1 minuto na Amazon CloudWatch habilitando o monitoramento detalhado. Aplicam-se cobranças adicionais. Para obter mais informações, consulte [Configurar monitoramento para instâncias do Auto Scaling \(p. 337\)](#).
 - 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.

Criar uma configuração de execução (AWS CLI)

Para criar uma configuração de execução usando a linha de comando

Você pode usar um dos comandos a seguir:

- [create-launch-configuration](#) (AWS CLI)
- [Novo-AS LaunchConfiguration \(\)](#) 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 oferece a opção de usar configurações de execução para configurar

as instâncias do Amazon EC2 em seus grupos do Auto Scaling para exigir o Instance Metadata Service Version 2 (IMDSv2), que é um método orientado a sessão para solicitar metadados de instância. Para obter detalhes sobre as vantagens do IMDSv2, consulte este artigo no blog da AWS sobre [melhorias na adição de defesa profunda ao serviço de metadados da instância 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 a AWS CLI ou um dos SDKs para configurar o IMDS, você deve usar a versão mais recente da AWS CLI ou o SDK para exigir o uso do IMDSv2.

Você pode configurar sua configuração de execução para:

- Exigir o uso do IMDSv2 ao solicitar metadados de 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 para instâncias do Linux.

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 \(p. 57\)](#).

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 da.
3. No painel de navegação esquerdo, em Grupos do Auto Scaling, escolha Grupos do Auto Scaling.
4. Escolha Configurações do execução na parte superior da página. Quando a confirmação for solicitada, escolha Exibir configurações de inicialização para confirmar que você deseja visualizar a página de 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. Só é possível especificar a condição para exigir IMDSv2 quando o endpoint HTTP estiver habilitado.
 - b. Em Metadata version (Versão dos metadados), você pode escolher 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 IMDSv1 e IMDSv2.
 - 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 do IMDSv2 em uma configuração de execução usando a AWS CLI

Use o [create-launch-configuration](#) comando a seguir com `--metadata-options set toHttpTokens=required`. Quando você especifica um valor para `HttpTokens`, você também deve definir `HttpEndpoint` como ativado. Como o cabeçalho de token seguro é definido como obrigatório para solicitações de recuperação de metadados, ele opta por exigir o uso do IMDSv2 na instância ao solicitar metadados de 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 seguinte [create-launch-configuration](#) comando para desativar o acesso aos metadados da instância. Você pode habilitar de novo o acesso posteriormente usando o [modify-instance-metadata-options](#) comando.

```
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

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 seus grupos do Auto Scaling para lançar modelos, consulte [Migre para lançar modelos \(p. 32\)](#).

O Amazon EC2 Auto Scaling fornece uma opção para criar uma configuração de execução usando os atributos de uma instância do EC2 em execução.

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.

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.

Você também pode [criar um grupo do Auto Scaling diretamente a partir de uma instância do EC2 \(p. 102\)](#). Quando você usa esse recurso, o Amazon EC2 Auto Scaling também cria automaticamente uma configuração de execução para você.

Os exemplos a seguir mostram como criar uma configuração de execução a partir de uma instância do EC2.

Exemplos

- [Criar uma configuração de execução usando uma instância do EC2 \(p. 54\)](#)
- [Criar uma configuração de execução a partir de uma instância e substituir os dispositivos de blocos \(AWS CLI\) \(p. 55\)](#)
- [Criar uma configuração de execução e substituir o tipo de instância \(AWS CLI\) \(p. 56\)](#)

Criar uma configuração de execução usando uma instância do EC2

Para criar uma configuração de execução usando os atributos de uma instância do EC2 existente, especifique o ID da instância.

Important

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

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

Você pode usar o console para criar uma configuração de execução e um grupo do Auto Scaling de uma instância do EC2 em execução e adicionar a instância ao novo grupo do Auto Scaling. Para obter mais informações, consulte [Anexar instâncias do EC2 a seu grupo do Auto Scaling \(p. 171\)](#).

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 seguinte [describe-launch-configurations](#) comando para descrever a configuração de execução e verificar se os atributos correspondem aos da instância.

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

Esta é uma resposta de exemplo.

```
{  
    "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.

Important

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

Criar uma configuração de execução e substituir os dispositivos de blocos

Use o seguinte [create-launch-configuration](#) comando para criar uma configuração de execução usando uma instância do EC2, mas com um mapeamento de dispositivo de bloco personalizado.

```
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 seguinte [describe-launch-configurations](#) comando para descrever a configuração de execução e verificar se ela usa seu mapeamento de dispositivo de bloco personalizado.

```
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.

Important

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

Criar uma configuração de execução e substituir o tipo de instância

Use o seguinte [create-launch-configuration](#) comando para criar uma configuração de execução usando uma instância do EC2, mas com um tipo de instância diferente ((at2.micro), t2.medium

```
aws autoscaling create-launch-configuration --launch-configuration-name my-lc-from-  
instance-changetype \
```

```
--instance-id i-a8e09d9c --instance-type t2.medium
```

Use o seguinte [describe-launch-configurations](#) comando 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 seus grupos do Auto Scaling para lançar modelos, consulte [Migre para lançar modelos \(p. 32\)](#).

Um grupo do Auto Scaling é associado a uma configuração de execução de cada vez, e você não pode modificar uma configuração de execução depois de criá-la. Para alterar a configuração de execução para um grupo do Auto Scaling, use uma configuração de execução existente como base para uma nova configuração de execução. Em seguida, atualize o grupo do Auto Scaling para usar a nova configuração de execução.

Depois que você alterar a configuração de execução para um grupo do Auto Scaling, todas as novas instâncias serão ativadas usando as novas opções de configuração, mas as instâncias existentes não serão afetadas. 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 encerramento \(p. 292\)](#).

Note

Você também pode substituir todas as instâncias no grupo do Auto Scaling para executar novas instâncias que usem a nova configuração de execução. Para obter mais informações, consulte [Substituir instâncias do Auto Scaling \(p. 108\)](#).

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. Na barra de navegação superior, selecione sua AWS região da.
3. No painel de navegação esquerdo, em Grupos do Auto Scaling, escolha Grupos do Auto Scaling.
4. Escolha Configurações do execução na parte superior da página. Quando a confirmação for solicitada, escolha Exibir configurações de inicialização para confirmar que você deseja visualizar a página de 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.
7. No painel de navegação esquerdo, em Grupos do Auto Scaling, escolha Grupos do Auto Scaling.
8. Marque a caixa de seleção ao lado do grupo do Auto Scaling.

Um painel dividido é aberto na parte inferior da página.

9. Na guia Details (Detalhes), escolha Launch configuration (Configuração de execução), Edit (Editar).
10. Em Launch Configuration (Configuração de execução), selecione a nova configuração de execução.
11. Quando terminar, escolha Update (Atualizar).

Para substituir a configuração de execução de um grupo do Auto Scaling (AWS CLI)

1. Descreva a configuração de execução atual usando o comando [describe-launch-configurations](#).
2. Crie uma configuração de execução usando o comando [create-launch-configuration](#).
3. Atualize a configuração de execução para o grupo do Auto Scaling usando o [update-auto-scaling-group](#) comando com o --launch-configuration-names parâmetro.

Para substituir a configuração de execução para um grupo do Auto Scaling (Tools for WindowsPowerShell)

1. Descreva a configuração de inicialização atual usando o LaunchConfiguration comando [Get-AS](#).
2. Crie uma nova configuração de inicialização usando o LaunchConfiguration comando [New-AS](#).
3. Atualize a configuração de execução para o grupo do Auto Scaling usando o AutoScalingGroup comando [Update-AS](#) com o parâmetro. -LaunchConfigurationName

Configurar a locação de instância com uma configuração de execução

A locação define como as instâncias do EC2 são distribuídas pelo hardware físico e afeta a definição de preço. Há três opções de locação disponíveis:

- Compartilhada (default): várias Contas da AWS podem compartilhar o mesmo hardware físico.

- Instância dedicada (**dedicated**): sua instância é executada em hardware de ocupante único.
- Host dedicado (**host**): sua instância é executada em um servidor físico com a capacidade de instância do EC2 totalmente dedicada ao uso, um servidor isolado com configurações que você pode controlar.

Este tópico descreve como iniciar instâncias dedicadas em seu grupo do Auto Scaling especificando configurações em uma configuração de execução. Para obter informações sobre preços e saber mais sobre instâncias dedicadas, consulte a página do produto [Instâncias dedicadas do Amazon EC2](#) e [Instâncias dedicadas](#) no Manual do usuário do Amazon EC2 para instâncias do Linux.

É possível configurar a locação para instâncias do EC2 usando uma configuração ou um modelo de execução. No entanto, o valor de locação host não pode ser usado com uma configuração de execução. Use somente os valores de locação **dedicated** ou **default**.

Important

Para usar um valor de locação host, é necessário usar um modelo de execução. Para obter mais informações, consulte [Criar um modelo de execução para um grupo do Auto Scaling \(p. 23\)](#). Antes de iniciar hosts dedicados, recomendamos que você se familiarize com a ativação e o gerenciamento de hosts dedicados usando o [AWS License Manager](#). Para obter mais informações, consulte o [Manual do usuário do License Manager](#).

Por padrão, quando você cria uma VPC, seu atributo de locação é definido como **default**. Nessa VPC, você pode ativar instâncias com um valor de locação **dedicated** para que elas sejam ativadas como instâncias de locação única. Caso contrário, elas são executadas como instâncias de locação compartilhada, por padrão. Se você definir o atributo de locação de uma VPC como **dedicated**, todas as instâncias ativadas na VPC são ativadas como instâncias de locação única.

Quando você cria uma configuração de execução, o valor padrão para locação de localização da instância é **null** e a locação da instância é controlada pelo atributo de locação da VPC. Você pode especificar a locação de posicionamento da instância para sua configuração de execução **default** ou **dedicated** usando o comando [create-launch-configuration](#) CLI com a **--placement-tenancy** opção.

A tabela a seguir resume a locação de localização das instâncias do Auto Scaling ativadas em uma VPC.

Locação da configuração de execução	Locação da VPC = default	Locação da VPC = dedicated
não especificado	instâncias de locação compartilhada	Dedicated Instances
default	instâncias de locação compartilhada	Dedicated Instances
dedicated	Dedicated Instances	Dedicated Instances

Para criar uma configuração de execução que crie instâncias dedicadas (AWS CLI)

Use o seguinte [create-launch-configuration](#) comando para criar uma configuração de execução que defina a locação da configuração de execução como **dedicated**.

```
aws autoscaling create-launch-configuration --launch-configuration-name my-launch-config --placement-tenancy dedicated --image-id ...
```

Você pode usar o seguinte [describe-launch-configurations](#) comando para verificar a locação de posicionamento da instância da configuração de execução.

```
aws autoscaling describe-launch-configurations --launch-configuration-names my-launch-config
```

O seguinte é a saída de exemplo de uma configuração de execução que cria instâncias dedicadas. O parâmetro `PlacementTenancy` só faz parte da saída desse comando quando você define explicitamente a locação de localização da instância.

```
{  
    "LaunchConfigurations": [  
        {  
            "UserData": null,  
            "EbsOptimized": false,  
            "PlacementTenancy": "dedicated",  
            "LaunchConfigurationARN": "arn:",  
            "InstanceMonitoring": {  
                "Enabled": true  
            },  
            "ImageId": "ami-b5a7ea85",  
            "CreatedTime": "2020-03-08T23:39:49.011Z",  
            "BlockDeviceMappings": [],  
            "KeyName": null,  
            "SecurityGroups": [],  
            "LaunchConfigurationName": "my-launch-config",  
            "KernelId": null,  
            "RamdiskId": null,  
            "InstanceType": "m3.medium"  
        }  
    ]  
}
```

Grupos do Auto Scaling

Note

Se você é novo nos grupos de Auto Scaling, você pode aprender mais no tutorial do [Conceitos básicos do Amazon EC2 Auto Scaling \(p. 15\)](#). Você começa criando um modelo de execução ou uma configuração de execução e, em seguida, use-o para criar um grupo do Auto Scaling no qual todas as instâncias têm os mesmos atributos de instância. Por exemplo, é possível definir os seguintes atributos de instância especificando-os como parte do modelo ou da configuração de execução: Imagem de máquina da Amazon (AMI), tipo de instância, dispositivos de armazenamento em bloco, par de chaves do SSH e grupos de segurança que controlam o tráfego de entrada e saída de uma instância.

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 do Auto Scaling \(p. 319\)](#).

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 [Escalar o tamanho do grupo do Auto Scaling \(p. 165\)](#).

Um grupo do Auto Scaling pode 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 modelo de execução.

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.

Índice

- [Crie grupos de Auto Scaling usando modelos de lançamento \(p. 62\)](#)
- [Crie grupos de Auto Scaling usando configurações de lançamento \(p. 99\)](#)
- [Atualizar um grupo de Auto Scaling \(p. 106\)](#)
- [Substituir instâncias do Auto Scaling \(p. 108\)](#)
- [Etiquetar grupos e instâncias do Auto Scaling \(p. 138\)](#)
- [Excluir infraestrutura do Auto Scaling \(p. 146\)](#)
- [Exemplos de criação e gerenciamento de grupos de Auto Scaling com o AWSSDKs \(p. 149\)](#)

Crie grupos de Auto Scaling usando modelos de lançamento

Se você criou um modelo de execução, pode criar um grupo de Auto Scaling que usa um modelo de execução como modelo de configuração para suas instâncias do EC2. O modelo de execução especifica informações como ID da AMI, tipo de instância, par de chaves, grupos de segurança e 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 \(p. 23\)](#).

Você deve ter permissões suficientes para criar um grupo de Auto Scaling. Você também deve ter permissões suficientes para criar a função vinculada ao serviço que o Amazon EC2 Auto Scaling usa para realizar ações em seu nome, caso ela ainda não exista. Para ver exemplos de políticas do IAM que um administrador pode usar como referência para conceder permissões a você, consulte [Exemplos de políticas baseadas em identidade \(p. 437\)](#) e [Suporte a modelo de execução \(p. 443\)](#).

Índice

- [Criar um grupo do Auto Scaling usando um modelo de execução \(p. 62\)](#)
- [Criar um grupo do Auto Scaling usando o assistente de execução do Amazon EC2 \(p. 64\)](#)
- [Grupos do Auto Scaling com vários tipos de instâncias e opções de compra \(p. 67\)](#)
- [Criar um grupo do Auto Scaling usando seleção de tipo de instância baseada em atributos \(p. 92\)](#)

Criar um grupo do Auto Scaling usando um modelo de execução

Ao criar um grupo do Auto Scaling, você deverá 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

- Você deve ter criado um modelo de execução. Para obter mais informações, consulte [Criar um modelo de execução para um grupo do Auto Scaling \(p. 23\)](#).

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 grupos do Auto Scaling 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 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 Choose instance launch options (Escolher as opções de execução da instância) em Network (Rede), para VPC, selecione uma VPC. O grupo do Auto Scaling deve ser criado na mesma VPC do grupo de segurança especificado no modelo de execução.
6. 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 \(p. 417\)](#).
7. 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 \(p. 67\)](#).

8. 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).
9. (Opcional) Na página Configure advanced options (Configurar opções avançadas), configure as seguintes opções e escolha Next (Próximo):
 - a. Para registrar suas instâncias do Amazon EC2 com um balanceador de carga, escolha um load balancer existente ou crie um novo. Para obter mais informações, consulte [Usar o Elastic Load Balancing para distribuir tráfego entre as instâncias no grupo do Auto Scaling \(p. 369\)](#). Para criar um novo balanceador de carga, siga o procedimento em [Configurar um Application Load Balancer ou Network Load Balancer pelo console do Amazon EC2 Auto Scaling \(p. 374\)](#).
 - b. (Opcional) Para verificações de integridade, tipos adicionais de verificação de saúde, selecione Ativar verificações de integridade do Elastic Load Balancing.
 - c. (Opcional) Para o período de carência da verificação de saúde, 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 \(p. 325\)](#).
 - d. Em Configurações adicionais, Monitoramento, escolha se deseja ativar a coleta de métricas CloudWatch do 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 [MonitorCloudWatchmétricas para seus grupos e instâncias do Auto Scaling \(p. 328\)](#).
 - e. 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 da instância aprimora 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 \(p. 200\)](#).

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. Em Desired capacity (Capacidade desejada), insira o número inicial de instâncias a serem executadas. Quando esse número é alterado para um valor fora dos limites de capacidade mínima ou máxima, é necessário atualizar os valores de Minimum capacity (Capacidade mínima) ou Maximum capacity (Capacidade máxima). Para obter mais informações, consulte [Definir limites de capacidade no grupo do Auto Scaling \(p. 166\)](#).
 - b. Para escalar automaticamente o tamanho do grupo do Auto Scaling, escolha Target tracking scaling policy (Política de escalabilidade com monitoramento do objetivo) e siga as instruções. Para obter mais informações, consulte [Políticas de escalabilidade com monitoramento do objetivo para o Amazon EC2 Auto Scaling \(p. 180\)](#).
 - c. 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 [Usar proteção de redução na escala na horizontal de instâncias \(p. 302\)](#).
11. (Opcional) Para receber notificações, em Add notification (Adicionar notificação), configure a notificação e, depois, escolha Next (Próximo). Para obter mais informações, consulte [Receber notificações do Amazon SNS quando o grupo do Auto Scaling escala \(p. 341\)](#).
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 \(p. 138\)](#).
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-AS AutoScalingGroup](#) () 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 executar 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, você deve 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 Região da AWS em que você planeja criar o grupo do Auto Scaling. Para mais informações, consulte [Create an AMI](#) (Criar uma AMI) no Guia do usuário do Amazon EC2 para instâncias do Linux.

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 sua AMI personalizada. Como alternativa, para configurar o modelo de inicialização do zero ou para obter mais descrição dos parâmetros que você pode configurar para seu modelo de inicialização, consulte [Criar seu modelo de execução \(console\) \(p. 23\)](#).

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 Região da AWS atual é exibida. Selecione uma região na qual iniciará o grupo do Auto Scaling.
3. No painel de navegação, escolha Instances (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 Application and OS Images (Amazon Machine Image) (Imagens de aplicações e sistemas operacionais [imagem de máquina da Amazon]), escolha Browse more AMIs (Procurar mais AMIs) para navegar pelo catálogo completo de AMIs.
 - c. Na página My AMIs (Minhas AMIs), localize a AMI criada anteriormente e escolha Select (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 hyperlink 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 lançamento.

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 \(p. 62\)](#).

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 Choose instance launch options (Escolher as opções de execução da instância) em Network (Rede), para VPC, selecione 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.

5. Para Availability Zones and subnets (Zonas de disponibilidade e sub-redes), selecione uma ou mais sub-redes na VPC especificada.
6. Selecione Próximo duas vezes para ir até a página Configure group size and scaling policies (Definir tamanho do grupo e políticas de escalabilidade).
7. Sob Group size (Tamanho do grupo), defina a Desired capacity (Capacidade desejada) (número inicial de instâncias a serem executadas imediatamente após a criação do grupo do Auto Scaling), Minimum capacity (Capacidade mínima) (número mínimo de instâncias) e Maximum capacity (Capacidade máxima) (número máximo de instâncias).
8. Escolha Skip to review (Ir para revisão).
9. 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 \(p. 461\)](#)
- [Solucionar problemas do Amazon EC2 Auto Scaling: problemas de AMI \(p. 467\)](#)
- [Solucionar problemas com as verificações de integridade do Amazon EC2 Auto Scaling \(p. 473\)](#)

Agora você pode anexar um平衡ador de carga na mesma região do grupo do Auto Scaling, se desejar. Para obter mais informações, consulte [Usar o Elastic Load Balancing para distribuir tráfego entre as instâncias no grupo do Auto Scaling \(p. 369\)](#).

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, você pode usar instâncias reservadas ou um Savings Plan para receber taxas de desconto da definição de preço normal de instância sob demanda. Todos esses fatores combinados ajudam você a otimizar sua economia de custos para instâncias do EC2 e a obter a escala e o desempenho desejados para seu aplicativo.

As etapas a seguir descrevem como criar o grupo do Auto Scaling:

- Especifique o modelo de execução para iniciar as instâncias.
- Escolha a VPC e as sub-redes para iniciar seu grupo do Auto Scaling.
- Escolha a opção de substituir os requisitos existentes do tipo de instância do modelo de inicialização por novos requisitos.
- Escolha e priorize manualmente seus tipos de instância. Por exemplo, você pode optar por priorizar os tipos de instância que podem se beneficiar do preço de desconto do Savings Plan ou da instância reservada para instâncias sob demanda.
- Especifique as porcentagens de instâncias sob demanda e de instâncias spot a serem iniciadas.
- 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.
- Especifique o tamanho do seu grupo, incluindo a capacidade desejada, capacidade mínima e capacidade máxima.

Aprimore a disponibilidade ao implantar sua aplicação em vários tipos de instâncias em execução em várias zonas de disponibilidade. Embora você possa usar um tipo de instância, é uma prática recomendada usar vários tipos de instância. Dessa forma, o Amazon EC2 Auto Scaling pode executar 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.

Índice

- [Estratégias de alocação \(p. 68\)](#)

- [Melhores práticas para instâncias spot \(p. 70\)](#)
- [Controlar a proporção de instâncias sob demanda \(p. 70\)](#)
- [Pré-requisitos \(p. 72\)](#)
- [Criar um grupo do Auto Scaling com instâncias spot e sob demanda \(console\) \(p. 72\)](#)
- [Criar um grupo do Auto Scaling com instâncias spot e sob demanda \(AWS CLI\) \(p. 74\)](#)
- [Verifique a configuração do grupo do Auto Scaling e as instâncias executadas \(AWS CLI\). \(p. 82\)](#)
- [Configurar substituições \(p. 82\)](#)

Estratégias de alocação

As seguintes estratégias de alocação determinam como o grupo do Auto Scaling cumpre sua capacidade sob demanda e spot a partir dos possíveis tipos de instância.

Primeiro, o Amazon EC2 Auto Scaling tenta equilibrar suas instâncias uniformemente em suas zonas de disponibilidade especificadas. Em seguida, ele inicia os tipos de instância de acordo com a estratégia de alocação especificada.

Instâncias spot

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

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, mas a capacidade oscila 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 respeita as prioridades de tipo de instância com base no melhor esforço, mas primeiro otimiza a capacidade. Essa é uma boa opção para workloads em que a possibilidade de interrupção deve ser minimizada, mas em que também a preferência por determinados tipos de instância for 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

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.

Como essa estratégia considera apenas o preço da instância e não a disponibilidade de capacidade, ela pode levar a altas taxas de interrupção.

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.

Para começar, recomendamos escolher a estratégia de alocação **price-capacity-optimized** e especificar um conjunto de tipos de instância apropriados para seu aplicativo. Além disso, é possível definir um intervalo de zonas de disponibilidade para que o Amazon EC2 Auto Scaling escolha ao iniciar instâncias.

Para obter mais informações sobre a **price-capacity-optimized** estratégia e os casos de uso em que ela é útil, consulte [Introdução à estratégia de price-capacity-optimized alocação para instâncias spot do EC2](#) no AWS blog.

Se preferir, você poderá especificar um preço máximo para as instâncias spot. Se você não especificar um preço máximo, o padrão será o preço sob demanda. No entanto, 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](#).

On-Demand Instances

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 de Auto Scaling preenche a capacidade sob demanda começando com `c5.large`, seguido por `c4.large` e por último `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 Plan, você pode 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 do Savings Plan. Para obter mais informações sobre Savings Plans, consulte o [Savings Plans User Guide](#) (Guia do usuário de Savings Plans).

Melhores práticas para instâncias spot

Antes de criar o grupo do Auto Scaling para solicitar instâncias spot, consulte [Melhores práticas para o EC2 Spot](#) no Manual do usuário do Amazon EC2 para instâncias do Linux. Use estas práticas recomendadas para planejar sua solicitação para que você possa provisionar o tipo de instância que deseja pelo preço mais baixo possível. Também recomendamos fazer o seguinte:

- 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. O preço máximo padrão é o preço sob demanda, e você paga apenas o preço spot pelas instâncias Spot iniciadas. Quando o preço Spot for inferior ao preço máximo, o cumprimento do seu pedido depende da disponibilidade. Para obter mais informações, consulte [Preços e economia](#) no Manual do usuário do Amazon EC2 para instâncias do Linux.
- Uma boa regra geral é ser flexível para pelo menos 10 tipos de instância para cada workload. Além disso, verifique se todas as zonas de disponibilidade estão configuradas para uso na VPC e selecionadas para a workload. Como a capacidade oscila independentemente para cada tipo de instância na zona de disponibilidade, é frequentemente possível obter maior capacidade computacional quando você tem flexibilidade de tipo 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, pois há menos demanda de clientes sob demanda.
- Habilite o rebalanceamento de capacidade. Quando você faz isso, o Amazon EC2 Auto Scaling tenta iniciar uma instância spot sempre que o serviço spot do Amazon EC2 notifica que uma instância spot está em risco elevado de interrupção. Depois de iniciar uma nova instância, ele então termina uma instância antiga. Para obter mais informações, consulte [Usar o rebalanceamento de capacidade para lidar com interrupções de spot do Amazon EC2 \(p. 346\)](#).
- Recomendamos que você use a estratégia de alocação price-capacity-optimized em vez da estratégia de alocação lowest-price para evitar a solicitação de instâncias spot de pools com grande chance de interrupção.
- Se você escolher a estratégia de alocação lowest-price e executar um serviço da Web, especifique um número alto de pools Spot, como N=10. Isso reduz o impacto das interrupções da instância Spot se um pool em uma das zonas de disponibilidade ficar temporariamente indisponível. Se você executar o processamento em lote ou outros aplicativos não essenciais, poderá especificar um número menor de pools Spot, como N=2. Isso provisona Instâncias Spot apenas dos pools Spot de preço mais baixo disponíveis por Zona de Disponibilidade.

O Amazon EC2 Auto Scaling tenta extrair instâncias spot do número N de pools que você especifica com base no melhor esforço. Se um pool ficar sem capacidade spot antes de atender à capacidade desejada, o Amazon EC2 Auto Scaling continuará 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.

Se você pretende especificar um preço máximo, use a AWS CLI ou um SDK para criar o grupo do Auto Scaling, mas seja cuidadoso. Se o seu preço máximo for inferior ao preço spot dos tipos de instâncias selecionados por você, as Instâncias spot não serão executadas.

Controlar a proporção de instâncias sob demanda

Você tem controle total sobre a proporção de instâncias do grupo do Auto Scaling que são iniciadas como instâncias sob demanda. Para garantir que você sempre tenha capacidade de instância, você pode

designar uma porcentagem do grupo para iniciar como instâncias sob demanda. Opcionalmente, você também pode designar um número base de instâncias sob demanda para começar. Se você designar uma capacidade básica de instâncias sob demanda, 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 expandido. 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. Você pode especificar qualquer número de 0 a 100 para a porcentagem sob demanda.

O Amazon EC2 Auto Scaling converte o percentual para o número equivalente de instâncias. Se o resultado criar um número fracionário, o Amazon EC2 Auto Scaling 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 de tamanho.

Exemplo: comportamento de escalabilidade

Distribuição de instâncias	Número total de instâncias em execução nas opções de compra			
	10	20	30	40
Exemplo 1				
On-Demand base: 10	10	10	10	10
On-Demand percentage above base: 50%	0	5	10	15
Spot percentage: 50%	0	5	10	15
Exemplo 2				
On-Demand base: 0	0	0	0	0
On-Demand percentage above base: 0%	0	0	0	0
Spot percentage: 100%	10	20	30	40
Exemplo 3				
On-Demand base: 0	0	0	0	0
On-Demand percentage above base: 60%	6	12	18	24
Spot percentage: 40%	4	8	12	16
Exemplo 4				
On-Demand base: 0	0	0	0	0

Distribuição de instâncias	Número total de instâncias em execução nas opções de compra			
On-Demand percentage above base: 100%	10	20	30	40
Spot percentage: 0%	0	0	0	0
Exemplo 5				
On-Demand base: 12	10	12	12	12
On-Demand percentage above base: 0%	0	0	0	0
Spot percentage: 100%	0	8	18	28

Pré-requisitos

- Criar um modelo de execução. Para obter mais informações, consulte [Criar um modelo de execução para um grupo do Auto Scaling \(p. 23\)](#).
- Verifique se o modelo de execução já não solicita instâncias spot.

Criar um grupo do Auto Scaling com instâncias spot e sob demanda (console)

Use o procedimento a seguir para escolher quais tipos de instância individual seu grupo pode iniciar. Se você preferir usar atributos de instância como critérios para selecionar tipos de instância, configure os requisitos de tipo de instância especificando o número de vCPUs e a memória de que você precisa. Para obter mais informações, consulte [Criar um grupo do Auto Scaling usando seleção de tipo de instância baseada em atributos \(p. 92\)](#). Você também pode especificar outros atributos, como tipo de armazenamento, interfaces de rede, fabricante da CPU e tipo de acelerador.

Para criar um grupo do Auto Scaling com instâncias spot e sob demanda

1. Abra o console do Amazon EC2 em <https://console.aws.amazon.com/ec2/> e escolha grupos do Auto Scaling 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. Para escolher uma VPC e as sub-redes nas quais você deseja iniciar as instâncias, faça o seguinte:
 - a. Na página Choose instance launch options (Escolher as opções de execução da instância) em Network (Rede), para VPC, selecione uma VPC. O grupo do Auto Scaling deve ser criado na mesma VPC do grupo de segurança especificado no modelo de execução.
 - b. 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 \(p. 417\)](#).
7. Para escolher manualmente quais tipos de instância individual seu grupo pode iniciar, 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.

Opcionalmente, 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 ponderação de instâncias para o Amazon EC2 Auto Scaling \(p. 86\)](#).

Note

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. Quando você especifica uma porcentagem para instâncias spot, pode marcar a caixa de seleção ao lado de Include On-Demand base capacity (Incluir capacidade básica sob demanda) e depois especificar 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. O preço mais baixo está oculto por padrão e só aparece quando você escolhe Mostrar todas as estratégias.

Note

Se você escolheu Preço mais baixo, insira o número de grupos com preços mais baixos para diversificar para os grupos com preços mais baixos. Se você escolher 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 Capacity rebalance (Rebalanceamento de capacidade), escolha se você deseja habilitar ou desabilitar o rebalanceamento de capacidade.

Se você escolher uma porcentagem para instâncias Spot, poderá usar o Rebalanceamento de capacidade para responder automaticamente quando suas instâncias Spot se aproximarem do encerramento de uma interrupção Spot. Para obter mais informações, consulte [Usar o rebalanceamento de capacidade para lidar com interrupções de spot do Amazon EC2 \(p. 346\)](#).

- h. Selecione Next (Avançar) duas vezes para ir para a página Configure group size and scaling policies (Configurar tamanho do grupo e políticas de escalação).
8. Na etapa Configure group size and scaling policies (Configurar o tamanho do grupo e as políticas de escalação), faça o seguinte:
 - a. Insira o tamanho inicial do grupo do Auto Scaling para a Desired capacity (Capacidade desejada) e atualize os limites de Minimum capacity (Capacidade mínima) e Maximum capacity (Capacidade máxima) conforme necessário. Para obter mais informações, consulte [Definir limites de capacidade no grupo do Auto Scaling \(p. 166\)](#).

Note

Por padrão, a capacidade desejada e os limites de capacidade mínimo e máximo são expressos como o número de instâncias. Se você atribuiu pesos aos seus tipos de instância, deve converter esses valores na mesma unidade de medida usada para atribuir pesos, como o número de vCPUs.

- b. (Opcional) Configure o grupo para dimensionar especificando uma política de dimensionamento de rastreamento de destino. Como alternativa, especifique essa política depois de criar o grupo. Para obter mais informações, consulte [Políticas de escalabilidade com monitoramento do objetivo para o Amazon EC2 Auto Scaling \(p. 180\)](#).
- c. (Opcional) Habilite a proteção de redução de instância, o que impede que seu grupo do Auto Scaling encerre instâncias durante a redução. Para obter mais informações, consulte [Usar proteção de redução na escala na horizontal de instâncias \(p. 302\)](#).
- d. Quando terminar, escolha Next (Próximo).
9. (Opcional) Para receber notificações quando a escala do grupo for aumentada ou reduzida, em Add notification (Adicionar notificação), configure a notificação e depois escolha Next (Avançar). Para obter mais informações, consulte [Receber notificações do Amazon SNS quando o grupo do Auto Scaling escala \(p. 341\)](#).
10. (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 \(p. 138\)](#).
11. Na página Review (Revisão), escolha Create Auto Scaling group (Criar grupo do Auto Scaling).

Criar um grupo do Auto Scaling com instâncias spot e sob demanda (AWS CLI)

As configurações de exemplo 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, será necessário 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 \(p. 75\)](#)
- [Exemplo 2: Iniciar instâncias spot usando a estratégia de alocação capacity-optimized-prioritized \(p. 77\)](#)
- [Exemplo 3: Iniciar instâncias spot usando a estratégia de alocação lowest-price diversificada em dois grupos \(p. 78\)](#)
- [Exemplo 4: Iniciar Instâncias spot usando a estratégia de alocação price-capacity-optimized \(p. 80\)](#)

Exemplo 1: Iniciar instâncias spot usando a estratégia de alocação capacity-optimized

O [create-auto-scaling-group](#) comando 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 executar 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
```

Veja a seguir um exemplo de arquivo config.json.

```
{  
    "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-group](#) comando 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 em vez de um arquivo JSON.

```
aws autoscaling create-auto-scaling-group --cli-input-yaml file://~/config.yaml
```

Veja a seguir um exemplo de arquivo config.yaml.

```
---  
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-group](#) comando 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 executar 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
```

Veja a seguir um exemplo de arquivo config.json.

```
{
    "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-group](#) comando 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 em vez de um arquivo JSON.

```
aws autoscaling create-auto-scaling-group --cli-input-yaml file://~/config.yaml
```

Veja a seguir um exemplo de arquivo config.yaml.

```
---
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-group](#) comando 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 executar 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
```

Veja a seguir um exemplo de arquivo `config.json`.

```
{
    "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-group](#) comando 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 em vez de um arquivo JSON.

```
aws autoscaling create-auto-scaling-group --cli-input-yaml file://~/config.yaml
```

Veja a seguir um exemplo de arquivo config.yaml.

```
---
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-group](#) comando 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 executar 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 executar 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
```

Veja a seguir um exemplo de arquivo config.json.

```
{
  "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-group](#) comando 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 em vez de um arquivo JSON.

```
aws autoscaling create-auto-scaling-group --cli-input-yaml file://~/config.yaml
```

Veja a seguir um exemplo de arquivo config.yaml.

```
---
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
```

Note

Para obter exemplos adicionais, consulte [Especificar um modelo de execução diferente para um tipo de instância \(p. 83\)](#), [Configurar ponderação de instâncias para o Amazon EC2 Auto Scaling \(p. 86\)](#) e [Usar o rebalanceamento de capacidade para lidar com interrupções de spot do Amazon EC2 \(p. 346\)](#).

Verifique a configuração do grupo do Auto Scaling e as instâncias executadas (AWS CLI).

Para verificar se seu grupo de Auto Scaling está configurado corretamente e se lançou instâncias, use o [describe-auto-scaling-groups](#) comando. Verifique se a política de instâncias mistas e a lista de sub-redes existem e estão configuradas corretamente. Se as instâncias tiverem sido iniciadas, você verá uma lista das instâncias e seus status. Para ver as atividades de escalabilidade resultantes do lançamento de instâncias, use o [describe-scaling-activities](#) comando. Você pode monitorar atividades de escalabilidade que estão em andamento e que foram concluídas recentemente. Para obter mais informações, consulte [Verificar uma ação de escalabilidade para um grupo do Auto Scaling \(p. 215\)](#).

Configurar substituições

As substituições são alterações feitas nas propriedades do seu modelo de execução. O Amazon EC2 Auto Scaling oferece suporte a substituições na propriedade tipo de instância. Dessa forma, você pode especificar vários tipos de instância. Você também pode definir vários modelos de execução para permitir que instâncias com diferentes arquiteturas de CPU (por exemplo, Arm e x86) sejam executadas no mesmo grupo do Auto Scaling. Para isso, você deve adicionar a estrutura `Overrides` à sua política de instâncias mistas.

A estrutura `Overrides` permite definir um conjunto de parâmetros que o Amazon EC2 Auto Scaling pode usar para iniciar instâncias, incluindo:

- `InstanceType`: o tipo de instância. 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.
- `LaunchTemplateSpecification`: (opcional) um modelo de execução para um tipo de instância individual. No momento, essa propriedade limita-se à AWS CLI ou a um SDK, e não está disponível no console.
- `WeightedCapacity`: (opcional) o número de unidades que uma instância provisionada deste tipo fornece para atender à capacidade desejada do grupo do Auto Scaling. Se você especificar um valor de peso para um tipo de instância, deverá especificar um valor de peso para todos eles.

Note

Como alternativa à especificação de uma lista de tipos de instância, é possível especificar um conjunto de atributos de instância a serem usados como critérios de seleção dos tipos de

instância usados pelo grupo do Auto Scaling. Isso é conhecido como seleção de tipo de instância baseada em atributos. Para mais informações, consulte [Using attribute-based instance type selection \(p. 92\)](#) (Usar a seleção de tipo de instância baseada em atributo).

Índice

- [Especificar um modelo de execução diferente para um tipo de instância \(p. 83\)](#)
- [Configurar ponderação de instâncias para o Amazon EC2 Auto Scaling \(p. 86\)](#)

Especificar um modelo de execução diferente para um tipo de instância

A estrutura `Overrides` permite definir um novo modelo de execução para tipos de instância individuais para um grupo do Auto Scaling novo ou existente. Por exemplo, se a arquitetura de um tipo de instância exigir uma AMI diferente do resto do grupo, você deverá especificar um modelo de execução com uma AMI compatível.

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 Graviton da AWS baseado na arquitetura Arm de 64 bits, enquanto as instâncias C5 e C5a são executadas em processadores Intel x86 de 64 bits. A AMI para instâncias C5 funciona em instâncias C5a e vice-versa, mas não em instâncias C6g. A propriedade `Overrides` permite incluir um modelo de execução diferente para instâncias C6g, enquanto ainda usa o mesmo modelo de execução para instâncias C5 e C5a.

Note

No momento, esse recurso estará disponível somente se você usar a AWS CLI ou um SDK, e não está disponível no console.

Índice

- [Adicionar ou alterar um modelo de execução para um tipo de instância \(AWS CLI\) \(p. 83\)](#)

Adicionar ou alterar um modelo de execução para um tipo de instância (AWS CLI)

O procedimento a seguir mostra como usar a AWS CLI para configurar um grupo do Auto Scaling para que um ou mais tipos de instância usem um modelo de execução diferente do resto do grupo.

Para criar e configurar um novo grupo do Auto Scaling

1. Crie um arquivo de configuração onde você especifica uma estrutura de política de instâncias mistas e inclua a estrutura `Overrides`.

Veja a seguir o conteúdo de um arquivo de configuração de exemplo formatado em JSON. 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.

```
{  
    "AutoScalingGroupName": "my-asg",  
    "MixedInstancesPolicy": {  
        "LaunchTemplate": {  
            "LaunchTemplateSpecification": {  
                "LaunchTemplateName": "my-launch-template-for-x86",  
                "Version": "$Latest"  
            },  
            "Overrides": [  
                {  
                    "InstanceType": "c6g.large",  
                    "Weight": 100  
                }  
            ]  
        }  
    }  
}
```

```
"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": [ ]  
}
```

2. Use o [create-auto-scaling-group](#) comando a seguir, referenciando o arquivo JSON como o único parâmetro para seu grupo de Auto Scaling.

```
aws autoscaling create-auto-scaling-group --cli-input-json file://~/config.json
```

Para alterar o modelo de execução de um tipo de instância em um grupo existente do Auto Scaling

- Use o [update-auto-scaling-group](#) comando a seguir para especificar um modelo de execução diferente para um tipo de instância passando a Overrides estrutura.

Quando essa alteração é feita, todas as novas instâncias que são iniciadas serão baseadas nas novas configurações, mas as instâncias existentes não serão afetadas. Para garantir que o grupo do Auto Scaling esteja usando as novas configurações, você pode substituir todas as instâncias do grupo iniciando uma atualização de instância ou usando o recurso de vida útil máxima da instância.

```
aws autoscaling update-auto-scaling-group --cli-input-json file://~/config.json
```

Veja a seguir um exemplo de arquivo config.json.

```
{  
    "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 verificar os modelos de execução de um grupo do Auto Scaling

- Use o [describe-auto-scaling-groups](#) comando a seguir para verificar e visualizar os modelos de execução atualmente especificados.

```
aws autoscaling describe-auto-scaling-groups --auto-scaling-group-name my-asg
```

Esta é uma resposta de exemplo.

```
{
    "AutoScalingGroups": [
        {
            "AutoScalingGroupName": "my-asg",
            "AutoScalingGroupARN": "arn",
            "MixedInstancesPolicy": {
                "LaunchTemplate": {
                    "LaunchTemplateSpecification": {
                        "LaunchTemplateId": "lt-0fb0e487336917fb2",
                        "LaunchTemplateName": "my-launch-template-for-x86",
                        "Version": "$Latest"
                    },
                    "Overrides": [
                        {
                            "InstanceType": "c6g.large",
                            "LaunchTemplateSpecification": {
                                "LaunchTemplateId": "lt-09d958b8fb2ba5bcc",
                                "LaunchTemplateName": "my-launch-template-for-arm",
                                "Version": "$Latest"
                            }
                        },
                        {
                            "InstanceType": "c5.large"
                        },
                        {
                            "InstanceType": "c5a.large"
                        }
                    ]
                },
                "InstancesDistribution": {
                    "OnDemandAllocationStrategy": "prioritized",
                    "OnDemandBaseCapacity": 1,
                    "OnDemandPercentageAboveBaseCapacity": 50,
                    "SpotAllocationStrategy": "capacity-optimized"
                }
            },
            "MinSize": 1,
            "MaxSize": 5,
            "DesiredCapacity": 3,
            "Instances": [
                {
                    "InstanceId": "i-07c63168522c0f620",
                    "InstanceType": "c5.large",
                    "AvailabilityZone": "us-west-2c",

```

```
"LifecycleState":"InService",
"HealthStatus":"Healthy",
"LaunchTemplate":{  
    "LaunchTemplateId":"lt-0fb0e487336917fb2",  
    "LaunchTemplateName":"my-launch-template-for-x86",  
    "Version":"1"  
},  
    "ProtectedFromScaleIn":false
},  
{
    "InstanceId":"i-0b7ff78be9896a2c2",
    "InstanceType":"c5.large",
    "AvailabilityZone":"us-west-2a",
    "LifecycleState":"InService",
    "HealthStatus":"Healthy",
    "LaunchTemplate":{  
        "LaunchTemplateId":"lt-0fb0e487336917fb2",  
        "LaunchTemplateName":"my-launch-template-for-x86",  
        "Version":"1"  
},  

```

Configurar ponderação de instâncias para o Amazon EC2 Auto Scaling

Ao configurar um grupo do Auto Scaling para iniciar vários tipos de instância, você tem a opção de definir o número de unidades de capacidade com que cada instância contribui para a capacidade desejada do grupo usando a ponderação de instâncias. Isso permite que você especifique o peso relativo de cada tipo de instância de modo que ele seja mapeado diretamente para a performance da aplicação. Você pode ponderar suas instâncias de acordo com as necessidades específicas do aplicativo, por exemplo, por núcleos (vCPUs) ou por memória (GiBs).

Por exemplo, digamos que você execute uma aplicação com uso intenso de computação que tenha melhor performance com pelo menos 8 vCPUs 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

Tipo de instância	vCPU	Memória (GiB)
c5.24xlarge	96	192

Por padrão, todos os tipos de instância são tratados como tendo o mesmo peso. 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.

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 executa as instâncias, seus pesos são considerados na capacidade desejada.

Índice

- [Preço por hora \(p. 87\)](#)
- [Considerações \(p. 88\)](#)
- [Adicionar ou modificar pesos para seu grupo do Auto Scaling \(p. 89\)](#)
- [Informações adicionais \(p. 92\)](#)

Preço por 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, Ohio) 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
c5.24xlarge	1,555 USD	1,555 USD	1,555 USD	4,08 USD

Com a ponderação de instâncias, você pode avaliar seus custos com base no que você usa por 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.

A maneira mais fácil de entender como o cálculo do preço por hora funciona com instâncias ponderadas é com um exemplo. Por exemplo, para facilitar o cálculo, digamos que você queira iniciar instâncias spot somente em us-east-1a. O preço por hora é capturado abaixo.

Exemplo: preço spot por unidade hora de exemplo

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

Considerações

Esta seção discute as principais considerações na implementação eficaz da ponderação de instâncias.

- Comece escolhendo alguns tipos de instância que reflitam os requisitos reais de performance da aplicação. Em seguida, decida quanto cada tipo de instância deve representar em relação à capacidade desejada do seu grupo do Auto Scaling especificando seus pesos. Os pesos se aplicam a instâncias atuais e futuras no grupo.
- Seja cauteloso ao escolher faixas muito grandes para seus pesos. 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. Se qualquer um dos tipos de instância tiver uma diferença de peso muito grande, isso poderá ter um efeito negativo na otimização contínua da relação custo/performance.
- O tamanho do grupo do Auto Scaling é medido em unidades de capacidade e não em instâncias. Por exemplo, se os seus pesos forem baseados em vCPUs, você deverá 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.

Com o peso de instâncias, os seguintes novos comportamentos são apresentados:

- A capacidade atual será a capacidade desejada ou acima dela. Como o Amazon EC2 Auto Scaling deseja provisionar instâncias até que a capacidade desejada seja totalmente atendida, um excedente pode ocorrer. 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 5 unidades restantes para atender a capacidade desejada, e o Amazon EC2 Auto Scaling provisionar uma c5.12xlarge, a capacidade desejada será excedida em 7 unidades.
- Quando o Amazon EC2 Auto Scaling provisiona instâncias para alcançar a capacidade desejada, a distribuição de instâncias entre zonas de disponibilidade e o respeito às estratégias de alocação para instâncias spot e sob demanda têm precedência sobre evitar excedentes.
- 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.

Observe o seguinte ao adicionar ou modificar pesos para grupos existentes:

- Ao adicionar pesos de instância a um grupo do Auto Scaling existente, você deve incluir quaisquer tipos de instância que já estejam em execução no grupo.
- Ao modificar pesos de instância existentes, o Amazon EC2 Auto Scaling iniciará ou terminará instâncias para alcançar sua capacidade desejada com base nos novos pesos.

- Se você remover um tipo de instância, todas as instâncias desse tipo em execução continuarão a ter seus últimos valores de peso atualizados, mesmo que o tipo de instância tenha sido removido.

Adicionar ou modificar pesos para seu grupo do Auto Scaling

É possível adicionar pesos a um grupo do Auto Scaling existente ou a um novo grupo do Auto Scaling ao criá-lo. Você também pode atualizar um grupo do Auto Scaling existente para definir novas opções de configuração (uso de spot/sob demanda, estratégia de alocação spot, tipos de instância). Se você alterar o número de instâncias spot ou sob demanda desejado, o Amazon EC2 Auto Scaling substituirá gradualmente as instâncias existentes para corresponder às novas opções de compra.

Antes de criar grupos do Auto Scaling usando a ponderação de instâncias, recomendamos que você se familiarize com a execução de grupos com vários tipos de instância. Para obter mais informações e exemplos adicionais, consulte [Grupos do Auto Scaling com vários tipos de instâncias e opções de compra \(p. 67\)](#).

Os exemplos a seguir mostram como usar a AWS CLI para adicionar pesos ao criar grupos do Auto Scaling e para adicionar ou modificar pesos de grupos do Auto Scaling existentes. Você pode configurar uma variedade de parâmetros em um arquivo JSON e, depois, referenciar o arquivo JSON como o único parâmetro para o seu grupo do Auto Scaling.

Para adicionar pesos a um grupo do Auto Scaling na criação

- Use o [create-auto-scaling-group](#) comando para criar um novo grupo de Auto Scaling. Por exemplo, o comando a seguir cria um novo grupo do Auto Scaling e adiciona peso de instância 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 de instância 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
```

Veja a seguir um exemplo de arquivo config.json.

```
{  
    "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 adicionar ou modificar pesos para um grupo do Auto Scaling existente

- Use o comando [update-auto-scaling-group](#) para adicionar ou modificar pesos. 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 de instância que correspondem à diferença de tamanho relativo (vCPUs) entre os tipos de instância (18, 24, 2, 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
```

Veja a seguir um exemplo de arquivo config.json.

```
{
    "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 de um grupo do Auto Scaling

- Use o comando [describe-auto-scaling-groups](#) a seguir para verificar os pesos.

```
aws autoscaling describe-auto-scaling-groups --auto-scaling-group-name my-asg
```

Esta é uma resposta de exemplo.

```
{  
    "AutoScalingGroups": [  
        {  
            "AutoScalingGroupName": "my-asg",  
            "AutoScalingGroupARN": "arn:",  
            "MixedInstancesPolicy": {  
                "LaunchTemplate": {  
                    "LaunchTemplateSpecification": {  
                        "LaunchTemplateId": "lt-0b97f1e282EXAMPLE",  
                        "LaunchTemplateName": "my-launch-template",  
                        "Version": "$Latest"  
                    },  
                    "Overrides": [  
                        {  
                            "InstanceType": "m4.16xlarge",  
                            "WeightedCapacity": "16"  
                        },  
                        {  
                            "InstanceType": "m5.24xlarge",  
                            "WeightedCapacity": "24"  
                        }  
                    ]  
                },  
                "InstancesDistribution": {  
                    "OnDemandAllocationStrategy": "prioritized",  
                    "OnDemandBaseCapacity": 0,  
                    "OnDemandPercentageAboveBaseCapacity": 0,  
                    "SpotAllocationStrategy": "capacity-optimized"  
                }  
            },  
            "MinSize": 160,  
            "MaxSize": 720,  
            "DesiredCapacity": 480,  
            "DefaultCooldown": 300,  
            "AvailabilityZones": [  
                "us-west-2a",  
                "us-west-2b",  
                "us-west-2c"  
            ],  
            "LoadBalancerNames": [],  
            "TargetGroupARNs": [],  
            "HealthCheckType": "EC2",  
            "HealthCheckGracePeriod": 0,  
            "Instances": [  
                {  
                    "InstanceId": "i-027327f0ace86f499",  
                    "InstanceType": "m5.24xlarge",  
                    "AvailabilityZone": "us-west-2a",  
                    "LifecycleState": "InService",  
                    "HealthStatus": "Healthy",  
                    "LaunchTemplate": {  
                        "LaunchTemplateId": "lt-0b97f1e282EXAMPLE",  
                        "LaunchTemplateVersion": "$Latest"  
                    }  
                }  
            ]  
        }  
    ]  
}
```

```
        "LaunchTemplateName": "my-launch-template",
        "Version": "7"
    },
    "ProtectedFromScaleIn": false,
    "WeightedCapacity": "24"
},
{
    "InstanceId": "i-0ec0d761cc134878d",
    "InstanceType": "m4.16xlarge",
    "AvailabilityZone": "us-west-2a",
    "LifecycleState": "Pending",
    "HealthStatus": "Healthy",
    "LaunchTemplate": {
        "LaunchTemplateId": "lt-0b97f1e282EXAMPLE",
        "LaunchTemplateName": "my-launch-template",
        "Version": "7"
    },
    "ProtectedFromScaleIn": false,
    "WeightedCapacity": "16"
},
...
]
}
```

Informações adicionais

Ponderação e estratégias de alocação de instâncias

As [estratégias de alocação \(p. 68\)](#) determinam de quais grupos de instâncias suas instâncias são provenientes. Quando você usa o recurso de ponderação da instância, as estratégias de alocação funcionam exatamente como em outros grupos do Auto Scaling. No entanto, há uma diferença crucial na forma como os grupos de instâncias são escolhidos quando você usa a estratégia `lowest-price`. Quando você escolhe `lowest-price` para sua estratégia de alocação, suas instâncias vêm dos grupos de instâncias com o menor preço por unidade em cada zona de disponibilidade.

Por exemplo, considere que você tem um grupo do Auto Scaling com vários tipos de instância com diferentes quantidades de vCPUs. 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.

Ponderação da instância e preço máximo de spot

Quando você cria o grupo do Auto Scaling usando a AWS CLI ou um SDK, é possível especificar o parâmetro `SpotMaxPrice`. Esse parâmetro determina o preço máximo que você estaria disposto a pagar por uma hora de instância spot e, por padrão, é definido como o preço sob demanda. Quando você usa o recurso de ponderação da instância, lembre-se de que o preço spot máximo reflete o preço máximo por unidade (por exemplo, preço por vCPU) ao invés do preço máximo de uma instância inteira.

Criar um grupo do Auto Scaling usando seleção de tipo de instância baseada em atributos

Ao criar um grupo de Auto Scaling, você deve especificar as informações para configurar o seguinte:

- O modelo de execução que especifica a AMI e um tipo de instância para as instâncias do Amazon EC2

- As zonas de disponibilidade e sub-redes VPC para as instâncias
- A capacidade desejada
- Os limites da capacidade máxima e mínima

Como alternativa à escolha manual de tipos de instância ao criar um [grupo de instâncias mistas \(p. 67\)](#), é possível 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.

Seu grupo de Auto Scaling ou seu modelo de execução especifica os atributos de sua instância. Esses atributos incluem a quantidade de memória e capacidade de computação necessária para os aplicativos que você planeja executar nas instâncias. Além disso, seu grupo de Auto Scaling ou seu modelo de execução especifica dois limites de proteção de preço para instâncias spot e sob demanda que você pode personalizar. Dessa forma, você pode impedir que o Amazon EC2 Auto Scaling execute tipos de instância mais caros se você não precisar deles.

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:

- 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.
- Com tantas opções disponíveis, encontrar os tipos de instância certos para o seu workload pode levar tempo. Ao especificar atributos de instância, você pode simplificar a seleção do tipo de instância ao configurar um grupo de instâncias mistas.
- 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.

Para obter informações sobre os parâmetros de configuração disponíveis para a seleção do tipo de instância com base em atributos ao criar um grupo de Auto Scaling usando a CLI ou um SDK, consulte [InstanceRequirements](#) na Referência da API do Amazon EC2 Auto Scaling. Para saber como declarar a seleção do tipo de instância com base em atributos ao criar um grupo de Auto Scaling usando AWS CloudFormation, consulte o trecho de exemplo na seção Trechos de [modelo de escalonamento automático](#) do Guia do usuário. AWS CloudFormation

Para saber mais sobre a seleção de tipo de instância baseada em atributos, consulte [Attribute-Based Instance Type Selection for EC2 Auto Scaling and EC2 Fleet](#) (Seleção de tipo de instância baseada em atributos para EC2 Auto Scaling e EC2 Fleet) no Blog AWS.

Índice

- [Considerações \(p. 93\)](#)
- [Pré-requisitos \(p. 95\)](#)
- [Usando a seleção de tipo de instância baseada em atributos \(p. 95\)](#)
- [Limitações \(p. 99\)](#)

Considerações

Ao usar a seleção de tipo de instância baseada em atributos, leve em consideração as seguintes coisas:

- Para a maioria dos workloads de uso geral, basta especificar o número de vCPUs e memória necessários. Para casos de uso avançados, você pode especificar atributos como tipo de armazenamento, interfaces de rede, fabricante da CPU e tipo de acelerador.
- Por padrão, o valor da capacidade desejada do grupo do Auto Scaling é definido como o número de instâncias. Opcionalmente, é possível especificar o tipo de capacidade desejado como o número de vCPUs ou a quantidade de memória ao usar a seleção de tipo de instância baseada em atributos. Então, quando o Amazon EC2 Auto Scaling inicia as instâncias, seu número de vCPUs ou quantidade de memória são considerados para a capacidade desejada. Ao criar seu grupo no console do Amazon EC2 Auto Scaling, essa configuração aparece na seção Group size (Tamanho do grupo) na página Configure group size and scaling policies (Definir políticas de escalabilidade e tamanho do grupo). Esse recurso é um substituto útil para o recurso [instance weighting \(p. 86\)](#) (ponderação da 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 [GetInstanceTypesFromInstanceRequirements](#) API do Amazon EC2 usando o AWS CLI ou um SDK. Transmite 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 for Linux Instances (Guia do usuário do Amazon EC2 para instâncias do Linux).

Entenda a proteção de preço

A proteção de preço é um recurso que protege seu grupo do Auto Scaling contra diferenças extremas de preços entre os tipos de instância. Ao criar um novo grupo do Auto Scaling ou atualizar um grupo do Auto Scaling existente com seleção de tipo de instância baseada em atributos, habilitamos a proteção de preço por padrão. Opcionalmente, você pode escolher seus limites de proteção de preço para instâncias spot e sob demanda. Quando você faz isso, o Amazon EC2 Auto Scaling não seleciona tipos de instância com preços superiores aos seus limites especificados. Os limites representam o que você está disposto a pagar, definido em termos de uma porcentagem acima de uma linha de base, ao invés de como valores absolutos. A linha de base é determinada pelo preço do tipo mais barato de instância M, C ou R da geração atual com os atributos especificados. Se seus atributos não corresponderem a nenhum tipo de instância M, C ou R, usaremos o tipo de instância com menor preço.

Se você não especificar um limite, os seguintes limites serão usados por padrão:

- Para instâncias sob demanda, o limite de proteção de preço é definido em 20%.
- Para instâncias spot, o limite de proteção de preço é definido em 100%.

Você pode atualizar esses valores ao criar seu grupo do Auto Scaling no console do Amazon EC2 Auto Scaling. Na página Choose instance launch options (Escolher opções de execução da instância), escolha o atributo de proteção de preço desejado na lista suspensa Additional instance attributes (Atributos adicionais da instância). Em seguida, digite ou escolha um valor para o atributo na caixa de texto. Você também pode atualizar esses valores posteriormente editando o grupo de Auto Scaling no console ou usando o AWS CLI ou um SDK.

Note

Se você definir Desired capacity type (Tipo de capacidade desejada) como vCPUs ou Memory GiB (Memória GiB), o limite de proteção de preço será aplicado com base no preço por vCPU ou por memória em vez do preço por instância.

Pré-requisitos

- Criar um modelo de execução. Para obter mais informações, consulte [Criar um modelo de execução para um grupo do Auto Scaling \(p. 23\)](#).
- Verifique se o modelo de execução já não solicita instâncias spot.

Usando a seleção de tipo de instância baseada em atributos

Use o procedimento a seguir para usar a seleção de tipo de instância baseada em atributo. Se você preferir escolher quais tipos de instância individuais seu grupo pode iniciar, consulte [Grupos do Auto Scaling com vários tipos de instâncias e opções de compra \(p. 67\)](#) para configurar os requisitos de tipo de instância escolhendo manualmente os tipos de instância.

As etapas a seguir descrevem como criar um grupo do Auto Scaling usando a seleção de tipo de instância baseada em atributos:

- Especifique o modelo de execução para iniciar as instâncias.
- Escolha a VPC e as sub-redes para iniciar seu grupo do Auto Scaling.
- Escolha a opção de substituir os requisitos existentes do tipo de instância do modelo de inicialização por novos requisitos.
- Especifique atributos de instância que correspondam aos requisitos de computação, como vCPUs, memória e armazenamento.
- Especifique as porcentagens de instâncias sob demanda e de instâncias spot a serem iniciadas.
- 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.
- Especifique o tamanho do grupo, incluindo a capacidade desejada, capacidade mínima, capacidade máxima e tipo de capacidade desejada, que define uma unidade de medida para a capacidade desejada.

Para criar um grupo do Auto Scaling usando seleção de tipo de instância baseada em atributos (console)

1. Abra o console do Amazon EC2 em <https://console.aws.amazon.com/ec2/> e escolha grupos do Auto Scaling 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. Para escolher a VPC e as sub-redes nas quais executar suas instâncias, faça o seguinte:

- a. Na página Choose instance launch options (Escolher as opções de execução da instância) em Network (Rede), para VPC, selecione uma VPC. O grupo do Auto Scaling deve ser criado na mesma VPC do grupo de segurança especificado no modelo de execução.
 - b. 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 \(p. 417\)](#).
7. Para configurar o grupo para usar a seleção de tipo de instância baseada em atributos, faça o seguinte:
- a. Para Instance type requirements (Requisitos de tipo de instância), selecione Override launch template (Substituir modelo de execução).

Note

Se você escolher um modelo de execução que já contenha seus requisitos, as configurações do modelo de execução, como vCPUs e memória, serão usadas automaticamente como atributos. Você pode atualizar esses atributos do console no Amazon EC2 Auto Scaling, a qualquer momento.

- b. Sob Specify instance attributes (Especificar os atributos da instância), comece inserindo seus requisitos de vCPUs e de memória.
 - vCPUs: insira o número mínimo e máximo de vCPUs para seus requisitos de computação. Selecione a caixa de seleção No minimum (Sem mínimo) ou No maximum (Sem máximo) para indicar ausência de limite nessa direção.
 - Memória (MiB): insira a quantidade mínima e máxima de memória, em MiB, para seus requisitos de computação. Selecione a caixa de seleção No minimum (Sem mínimo) ou No maximum (Sem máximo) para indicar ausência de limite nessa direção.
- c. (Opcional) Para Additional instance attributes (Atributos de instância adicionais) e selecione Add attribute (Adicionar atributo) para expressar seus requisitos de computação mais detalhados. Os atributos e valores escolhidos aqui limitam ainda mais quais tipos de instância podem ser iniciados.

Para obter mais informações sobre todos os atributos suportados, consulte a Referência [InstanceRequirements](#) API do Amazon EC2 Auto Scaling.

- d. Em Preview matching instance types (Previsualizar os tipos de instância correspondentes), visualize os tipos de instância que correspondem aos requisitos de computação especificados, como vCPUs, memória e armazenamento.
- e. 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.
- f. Quando você especifica uma porcentagem para instâncias spot, pode marcar a caixa de seleção ao lado de Include On-Demand base capacity (Incluir capacidade básica sob demanda) e depois especificar 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. O preço mais baixo está oculto por padrão e só aparece quando você escolhe Mostrar todas as estratégias.

Note

Se você escolheu Preço mais baixo, insira o número de grupos com preços mais baixos para diversificar para os grupos com preços mais baixos.

- i. Em Capacity rebalance (Rebalanceamento de capacidade), escolha se você deseja habilitar ou desabilitar o rebalanceamento de capacidade.

Se você escolher uma porcentagem para instâncias Spot, poderá usar o Rebalanceamento de capacidade para responder automaticamente quando suas instâncias Spot se aproximarem do encerramento de uma interrupção Spot. Para obter mais informações, consulte [Usar o rebalanceamento de capacidade para lidar com interrupções de spot do Amazon EC2 \(p. 346\)](#).

- j. Selecione Next (Avançar) duas vezes para ir para a página Configure group size and scaling policies (Configurar tamanho do grupo e políticas de escalação).
8. Na etapa Configure group size and scaling policies (Configurar o tamanho do grupo e as políticas de escalação), faça o seguinte:
 - a. Se você não quiser que a capacidade desejada seja medida em instâncias, mas em outras unidades, escolha a opção apropriada em Desired capacity type (Tipo de capacidade desejada). As opções compatíveis são Units (Unidades), vCPUs e Memory GiB (GiBs de memória). Por padrão, o Amazon EC2 Auto Scaling especifica Units (Unidades), o que quer dizer número de instâncias.
 - b. Insira o tamanho inicial do grupo do Auto Scaling para a Desired capacity (Capacidade desejada) e atualize os limites de Minimum capacity (Capacidade mínima) e Maximum capacity (Capacidade máxima) conforme necessário. Para obter mais informações, consulte [Definir limites de capacidade no grupo do Auto Scaling \(p. 166\)](#).
 - c. (Opcional) Configure o grupo para dimensionar especificando uma política de dimensionamento de rastreamento de destino. Como alternativa, especifique essa política depois de criar o grupo. Para obter mais informações, consulte [Políticas de escalabilidade com monitoramento do objetivo para o Amazon EC2 Auto Scaling \(p. 180\)](#).
 - d. (Opcional) Habilite a proteção de redução de instância, o que impede que seu grupo do Auto Scaling encerre instâncias durante a redução. Para obter mais informações, consulte [Usar proteção de redução na escala na horizontal de instâncias \(p. 302\)](#).
 - e. Quando terminar, escolha Next (Próximo).
9. (Opcional) Para receber notificações quando a escala do grupo for aumentada ou reduzida, em Add notification (Adicionar notificação), configure a notificação e depois escolha Next (Avançar). Para obter mais informações, consulte [Receber notificações do Amazon SNS quando o grupo do Auto Scaling escala \(p. 341\)](#).
10. (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 \(p. 138\)](#).
11. Na página Review (Revisão), escolha Create Auto Scaling group (Criar grupo do Auto Scaling).

Exemplo: criar um grupo do Auto Scaling usando seleção de tipo de instância baseada em atributos (AWS CLI)

Para criar um grupo de Auto Scaling com seleção de tipo de instância baseada em atributos usando a linha de comando, você pode usar o comando a seguir. [create-auto-scaling-group](#)

Os seguintes atributos de instância são especificados:

- VcpuCount: os tipos de instância devem ter um mínimo de quatro e um máximo de oito vCPUs.
- 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 vCPUs 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-group](#) comando 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 em vez de um arquivo JSON.

```
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 vCPUs 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.

Limitações

- Você pode configurar a seleção de tipo de instância baseada em atributos apenas para grupos do Auto Scaling que usam um modelo de execução.
- Se você tiver um grupo do Auto Scaling existente e planeja substituir os tipos de instância pelos atributos de instância necessários, sua estratégia de alocação sob demanda deve ser lowest-price. Para usar a estratégia de alocação prioritized, você deve continuar adicionando e priorizando manualmente seus tipos de instância. Além disso, sua estratégia de alocação Spot deve ser price-capacity-optimized, capacity-optimized ou lowest-price. Para usar a estratégia de alocação capacity-optimized-prioritized, você deve adicionar e priorizar manualmente seus tipos de instância.

Crie grupos de Auto Scaling usando configurações de lançamento

Important

As configurações de inicialização não adicionam mais suporte para novos tipos de instância do Amazon EC2 lançados após 31 de dezembro de 2022. Além disso, todas as novas contas criadas em ou após 1 de junho de 2023 não terá a opção de criar novas configurações de inicialização por meio do console. No entanto, API, CLI e CloudFormation acesso estará disponível para novas contas criadas entre 1 de junho de 2023 e 31 de dezembro de 2023 para oferecer suporte aos clientes com casos de uso de automação. Novas contas criadas em ou depois 1 de janeiro de 2024 não poderão criar novas configurações de lançamento usando o console, a API, a CLI e CloudFormation. Para obter informações sobre como migrar seus grupos do Auto Scaling para lançar modelos, consulte [Migre para lançar modelos \(p. 32\)](#).

Se você criou uma configuração de execução ou uma instância do EC2, você pode criar um grupo de Auto Scaling que usa 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 blocos para suas instâncias. Para obter informações sobre como criar configurações de inicialização, consulte [Criar uma configuração de execução \(p. 49\)](#).

Você deve ter permissões suficientes para criar um grupo de Auto Scaling. Você também deve ter permissões suficientes para criar a função vinculada ao serviço que o Amazon EC2 Auto Scaling usa para realizar ações em seu nome, caso ela ainda não exista. Para ver exemplos de políticas do IAM que um

administrador pode usar como referência para conceder permissões a você, consulte [Exemplos de políticas baseadas em identidade \(p. 437\)](#).

Índice

- [Criar um grupo do Auto Scaling usando uma configuração de execução \(p. 100\)](#)
- [Criar um grupo do Auto Scaling usando parâmetros de uma instância existente \(p. 102\)](#)

Criar um grupo do Auto Scaling usando uma configuração de execução

Important

As configurações de inicialização não adicionam mais suporte para novos tipos de instância do Amazon EC2 lançados após 31 de dezembro de 2022. Além disso, todas as novas contas criadas em ou após 1º de junho de 2023 não terão a opção de criar novas configurações de lançamento por meio do console. No entanto, a API, a CLI e o CloudFormation acesso estarão disponíveis para novas contas criadas entre 1º de junho de 2023 e 31 de dezembro de 2023 para oferecer suporte aos clientes com casos de uso de automação. Novas contas criadas em ou após 1º de janeiro de 2024 não poderão criar novas configurações de lançamento usando o console, a API, a CLI e o CloudFormation. Para obter informações sobre como migrar seus grupos do Auto Scaling para lançar modelos, consulte [Migre para lançar modelos \(p. 32\)](#).

Ao criar um grupo do Auto Scaling, você deverá 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 mínima e máxima.

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 \(p. 57\)](#).

Pré-requisitos

- É necessário ter criado uma configuração de execução. Para obter mais informações, consulte [Criar uma configuração de execução \(p. 49\)](#).

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 grupos do Auto Scaling no painel de navegação.
2. Na barra de navegação na parte superior da tela, selecione a mesma Região da AWS usada ao criar a configuração 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 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 (Definir configurações) Configure instance launch options (Configurar as opções de execução da instância) sob Rede, para VPC, selecione uma VPC. O grupo do Auto Scaling deve ser criado na mesma VPC do grupo de segurança especificado na configuração de execução.
 7. 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 \(p. 417\)](#).
 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. Para registrar suas instâncias do Amazon EC2 com um balanceador de carga, escolha um load balancer existente ou crie um novo. Para obter mais informações, consulte [Usar o Elastic Load Balancing para distribuir tráfego entre as instâncias no grupo do Auto Scaling \(p. 369\)](#). Para criar um novo balanceador de carga, siga o procedimento em [Configurar um Application Load Balancer ou Network Load Balancer pelo console do Amazon EC2 Auto Scaling \(p. 374\)](#).
 - b. (Opcional) Para verificações de integridade, tipos adicionais de verificação de saúde, selecione Ativar verificações de integridade do Elastic Load Balancing.
 - c. (Opcional) Para o período de carência da verificação de saúde, 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 \(p. 325\)](#).
 - d. Em Configurações adicionais, Monitoramento, escolha se deseja ativar a coleta de métricas CloudWatch do 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 [MonitorCloudWatchmétricas para seus grupos e instâncias do Auto Scaling \(p. 328\)](#).
 - e. 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 da instância aprimora 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 \(p. 200\)](#).
 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. Em Desired capacity (Capacidade desejada), insira o número inicial de instâncias a serem executadas. Quando esse número é alterado para um valor fora dos limites de capacidade mínima ou máxima, é necessário atualizar os valores de Minimum capacity (Capacidade mínima) ou Maximum capacity (Capacidade máxima). Para obter mais informações, consulte [Definir limites de capacidade no grupo do Auto Scaling \(p. 166\)](#).
 - b. Para escalar automaticamente o tamanho do grupo do Auto Scaling, escolha Target tracking scaling policy (Política de escalabilidade com monitoramento do objetivo) e siga as instruções. Para obter mais informações, consulte [Políticas de escalabilidade com monitoramento do objetivo para o Amazon EC2 Auto Scaling \(p. 180\)](#).
 - c. 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 [Usar proteção de redução na escala na horizontal de instâncias \(p. 302\)](#).
 11. (Opcional) Para receber notificações, em Add notification (Adicionar notificação), configure a notificação e, depois, escolha Next (Próximo). Para obter mais informações, consulte [Receber notificações do Amazon SNS quando o grupo do Auto Scaling escala \(p. 341\)](#).

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 \(p. 138\)](#).
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-AS AutoScalingGroup](#) () AWS Tools for Windows PowerShell

Criar um grupo do Auto Scaling usando parâmetros de uma instância existente

Important

Este tópico só se aplica à criação de um grupo do Auto Scaling usando a configuração de execução. Fornecemos informações sobre configurações de execução para clientes que ainda não migraram das configurações de execução para os 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 \(p. 64\)](#).

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 de 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

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 \(p. 51\)](#).
- 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.

O novo grupo do Auto Scaling inicia instâncias na mesma zona de disponibilidade, VPC e sub-rede na qual a instância identificada está localizada.

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.

Pré-requisitos

A instância EC2 deve atender aos seguintes critérios:

- A instância é a sub-rede e a zona de disponibilidade na qual você deseja criar o grupo do Auto Scaling.
- 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.

Criar um grupo do Auto Scaling com base em uma instância do EC2 (console)

Para criar um grupo do Auto Scaling a partir de uma instância do EC2

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) e selecione uma instância.
3. Escolha Actions (Ações), Instance settings (Configurações da instância), Attach to Auto Scaling Group (Anexar ao grupo do Auto Scaling).
4. Na página Attach to Auto Scaling group (Anexar ao grupo do Auto Scaling), em Auto Scaling Group (Grupo do Auto Scaling), insira um nome para o grupo e escolha Attach (Anexar).

Depois que a instância for anexada, ela será considerada parte do grupo do Auto Scaling. O novo grupo do Auto Scaling é criado usando uma nova configuração de execução com o mesmo nome que você especificou para o grupo do Auto Scaling. O grupo do Auto Scaling tem uma capacidade desejada e um tamanho máximo de 1.

5. (Opcional) Para editar as configurações do grupo do Auto Scaling, no painel de navegação, em Auto Scaling, escolha Auto Scaling Groups (Grupos do Auto Scaling). Marque a caixa de seleção ao lado do novo grupo do Auto Scaling, escolha o botão Edit (Editar) que está acima da lista de grupos, altere as configurações conforme necessário e escolha Update (Atualizar).

Crie um grupo do Auto Scaling a partir de uma instância do EC2 (AWS CLI).

Neste exercício, demonstramos como usar a AWS 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, você deve 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-group](#) comando a seguir para criar um grupo de Auto Scaling, my-asg-from-instance, a partir da instância do EC2 i-0e69cc3f05f825f4f.

```
aws autoscaling create-auto-scaling-group --auto-scaling-group-name my-asg-from-instance \
--instance-id i-0e69cc3f05f825f4f --min-size 1 --max-size 2 --desired-capacity 2
```

Para verificar se seu grupo do Auto Scaling executou instâncias

- Use o seguinte [describe-auto-scaling-groups](#) comando para verificar se o grupo do Auto Scaling foi criado com êxito.

```
aws autoscaling describe-auto-scaling-groups --auto-scaling-group-name my-asg-from-instance
```

O exemplo de resposta 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-06905f55584de02da",
          "HealthStatus": "InService",
          "LifecycleTransition": "autoscaling:EC2_INSTANCE_LAUNCH",
          "LastSuccessfulHealthCheck": "2018-05-15T18:45:00Z",
          "LastUnsuccessfulHealthCheck": null,
          "LastSuccessfulCheckTime": "2018-05-15T18:45:00Z",
          "LastUnsuccessfulCheckTime": null
        }
      ]
    }
  ]
}
```

```
"InstanceType":"t2.micro",
"AvailabilityZone":"us-west-2a",
"LifecycleState":"InService",
"HealthStatus":"Healthy",
"LaunchConfigurationName":"my-asg-from-instance",
"ProtectedFromScaleIn":false
},
{
  "InstanceId": "i-087b42219468eacde",
  "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-6bea5f06",
"EnabledMetrics": [ ],
"Tags": [ ],
"TerminationPolicies": [
  "Default"
],
"NewInstancesProtectedFromScaleIn": false,
"ServiceLinkedRoleARN": "arn"
}
]
```

Para visualizar a configuração de execução

- Use o [describe-launch-configurations](#) comando a seguir para visualizar os detalhes da configuração de execução.

```
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-0528a5175983e7f28",
      "KeyName": "my-key-pair-uswest2",
      "SecurityGroups": [
        "sg-05eaec502fcadad2e"
      ],
      "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-0e69cc3f05f825f4f`.

```
aws ec2 terminate-instances --instance-ids i-0e69cc3f05f825f4f
```

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.

Atualizar um grupo de Auto Scaling

Você pode atualizar a maioria dos detalhes do seu grupo de Auto Scaling. Você não pode atualizar o nome de um grupo de Auto Scaling nem alterar seu Região da AWS.

Para atualizar um grupo de Auto Scaling (console)

1. Abra o console do Amazon EC2 em <https://console.aws.amazon.com/ec2/> e escolha grupos do Auto Scaling no painel de navegação.
2. Escolha seu grupo de Auto Scaling para exibir informações sobre o grupo, com guias para Detalhes, Atividade, Escalabilidade automática, Gerenciamento de instâncias, Monitoramento, e Atualização da instância.
3. Escolha as guias para as áreas de configuração nas quais você está interessado e atualize as configurações conforme necessário. Para cada configuração que você edita, escolha Atualizar para salvar suas alterações na configuração do grupo Auto Scaling.
 - Detalhes abertos

Essas são as configurações gerais do seu grupo de Auto Scaling. Você pode editá-los e gerenciá-los da mesma forma que durante a criação do grupo do Auto Scaling.

O Configurações avançadas A seção tem algumas opções que não estão disponíveis ao criar o grupo, como [políticas de rescisão \(p. 295\)](#), [tempo de recarga \(p. 205\)](#), [processos suspensos \(p. 312\)](#), [evídua útil máxima da instância \(p. 135\)](#). Você também pode visualizar, mas não editar, o grupo de posicionamento e [função vinculada ao serviço \(p. 432\)](#) do grupo Auto Scaling.

Se o grupo estiver associado aos recursos do Elastic Load Balancing, consulte [Adicionar e remover zonas de disponibilidade \(p. 377\)](#) antes de mudar 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.

- Atividade aberta
 - Notificações de atividades—[Notificações do Amazon SNS \(p. 341\)](#)
- Escalabilidade automática aberta
 - Políticas de escalabilidade dinâmica—[Políticas de escalabilidade dinâmica \(p. 178\)](#)
 - Políticas de escalabilidade preditiva—[Políticas de escalabilidade preditiva \(p. 222\)](#)

- Ações programadas—[Ações programadas \(p. 247\)](#)
- Gerenciamento de instâncias abas
 - Ganchos de ciclo de vida—[Ganchos de ciclo de vida \(p. 252\)](#)
 - Piscina aquecida—[Piscinas quentes \(p. 279\)](#)
- Monitoramento abas
 - Há apenas uma única opção nessa guia, que permite ativar ou desativar[CloudWatch coleção de métricas de grupo \(p. 332\)](#).

Para atualizar um grupo de Auto Scaling usando a linha de comando

Você pode usar um dos comandos a seguir:

- [update-auto-scaling-group](#) (AWS CLI)
- [Atualize-asAutoScalingGroup](#)(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 de Auto Scaling, todas as novas instâncias receberão a configuração atualizada. As instâncias existentes continuam sendo executadas com a configuração com a qual foram lançadas originalmente. Para aplicar suas alterações às instâncias existentes, você tem as seguintes opções:

- Inicie uma atualização de instância para substituir as instâncias mais antigas. Para ter mais informações, consulte [Substituir instâncias do Auto Scaling \(p. 108\)](#) e [Atualizar um grupo de alta atividade \(p. 283\)](#).
- Aguarde até que as atividades de escalabilidade substituam gradualmente as instâncias mais antigas por instâncias mais novas, com base em suas[políticas de rescisão \(p. 292\)](#).
- Encerre-os manualmente para que sejam substituídos pelo seu grupo de Auto Scaling.

Note

Você pode alterar os seguintes atributos da instância especificando-os como parte do modelo de execução ou da configuração de execução:

- Imagem de máquina da Amazon (AMI)
- dispositivos de bloqueio
- par de chaves
- tipo de instância
- security groups
- dados do usuário
- monitoramento
- Perfil de instância do IAM
- locação de colocação
- kernel
- disco RAM
- se a instância tem um endereço IP público

Substituir instâncias do Auto Scaling

O Amazon EC2 Auto Scaling oferece recursos que permitem substituir instâncias depois de atualizar seu grupo de Auto Scaling. O Amazon EC2 Auto Scaling 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.
- 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 houver alguma AmazonCloudWatchLogs alarmes que você especificar vão para o ALARMEstado.
- Limitar a vida útil das instâncias para fornecer versões de software consistentes e configurações de instância em todo o grupo do Auto Scaling.

Índice

- [Substituir instâncias do Auto Scaling com base em uma atualização de instância \(p. 108\)](#)
- [Substituir instâncias do Auto Scaling com base na vida útil máxima da instância \(p. 135\)](#)

Substituir instâncias do Auto Scaling com base em uma atualização de instância

Você pode usar uma atualização de instância para atualizar as instâncias em seu grupo do Auto Scaling em vez de substituir manualmente instâncias algumas de cada vez. Isso pode ser útil quando uma alteração de configuração requer a substituição de instâncias e você tem um grande número de instâncias no grupo do Auto Scaling.

Uma atualização de instância pode ser útil quando há uma nova Imagem de máquina da Amazon (AMI) ou um novo script de dados de usuário. Para usar uma atualização de instância, crie primeiro um novo modelo de execução que especifique a nova AMI ou o novo script de dados do usuário. Em seguida, inicie uma atualização de instância para começar a atualizar imediatamente as instâncias no grupo.

Uma atualização de instância também pode ser útil quando você está migrando seus grupos de Auto Scaling das configurações de execução para os modelos de execução. Primeiro, copie suas configurações de lançamento para novos modelos de lançamento. Em seguida, inicie uma atualização de instância que especifique o modelo de execução como parte da configuração desejada para começar a atualizar as instâncias no grupo imediatamente. Para obter mais informações sobre a migração para modelos de lançamento, consulte [Migre para lançar modelos \(p. 32\)](#).

Índice

- [Como funcionam \(p. 109\)](#)
- [Conceitos e termos fundamentais \(p. 109\)](#)

- [Compatibilidade de tipo de instância \(p. 111\)](#)
- [Limitações \(p. 111\)](#)
- [Iniciar uma atualização de instância \(p. 111\)](#)
- [Entender os valores padrão de uma atualização de instância \(p. 117\)](#)
- [Verificar o status de uma atualização de instância \(p. 119\)](#)
- [Cancelar uma atualização de instância \(p. 121\)](#)
- [Desfazer alterações com uma reversão \(p. 122\)](#)
- [Usar uma atualização de instância com opção de ignorar correspondência \(p. 125\)](#)
- [Adicionar pontos de verificação a uma atualização de instância \(p. 132\)](#)

Como funcionam

Para atualizar seu grupo de Auto Scaling com um novo modelo de lançamento, você normalmente executa as seguintes ações:

- Crie um novo modelo de execução para seu grupo do Auto Scaling. Para obter mais informações, consulte [Criar um modelo de execução para um grupo do Auto Scaling \(p. 23\)](#).
- Você configura a porcentagem mínima de integridade, os pontos de verificação opcionais e quaisquer outras configurações conforme necessário, especifica uma configuração desejada que inclua seu modelo de execução e inicia uma atualização da instância. Opcionalmente, a configuração desejada pode especificar se deve aplicar uma [política de instâncias mistas \(p. 67\)](#).
- O Amazon EC2 Auto Scaling começa a executar uma substituição contínua das instâncias. Ele retira um conjunto de instâncias de serviço, termina-as e, em seguida, inicia um conjunto de instâncias com a nova configuração. Depois, ele aguarda até que as instâncias sejam aprovadas em suas verificações de integridade e concluem o aquecimento antes de começar a substituir outras instâncias.
- Após um determinado percentual do grupo ser substituído, um ponto de verificação é atingido. Sempre que há um ponto de verificação, o Amazon EC2 Auto Scaling interrompe temporariamente a substituição de instâncias e envia uma notificação. Em seguida, ele aguarda o tempo especificado antes de continuar. Depois de receber a notificação, você poderá verificar se suas novas instâncias estão funcionando conforme o esperado.
- Depois que a atualização da instância for bem-sucedida, as configurações do grupo do Auto Scaling serão atualizadas automaticamente com o modelo de execução que você especificou no início da operação.

Conceitos e termos fundamentais

Antes de começar, familiarize-se com os seguintes conceitos e termos fundamentais da atualização de instâncias:

Percentual mínimo de integridade

Como parte do início de uma atualização de instância, especifique o percentual mínimo de integridade a ser mantido em todos os momentos. Esta é a quantidade de capacidade em um grupo do Auto Scaling que deve passar em suas [verificações de integridade \(p. 319\)](#) durante uma atualização de instância para que a atualização possa continuar. Por exemplo, se a porcentagem mínima de integridade for 90%, 10% será a porcentagem da capacidade que será encerrada e 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 conseguir iniciar nenhuma instância saudável, ela acabará falhando, deixando os outros 90% do grupo intocados. Se as novas instâncias permanecerem saudáveis e terminarem o período de aquecimento, o Amazon EC2 Auto Scaling poderá continuar substituindo outras instâncias.

A atualização de instância pode substituir uma instância por vez, várias por vez ou todas de uma vez. Para substituir uma instância de cada vez, estabeleça um percentual mínimo de integridade igual a 100%. Para substituir todas de uma vez, defina uma porcentagem mínima íntegra de 0%.

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 do Auto Scaling \(p. 319\)](#). Para garantir que essas verificações de saúde comecem o mais rápido possível, não defina o período de carência da verificação de saúde do grupo muito alto, mas alto o suficiente para que suas verificações de saúde do Elastic Load Balancing determinem se um alvo 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 \(p. 325\)](#).

Aquecimento da instância

O aquecimento da instância é o período de tempo a partir do qual o estado de uma nova instância muda para InService até quando se considera que a inicialização foi 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 seu aplicativo ainda precisa de algum tempo de inicialização antes de responder às solicitações.

É possível reduzir o valor do período de aquecimento caso tenha usado um gancho do ciclo de vida para preparar novas instâncias para uso. Para obter mais informações, consulte [Ganchos do ciclo de vida do Amazon EC2 Auto Scaling \(p. 252\)](#).

Somente especifique o aquecimento da instância para uma atualização da instância quando você não tiver ativado o aquecimento padrão da instância ou se precisar substituí-lo. O aquecimento da instância funciona da mesma forma que o aquecimento padrão da instância. Portanto, as mesmas considerações de escala se aplicam. Para obter mais informações, consulte [Definir o aquecimento padrão da instância para um grupo do Auto Scaling \(p. 200\)](#).

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 um EventBridge para enviar os eventos a um alvo, como o Amazon SNS, para ser notificado quando um ponto de verificação for atingido. 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.

Important

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 \(p. 117\)](#).

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

- Instâncias terminadas antes de iniciar: quando há apenas uma instância no grupo do Auto Scaling, iniciar uma atualização de instância pode resultar em uma interrupção. Isso ocorre porque o Amazon EC2 Auto Scaling termina uma instância e inicia outra.
- Duração total: o tempo máximo que uma atualização de instância pode permanecer ativamente substituindo instâncias é 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 ponderação de instâncias para o Amazon EC2 Auto Scaling \(p. 86\)](#). 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 horizontalmente, 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.

Iniciar uma atualização de instância

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 \(p. 122\)](#).

Os procedimentos a seguir ajudam a iniciar uma atualização de instância usando o AWS Management Console ou a 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 \(p. 67\)](#) para seu grupo do Auto Scaling. Se você já definiu uma política de instâncias mistas, consulte [Iniciar uma atualização de instância no console \(grupo de instâncias mistas\) \(p. 114\)](#) 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 grupos do Auto Scaling 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. Em Minimum healthy percentage (Percentual mínimo de integridade), insira o percentual do grupo do Auto Scaling que deve permanecer íntegro durante uma atualização de instância. Aumentar percentual de integridade mínima para 100 limita a taxa de substituição a uma instância de cada vez. Por outro lado, definir como 0% faz com que todas as instâncias sejam substituídas ao mesmo tempo.
5. Para Aquecimento da instância, insira o número de segundos a partir do momento em que o estado de uma nova instância muda para InService até quando terminar de inicializar. 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. Para obter mais informações sobre como configurar o aquecimento de instâncias, consulte [Definir o aquecimento padrão da instância para um grupo do Auto Scaling \(p. 200\)](#).

6. (Opcional) Em Checkpoints (Pontos de verificação), escolha Enable checkpoints (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 \(console\) \(p. 133\)](#) para obter etapas adicionais.

7. Habilitar ou desativar Skip matching (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 habilitar a correspondência por ignorar, você pode definir um novo modelo de lançamento ou uma nova versão do modelo de lançamento em vez de usar o existente. Faça isso no [Configuração desejada](#) e [Iniciar atualização da instância](#) na página.

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 Desired configuration (Configuração desejada). Não há suporte para ignorar a correspondência com uma configuração de inicialização.

8. 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 temporariamente instâncias do grupo do Auto Scaling \(p. 308\)](#).

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.

9. 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 [Usar proteção de redução na escala na horizontal de instâncias \(p. 302\)](#).

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.

10. (Opcional) Para CloudWatchalararme, escolha Ativar CloudWatchalarmese, em seguida, escolha um ou mais alarmes. CloudWatchos alarmes podem ser usados para identificar quaisquer problemas e falhar na operação se um alarme for acionado ALARMestado. Para obter mais informações, consulte [Iniciar uma atualização de instância com reversão automática \(p. 123\)](#).
11. (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 Use code editor (Usar editor de código) em vez de Use console interface (Usar a interface do console). O procedimento a seguir explica como fazer seleções usando a interface do console.

- a. Para Update launch template (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 já criou um novo modelo de execução ou uma nova versão de um modelo de execução, marque essa 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 executar 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 \(p. 67\)](#).

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.

12. (Opcional) Em Configurações de reversão, escolha Habilitar reversão automática para reverter automaticamente a atualização de instância em caso de falha.

Essa configuração não pode ser habilitada quando o grupo do Auto Scaling não atende aos pré-requisitos para uso de reversões.

Para obter mais informações, consulte [Desfazer alterações com uma reversão \(p. 122\)](#).

13. 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 \(p. 111\)](#).

14. Quando estiver satisfeito com suas seleções de atualização de 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 \(p. 67\)](#). Se você não definiu ainda uma política de instâncias mistas para seu grupo, consulte [Iniciar uma atualização de instância no console \(procedimento básico\) \(p. 112\)](#) 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 grupos do Auto Scaling 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. Em Minimum healthy percentage (Percentual mínimo de integridade), insira o percentual do grupo do Auto Scaling que deve permanecer íntegro durante uma atualização de instância. Aumentar percentual de integridade mínima para 100 limita a taxa de substituição a uma instância de cada vez. Por outro lado, definir como 0% faz com que todas as instâncias sejam substituídas ao mesmo tempo.
5. Para Aquecimento da instância, insira o número de segundos a partir do momento em que o estado de uma nova instância muda para InService até quando terminar de inicializar. 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, entre outros.). 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 (a menos que deseje substituir o padrão). Para obter mais informações sobre como configurar o aquecimento de instâncias, consulte [Definir o aquecimento padrão da instância para um grupo do Auto Scaling \(p. 200\)](#).

6. (Opcional) Em Checkpoints (Pontos de verificação), escolha Enable checkpoints (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 \(console\) \(p. 133\)](#) para obter etapas adicionais.

7. Habilitar ou desativar Skip matching (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 habilitar a correspondência por ignorar, você pode definir um novo modelo de lançamento ou uma nova versão do modelo de lançamento em vez de usar o existente. Faça isso na Configuração desejada se você iniciar a atualização da instância na página.

8. 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 temporariamente instâncias do grupo do Auto Scaling \(p. 308\)](#).

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.

9. 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 [Usar proteção de redução na escala na horizontal de instâncias \(p. 302\)](#).

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.

10. (Opcional) Para CloudWatchalarms, escolha Ativar CloudWatchalarmes, em seguida, escolha um ou mais alarmes. CloudWatch alarms podem ser usados para identificar quaisquer problemas e falhar na operação se um alarme for acionado ALARMestado. Para obter mais informações, consulte [Iniciar uma atualização de instância com reversão automática \(p. 123\)](#).
11. 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 Use code editor (Usar editor de código) em vez de Use console interface (Usar a interface do console). O procedimento a seguir explica como fazer seleções usando a interface do console.

- a. Para Update launch template (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 já criou um novo modelo de execução ou uma nova versão de um modelo de execução, marque essa 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 \(p. 67\)](#).

Warning

Recomendamos não desmarcar essa caixa de seleção. Apenas a desmarcação 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.

12. (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 não pode ser habilitada quando o grupo do Auto Scaling não atende aos pré-requisitos para uso de reversões.

Para obter mais informações, consulte [Desfazer alterações com uma reversão \(p. 122\)](#).

13. 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 \(p. 111\)](#).

Quando estiver satisfeito com suas seleções de atualização de instância, escolha Iniciar atualização da instância.

Iniciar uma atualização de instância (AWS CLI)

Para iniciar uma atualização de instância

Use o seguinte `start-instance-refresh` comando para iniciar a atualização de uma instância a partir do AWS CLI. 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 \(p. 117\)](#).

Exemplos de resultado:

```
{  
    "InstanceRefreshId": "08b91cf7-8fa6-48af-b6a6-d227f40f1b9b"  
}
```

Entender os valores padrão de uma atualização de instância

Antes de iniciar uma atualização da instância, você pode personalizar várias preferências que afetam a atualização da instância. Alguns padrões de preferência são diferentes dependendo se você usa o console ou a linha de comando (AWS CLI ou AWSSDK).

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
CloudWatchalarme	Desativado (nulo)	Desabilitado
Reversão automática	Desabilitado (<code>false</code>)	Desabilitado
Pontos de verificação	Desabilitado (<code>false</code>)	Desabilitado
Atraso no ponto de verificação	1 hora (3600 segundos)	1 hora
Aquecimento da instância	O aquecimento de instâncias padrão (p. 200) , se estiver definido, ou o período de carência da verificação de integridade (p. 325) , se não estiver.	O aquecimento de instâncias padrão (p. 200) , se estiver definido, ou o período de carência da verificação de integridade (p. 325) , se não estiver.
Percentual mínimo de integridade	90	90
Instâncias protegidas contra redução da escala na horizontal	Wait	Ignorar
Ignorar correspondência	Desabilitado (<code>false</code>)	Habilitado
Instâncias em espera	Wait	Ignorar

Segue uma descrição de cada configuração:

CloudWatchalarme (**AlarmSpecification**)

O CloudWatchspecificação de alarme. CloudWatchos alarmes podem ser usados para identificar quaisquer problemas e falhar na operação se um alarme for acionadoALARMestado. Para obter mais informações, consulte [Iniciar uma atualização de instância com reversão automática \(p. 123\)](#).

Reversão automática (**AutoRollback**)

Controla se o Amazon EC2 Auto Scaling reverte o grupo Auto Scaling para sua configuração anterior se a atualização da instância falhar. Para obter mais informações, consulte [Desfazer alterações com uma reversão \(p. 122\)](#).

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 \(p. 132\)](#).

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 \(p. 132\)](#).

Aquecimento da instância (**InstanceWarmup**)

Um período de tempo, em segundos, durante o qual o Amazon EC2 Auto Scaling espera até que uma nova instância tenha sido inicializada antes de prosseguir com a substituição da 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 \(p. 200\)](#).

Percentual mínimo de integridade (**MinHealthyPercentage**)

A porcentagem da capacidade desejada do grupo do Auto Scaling que deve passar nas verificações de integridade do grupo antes que a atualização possa continuar. Para obter mais informações sobre essas verificações de integridade, consulte [Verificações de integridade para instâncias do Auto Scaling \(p. 319\)](#).

Instâncias protegidas contra redução da escala na horizontal (**ScaleInProtectedInstances**)

Controla o que o Amazon EC2 Auto Scaling faz se forem encontradas instâncias protegidas contra escalabilidade. Para obter mais informações sobre essas instâncias, consulte [Usar proteção de redução na escala na horizontal de instâncias \(p. 302\)](#).

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 \(p. 125\)](#).

Instâncias em espera (**StandbyInstances**)

Controla o que o Amazon EC2 Auto Scaling faz se forem encontradas instâncias em Standby estado. Para obter mais informações sobre essas instâncias, consulte [Remover temporariamente instâncias do grupo do Auto Scaling \(p. 308\)](#).

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.

Verificar o status de uma atualização de instância

Depois que uma atualização de instância for iniciada, você poderá obter o status usando o AWS Management Console ou a AWS CLI.

Tip

No procedimento a seguir, você visualiza as seções Instance refresh history (Histórico de atualizações da instância), Activity history (Histórico de atividades) e Instances (Instâncias) do grupo do Auto Scaling. 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 grupos do Auto Scaling 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 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. 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. Na coluna Descrição você pode encontrar o ID da instância.
5. (Opcional) Caso tenha muitas atividades de escalabilidade, você poderá ver mais delas escolhendo o ícone > na parte superior do histórico de atividades.
6. Na guia Gerenciamento de instâncias, em Instâncias, é possível verificar se as instâncias foram executadas com êxito. Inicialmente, suas instâncias estão em um estado pendente enquanto aguardam a conclusão de qualquer ação definida pelos ganchos do ciclo de vida. Depois disso, a instância é adicionada ao grupo Auto Scaling e seu estado é InService. A coluna Health Status (Status de integridade) mostra o resultado das verificações de integridade em suas instâncias.

Como verificar o status de uma atualização de instância (AWS CLI)

Veja as atualizações de instância para um grupo de Auto Scaling usando o seguinte [describe-instance-refreshes](#) comando.

```
aws autoscaling describe-instance-refreshes --auto-scaling-group-name my-asg
```

Exemplos de resultado:

```
{  
    "InstanceRefreshes": [  
        {  
            "InstanceRefreshId": "08b91cf7-8fa6-48af-b6a6-d227f40f1b9b",  
            "AutoScalingGroupName": "my-asg",  
            "Status": "InProgress",  
            "StatusReason": "Waiting for instances to warm up before continuing. For example: 0e69cc3f05f825f4f is warming up.",  
            "EndTime": "2023-03-23T16:42:55Z",  
            "PercentageComplete": 0,  
            "InstancesToUpdate": 0,  
            "Preferences": {  
                "MinHealthyPercentage": 100,  
                "InstanceWarmup": 300,  
                "CheckpointPercentages": [  
                    50  
                ],  
                "CheckpointDelay": 3600,  
                "SkipMatching": false,  
                "AutoRollback": true,  
                "ScaleInProtectedInstances": "Ignore",  
                "StandbyInstances": "Ignore"  
            }  
        },  
        {  
            "InstanceRefreshId": "dd7728d0-5bc4-4575-96a3-1b2c52bf8bb1",  
            "AutoScalingGroupName": "my-asg",  
            "Status": "Successful",  
            "StatusReason": "All instances have been successfully updated.",  
            "EndTime": "2023-03-23T16:42:55Z",  
            "PercentageComplete": 100,  
            "InstancesToUpdate": 0,  
            "Preferences": {  
                "MinHealthyPercentage": 100,  
                "InstanceWarmup": 300,  
                "CheckpointPercentages": [  
                    50  
                ],  
                "CheckpointDelay": 3600,  
                "SkipMatching": false,  
                "AutoRollback": true,  
                "ScaleInProtectedInstances": "Ignore",  
                "StandbyInstances": "Ignore"  
            }  
        }  
    ]  
}
```

```
"AutoScalingGroupName": "my-asg",
"Status": "Successful",
"EndTime": "2022-06-02T16:53:37Z",
"PercentageComplete": 100,
"InstancesToUpdate": 0,
"Preferences": {
    "MinHealthyPercentage": 90,
    "InstanceWarmup": 300,
    "SkipMatching": true,
    "AutoRollback": true,
    "ScaleInProtectedInstances": "Ignore",
    "StandbyInstances": "Ignore"
}
}
```

Status de atualização de instância

Quando uma atualização de instância é iniciada, ela entra no status Pending. Passa de Pendente para InProgress até chegar bem sucedido, Falhou, Cancelado, RollbackSuccessful, ou RollbackFailed.

A atualização de instância pode ter os seguintes status:

Status	Descrição
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.
Failed	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.

Cancelar uma atualização de instância

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 \(p. 122\)](#).

Tópicos

- [Cancelar uma atualização de instância \(console\) \(p. 122\)](#)
- [Cancelar uma atualização de instância \(AWS CLI\) \(p. 122\)](#)

CANCELAR UMA ATUALIZAÇÃO DE INSTÂNCIA (CONSOLE)

Para cancelar uma atualização de instância

1. Abra o console do Amazon EC2 em <https://console.aws.amazon.com/ec2/> e escolha grupos do Auto Scaling 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 Confirm (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)

Para cancelar uma atualização de instância

Use o [cancel-instance-refresh](#) comando do AWS CLI e forneça o nome do grupo Auto Scaling.

```
aws autoscaling cancel-instance-refresh --auto-scaling-group-name my-asg
```

Exemplos de resultado:

```
{  
    "InstanceRefreshId": "08b91cf7-8fa6-48af-b6a6-d227f40f1b9b"  
}
```

DEFINIR ALTERAÇÕES COM UMA REVERSÃO

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 houver algum CloudWatch alarms que você especificar vão para o ALARMEstado.

Índice

- [Considerações \(p. 122\)](#)
- [Iniciar manualmente uma reversão \(p. 123\)](#)
- [Iniciar uma atualização de instância com reversão automática \(p. 123\)](#)

CONSIDERAÇÕES

As seguintes considerações se aplicam ao usar uma reversão:

- A opção de reversão só está disponível se você especificar uma configuração desejada como parte do início de uma atualização da instância.
- Você só pode reverter para uma versão anterior de um modelo de lançamento se a versão for uma versão numerada específica. A opção de reversão não estará disponível se o grupo Auto Scaling estiver configurado para usar o \$Latest ou \$Default versão do modelo de lançamento.
- Você também não pode reverter para um modelo de execução configurado para usar um alias de AMI do AWS Systems Manager Armazenamento de parâmetros.
- A configuração que você salvou pela última vez no grupo Auto Scaling deve estar em um estado estável. Se não estiver em um estado estável, o fluxo de trabalho de reversão ainda ocorrerá, mas acabará falhando. Até você resolver o problema, o grupo do Auto Scaling poderá estar em um estado de falha e não conseguir mais executar instâncias com êxito. Isso pode afetar a disponibilidade do serviço ou da aplicação.

Iniciar manualmente uma reversão

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 grupos do Auto Scaling 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 Confirm (Confirmar).

AWS CLI

Para iniciar manualmente a reversão de uma atualização de instância (AWS CLI)

Use o `rollback-instance-refresh` comando do AWS CLI e forneça o nome do grupo Auto Scaling.

```
aws autoscaling rollback-instance-refresh --auto-scaling-group-name my-asg
```

Exemplos de resultado:

```
{  
    "InstanceRefreshId": "08b91cf7-8fa6-48af-b6a6-d227f40f1b9b"  
}
```

Tip

Se esse comando lançar um erro, verifique se você atualizou a 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 uma Amazon CloudWatch alarme entra no ALARM estado.

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 na inicialização do EC2, verificações de integridade mal configuradas ou

por não ignorar ou permitir o encerramento de instâncias que estão em Standby estado ou protegido da escala em.

Especificando CloudWatch alarms são opcionais. 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 a [Amazônia CloudWatch Guia do usuário](#). Usando métricas do Elastic Load Balancing como exemplo, se você usar um Application Load Balancer, poderá usar o HTTPCode_ELB_5XX_Count e HTTPCode_ELB_4XX_Count métricas.

Considerações

- Se você especificar um CloudWatch alarme, mas não habilite a reversão automática e o estado do alarme vai para ALARM, a atualização da instância falha sem ser revertida.
- Você pode escolher no máximo 10 alarmes ao iniciar uma atualização de 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 recebe um erro ao tentar iniciar a atualização da instância.
- Ao criar um alarme para o Amazon EC2 Auto Scaling usar, 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 as instâncias até que novos pontos de dados sejam encontrados. Para forçar o alarme a manter o anterior ALARM ou OK estado, você pode optar por ignorar os dados ausentes em vez disso. Para obter mais informações, consulte [Configurando como os alarmes tratam os dados perdidos](#) na Amazônia CloudWatch Guia do usuário.

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 grupos do Auto Scaling no painel de navegação.
2. Marque a caixa de seleção ao lado do grupo 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. Siga o [Iniciar uma atualização de instância \(console\) \(p. 112\)](#) procedimento e defina as configurações de atualização da instância conforme necessário.
5. (Opcional) Abaixo Atualizar configurações, para CloudWatch alarme, escolha Ativar CloudWatch alarmes, em seguida, escolha um ou mais alarmes para identificar quaisquer problemas e falhar na operação se um alarme for acionado no ALARM ou OK estado.
6. Abaixo Configurações de reversão, escolha Ativar reversão automática para reverter automaticamente uma atualização de instância com falha para a configuração que você salvou pela última vez no grupo de Auto Scaling antes de iniciar a atualização da instância.
7. Revise suas seleções e, em seguida, escolha Iniciar atualização da instância.

AWS CLI

Para iniciar uma atualização de instância com reversão automática (AWS CLI)

Use o `start-instance-refresh` comando e especifique `true` para o `AutoRollback` opção no `Prefeferences`.

O exemplo a seguir mostra como iniciar uma atualização de instância que será revertida automaticamente se algo falhar. Substitua os *italicized* valores de parâmetros com os 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 uma especificação CloudWatch alarme está no ALARM estado, especifique o AlarmSpecification opção no Preferences se forneça o nome do alarme, como no exemplo a seguir. Substitua os *italicized* valores de parâmetros com os 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 lançar um erro, verifique se você atualizou a 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.
- Durante uma atualização de instância com a opção de ignorar correspondência, se for especificado \$Default ou \$Latest para o modelo de execução, o Amazon EC2 Auto Scaling reavaliará a versão do modelo de execução à medida que cada instância for substituída. Portanto, se você criar uma nova versão de seu modelo de execução enquanto as instâncias ainda estiverem sendo substituídas, o Amazon EC2 Auto Scaling poderá acabar usando a nova versão do modelo de execução. Isso pode fazer com que instâncias substituídas anteriormente sejam substituídas novamente, pois o Amazon EC2 Auto Scaling aplicará a nova versão em todo o grupo.

Essa seção da AWS CLI contém exemplos para iniciar uma atualização de instância com a opção ignorar correspondência habilitada. Para obter instruções sobre como usar o console, consulte [Iniciar uma atualização de instância \(console\) \(p. 112\)](#).

Ignorar correspondência (procedimento básico)

Siga as etapas desta seção para usar a 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 `ocreate-launch-template` comando 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 criação e gerenciamento de modelos de execução com a AWS Command Line Interface \(AWS CLI\) \(p. 37\)](#).

2. Use `start-instance-refresh` comando para iniciar o fluxo de trabalho de substituição da instância e aplicar seu novo modelo de execução 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 do Auto Scaling com uma [política de instâncias mistas \(p. 67\)](#), siga as etapas desta seção para usar a AWS CLI para iniciar uma atualização de instância com a opção de ignorar correspondência. 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-template](#) comando 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 criação e gerenciamento de modelos de execução com a AWS Command Line Interface \(AWS CLI\) \(p. 37\)](#).

2. Para ver a política de instâncias mistas existente para seu grupo de Auto Scaling, execute o [describe-auto-scaling-groups](#) comando. 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 `start-instance-refresh` comando para iniciar o fluxo de trabalho de substituição da instância e aplicar seu novo modelo de execução 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-groups`](#) comando. 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": 0,
                    "SpotAllocationStrategy": "prioritized",
                    "SpotBaseCapacity": 0,
                    "SpotPercentageAboveBaseCapacity": 0
                }
            }
        }
    ]
}
```

```
        "OnDemandBaseCapacity":1,  
        "OnDemandPercentageAboveBaseCapacity":50,  
        "SpotAllocationStrategy":"price-capacity-optimized"  
    }  
},  
"MinSize":1,  
"MaxSize":5,  
"DesiredCapacity":4,  
...  
]  
]  
}
```

2. Use o [start-instance-refresh](#) comando para iniciar o fluxo de trabalho de substituição da instância 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 \(p. 68\)](#) 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  
    }  
},  
"MinSize":1,  
"MaxSize":5,  
"DesiredCapacity":4,  
...  
]  
]  
}
```

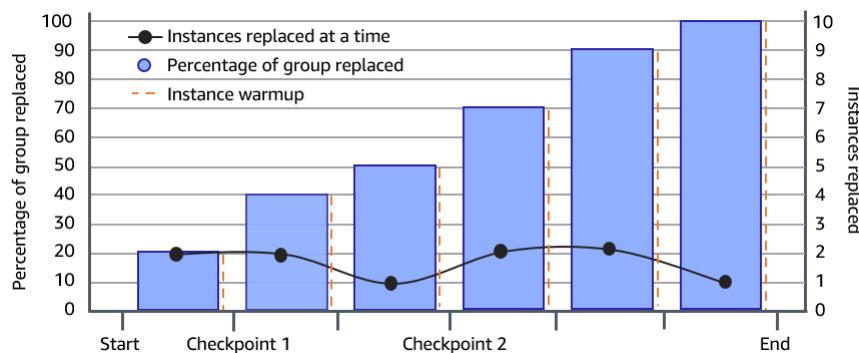
}

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.

O Amazon EC2 Auto Scaling emite eventos para cada ponto de verificação. Você pode adicionar um EventBridge regra para enviar os eventos para um alvo como o Amazon SNS. Assim, você é notificado quando pode executar as verificações necessárias. Para obter mais informações, consulte [Criar EventBridge regras para eventos de atualização de instância \(p. 411\)](#).

Considere o seguinte grupo de Auto Scaling que tem 10 instâncias. As porcentagens do posto de controle são [20, 50, 100] e a porcentagem mínima de saúde é de 80%. Para manter a porcentagem mínima de integridade, somente duas instâncias podem ser substituídas por vez. 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 permite. O diagrama a seguir resume o processo de substituição de instâncias antes que um ponto de verificação seja atingido.



No exemplo acima, há um período de aquecimento da instância para cada nova instância iniciada. Você também pode ter um gancho de ciclo de vida que coloca uma instância em estado de espera e, em seguida, executa uma ação personalizada enquanto ela é iniciada ou encerrada.

Índice

- [Considerações \(p. 132\)](#)
- [Habilitar pontos de verificação \(console\) \(p. 133\)](#)
- [Habilitar pontos de verificação \(AWS CLI\) \(p. 134\)](#)

Considerações

Mantenha as seguintes considerações em mente ao usar pontos de verificação:

- Um ponto de verificação é atingido quando uma porcentagem especificada do número total de instâncias no grupo Auto Scaling é substituída. A porcentagem é a porcentagem mínima do grupo de Auto Scaling que você deseja que o ponto de verificação represente. 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 posto de controle 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.

Por exemplo, suponha que seu grupo do Auto Scaling tenha 10 instâncias. 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%. Seu grupo faz 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%.

- 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.

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

- Abra o console do Amazon EC2 em <https://console.aws.amazon.com/ec2/> e escolha grupos do Auto Scaling no painel de navegação.
- 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).

- 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 suas seleções de atualização de instância, escolha Iniciar atualização da instância.

Habilitar pontos de verificação (AWS CLI)

Para iniciar uma atualização de instância com pontos de verificação habilitados usando a 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 faz a transição para Successful depois que essa porcentagem for substituída com sucesso e cada instância for considerada concluída.

Para criar vários pontos de verificação

Para criar vários pontos de verificação, use o exemplo a seguir `start-instance-refresh` comando. 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 exemplo a seguir[start-instance-refresh](#) comando. 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 exemplo a seguir[start-instance-refresh](#) comando. 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.

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 gerenciar a taxa de substituição, você pode fazer o seguinte:

- Defina um período mais longo para o limite de tempo de vida máximo da instância. Isso espaça as substituições, o que é útil para grupos que têm um grande número de instâncias a substituir.
- Adicione mais tempo entre determinadas substituições usando a proteção de instâncias. Isso impede temporariamente que instâncias individuais no grupo do Auto Scaling sejam substituídas. Quando estiver pronto para substituir essas instâncias, remova a proteção de instâncias de cada instância individual. Para obter mais informações, consulte [Usar proteção de redução na escala na horizontal de instâncias \(p. 302\)](#).

Note

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 AMI de uma versão diferente da aplicação, essa versão da aplicação será implantada automaticamente.

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 grupos do Auto Scaling 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. Escolha 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 a AWS CLI para configurar o tempo de vida máximo da instância para grupos do Auto Scaling novos ou existentes.

Para novos grupos de Auto Scaling, use o [create-auto-scaling-group](#) comando.

```
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": "$Latest"  
    },  
    "MinSize": 1,  
    "MaxSize": 5,  
    "MaxInstanceLifetime": 2592000,  
    "VPCZoneIdentifier": "subnet-5ea0c127,subnet-6194ea3b,subnet-c934b782",  
    "Tags": []  
}
```

Para grupos existentes de Auto Scaling, use o [update-auto-scaling-group](#) comando.

```
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
```

Esta é uma resposta de exemplo.

```
{  
    "AutoScalingGroups": [  
        {  
            "AutoScalingGroupName": "my-asg",  
            "AutoScalingGroupARN": "arn",  
            "LaunchTemplate": {  
                "LaunchTemplateId": "lt-0b97f1e282EXAMPLE",  
                "LaunchTemplateName": "my-launch-template",  
                "Version": "$Latest"  
            },  
            "MinSize": 1,  
            "MaxSize": 5,  
            "DesiredCapacity": 1,  
            "DefaultCooldown": 300,  
            "AvailabilityZones": [  
                "us-west-2a",  
                "us-west-2b",  
                "us-west-2c"  
            ],  
            "LoadBalancerNames": [],  
            "TargetGroupARNs": [],  
            "HealthCheckType": "EC2",  
            "HealthCheckGracePeriod": 0,  
            "Instances": [  
                {  
                    "InstanceId": "i-04d180b9d5fc578fc",  
                    "InstanceType": "t2.small",  
                    "AvailabilityZone": "us-west-2b",  
                    "LifecycleState": "Pending",  
                    "HealthStatus": "Healthy",  
                    "LaunchTemplate": {  
                        "LaunchTemplateId": "lt-0b97f1e282EXAMPLE",  
                        "Version": "$Latest"  
                    }  
                }  
            ]  
        }  
    ]  
}
```

```
        "LaunchTemplateName": "my-launch-template",
        "Version": "7"
    },
    "ProtectedFromScaleIn": false
}
],
"CreatedTime": "2019-11-14T22:56:15.487Z",
"SuspendedProcesses": [],
"VPCZoneIdentifier": "subnet-5ea0c127,subnet-6194ea3b,subnet-c934b782",
"EnabledMetrics": [],
"Tags": [],
"TerminationPolicies": [
    "Default"
],
"NewInstancesProtectedFromScaleIn": false,
"ServiceLinkedRoleARN": "arn",
"MaxInstanceLifetime": 2592000
}
]
```

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.
- Instâncias terminadas antes de iniciar: quando há apenas uma instância no grupo do Auto Scaling, o recurso de tempo de vida máximo da instância pode resultar em uma interrupção porque o Amazon EC2 Auto Scaling termina uma instância e, em seguida, inicia uma nova.

Etiquetar grupos e instâncias do Auto Scaling

Uma tag é um rótulo de atributo personalizado que você ou a AWS atribui a um recurso da AWS. Cada tag 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 você a fazer o seguinte:

- Monitorar seus custos da AWS. Você pode ativar essas tags no painel do AWS Billing and Cost Management. A 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 \(p. 142\)](#).
- 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 \(p. 144\)](#).
- Identificar e organizar seus recursos da AWS. Muitos Serviços da AWS são compatíveis com marcação, permitindo que você atribua a mesma tag a recursos de diferentes serviços 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 \(p. 23\)](#).

Você pode criar e gerenciar tags do pelo AWS Management Console, AWS CLI ou SDKs.

Índice

- [Restrições de nomeação e uso de tags \(p. 139\)](#)
- [Ciclo de vida de marcação de instâncias do EC2 \(p. 139\)](#)
- [Marcar seus grupos do Auto Scaling \(p. 140\)](#)
- [Excluir tags \(p. 142\)](#)
- [Etiquetas para segurança \(p. 142\)](#)
- [Controlar o acesso usando etiquetas \(p. 143\)](#)
- [Usar etiquetas para filtrar grupos do Auto Scaling \(p. 144\)](#)

Restrições de nomeação e uso de tags

As restrições básicas a seguir se aplicam às 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 comprimento máximo do valor é 256 caracteres Unicode.
- As chaves e os valores de tags 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 prefixo aws : no nome nem no valor de suas tags, pois ele é reservado para uso da AWS. 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.

Índice

- [Adicionar ou modificar tags \(console\) \(p. 140\)](#)
- [Adicionar ou modificar tags \(AWS CLI\) \(p. 140\)](#)

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 grupos do Auto Scaling 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 a AWS CLI para adicionar tags ao criar grupos do Auto Scaling e para adicionar ou modificar tags de grupos do Auto Scaling existentes.

Para marcar um grupo do Auto Scaling na criação

Use o `create-auto-scaling-group` comando para criar um novo grupo de Auto Scaling e adicionar uma tag, por exemplo, `environment=production`, para o 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
```

Esta é uma resposta de exemplo.

```
{  
    "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
```

Esta é uma resposta de exemplo.

```
{  
    "AutoScalingGroups": [  
        {  
            "AutoScalingGroupARN": "arn",  
            "HealthCheckGracePeriod": 0,  
            "SuspendedProcesses": [],  
            "DesiredCapacity": 1,  
            "Tags": [  
                {  
                    "ResourceType": "auto-scaling-group",  
                    "ResourceId": "my-asg",  
                    "PropagateAtLaunch": true,  
                    "Value": "production",  
                    "Key": "environment"  
                }  
            ]  
        }  
    ]  
}
```

```
    ],
    "EnabledMetrics": [],
    "LoadBalancerNames": [],
    "AutoScalingGroupName": "my-asg",
    ...
}
]
```

Excluir tags

Você pode excluir uma tag associada a seu grupo do Auto Scaling a qualquer momento.

Índice

- [Excluir tags \(console\) \(p. 142\)](#)
- [Excluir tags \(AWS CLI\) \(p. 142\)](#)

Excluir tags (console)

Para excluir uma tag

1. Abra o console do Amazon EC2 em <https://console.aws.amazon.com/ec2/> e escolha grupos do Auto Scaling 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. Escolha 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"
```

Você deve 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.

```
{  
    "Version": "2012-10-17",  
    "Statement": [  
        {  
            "Effect": "Deny",  
            "Action": [  
                "autoscaling:CreateAutoScalingGroup",  
                "autoscaling:UpdateAutoScalingGroup",  
                "autoscaling:DeleteAutoScalingGroup"  
            ],  
            "Resource": "*",  
            "Condition": {  
                "StringEquals": {"autoscaling:ResourceTag/environment": "production"}  
            }  
        }  
    ]  
}
```

Para obter mais exemplos, consulte [Exemplos de políticas baseadas em identidade do Amazon EC2 Auto Scaling \(p. 437\)](#).

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.

Por exemplo, você pode criar uma política do IAM que permita remover somente a tag com a chave **temporary** dos grupos do Auto Scaling.

```
{  
    "Version": "2012-10-17",  
    "Statement": [  
        {  
            "Effect": "Allow",  
            "Action": "autoscaling:DeleteTags",  
            "Resource": "*",  
            "Condition": {  
                "ForAllValues:StringEquals": { "aws:TagKeys": ["temporary"] }  
            }  
        }  
    ]  
}
```

Para obter mais exemplos, consulte [Exemplos de políticas baseadas em identidade do Amazon EC2 Auto Scaling \(p. 437\)](#).

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 executadas. Para exemplos que controlam o acesso a tags em instâncias do EC2, consulte [Exemplo: marcação de recursos](#) no Manual do usuário do Amazon EC2 para instâncias do Linux.

Usar etiquetas para filtrar grupos do Auto Scaling

Os exemplos a seguir mostram como usar filtros com o comando `describe-auto-scaling-groups` para descrever grupos de Auto Scaling com tags específicas. A filtragem por etiquetas é limitada à AWS CLI ou 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
```

Esta é uma resposta de exemplo.

```
{
    "AutoScalingGroups": [
        {
            "AutoScalingGroupARN": "arn",
            "HealthCheckGracePeriod": 0,
            "SuspendedProcesses": [],
            "DesiredCapacity": 1,
            "Tags": [
                {
                    "ResourceType": "auto-scaling-group",
                    "ResourceId": "my-asg",
                    "PropagateAtLaunch": true,
                    "Value": "production",
                    "Key": "environment"
                }
            ],
            "EnabledMetrics": [],
            "LoadBalancerNames": [],
            "AutoScalingGroupName": "my-asg",
            ...
        }
    ]
}
```

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
```

Também é possível filtrar a saída AWS CLI usando a opção `--query`. O exemplo a seguir mostra como limitar saída AWS CLI para o comando anterior apenas para o 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]"
```

Esta é uma resposta de exemplo.

```
[  
  {  
    "AutoScalingGroupName": "my-asg",  
    "MinSize": 0,  
    "MaxSize": 10,  
    "DesiredCapacity": 1  
  }  
  ...  
]
```

Para obter mais informações sobre filtragem, consulte [Filtragem da saída da AWS CLI](#) no Guia do usuário da AWS Command Line Interface.

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
```

Excluir infraestrutura do Auto Scaling

Para excluir completamente sua infraestrutura de escalabilidade, execute as tarefas a seguir.

Tarefas

- [Excluir seu grupo do Auto Scaling \(p. 146\)](#)
- [\(Opcional\) Excluir a configuração de execução \(p. 147\)](#)
- [\(Opcional\) Excluir o modelo de execução \(p. 147\)](#)
- [\(Opcional\) Excluir o平衡ador de carga e grupos de destino \(p. 148\)](#)
- [\(Opcional\) Excluir CloudWatchalarms \(p. 149\)](#)

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 grupos do Auto Scaling no painel de navegação.

2. Marque a caixa de seleção ao lado do seu grupo de 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 Desired (Desejado), Min (Mínimo) e Max (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 seguinte [delete-auto-scaling-group](#) comando 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 esquerdo, em Escalabilidade automática, escolha Grupos de Auto Scaling.
3. Escolha Configurações de lançamento próximo ao topo da página. Quando solicitada a confirmação, escolha Exibir configurações de lançamento para confirmar que você deseja visualizar as configurações de lançamento na página.
4. Selecione sua configuração de lançamento e escolha Ações, Excluir configuração de lançamento.
5. Quando a confirmação for solicitada, escolha Delete (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 Yes, Delete (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 Yes, Delete (Sim, excluir).

Para excluir o balanceador de carga associado ao grupo do Auto Scaling (AWS CLI)

Para平衡adores de carga de aplicativos e平衡adores de carga de rede, use o seguinte [delete-load-balancer](#) e [delete-target-group](#) comando.

```
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平衡adores de carga clássicos, use o seguinte [delete-load-balancer](#) comando.

```
aws elb delete-load-balancer --load-balancer-name my-load-balancer
```

(Opcional) ExcluirCloudWatchalarms

Para excluir oCloudWatchalarms associados ao seu grupo de Auto Scaling, conclua as etapas a seguir. Por exemplo, você pode ter alarmes associados a políticas de escalonamento de etapas ou de escalabilidade simples.

Note

A exclusão de um grupo de Auto Scaling exclui automaticamente oCloudWatchalarms que o Amazon EC2 Auto Scaling gerencia para uma política de escalabilidade de rastreamento de alvos.

Você pode pular essa etapa se seu grupo de Auto Scaling não estiver associado a nenhumCloudWatchalarms, ou se você quiser manter os alarmes para uso futuro.

Para excluir os alarmes do CloudWatch (console)

1. Abra o console do CloudWatch 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 Delete (Excluir).

Para excluir os alarmes do CloudWatch (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
```

Exemplos de criação e gerenciamento de grupos de Auto Scaling com oAWSSDKs

Você pode criar um grupo de Auto Scaling usando oAWS Management Console, oAWS CLI, eAWSSDK eAWS CloudFormation.

Os exemplos de código a seguir mostram como criar, atualizar e excluir um grupo de Auto Scaling em sua linguagem de programação compatível favorita usando oAWSSDKs.

Índice

- [Crie um grupo de Auto Scaling usando umAWSSDK \(p. 149\)](#)
- [Atualize um grupo de Auto Scaling usando umAWSSDK \(p. 155\)](#)
- [Exclua um grupo de Auto Scaling usando umAWSSDK \(p. 159\)](#)
- [Recursos relacionados \(p. 163\)](#)

Crie um grupo de Auto Scaling usando umAWSSDK

Os exemplos de código a seguir mostram como criar um grupo de Auto Scaling.

.NET

AWS SDK for .NET

Note

Há mais sobreGitHub. Encontre o exemplo completo e saiba como configurar e executar no [Repositório de exemplos de código da AWS](#).

```
/// <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
    };

    var response = await
    _amazonAutoScaling.CreateAutoScalingGroupAsync(request);
    Console.WriteLine($"{groupName} Auto Scaling Group created");
    return response.HttpStatusCode == System.Net.HttpStatusCode.OK;
}
```

- Para obter detalhes da API, consulte[CreateAutoScalingGroup](#)emAWS SDK for .NETReferência da API.

C++

SDK para C++

Note

Há mais sobreGitHub. Encontre o exemplo completo e saiba como configurar e executar no [Repositório de exemplos de código da AWS](#).

```
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().GetErrorCode() ==
    Aws::AutoScaling::AutoScalingErrors::ALREADY_EXISTSFAULT) {
    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 [CreateAutoScalingGroup](#) em AWS SDK for C++ Referência da API.

Java

SDK para Java 2.x

Note

Há mais sobre GitHub. Encontre o exemplo completo e saiba como configurar e executar no [Repositório de exemplos de código da AWS](#).

```
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 [CreateAutoScalingGroup](#) em AWS SDK for Java 2.x Referência da API.

Kotlin

SDK para Kotlin

Note

Essa documentação é de pré-lançamento para um recurso em versão de pré-visualização. Está sujeita a alteração.

Note

Há mais sobre GitHub. Encontre o exemplo completo e saiba como configurar e executar no [Repositório de exemplos de código da AWS](#).

```
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[CreateAutoScalingGroup](#)em AWSSDK para referência da API Kotlin.

PHP

SDK para PHP

Note

Há mais sobreGitHub. Encontre o exemplo completo e saiba como configurar e executar no [Repositório de exemplos de código da AWS](#).

```
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[CreateAutoScalingGroup](#)em AWS SDK for PHP Referência da API.

Python

SDK para Python (Boto3).

Note

Há mais sobreGitHub. Encontre o exemplo completo e saiba como configurar e executar no [Repositório de exemplos de código da AWS](#).

```
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, group_zones, launch_template_name, min_size,
        max_size):
        """
        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.
        """
        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
            )
        except ClientError as err:
            logger.error(
                "Couldn't create group %s. Here's why: %s: %s",
                group_name,
                err.response['Error']['Code'], err.response['Error']['Message'])
            raise
```

- Para obter detalhes da API, consulte [CreateAutoScalingGroup](#) em AWS Referência da API SDK para Python (Boto3).

Rust

SDK para Rust

Note

Esta documentação destina-se a um SDK na versão de pré-visualização. O SDK está sujeito a alterações e não deve ser usado em ambientes de produção.

Note

Há mais sobre GitHub. Encontre o exemplo completo e saiba como configurar e executar no [Repositório de exemplos de código da AWS](#).

```
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 [CreateAutoScalingGroup](#) em AWSSDK para referência da API Rust.

Atualize um grupo de Auto Scaling usando umAWSSDK

Os exemplos de código a seguir mostram como atualizar a configuração de um grupo de Auto Scaling.

.NET

AWS SDK for .NET

Note

Há mais sobre GitHub. Encontre o exemplo completo e saiba como configurar e executar no [Repositório de exemplos de código da AWS](#).

```
/// <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[UpdateAutoScalingGroup](#)emAWS SDK for .NETReferência da API.

C++

SDK para C++

Note

Há mais sobreGitHub. Encontre o exemplo completo e saiba como configurar e executar no [Repositório de exemplos de código da AWS](#).

```
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[UpdateAutoScalingGroup](#)emAWS SDK for C++Referência da API.

Java

SDK para Java 2.x

Note

Há mais sobreGitHub. Encontre o exemplo completo e saiba como configurar e executar no [Repositório de exemplos de código da AWS](#).

```
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 [UpdateAutoScalingGroup](#) em AWS SDK for Java 2.x Referência da API.

Kotlin

SDK para Kotlin

Note

Essa documentação é de pré-lançamento para um recurso em versão de pré-visualização. Está sujeita a alteração.

Note

Há mais sobre GitHub. Encontre o exemplo completo e saiba como configurar e executar no [Repositório de exemplos de código da AWS](#).

```
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[UpdateAutoScalingGroup](#)em AWSSDK para referência da API Kotlin.

PHP

SDK para PHP

Note

Há mais sobreGitHub. Encontre o exemplo completo e saiba como configurar e executar no [Repositório de exemplos de código da AWS](#).

```
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[UpdateAutoScalingGroup](#)em AWS SDK for PHPReferência da API.

Python

SDK para Python (Boto3).

Note

Há mais sobreGitHub. Encontre o exemplo completo e saiba como configurar e executar no [Repositório de exemplos de código da AWS](#).

```
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, **kwargs):  
    """  
    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.  
    """  
    try:  
        self.autoscaling_client.update_auto_scaling_group(  
            AutoScalingGroupName=group_name, **kwargs)  
    except ClientError as err:  
        logger.error(  
            "Couldn't update group %s. Here's why: %s: %s", group_name,  
            err.response['Error']['Code'], err.response['Error']['Message'])  
        raise
```

- Para obter detalhes da API, consulte [UpdateAutoScalingGroup](#) em AWS Referência da API SDK para Python (Boto3).

Rust

SDK para Rust

Note

Esta documentação destina-se a um SDK na versão de pré-visualização. O SDK está sujeito a alterações e não deve ser usado em ambientes de produção.

Note

Há mais sobre [GitHub](#). Encontre o exemplo completo e saiba como configurar e executar no [Repositório de exemplos de código da AWS](#).

```
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 [UpdateAutoScalingGroup](#) em AWSSDK para referência da API Rust.

Exclua um grupo de Auto Scaling usando um AWSSDK

Os exemplos de código a seguir mostram como excluir um grupo de Auto Scaling.

.NET

AWS SDK for .NET

Note

Há mais sobre GitHub. Encontre o exemplo completo e saiba como configurar e executar no [Repositório de exemplos de código da AWS](#).

```
/// <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 [DeleteAutoScalingGroup](#) em AWS SDK for .NET Referência da API.

C++

SDK para C++

Note

Há mais sobre GitHub. Encontre o exemplo completo e saiba como configurar e executar no [Repositório de exemplos de código da AWS](#).

```
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[DeleteAutoScalingGroup](#)emAWS SDK for C++Referência da API.

Java

SDK para Java 2.x

Note

Há mais sobreGitHub. Encontre o exemplo completo e saiba como configurar e executar no [Repositório de exemplos de código da AWS](#).

```
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[DeleteAutoScalingGroup](#)emAWS SDK for Java 2.xReferência da API.

Kotlin

SDK para Kotlin

Note

Essa documentação é de pré-lançamento para um recurso em versão de pré-visualização. Está sujeita a alteração.

Note

Há mais sobreGitHub. Encontre o exemplo completo e saiba como configurar e executar no [Repositório de exemplos de código da AWS](#).

```
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[DeleteAutoScalingGroup](#)emAWSSDK para referência da API Kotlin.

PHP

SDK para PHP

Note

Há mais sobreGitHub. Encontre o exemplo completo e saiba como configurar e executar no [Repositório de exemplos de código da AWS](#).

```
public function deleteAutoScalingGroup($autoScalingGroupName)
{
    return $this->autoScalingClient->deleteAutoScalingGroup([
        'AutoScalingGroupName' => $autoScalingGroupName,
        'ForceDelete' => true,
    ]);
}
```

- Para obter detalhes da API, consulte[DeleteAutoScalingGroup](#)emAWS SDK for PHPReferência da API.

Python

SDK para Python (Boto3).

Note

Há mais sobreGitHub. Encontre o exemplo completo e saiba como configurar e executar no [Repositório de exemplos de código da AWS](#).

```
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 delete_group(self, group_name):
    """
    Deletes an Auto Scaling group. All instances must be stopped before the
    group can be deleted.

    :param group_name: The name of the group to delete.
    """
    try:
        self.autoscaling_client.delete_auto_scaling_group(
            AutoScalingGroupName=group_name)
    except ClientError as err:
        logger.error(
            "Couldn't delete group %s. Here's why: %s: %s",
            group_name,
            err.response['Error']['Code'],
            err.response['Error']['Message'])
        raise
```

- Para obter detalhes da API, consulte[DeleteAutoScalingGroup](#)emAWSReferência da API SDK para Python (Boto3).

Rust

SDK para Rust

Note

Esta documentação destina-se a um SDK na versão de pré-visualização. O SDK está sujeito a alterações e não deve ser usado em ambientes de produção.

Note

Há mais sobreGitHub. Encontre o exemplo completo e saiba como configurar e executar no [Repositório de exemplos de código da AWS](#).

```
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[DeleteAutoScalingGroup](#)emAWSSDK para referência da API Rust.

Recursos relacionados

Para obter mais exemplos e informações sobre outras propriedades que você pode usar ao criar grupos de Auto Scaling, consulte os recursos a seguir.

- [AWS SDK for .NET](#)
- [AWS SDK for C++](#)
- [AWS SDK for Go](#)
- [AWS SDK para Java V2](#)
- [AWSSDK paraJavaScript](#)
- [AWS SDK para PHP V3](#)
- [AWS SDK para Python](#)
- [AWS SDK for Ruby V3](#)

Para obter mais exemplos e informações sobre outras propriedades que você pode usar ao atualizar grupos do Auto Scaling, consulte os recursos a seguir.

- [AWS SDK for .NET](#)
- [AWS SDK for C++](#)
- [AWS SDK for Go](#)
- [AWS SDK para Java V2](#)
- [AWSSDK paraJavaScript](#)
- [AWS SDK para PHP V3](#)
- [AWS SDK para Python](#)
- [AWS SDK for Ruby V3](#)

Escalar o tamanho do grupo do Auto Scaling

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 função 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 paga apenas 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.

Índice

- [Opções de escalabilidade \(p. 165\)](#)
- [Definir limites de capacidade no grupo do Auto Scaling \(p. 166\)](#)
- [Manter um número fixo de instâncias em seu grupo do Auto Scaling \(p. 168\)](#)
- [Escalabilidade manual para o Amazon EC2 Auto Scaling \(p. 168\)](#)
- [Escalabilidade dinâmica para o Amazon EC2 Auto Scaling \(p. 178\)](#)
- [Escala preditiva para o Amazon EC2 Auto Scaling \(p. 222\)](#)
- [Escalabilidade programada para o Amazon EC2 Auto Scaling \(p. 247\)](#)
- [Ganchos do ciclo de vida do Amazon EC2 Auto Scaling \(p. 252\)](#)
- [Grupos de alta atividade do Amazon EC2 Auto Scaling \(p. 279\)](#)
- [Controlar quais instâncias do Auto Scaling serão terminadas durante uma redução de escala na horizontal \(p. 292\)](#)
- [Remover temporariamente instâncias do grupo do Auto Scaling \(p. 308\)](#)
- [Suspender e retomar um processo para um grupo do Auto Scaling \(p. 312\)](#)

Opções de escalabilidade

O Amazon EC2 Auto Scaling fornece várias formas de escalar seu grupo do Auto Scaling.

Manter níveis de instâncias atuais em todos os momentos

Você pode configurar seu grupo do Auto Scaling para manter um número especificado de instâncias em execução a todo momento. Para manter os níveis de instâncias atuais, o Amazon EC2 Auto Scaling executa uma verificação de integridade periódica nas instâncias em execução em um grupo do Auto Scaling. Quando o Amazon EC2 Auto Scaling localiza uma instância não íntegra, ele termina essa

instância e inicia uma nova. Para obter mais informações, consulte [Manter um número fixo de instâncias em seu grupo do Auto Scaling \(p. 168\)](#).

Dimensionar manualmente

A escalabilidade manual é a maneira mais básica para escalar seus recursos, onde você especifica apenas a alteração na capacidade máxima, mínima ou desejada de seu grupo do Auto Scaling. O Amazon EC2 Auto Scaling gerencia o processo de criação ou término de instâncias para manter a capacidade atualizada. Para obter mais informações, consulte [Escalabilidade manual para o Amazon EC2 Auto Scaling \(p. 168\)](#).

Escala baseada em uma programação

Escalabilidade por programação significa que as ações de escalabilidade são executadas automaticamente como uma função de 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 \(p. 247\)](#).

Escala com base em 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 reduções em resposta a mudanças nas condições, quando você não sabe quando essas condições mudarão. Você pode configurar o Amazon EC2 Auto Scaling para responder por você. Para obter mais informações, consulte [Escalabilidade dinâmica para o Amazon EC2 Auto Scaling \(p. 178\)](#).

Usar a escalabilidade preditiva

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 \(p. 222\)](#).

Definir limites de capacidade no grupo do Auto Scaling

Os limites de capacidade representam os tamanhos mínimo e máximo de 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 do Auto Scaling não pode diminuir sua capacidade desejada abaixo do limite de tamanho mínimo.

- Maximum capacity (Capacidade máxima): representa o tamanho máximo do grupo. Quando as políticas de escalabilidade estão definidas, um grupo do Auto Scaling não pode aumentar sua capacidade desejada acima do limite de tamanho máximo.

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 Details (Detalhes), visualize ou altere as configurações atuais de capacidade mínima, máxima e desejada. Para obter mais informações, consulte [Alterar o tamanho do grupo do Auto Scaling \(console\) \(p. 168\)](#).

Acima de painel Details (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.

Auto Scaling groups (1/1)								
		Name	Launch template...	Instances	Status	Desired...	Min	Max
<input checked="" type="checkbox"/>	my-asg	my_template	Version Def	0	Updating capacity	1	0	1

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 instâncias que estão em execução no momento aa guia Instance management (Gerenciamento de instâncias), em Instances (Instâncias).

Note

Você pode usar o Service Quotas para atualizar os limites de capacidade total para instâncias do EC2 e de outros recursos em sua Conta da AWS. No console do Service Quotas, você pode visualizar 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 do Service Quotas.

Manter um número fixo de instâncias em seu grupo do Auto Scaling

O Amazon EC2 Auto Scaling permite que você configure um grupo do Auto Scaling para manter um tamanho fixo. Em seguida, é possível escolher se ajusta a capacidade desejada do grupo ou adiciona ou remove manualmente instâncias do Amazon EC2 do grupo para processar as alterações de tráfego em sua aplicação.

Se for necessário um número fixo de instâncias, isso poderá ser obtido com a definição do mesmo valor para as capacidades mínima, máxima e desejada. Depois de você ter criado seu grupo do Auto Scaling, o grupo começa iniciando instâncias suficientes para atender à sua capacidade desejada. Se não houver outras condições de escalabilidade anexadas ao grupo do Auto Scaling, o grupo sempre manterá esse número de instâncias em execução.

Seu grupo do Auto Scaling continua a manter um número fixo de instâncias, mesmo que uma instância se torne não íntegra. O Amazon EC2 Auto Scaling monitora a integridade de cada instância do Auto Scaling. Ao encontrar uma instância que não está mais íntegra, ele encerra essa instância e executa uma nova. As instâncias podem falhar em uma verificação de integridade por vários motivos. Para obter mais informações, consulte [Verificações de integridade para instâncias do Auto Scaling \(p. 319\)](#).

Escalabilidade manual para o Amazon EC2 Auto Scaling

A qualquer momento, você pode alterar o tamanho de um grupo do Auto Scaling manualmente. Atualize a capacidade desejada do grupo do Auto Scaling ou atualize as instâncias que estão conectadas ao grupo do Auto Scaling. A escalabilidade manual do seu grupo pode ser útil quando a escalabilidade automática não é necessária ou quando você precisa manter a capacidade em um número fixo de instâncias.

Alterar o tamanho do grupo do Auto Scaling (console)

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.

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

1. Abra o console do Amazon EC2 em <https://console.aws.amazon.com/ec2/> e escolha grupos do Auto Scaling 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, se o valor atual for 1, insira 2.

A capacidade desejada deve ser menor ou igual ao tamanho máximo do grupo. Se o novo valor para Desired capacity (Capacidade desejada) for maior que Maximum capacity (Capacidade máxima), será necessário atualizar Maximum capacity (Capacidade máxima).

5. Quando terminar, escolha Atualizar.

Agora, verifique se o grupo do Auto Scaling iniciou uma instância adicional.

Para verificar se o tamanho do grupo do Auto Scaling foi alterado

1. Na guia Activity (Atividade), em Activity history (Histórico de atividades), a coluna Status mostra o status atual de sua instância. Use o botão de atualização até ver o status da sua instância ser alterado para Successful (Bem-sucedido). Isso indica que seu grupo do Auto Scaling iniciou com êxito uma nova instância.

Note

Se a instância não for executada, será possível encontrar dicas de solução de problemas em [Solucionar problemas do Amazon EC2 Auto Scaling \(p. 459\)](#).

2. Na guia Instance management (Gerenciamento de instâncias), em Instances (Instâncias), a coluna Lifecycle (Ciclo de vida) exibe o estado das suas instâncias. Demora um pouco para iniciar uma instância. Depois que a instância é iniciada, seu estado muda para InService. Você pode ver que seu grupo do Auto Scaling iniciou 1 nova instância, e que ela está no estado InService.

Alterar o tamanho do grupo do Auto Scaling (AWS CLI)

Quando você altera a capacidade desejada de seu grupo do Auto Scaling, o Amazon EC2 Auto Scaling gerencia o processo de início ou término de instâncias para manter o novo tamanho do grupo. O comportamento padrão é não aguardar que o período de desaquecimento padrão seja concluído, mas é possível substituir o padrão e aguardar a conclusão do período de desaquecimento. Para obter mais informações, consulte [Desaquecimento de escalabilidade para o Amazon EC2 Auto Scaling \(p. 205\)](#).

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-capacity](#) comando 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.

```
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-groups](#) comando 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": [  
        {  
            "AutoScalingGroupARN": "arn:",  
            "ServiceLinkedRoleARN": "arn:",  
            "TargetGroupARNS": [],  
            "SuspendedProcesses": [],  
            "LaunchTemplate": {  
                "LaunchTemplateName": "my-launch-template",  
                "Version": "1",  
                "LaunchTemplateId": "lt-050555ad16a3f9c7f"  
            },  
            "Tags": [],  
            "EnabledMetrics": [],  
            "LoadBalancerNames": [],  
            "AutoScalingGroupName": "my-asg",  
            "DefaultCooldown": 300,  
            "MinSize": 1,  
            "Instances": [  
                {  
                    "ProtectedFromScaleIn": false,  
                    "AvailabilityZone": "us-west-2a",  
                    "LaunchTemplate": {  
                        "LaunchTemplateName": "my-launch-template",  
                        "Version": "1",  
                        "LaunchTemplateId": "lt-050555ad16a3f9c7f"  
                    },  
                    "InstanceId": "i-05b4f7d5be44822a6",  
                    "InstanceType": "t2.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": "t2.micro",  
                    "HealthStatus": "Healthy",  
                    "LifecycleState": "InService"  
                }  
            ],  
            "MaxSize": 5,  
            "VPCZoneIdentifier": "subnet-c87f2be0",  
            "HealthCheckGracePeriod": 300,  
            "TerminationPolicies": [  
                "Default"  
            ],  
            "CreatedTime": "2019-03-18T23:30:42.611Z",  
            "AvailabilityZones": [  
                "us-west-2a"  
            ],  
            "HealthCheckType": "EC2",  
            "HealthCheckGracePeriod": 300  
        }  
    ]  
}
```

```
        "NewInstancesProtectedFromScaleIn": false,  
        "DesiredCapacity": 2  
    }  
}
```

Observe que DesiredCapacity mostra o novo valor. Seu grupo do Auto Scaling iniciou uma instância adicional.

Anexar instâncias do EC2 a seu grupo do Auto Scaling

O Amazon EC2 Auto Scaling oferece a opção de anexar uma ou mais instâncias do EC2 aos seu grupo existente do Auto Scaling. Depois que uma instância é anexada, ela é considerada parte do grupo do Auto Scaling.

Ao anexar as instâncias, considere o seguinte:

- Quando você anexa instâncias, a capacidade desejada do grupo aumente de acordo com o número de instâncias que estão sendo anexadas. Se o número de instâncias que estão sendo conectadas mais a capacidade desejada exceder o tamanho máximo do grupo, a solicitação falha.
- Se você anexar uma instância a um grupo do Auto Scaling que tenha um grupo de destino de平衡ador 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 em seu 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.

Anexar uma instância (console)

Siga o procedimento a seguir para desanexar uma instância de seu grupo do Auto Scaling.

Para anexar uma instância a um grupo do Auto Scaling existente

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 uma instância que atenda aos critérios.

Anexar uma instância (AWS CLI)

Siga o procedimento a seguir para desanexar uma instância de seu grupo do Auto Scaling.

Os exemplos usam um grupo do Auto Scaling com a seguinte configuração:

- Nome do grupo do Auto Scaling = my-asg
- Tamanho mínimo = 1
- Tamanho máximo = 5
- Capacidade desejada = 2

Para anexar uma instância a um grupo do Auto Scaling

1. Descreva um grupo específico de Auto Scaling usando o [describe-auto-scaling-groups](#) comando a seguir.

```
aws autoscaling describe-auto-scaling-groups --auto-scaling-group-names my-asg
```

O exemplo de resposta a seguir mostra que a capacidade desejada é 2 e que o grupo tem duas instâncias em execução.

```
{  
    "AutoScalingGroups": [  
        {  
            "AutoScalingGroupARN": "arn",  
            "ServiceLinkedRoleARN": "arn",  
            "TargetGroupARNS": [],  
            "SuspendedProcesses": [],  
            "LaunchTemplate": {  
                "LaunchTemplateName": "my-launch-template",  
                "Version": "1",  
                "LaunchTemplateId": "lt-050555ad16a3f9c7f"  
            },  
            "Tags": [],  
            "EnabledMetrics": [],  
            "LoadBalancerNames": [],  
            "AutoScalingGroupName": "my-asg",  
            "DefaultCooldown": 300,  
            "MinSize": 1,  
            "Instances": [  
                {  
                    "ProtectedFromScaleIn": false,  
                    "AvailabilityZone": "us-west-2a",  
                    "LaunchTemplate": {  
                        "LaunchTemplateName": "my-launch-template",  
                        "Version": "1",  
                        "LaunchTemplateId": "lt-050555ad16a3f9c7f"  
                    },  
                    "InstanceId": "i-05b4f7d5be44822a6",  
                    "InstanceType": "t2.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": "t2.micro",
        "HealthStatus": "Healthy",
        "LifecycleState": "InService"
    }
],
"MaxSize": 5,
"VPCZoneIdentifier": "subnet-c87f2be0",
"HealthCheckGracePeriod": 300,
"TerminationPolicies": [
    "Default"
],
"CreatedTime": "2019-03-18T23:30:42.611Z",
"AvailabilityZones": [
    "us-west-2a"
],
"HealthCheckType": "EC2",
"NewInstancesProtectedFromScaleIn": false,
"DesiredCapacity": 2
}
]
```

2. Anexe uma instância ao grupo do Auto Scaling usando o seguinte comando [attach-instances](#).

```
aws autoscaling attach-instances --instance-ids i-0787762faf1c28619 --auto-scaling-group-name my-asg
```

3. Para verificar se a instância está associada, use o seguinte comando [describe-auto-scaling-groups](#):

```
aws autoscaling describe-auto-scaling-groups --auto-scaling-group-names my-asg
```

O exemplo de resposta a seguir mostra que a capacidade desejada aumentou em 1 instância (para a nova capacidade de 3) e que há uma nova instância, i-0787762faf1c28619,

```
{
    "AutoScalingGroups": [
        {
            "AutoScalingGroupARN": "arn",
            "ServiceLinkedRoleARN": "arn",
            "TargetGroupARNs": [],
            "SuspendedProcesses": [],
            "LaunchTemplate": {
                "LaunchTemplateName": "my-launch-template",
                "Version": "1",
                "LaunchTemplateId": "lt-050555ad16a3f9c7f"
            },
            "Tags": [],
            "EnabledMetrics": [],
            "LoadBalancerNames": [],
            "AutoScalingGroupName": "my-asg",
            "DefaultCooldown": 300,
            "MinSize": 1,
            "Instances": [
                {
                    "ProtectedFromScaleIn": false,
                    "AvailabilityZone": "us-west-2a",
                    "LaunchTemplate": {
                        "LaunchTemplateName": "my-launch-template",
                        "Version": "1",
                        "LaunchTemplateId": "lt-050555ad16a3f9c7f"
                    },

```

```
        "InstanceId": "i-05b4f7d5be44822a6",
        "HealthStatus": "Healthy",
        "LifecycleState": "Pending"
    },
    {
        "ProtectedFromScaleIn": false,
        "AvailabilityZone": "us-west-2a",
        "LaunchTemplate": {
            "LaunchTemplateName": "my-launch-template",
            "Version": "1",
            "LaunchTemplateId": "lt-050555ad16a3f9c7f"
        },
        "InstanceId": "i-0c20ac468fa3049e8",
        "HealthStatus": "Healthy",
        "LifecycleState": "InService"
    },
    {
        "ProtectedFromScaleIn": false,
        "AvailabilityZone": "us-west-2a",
        "LaunchTemplate": {
            "LaunchTemplateName": "my-launch-template",
            "Version": "1",
            "LaunchTemplateId": "lt-050555ad16a3f9c7f"
        },
        "InstanceId": "i-0787762faf1c28619",
        "HealthStatus": "Healthy",
        "LifecycleState": "InService"
    }
],
"MaxSize": 5,
"VPCZoneIdentifier": "subnet-c87f2be0",
"HealthCheckGracePeriod": 300,
"TerminationPolicies": [
    "Default"
],
"CreatedTime": "2019-03-18T23:30:42.611Z",
"AvailabilityZones": [
    "us-west-2a"
],
"HealthCheckType": "EC2",
"NewInstancesProtectedFromScaleIn": false,
"DesiredCapacity": 3
}
]
```

Desvincular instâncias do EC2 do seu grupo do Auto Scaling

É possível remover (desvincular) de um grupo do Auto Scaling uma instância que esteja no estado InService. Depois que as instâncias é desvinculada, você pode gerenciá-la de forma independente do resto do grupo do Auto Scaling. Ao dissociar uma instância, você pode:

- Mova uma instância para fora de um grupo do Auto Scaling e anexe-a a um grupo diferente. Para obter mais informações, consulte [Anexar instâncias do EC2 a seu grupo do Auto Scaling \(p. 171\)](#).
- Teste um grupo do Auto Scaling criando-o por meio de instâncias existentes em execução em sua aplicação. Depois disso, é possível desvincular essas instâncias do grupo do Auto Scaling quando seus testes forem concluídos.

Ao desanexar as instâncias, considere o seguinte:

- Se o número de instâncias que você está desanexando reduzir o tamanho do grupo do Auto Scaling para abaixo da capacidade mínima, você precisará reduzir a capacidade mínima do grupo para desvincular as instâncias.
- Ao desvincular instâncias, você tem a opção de diminuir a capacidade desejada para o grupo do Auto Scaling de acordo com o número de instâncias que estão sendo desvinculadas. Se você optar por não reduzir a capacidade, o Amazon EC2 Auto Scaling iniciará novas instâncias para substituir as desvinculadas. Se você reduzir a capacidade, mas desvincular várias instâncias da mesma zona de disponibilidade, o Amazon EC2 Auto Scaling poderá rebalancear as Zonas de disponibilidade, a menos que você suspenda o processo AZRebalance. Para obter mais informações, consulte [Suspender e retomar um processo para um grupo do Auto Scaling \(p. 312\)](#).
- 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.

Note

Se você estiver desanexando instâncias que estão no estado Standby, adote cautela. A tentativa de desanexar instâncias após colocá-las no estado Standby pode fazer com que outras instâncias sejam encerradas inesperadamente.

Desvincular instâncias (console)

Use o procedimento a seguir para desvincular uma instância de seu grupo do Auto Scaling.

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 grupos do Auto Scaling 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 remoção da instância especificada do grupo Auto Scaling e escolha Desanexar instância.

Desvincular instâncias (AWS CLI)

Use o procedimento a seguir para desvincular uma instância de seu grupo do Auto Scaling.

Os exemplos usam um grupo do Auto Scaling com a seguinte configuração:

- Nome do grupo do Auto Scaling = my-asg
- Tamanho mínimo = 1
- Tamanho máximo = 5
- Capacidade desejada = 4

Para desvincular uma instância de um grupo do Auto Scaling existente

1. Liste as instâncias atuais usando o seguinte comando [describe-auto-scaling-instances](#):

```
aws autoscaling describe-auto-scaling-instances
```

O exemplo de resposta a seguir mostra que o grupo tem quatro instâncias em execução.

```
{  
    "AutoScalingInstances": [  
        {  
            "ProtectedFromScaleIn": false,  
            "AvailabilityZone": "us-west-2a",  
            "LaunchTemplate": {  
                "LaunchTemplateName": "my-launch-template",  
                "Version": "1",  
                "LaunchTemplateId": "lt-050555ad16a3f9c7f"  
            },  
            "InstanceId": "i-05b4f7d5be44822a6",  
            "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",  
            "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",  
            "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",  
            "AutoScalingGroupName": "my-asg",  
            "HealthStatus": "HEALTHY",  
            "LifecycleState": "InService"  
        }  
    ]
```

```
}
```

2. Desvincule uma instância e reduza a capacidade desejada usando o seguinte comando [detach-instances](#).

```
aws autoscaling detach-instances --instance-ids i-05b4f7d5be44822a6 \
--auto-scaling-group-name my-asg --should-decrement-desired-capacity
```

3. Verifique se a instância está desanexada usando o seguinte comando [describe-auto-scaling-instances](#).

```
aws autoscaling describe-auto-scaling-instances
```

O exemplo de resposta a seguir mostra que agora há três instâncias em execução.

```
{
    "AutoScalingInstances": [
        {
            "ProtectedFromScaleIn": false,
            "AvailabilityZone": "us-west-2a",
            "LaunchTemplate": {
                "LaunchTemplateName": "my-launch-template",
                "Version": "1",
                "LaunchTemplateId": "lt-050555ad16a3f9c7f"
            },
            "InstanceId": "i-0c20ac468fa3049e8",
            "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",
            "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",
            "AutoScalingGroupName": "my-asg",
            "HealthStatus": "HEALTHY",
            "LifecycleState": "InService"
        }
    ]
}
```

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.

Se você estiver ajustando a escalabilidade com base em uma métrica que aumenta ou diminui proporcionalmente ao número de instâncias em um grupo do Auto Scaling, recomendamos que use políticas de escalabilidade com monitoramento de objetivo. Caso contrário, recomendamos que você use as políticas de escalabilidade em etapas.

Com o monitoramento de objetivos, um grupo do Auto Scaling reduz a escala horizontalmente em proporção direta à carga real em seu aplicativo. 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.

Por padrão, novos grupos do Auto Scaling começam sem nenhuma política de escalabilidade. Quando você usa um grupo do Auto Scaling sem nenhuma forma de escalabilidade dinâmica, ele não realiza escalabilidade de maneira autônoma, a menos que você configure a escalabilidade programada ou a escalabilidade preditiva.

Índice

- [Como funcionam as políticas de escalabilidade dinâmica \(p. 178\)](#)
- [Várias políticas de escalabilidade dinâmica \(p. 179\)](#)
- [Políticas de escalabilidade com monitoramento do objetivo para o Amazon EC2 Auto Scaling \(p. 180\)](#)
- [Políticas de escalabilidade simples e em etapas do Amazon EC2 Auto Scaling \(p. 190\)](#)
- [Definir valores padrão para aquecimento de instância ou desaquecimento de escalabilidade \(p. 200\)](#)
- [Escalabilidade baseada no Amazon SQS \(p. 210\)](#)
- [Verificar uma ação de escalabilidade para um grupo do Auto Scaling \(p. 215\)](#)
- [Desabilitar uma política de escalabilidade para um grupo do Auto Scaling \(p. 216\)](#)
- [Excluir uma política de escalabilidade \(p. 218\)](#)
- [Exemplo de políticas de escalabilidade para a AWS Command Line Interface \(AWS CLI\) \(p. 220\)](#)

Como funcionam as políticas de escalabilidade dinâmica

Uma política de escalabilidade dinâmica instrui o Amazon EC2 Auto Scaling a rastrear uma CloudWatch métrica 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 \(p. 86\)](#) 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 viole 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 ($25\% \text{ de } 8 = 2$). 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

Para criar uma política de escalabilidade de rastreamento de metas, você especifica uma CloudWatch métrica da Amazon e um valor alvo que representa a utilização média ideal ou o nível de produtividade do seu aplicativo. Em seguida, o Amazon EC2 Auto Scaling poderá aumentar a escala na horizontal do seu grupo (adicionar mais instâncias) para processar picos de tráfego, e reduzir a escala na horizontal do seu grupo (executar menos instâncias) para reduzir custos durante períodos de baixa utilização ou throughput.

Por exemplo, digamos que você tenha um aplicativo que seja executado em duas instâncias e queira que a utilização de CPU do grupo do Auto Scaling permaneça em cerca de 50% quando a carga no aplicativo 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, seu grupo do Auto Scaling dimensionará o número de instâncias para manter o valor efetivo da métrica em ou perto de 50%.

Tópicos

- [Várias políticas de escalabilidade de monitoramento de objetivo \(p. 181\)](#)
- [Considerações \(p. 181\)](#)
- [Escolher métricas \(p. 182\)](#)
- [Definir valor de objetivo \(p. 183\)](#)

- [Definir o tempo de aquecimento da instância \(p. 183\)](#)
- [Criar uma política de escalabilidade com monitoramento do objetivo \(console\) \(p. 184\)](#)
- [Criar uma política de escalabilidade com monitoramento do objetivo \(AWS CLI\) \(p. 185\)](#)
- [Crie uma política de escalabilidade de rastreamento de destino para Amazon EC2 Auto Scaling usando matemática em métricas \(p. 187\)](#)

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.

A intenção do Amazon EC2 Auto Scaling é sempre priorizar a disponibilidade, portanto, seu comportamento será diferente dependendo se as políticas de monitoramento do objetivo estão prontas para aumentar ou reduzir a escala na horizontal. Ele vai aumentar a escala na horizontal do grupo do Auto Scaling se qualquer uma das políticas com monitoramento do objetivo estiverem prontas para aumentar a escala, mas vai reduzir a escala na horizontal somente se todas as políticas com monitoramento do objetivo (com a parte de redução habilitada) estiverem prontas para reduzir a escala.

Considerações

As considerações a seguir são aplicáveis ao trabalhar com políticas de escalabilidade com monitoramento de objetivo:

- Não crie, edite nem exclua os CloudWatch alarmes usados com uma política de escalabilidade de rastreamento de metas. O Amazon EC2 Auto Scaling cria e gerencia os CloudWatch alarmes associados às suas políticas de escalabilidade de rastreamento de alvos e os exclui quando não são mais necessários.
- 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 que seu grupo do Auto Scaling tenha a escala reduzida na horizontal imediatamente após o término de uma workload, é possível desabilitar a parte de redução da escala da política. Isso proporciona a flexibilidade de usar o método de redução da escala na horizontal que melhor atenda às suas necessidades quando a utilização estiver baixa. Para garantir que a redução da escala na horizontal ocorra o mais rápido possível, recomendamos não usar uma política simples de escalabilidade para evitar a adição de um período de esfriamento.
- Se houver pontos de dados faltando na métrica, isso fará com que o estado de alarme do CloudWatch 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.
- 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, vamos supor que você defina um valor de objetivo de 50% para a utilização da CPU, e o seu grupo do Auto Scaling excede 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 remover 1,5 instância aumenta a utilização da CPU para acima de 50%, removeremos apenas uma instância.

- 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.

Escolher métricas

É possível criar políticas de escalabilidade de rastreamento de destino com métricas predefinidas ou personalizadas.

Ao criar uma política de escalabilidade de rastreamento de metas com um tipo de métrica predefinido, você escolhe uma métrica da seguinte lista de métricas predefinidas:

- ASGAverageCPUUtilization: média de utilização da CPU do grupo do Auto Scaling.
- ASGAverageNetworkIn: número médio de bytes recebidos por uma única instância em todas as interfaces de rede.
- ASGAverageNetworkOut: número médio de bytes enviados de uma única instância em todas as interfaces de rede.
- ALBRequestCountPerTarget: quantidade média de solicitações do Application Load Balancer por destino.

Important

Outras informações valiosas sobre as métricas de utilização da CPU, E/S de rede e contagem de solicitações do Application Load Balancer por alvo podem ser encontradas no tópico [Listar CloudWatch as métricas disponíveis para suas instâncias no Guia do usuário do Amazon EC2 para instâncias Linux](#) e as CloudWatch 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. É necessário usar a AWS CLI ou um SDK para criar uma política com monitoramento do objetivo com uma métrica personalizada.

Lembre-se do seguinte ao escolher uma métrica:

- Recomendamos que você use somente métricas que estejam disponíveis em intervalos de um minuto para ajudá-lo a escalar mais rapidamente em resposta às mudanças de utilização. O rastreamento de destino avaliará as métricas agregadas com uma granularidade de um minuto para todas as métricas predefinidas e personalizadas, mas a métrica subjacente talvez publique os dados com menos frequência. Por exemplo, todas as métricas do Amazon EC2 são enviadas em intervalos de cinco minutos, por padrão, mas podem ser configuradas para um minuto (o que é conhecido como monitoramento detalhado). Essa escolha depende dos serviços individuais. A maioria tenta usar o menor intervalo possível. Para obter informações sobre como ativar o monitoramento detalhado, consulte [Configurar monitoramento para instâncias do Auto Scaling \(p. 337\)](#).
- Nem todas as métricas personalizadas funcionam para rastreamento de destino. A métrica deve ser de utilização válida e descrever o quanto uma instância está ocupada. 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 [Escalabilidade baseada no Amazon SQS \(p. 210\)](#).
- 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.
- Quando uma métrica emite valores reais de 0 para CloudWatch (por exemplo, ALBRequestCountPerTarget), um grupo de Auto Scaling pode escalar para 0 quando não há tráfego em seu aplicativo. 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.
- Ao criar uma política de escalabilidade de rastreamento de destino com uma métrica personalizada, é possível usar matemática em métricas para combinar métricas. Para obter mais informações, consulte [Crie uma política de escalabilidade de rastreamento de destino para Amazon EC2 Auto Scaling usando matemática em métricas \(p. 187\)](#).

Definir valor de objetivo

Ao criar uma política de escalabilidade com monitoramento de objetivo, você deve 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 seu aplicativo 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 aquecer, se ocorrer um evento de aumento de escala, 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 do aquecimento de instância padrão definido para o grupo. Se o aquecimento padrão da instância for nulo, ele voltará ao valor do tempo de recarga padrão. Recomendamos usar o aquecimento padrão da instância para facilitar a atualização de todas as políticas de escalabilidade quando o tempo de aquecimento mudar. Para obter mais informações, consulte [Definir o aquecimento padrão da instância para um grupo do Auto Scaling \(p. 200\)](#).

Criar uma política de escalabilidade com monitoramento do objetivo (console)

Você pode optar por configurar uma política de escalabilidade com monitoramento do objetivo em um grupo do Auto Scaling enquanto o cria ou após o grupo do Auto Scaling ser criado.

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).

Para criar um grupo do Auto Scaling com uma política de escalabilidade com monitoramento do objetivo

1. Abra o console do Amazon EC2 em <https://console.aws.amazon.com/ec2/> e escolha grupos do Auto Scaling 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 Group size (Tamanho do grupo), especifique o intervalo no qual você deseja escalar atualizando a capacidade mínima e a capacidade máxima. Essas duas configurações permitem que seu grupo do Auto Scaling seja escalado dinamicamente. O Amazon EC2 Auto Scaling escala seu grupo no intervalo de valores especificados pela capacidade mínima e capacidade máxima.
5. Em Scaling policies (Políticas de escalabilidade), escolha Target tracking scaling policy (Política de escalabilidade com monitoramento do objetivo).
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 Instances need (As instâncias precisam), atualize o valor de aquecimento de instâncias 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 grupos do Auto Scaling 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. Verifique se a capacidade mínima e a capacidade máxima estão definidas adequadamente. Por exemplo, se seu grupo já estiver em seu tamanho máximo, especifique um novo máximo para expandir. O Amazon EC2 Auto Scaling não escala seu grupo abaixo da capacidade mínima ou acima da capacidade máxima. Para atualizar o grupo, na guia Detalhes, altere as configurações atuais de capacidade mínima e máxima.
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. Para o tipo de política, mantenha o padrão de escalonamento de rastreamento do Target.
 - 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.
6. Escolha Create (Criar).
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 Instances need (As instâncias precisam), atualize o valor de aquecimento de instâncias 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.

Criar uma política de escalabilidade com monitoramento do objetivo (AWS CLI)

Use a AWS CLI da seguinte maneira para configurar políticas de escalabilidade com monitoramento do objetivo para o grupo do Auto Scaling.

Tarefas

- [Etapa 1: Criar um grupo do Auto Scaling \(p. 185\)](#)
- [Etapa 2: Criar uma política de escalabilidade com monitoramento do objetivo \(p. 186\)](#)

Etapa 1: Criar um grupo do Auto Scaling

Use o `create-auto-scaling-group` comando para criar um grupo de Auto Scaling nomeado `my-asg` usando o modelo `my-template` de execução. Se você não tiver um modelo de inicialização, consulte [Exemplos da AWS CLI para trabalhar com modelos de execução \(p. 37\)](#).

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

```
--vpc-zone-identifier "subnet-5ea0c127,subnet-6194ea3b,subnet-c934b782" \  
--max-size 5 --min-size 1
```

Etapa 2: Criar uma política de escalabilidade com monitoramento do objetivo

Depois de criar o grupo do Auto Scaling, é possível criar uma política de escalabilidade com monitoramento do objetivo que instrua o Amazon EC2 Auto Scaling a aumentar e diminuir dinamicamente o número de instâncias do EC2 em execução no grupo quando a carga na aplicação for alterada.

Exemplo: arquivo de configuração de rastreamento de destino

Veja a seguir, um exemplo de configuração de rastreamento de destino que mantém a utilização média da CPU em 40%. Salve esta configuração em um arquivo chamado config.json.

```
{  
    "TargetValue": 40.0,  
    "PredefinedMetricSpecification":  
    {  
        "PredefinedMetricType": "ASGAverageCPUUtilization"  
    }  
}
```

Para obter mais informações, consulte [PredefinedMetricSpecification](#)o Amazon EC2 Auto Scaling API Reference.

Se preferir, você poderá personalizar a métrica usada para a escalabilidade criando uma especificação de métrica personalizada e adicionando valores para cada parâmetro do CloudWatch. Veja a seguir um exemplo de configuração de rastreamento de destino que mantém a utilização média da métrica especificada em 40%.

```
{  
    "TargetValue":40.0,  
    "CustomizedMetricSpecification":{  
        "MetricName":"MyUtilizationMetric",  
        "Namespace":"MyNamespace",  
        "Dimensions": [  
            {  
                "Name":"MyOptionalMetricDimensionName",  
                "Value":"MyOptionalMetricDimensionValue"  
            }  
        ],  
        "Statistic":"Average",  
        "Unit":"Percent"  
    }  
}
```

Para obter mais informações, consulte [CustomizedMetricSpecification](#)o Amazon EC2 Auto Scaling API Reference.

Exemplo: cpu40- target-tracking-scaling-policy

Use o [put-scaling-policy](#) comando, junto com o config.json arquivo que você criou anteriormente, para criar uma política de escalabilidade denominada cpu40-target-tracking-scaling-policy que mantenha a utilização média da CPU do grupo de Auto Scaling em 40 por cento.

```
aws autoscaling put-scaling-policy --policy-name cpu40-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 ARNs e os nomes dos dois alarmes do CloudWatch criados em seu nome.

```
{  
    "PolicyARN": "arn:aws:autoscaling:region:account-id:scalingPolicy:228f02c2-c665-4bfd-  
aac-8b04080bea3c:autoScalingGroupName/my-asg:policyName/cpu40-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"  
        }  
    ]  
}
```

Crie uma política de escalabilidade de rastreamento de destino para Amazon EC2 Auto Scaling usando matemática em métricas

Ao usar a matemática de métricas, você pode consultar várias métricas do CloudWatch e usar expressões matemáticas para criar novas séries temporais de acordo com essas métricas. Você pode visualizar as séries temporais resultantes no console do CloudWatch e adicioná-las aos painéis. Para obter mais informações sobre matemática métrica, consulte [Usar 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](#).

Note

Você pode criar uma política de escalabilidade de rastreamento de destino usando matemática em métricas somente se usar a AWS CLI ou um SDK. Este recurso ainda não está disponível no console e no AWS CloudFormation.

Exemplo: lista de pendências da fila do por instância

Para calcular a lista de pendências da fila do 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 [Escalabilidade baseada no Amazon SQS \(p. 210\)](#).

A lógica da expressão é a seguinte:

`sum of (number of messages in the queue)/(number of InService instances)`

Então, as informações da sua métrica do CloudWatch são as seguintes:

ID	Métrica do CloudWatch	Estatística	Período
m1	ApproximateNumberOfMessagesVisible		1 minuto
m2	GroupInServiceInstances	Média	1 minuto

1 minuto

ID	Expressão
e1	(m1)/(m2)

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 os *valores substituíveis em itálico* pelos valores que são apropriados para sua aplicação.

```
{
    "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": "Sum"
                },
                "ReturnData": false
            }
        ]
    }
}
```

```
        "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 [TargetTrackingConfiguration](#) no Amazon EC2 Auto Scaling API Reference.

Note

A seguir estão 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.](#)
2. Para criar essa política, execute o [put-scaling-policy](#) comando 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 os ARNs dos dois CloudWatch alarmes criados em seu nome.

```
{
    "PolicyARN": "arn:aws:autoscaling:region:account-id:scalingPolicy:228f02c2-c665-4bfd-aaac-8b04080bea3c:autoScalingGroupName/my-asg:policyName/sqs-backlog-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"
        }
    ]
}
```

}

Note

Se esse comando lançar um erro, verifique se você atualizou a AWS CLI localmente para a versão mais recente.

Políticas de escalabilidade simples e em etapas do Amazon EC2 Auto Scaling

Com o escalonamento por etapas e o escalonamento simples, você escolhe métricas de escalabilidade e valores limite para os CloudWatch alarmes que invocam o processo de escalonamento. Você também define como seu grupo do Auto Scaling deve ser escalado quando um limite for violado para um número especificado de períodos de avaliação.

É altamente recomendável usar uma política de dimensionamento com monitoramento do objetivo para escalar de acordo com uma métrica, como a média de utilização da CPU ou a métrica RequestCountPerTarget do Application Load Balancer. 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 \(p. 180\)](#). 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 demanda atinge um determinado nível.

Índice

- [Diferenças entre políticas de escalabilidade simples e políticas de escalabilidade em etapas \(p. 190\)](#)
- [Ajustes em etapas para escalabilidade em etapas \(p. 191\)](#)
- [Tipos de ajuste da escalabilidade \(p. 193\)](#)
- [Aquecimento da instância \(p. 194\)](#)
- [Crie CloudWatch alarmes para os limites métricos altos e baixos \(console\) \(p. 194\)](#)
- [Criar políticas de escalabilidade em etapas \(console\) \(p. 196\)](#)
- [Crie políticas de escalabilidade e CloudWatch alarmes \(\) AWS CLI \(p. 197\)](#)
 - [Etapa 1: Criar um grupo do Auto Scaling \(p. 198\)](#)
 - [Etapa 2: Criar políticas de escalabilidade \(p. 198\)](#)
 - [Políticas de escalabilidade em etapas \(p. 198\)](#)
 - [Políticas de escalabilidade simples \(p. 199\)](#)
 - [Etapa 3: criar CloudWatch alarmes para os limites métricos altos e baixos \(p. 199\)](#)

Diferenças entre políticas de escalabilidade simples e políticas de escalabilidade em etapas

Políticas de escalabilidade simples e políticas de escalabilidade em etapas são duas das opções de escalabilidade dinâmica disponíveis para uso. Ambas exigem que você crie alarmes do CloudWatch para as políticas de escalabilidade. Ambas exigem que você especifique os limites altos e baixos para os alarmes. Ambas exigem que você defina se deseja adicionar ou remover instâncias, e quantas delas, ou defina o grupo para um tamanho exato.

A principal diferença entre os tipos de política são os ajustes em etapas que você obtém com políticas de escalabilidade em etapas. Quando os ajustes em etapas são aplicados, e eles aumentam ou diminuem a

capacidade atual do seu grupo do Auto Scaling, os ajustes variam de acordo com o tamanho da violação do alarme.

Na maioria dos casos, as políticas de escalabilidade em etapas são uma escolha melhor do que as políticas de escalabilidade simples, mesmo que você tenha apenas um único ajuste de escalabilidade.

O principal problema com a escalabilidade simples é que, depois que uma ação de escalabilidade é iniciada, a política deve aguardar a conclusão da substituição da ação de escalabilidade ou da verificação de integridade e o fim do [período de desaquecimento \(p. 205\)](#) para responder a outros alarmes. Os desaquecimentos ajudam a evitar a inicialização de ações de escalabilidade adicionais antes que os efeitos de atividades anteriores sejam visíveis.

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. Portanto, todos os alarmes que são violados são avaliados pelo Amazon EC2 Auto Scaling à medida que recebe as mensagens de alarme.

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.

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

O CloudWatch agrupa pontos de dados de métricas com base na estatística da métrica associada ao alarme do CloudWatch. Quando o alarme é violado, a política de dimensionamento apropriada é invocada. O Amazon EC2 Auto Scaling aplica o tipo de agregação aos pontos de dados métricos mais recentes de CloudWatch (em oposição aos dados métricos brutos). Ele compara esse valor de métrica agragada 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 um grupo do Auto Scaling que tenha uma capacidade atual e uma capacidade desejada de 10. Você tem um alarme do CloudWatch com um limite de violação de 50% e políticas de expansão e redução. 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:

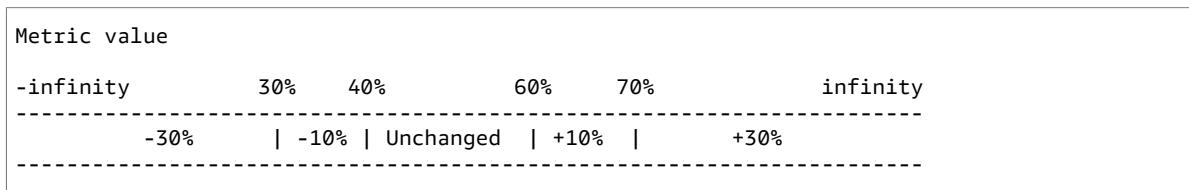
Exemplo: ajustes em etapas para política de expansão

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.



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 (adicone 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 estiver usando o AWS Management Console, você especificará os limites superior e inferior como valores absolutos. Se estiver usando a AWS CLI ou um SDK, especifique os limites máximo e mínimo relativos 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 arredondados 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 \(p. 86\)](#) 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. Você deve 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 mesmo 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, digamos que você crie uma política com duas etapas. A primeira etapa adiciona 10% quando a métrica chega a 60 e a segunda etapa adiciona 30% quando a métrica chega a 70%. Seu grupo de Auto Scaling tem uma capacidade desejada e atual de 10. A capacidade desejada e a atual não mudam enquanto o valor métrico agregado é menor que 60. Suponha que a métrica chegue a 60, então 1 instância seja adicionada (10% 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 ainda 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 aquecer, se ocorrer um evento de aumento de escala, 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 do aquecimento de instância padrão definido para o grupo. Se o aquecimento padrão da instância for nulo, ele voltará ao valor do tempo de recarga padrão. Recomendamos usar o aquecimento padrão da instância para facilitar a atualização de todas as políticas de escalabilidade quando o tempo de aquecimento mudar. Para obter mais informações, consulte [Definir o aquecimento padrão da instância para um grupo do Auto Scaling \(p. 200\)](#).

Crie CloudWatch alarmes para os limites métricos altos e baixos (console)

Você pode usar o CloudWatch console para criar dois alarmes, um para redução de escala (métrica alta) e outro para ampliação (métrica baixa). Quando o limite especificado de um alarme for violado, por exemplo, porque o tráfego aumentou, o alarme entrará no estado de ALARME. Essa alteração de estado invoca a política de dimensionamento associada ao alarme. Então, a política instrui o Amazon EC2 Auto Scaling sobre como responder quando o alarme for violado, por exemplo, adicionando ou removendo um número especificado de instâncias.

Exemplo: Para criar um CloudWatch alarme para o limite máximo da métrica

1. Abra o console do CloudWatch 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 Select metric (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 Specify metric and conditions (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 Threshold type (Tipo de limite), escolha Static (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 de aumentar a escala horizontalmente (alarme superior), certifique-se de não escolher um valor menor que ou igual ao limite. 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 os dados perdidos no Guia](#) do CloudWatchusuário da Amazon.

9. Escolha Próximo.

A página Configure actions (Configurar ações) é exibida.

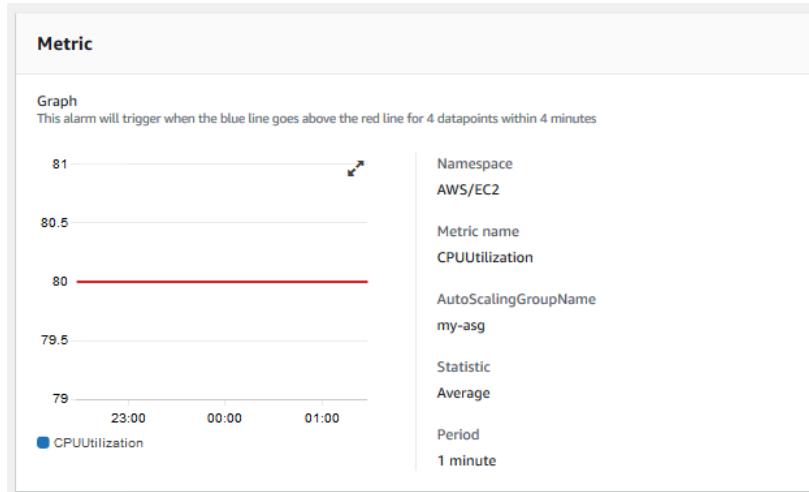
10. Em Notification (Notificação), selecione um tópico do Amazon SNS para notificar quando o alarme estiver no estado ALARM, OK ou INSUFFICIENT_DATA.

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 Create alarm (Criar alarme).

Exemplo: CloudWatch alarme que é violado se a CPU ultrapassar o limite de 80% por 4 minutos



Criar políticas de escalabilidade em etapas (console)

Você pode optar por configurar as políticas de dimensionamento em etapas em um grupo do Auto Scaling depois que o grupo é criado.

O procedimento a seguir mostra como usar o console do Amazon EC2 Auto Scaling para criar duas políticas de escalabilidade em etapas: uma política de aumento de escala na horizontal que aumenta a capacidade do grupo em 30% e uma política de redução de escala na horizontal que diminui a capacidade do grupo para duas instâncias.

Enquanto estiver configurando a política de escalabilidade, é possível criar alarmes ao mesmo tempo. Como alternativa, é possível usar os alarmes criados no console do CloudWatch, conforme descrito na seção anterior.

Como criar uma política de escalabilidade em etapas para expansão

1. Abra o console do Amazon EC2 em <https://console.aws.amazon.com/ec2/> e escolha grupos do Auto Scaling 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. Verifique se os limites de tamanho mínimo e máximo estão definidos adequadamente. Por exemplo, se seu grupo já estiver em seu tamanho máximo, você precisará especificar um novo máximo para expandir. O Amazon EC2 Auto Scaling não escala seu grupo abaixo da capacidade mínima ou acima da capacidade máxima. Para atualizar o grupo, na guia Detalhes, altere as configurações atuais de capacidade mínima e máxima.
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 uma política de expansão (aumento da capacidade), faça o seguinte:
 - a. Em Policy type (Tipo de política), escolha Step scaling (Escalabilidade em etapas).
 - b. Especifique um nome para a política.
 - c. Para CloudWatchalarme, escolha seu alarme. Se você ainda não criou um alarme, escolha Criar um CloudWatch alarme e conclua as etapas 4 a 14 [Crie CloudWatch alarmes para os limites métricos altos e baixos \(console\) \(p. 194\)](#) para criar um alarme que monitore a utilização da CPU. Defina o limite do alarme como maior ou igual a 80%.

- d. 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, escolha Add, insira 30 no campo seguinte 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.

- e. 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.
- f. 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).
- g. (Opcional) Em Instances need (As instâncias precisam), atualize o valor de aquecimento de instâncias conforme necessário.

6. Escolha Create (Criar).

Como criar uma política de escalabilidade em etapas para redução

1. Escolha Create dynamic scaling policy (Criar política de escalabilidade dinâmica) para continuar de onde parou depois de criar uma política para aumento de escala na horizontal.
2. Para definir uma política para redução (diminuição da capacidade), faça o seguinte:
 - a. Em Policy type (Tipo de política), escolha Step scaling (Escalabilidade em etapas).
 - b. Especifique um nome para a política.
 - c. Para CloudWatchalarme, escolha seu alarme. Se você ainda não criou um alarme, escolha Criar um CloudWatch alarme e conclua as etapas 4 a 14 [Crie CloudWatch alarmes para os limites métricos altos e baixos \(console\) \(p. 194\)](#) para criar um alarme que monitore a utilização da CPU. Defina o limite de alarme como menor ou igual a 40%.
 - d. 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, escolha Remove, insira 2 no campo seguinte e escolha capacity units. Por padrão, o limite superior desse ajuste em etapas é o limite do alarme, e o limite inferior é negativo (-) infinito.

- e. 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.
3. Escolha Create (Criar).

Crie políticas de escalabilidade e CloudWatch alarmes () AWS CLI

Use a AWS CLI da seguinte maneira para configurar políticas de escalabilidade em etapas ou simples para o grupo do Auto Scaling.

Tarefas

- [Etapa 1: Criar um grupo do Auto Scaling \(p. 198\)](#)
- [Etapa 2: Criar políticas de escalabilidade \(p. 198\)](#)
- [Etapa 3: criar CloudWatch alarmes para os limites métricos altos e baixos \(p. 199\)](#)

Etapa 1: Criar um grupo do Auto Scaling

Use o [create-auto-scaling-group](#) comando a seguir para criar um grupo de Auto Scaling chamado my-asg usando o modelo my-template de execução. Se você não tiver um modelo de inicialização, consulte [Exemplos da AWS CLI para trabalhar com modelos de execução \(p. 37\)](#).

```
aws autoscaling create-auto-scaling-group --auto-scaling-group-name my-asg \
--launch-template LaunchTemplateName=my-template,Version='2' \
--vpc-zone-identifier "subnet-5ea0c127,subnet-6194ea3b,subnet-c934b782" \
--max-size 5 --min-size 1
```

Etapa 2: Criar políticas de escalabilidade

Você pode criar políticas de escalabilidade simples ou em etapas que informem ao grupo do Auto Scaling o que fazer quando a carga na aplicação for alterada.

Políticas de escalabilidade em etapas

Exemplo: my-step-scale-out -policy

Use o seguinte comando [put-scaling-policy](#) para criar uma política de escalabilidade em etapas chamada my-step-scale-out-policy, com um tipo de ajuste PercentChangeInCapacity que aumenta a capacidade do grupo com base nos seguintes ajustes em etapas (supondo um limite de alarme do CloudWatch 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. Ele é necessário para criar um alarme do CloudWatch para a política.

```
{
  "PolicyARN": "arn:aws:autoscaling:region:123456789012:scalingPolicy:4ee9e543-86b5-4121-
b53b-aa4c23b5bbcc:autoScalingGroupName/my-asg:policyName/my-step-scale-in-policy"
}
```

Exemplo: my-step-scale-in -policy

Use o [put-scaling-policy](#) comando a seguir para criar uma política de escalabilidade de etapas chamada my-step-scale-in-policy, com um tipo de ajuste ChangeInCapacity que diminui a capacidade do grupo em 2 instâncias quando o CloudWatch alarme associado viola o valor métrico baixo do limite.

```
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  
}
```

Políticas de escalabilidade simples

Como alternativa, você pode criar políticas de escalabilidade simples usando os seguintes comandos da CLI em vez de comandos anteriores da CLI. Lembre-se de que um período de desaquecimento será implementado devido ao uso de políticas de escalabilidade simples.

Exemplo: my-simple-scale-out -policy

Use o [put-scaling-policy](#) comando a seguir para criar uma política de escalabilidade simples chamada **my-simple-scale-out-policy**, com um tipo de ajuste **PercentChangeInCapacity** que aumenta a capacidade do grupo em 30% quando o CloudWatch alarme associado viola o valor do limite máximo 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.

Exemplo: my-simple-scale-in -policy

Use o [put-scaling-policy](#) comando a seguir para criar uma política de escalabilidade simples chamada **my-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 do limite baixo 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.

Etapa 3: criar CloudWatch alarmes para os limites métricos altos e baixos

Na etapa 2, você criou políticas de dimensionamento que forneciam instruções para o Amazon EC2 Auto Scaling sobre como aumentar a escala horizontalmente e reduzir a escala horizontalmente quando o valor de uma métrica estivesse aumentando ou 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.

Exemplo: AddCapacity

Use o CloudWatch [put-metric-alarm](#) comando 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 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-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
```

Exemplo: RemoveCapacity

Use o CloudWatch [put-metric-alarm](#) comando a seguir para criar um alarme que diminua o tamanho do grupo de Auto Scaling com base no valor médio do limite 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
```

Definir valores padrão para aquecimento de instância ou desaquecimento de escalabilidade

O aquecimento de instância padrão e o desaquecimento padrão são propriedades do grupo do Auto Scaling que podem ser usadas para ajustar a performance de escalabilidade.

Índice

- [Definir o aquecimento padrão da instância para um grupo do Auto Scaling \(p. 200\)](#)
- [Desaquecimento de escalabilidade para o Amazon EC2 Auto Scaling \(p. 205\)](#)
- [Configurações de aquecimento e desaquecimento disponíveis \(p. 207\)](#)

Definir o aquecimento padrão da instância para um grupo do Auto Scaling

CloudWatch coleta e agrupa 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 InService estado de espera antes de contribuir com dados de uso para as métricas agregadas. 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.

O aquecimento padrão da instância não está configurado e não é habilitado por padrão. Para otimizar o desempenho de suas políticas de rastreamento de metas e escalonamento de etapas, é altamente recomendável que você ative o aquecimento padrão da instância.

Além de afetar a performance da escalabilidade, o aquecimento de instâncias padrão influencia o tempo total necessário para substituir uma instância durante uma atualização de instância.

Índice

- [Considerações sobre escalabilidade de desempenho \(p. 201\)](#)
- [Escolha o tempo padrão de aquecimento da instância \(p. 201\)](#)
- [Habilitar o aquecimento de instância padrão para um grupo \(p. 202\)](#)
- [Verificar o aquecimento de instância padrão para um grupo \(p. 204\)](#)
- [Encontre políticas de escalabilidade com um tempo de aquecimento da instância previamente definido \(p. 204\)](#)
- [Limpe o aquecimento da instância definido anteriormente para uma política de escalabilidade \(p. 205\)](#)

Considerações sobre escalabilidade de desempenho

A maioria dos aplicativos se beneficiará de ter um tempo de aquecimento de instância padrão que se aplica a todos os recursos, em vez de tempos de aquecimento diferentes para recursos diferentes. Por exemplo, se você não definir um aquecimento de instância padrão, o recurso 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 metas e escalonamento de etapas, elas usarão o valor definido para o tempo de recarga padrão como o tempo de aquecimento padrão. Se você tiver alguma política de escalabilidade preditiva, ela não tem tempo de aquecimento padrão.

Enquanto as instâncias estão se aquecendo, suas políticas de escalabilidade dinâmica só se expandem se o valor métrico das instâncias que não estão se aquecendo for maior do que o limite máximo de alarme da política (ou a meta de utilização de uma política de escalabilidade de rastreamento de metas). Se a demanda aumentar, a intenção será reduzir a escala na horizontal de forma conservadora para proteger a disponibilidade de sua aplicação. Isso bloqueia a escalabilidade dinâmica até que as instâncias terminem de aquecer.

Embora reduzamos a escala, também não consideramos os casos que estão se aquecendo como parte da capacidade atual do grupo. 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). Para obter mais informações, consulte a [Aquecimento da instância \(p. 194\)](#) seção no tópico de escalonamento de etapas.

Se o aquecimento padrão da instância não estiver ativado, o 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 a carga de trabalho 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 políticas de escalabilidade preditiva, as instâncias começam a contribuir com as métricas agregadas imediatamente. Se essas instâncias tiverem maior uso de recursos na inicialização, a adição de instâncias poderá fazer com que as métricas agregadas aumentem. Dependendo do tempo necessário para que o uso se estabilize, isso pode afetar qualquer política de escalabilidade dinâmica que use essas métricas. Se o limite alto de alarme de uma política de escalonamento dinâmico for violado, o grupo aumentará de tamanho novamente. Enquanto as novas instâncias estiverem se aquecendo, as atividades de expansão serão bloqueadas.

Escolha o tempo padrão de aquecimento da instância

A chave para definir o aquecimento padrão da instância é determinar por quanto tempo suas instâncias precisam concluir a inicialização e para que o consumo de recursos se estabilize após atingirem o estado.

InService Ao escolher o tempo de aquecimento da instância, busque um equilíbrio ideal entre coletar dados de uso para tráfego legítimo e minimizar a coleta de dados associada a picos temporários de uso na inicialização.

Suponha que você tenha um grupo de Auto Scaling conectado a um平衡ador de carga do Elastic Load Balancing. Quando as novas instâncias terminam de ser iniciadas, elas são registradas no平衡ador de carga antes de entrarem no **InService** estado. 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 da instância fornece o tempo de atraso necessário para que o consumo de recursos se estabilize.

Important

Se você não tiver certeza de quanto tempo precisa, comece com 300 segundos e diminua-o ou aumente-o gradualmente até obter o melhor desempenho de escalabilidade para seu aplicativo. Você provavelmente precisará experimentar até 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 previamente definido \(p. 204\)](#).

Dependendo de seu grupo usar algum gancho de ciclo de vida, talvez você consiga usar um tempo de aquecimento mais curto.

Se	Então		
Seu grupo de Auto Scaling tem um gancho de ciclo de vida que atrasa a colocação em serviço das instâncias até que elas terminem de inicializar.	Você pode definir o aquecimento padrão da instância para uma duração menor.		
Seu grupo de Auto Scaling não tem um gancho de ciclo de vida que inicialize as instâncias antes de serem colocadas em serviço.	Você pode especificar uma duração maior para o aquecimento padrão da instância, que inclui o tempo que suas instâncias precisam para concluir a inicialização.		

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 ativar o recurso padrão de aquecimento da instância, você não precisa mais especificar valores para parâmetros de aquecimento para os seguintes recursos:

- [Atualização de instância \(p. 109\)](#)
- [Escalabilidade de rastreamento de alvos \(p. 183\)](#)
- [Escalabilidade por etapas \(p. 194\)](#)

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-group` comando a seguir cria um grupo de Auto Scaling com o nome my-asg e ativa o aquecimento padrão da instância com um valor de 120 segundos.

```
aws autoscaling create-auto-scaling-group --auto-scaling-group-name my-asg --default-instance-warmup 120 ...
```

Tip

Se esse comando lançar um erro, verifique se você atualizou a 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 grupos do Auto Scaling 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. Escolha Update (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-group` comando para ativar o aquecimento padrão da instância com um valor de 120 segundos para um grupo existente de Auto Scaling chamado my-asg.

```
aws autoscaling update-auto-scaling-group --auto-scaling-group-name my-asg --default-instance-warmup 120
```

Tip

Se esse comando lançar um erro, verifique se você atualizou a AWS CLI localmente para a versão mais recente.

Verificar o aquecimento de instância padrão para um grupo

Para verificar o aquecimento padrão da instância para um grupo do Auto Scaling (AWS CLI)

Use o comando [describe-auto-scaling-groups](#).

```
aws autoscaling describe-auto-scaling-groups --auto-scaling-group-name my-asg
```

Esta é uma resposta de exemplo.

```
{  
    "AutoScalingGroups": [  
        {  
            "AutoScalingGroupName": "my-asg",  
            "AutoScalingGroupARN": "arn:",  
            ...  
            "DefaultInstanceWarmup": 120  
        }  
    ]  
}
```

Encontre políticas de escalabilidade com um tempo de aquecimento da instância previamente definido

Para identificar se você tem políticas que têm seu próprio horário de aquecimentoEstimatedInstanceWarmup, execute o seguinte comando [describe-policies](#) usando o AWS CLI Substitua *my-asg* pelo nome do seu grupo de Auto Scaling.

```
aws autoscaling describe-policies --auto-scaling-group-name my-asg  
--query 'ScalingPolicies[?EstimatedInstanceWarmup!=`null`]'
```

A seguir está um exemplo de saída.

```
[  
    {  
        "AutoScalingGroupName": "my-asg",  
        "PolicyName": "cpu40-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": 40.0,  
                "DisableScaleIn": false  
            },  
            "TargetValue": 40.0,  
            "DisableScaleIn": false  
        ],  
        "TargetTrackingConfiguration": {  
            "PredefinedMetricSpecification": {  
                "PredefinedMetricType": "ASGAverageCPUUtilization"  
            },  
            "TargetValue": 40.0,  
            "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 ou AWS SDKs. AWS CLI Esta seção aborda as etapas do console. Se você usar os AWS SDKs AWS CLI ou, 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 grupos do Auto Scaling 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 Escalabilidade automática, em Políticas de escalonamento dinâmico, escolha a política na qual você está interessado e, em seguida, escolha Ações, Editar.
4. Para a necessidade de instâncias, limpe o valor de aquecimento da instância para usar o valor padrão de aquecimento da instância em vez disso.
5. Escolha Update (Atualizar).

Desaquecimento de escalabilidade para o Amazon EC2 Auto Scaling

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 desaquecimento é impedir que o grupo do Auto Scaling inicie ou encerre outras instâncias antes que os efeitos de atividades anteriores sejam visíveis.

Important

Como prática recomendada, recomendamos não usar políticas de escalabilidade simples e desaquecimento de escalabilidade.

Na maioria dos casos, uma política de dimensionamento com monitoramento do objetivo 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 \(p. 180\)](#) em escalabilidade simples ou escalabilidade em etapas.

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.

Índice

- [Considerações \(p. 206\)](#)
- [Ganchos do ciclo de vida causam mais atrasos \(p. 206\)](#)
- [Alterar o período de desaquecimento padrão \(p. 207\)](#)
- [Definir um período de desaquecimento para políticas de escalabilidade simples específicas \(p. 207\)](#)

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 \(p. 200\)](#).
- 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. Porém, é possível substituir esse comportamento e respeitar o desaquecimento padrão ao usar a 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平衡ador 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 causam mais atrasos

Caso um [gancho do ciclo de vida \(p. 252\)](#) 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). Ele não espera pela duração do gancho. 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-groups](#) comando.

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 políticas 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-policy](#) comando. Para obter mais informações, consulte [Políticas de escalabilidade simples e em etapas](#) (p. 190).

Configurações de aquecimento e desaquecimento disponíveis

Para ajudar a otimizar a performance de escalabilidade, escolha as configurações de aquecimento e desaquecimento apropriadas para o grupo do Auto Scaling.

Recomendamos usar a configuração `DefaultInstanceWarmup`, que unifica todas as configurações de aquecimento e desaquecimento. Se necessário, você também pode usar a configuração `HealthCheckGracePeriod`.

As outras configurações disponíveis não devem ser usadas ao mesmo tempo que a configuração `DefaultInstanceWarmup`. Por exemplo, `EstimatedInstanceWarmup` e `InstanceWarmup` não são necessários quando `DefaultInstanceWarmup` já está definido, e `DefaultCooldown` e `Cooldown` só são necessários quando você usa políticas de escalabilidade simples. (Como prática recomendada, recomendamos políticas de monitoramento do objetivo e escalabilidade de etapas em vez de políticas de escalabilidade simples.) Porém, continuaremos oferecendo suporte a todas essas configurações para que você possa escolher as configurações apropriadas para seu caso de uso.

DefaultInstanceWarmup (Recomendado)

Operação da API: [CreateAutoScalingGroup](#), [UpdateAutoScalingGroup](#)

O tempo decorrido, em segundos, até a inicialização de uma nova instância ser concluída e o consumo de recursos ficar estável após entrar no estado `InService`.

Durante uma atualização de instância, o Amazon EC2 Auto Scaling aguarda o período de aquecimento após a substituição de uma instância antes de passar para a substituição da próxima instância. O Amazon EC2 Auto Scaling também aguarda o período de aquecimento antes de agregar as métricas de novas instâncias às instâncias existentes nas métricas da Amazon que são usadas para escalabilidade, resultando em CloudWatch dados de uso mais confiáveis. Para obter mais informações, consulte [Definir o aquecimento padrão da instância para um grupo do Auto Scaling \(p. 200\)](#).

É possível reduzir o valor do período de aquecimento caso tenha usado um gancho do ciclo de vida para preparar novas instâncias para uso. Para obter mais informações, consulte [Ganchos do ciclo de vida do Amazon EC2 Auto Scaling \(p. 252\)](#).

Important

Para gerenciar várias configurações de aquecimento no nível do grupo, recomendamos configurar o aquecimento de instância padrão, mesmo que esteja definido como 0 segundo. Para remover um valor definido anteriormente, inclua a propriedade, mas especifique o valor -1. No entanto, é altamente recomendável manter o aquecimento de instância padrão habilitado especificando um valor de 0 ou outro valor nominal.

Padrão: nenhum

EstimatedInstanceWarmup

Não será necessário se `DefaultInstanceWarmup` estiver definido.

Operação da API: [PutScalingPolicy](#)

O tempo estimado, em segundos, até que uma instância recém-lançada possa contribuir com as CloudWatch métricas. Esse período de aquecimento se aplica a instâncias executadas por causa de uma política específica de monitoramento do objetivo ou escalabilidade de etapas. Quando um período de aquecimento é especificado aqui, ele substitui o aquecimento da instância padrão. Para ter mais informações, consulte [Políticas de escalabilidade de rastreamento de destino \(p. 180\)](#) e [Políticas de escalabilidade simples e em etapas \(p. 190\)](#).

Default (Padrão): se não for definido, o valor desse parâmetro assumirá como padrão o valor do aquecimento de instância padrão definido para o grupo. Se o aquecimento da instância padrão for nulo, ele retornará ao valor de desaquecimento padrão.

InstanceWarmup

Não será necessário se `DefaultInstanceWarmup` estiver definido.

Operação da API: [StartInstanceRefresh](#)

Um período, em segundos, em que a atualização de uma instância aguarda antes de passar para a substituição da próxima instância depois que uma nova instância entra no estado `InService`. Especificar um período de aquecimento ao iniciar uma atualização de instância substitui o aquecimento de instância padrão, mas apenas para a atualização da instância atual. Para obter mais informações, consulte [Substituir instâncias do Auto Scaling com base em uma atualização de instância \(p. 108\)](#).

Default (Padrão): se não for definido, o valor desse parâmetro assumirá como padrão o valor do aquecimento de instância padrão definido para o grupo. Se o aquecimento de instância padrão for nulo, ele voltará ao valor do período de carência da verificação de integridade.

DefaultCooldown

Só será necessário se você usar políticas de escalabilidade simples.

Operação da API: [CreateAutoScalingGroup](#), [UpdateAutoScalingGroup](#)

A quantidade de tempo, em segundos, entre uma ação de escalabilidade que termina e outra que começa por causa de políticas de escalabilidade simples. Para obter mais informações, consulte [Desaquecimento de escalabilidade para o Amazon EC2 Auto Scaling \(p. 205\)](#).

Default (Padrão): 300 segundos

Cooldown

Só será necessário se você usar políticas de escalabilidade simples.

Operação da API: [PutScalingPolicy](#)

Um período de desaquecimento, em segundos, que se aplica a uma política de escalabilidade simples específica. (As políticas de escalabilidade simples não são mais recomendadas. Como prática recomendada, recomendamos políticas de monitoramento do objetivo e escalabilidade de etapas.) Quando um período de desaquecimento é especificado aqui, ele substitui o desaquecimento padrão. Para obter mais informações, consulte [Desaquecimento de escalabilidade para o Amazon EC2 Auto Scaling \(p. 205\)](#).

Padrão: nenhum

HealthCheckGracePeriod

Operação da API: [CreateAutoScalingGroup](#), [UpdateAutoScalingGroup](#)

O tempo em segundos que o Amazon EC2 Auto Scaling aguardará antes de verificar o status de integridade de uma instância do EC2 que entrou em serviço e marcá-la como não íntegra por causa de falha no Elastic Load Balancing ou verificação de integridade personalizada. Isso será útil se suas instâncias não passarem imediatamente nessas verificações de integridade depois de entrarem no estado `InService`. Para obter mais informações, consulte [Definir um período de carência da verificação de integridade para um grupo do Auto Scaling \(p. 325\)](#).

Será possível definir o valor do período de carência da verificação de integridade como 0, se você usar um gancho do ciclo de vida para iniciar a instância. Se o grupo do Auto Scaling estiver atrás de um平衡ador de carga, considere a possibilidade de adicionar um gancho do ciclo de vida ao grupo para garantir que as instâncias estejam prontas para fornecer tráfego antes de serem registradas no balanceador de carga no final do ganho do ciclo de vida. Para obter mais informações, consulte [Ganchos do ciclo de vida do Amazon EC2 Auto Scaling \(p. 252\)](#).

Padrão: 300 segundos para um grupo do Auto Scaling criado por você no AWS Management Console. 0 segundo para um grupo do Auto Scaling criado por você usando a AWS CLI ou um SDK.

Escalabilidade baseada no Amazon SQS

Important

As informações e etapas a seguir mostram como calcular o acúmulo de filas do Amazon SQS por instância usando o atributo `ApproximateNumberOfMessages` que é usado antes de publicá-lo como uma métrica personalizada para o 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 [Crie uma política de escalabilidade de rastreamento de destino para Amazon EC2 Auto Scaling usando matemática em métricas \(p. 187\)](#).

Esta seção mostra como dimensionar o 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.

Índice

- [Usar o monitoramento do objetivo com a métrica correta \(p. 210\)](#)
- [Limitações e pré-requisitos \(p. 212\)](#)
- [Configurar escalabilidade baseada no Amazon SQS \(p. 212\)](#)
- [Amazon SQS e proteção contra redução de escala na horizontal de instâncias \(p. 214\)](#)

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 \(p. 182\)](#).

O problema de usar uma métrica do CloudWatch Amazon SQS, como `ApproximateNumberOfMessagesVisible` para rastreamento de alvos, 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.

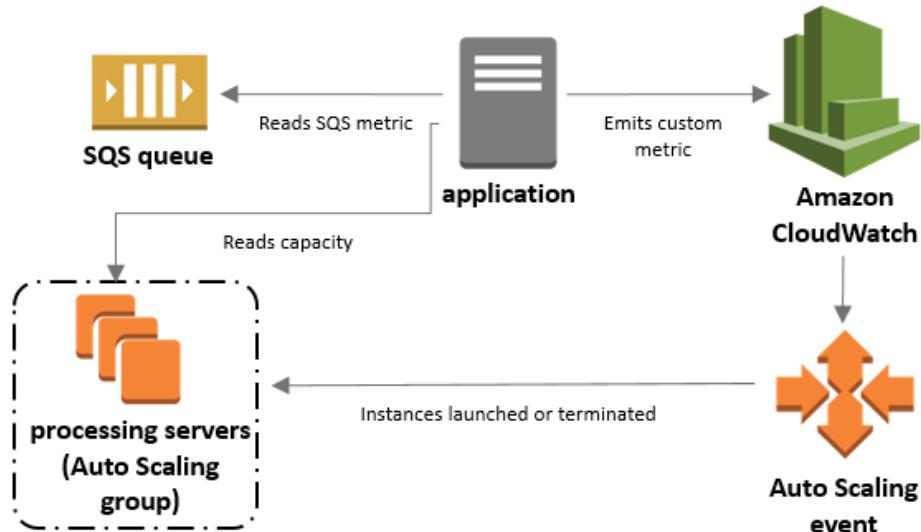
Important

Para reduzir custos, lembre-se de usar a matemática em métricas. Para obter mais informações, consulte [Crie uma política de escalabilidade de rastreamento de destino para Amazon EC2 Auto Scaling usando matemática em métricas \(p. 187\)](#).

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 com monitoramento do objetivo que configura seu grupo do Auto Scaling para escalar com base na métrica personalizada e em um valor de objetivo definido. Os alarmes do CloudWatch invocam a política de escalabilidade.

O diagrama a seguir ilustra a arquitetura dessa configuração.



Limitações e pré-requisitos

Para usar essa configuração, é necessário estar ciente das seguintes limitações:

- Você deve usar o AWS CLI ou um SDK para publicar sua métrica personalizada. CloudWatch Depois, é possível monitorar a métrica com o AWS Management Console.
- O console do Amazon EC2 Auto Scaling não oferece suporte a políticas de escalabilidade com monitoramento do objetivo que usam métricas personalizadas. Você deve usar a AWS CLI ou um SDK para especificar uma métrica personalizada para sua política de escalabilidade.

As seções a seguir orientam como usar a AWS CLI para as tarefas que você precisa executar. Por exemplo, para obter dados métricos que refletem o uso atual da fila, você usa o comando SQS. [get-queue-attributes](#) A CLI deve estar [instalada](#) e [configurada](#).

Antes de começar, é necessário ter uma fila do Amazon SQS para usar. Nas seções 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](#).

Configurar escalabilidade baseada no Amazon SQS

Tarefas

- [Etapa 1: criar uma métrica CloudWatch personalizada \(p. 212\)](#)
- [Etapa 2: Criar uma política de escalabilidade com monitoramento do objetivo \(p. 213\)](#)
- [Etapa 3: Testar sua política de escalabilidade \(p. 214\)](#)

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 conta da AWS. Depois, calcule a métrica de backlog por instância, conforme recomendado em uma seção anterior. Por fim, publique esse número no CloudWatch a uma 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 da lista de pendências por instância e publicá-lo em uma métrica CloudWatch personalizada. Ao publicar uma métrica personalizada, você especifica o nome, o namespace, a unidade, o valor e zero ou mais dimensões da métrica. Uma dimensão consiste em um nome de dimensão e um valor de dimensão.

Para publicar sua métrica personalizada, substitua os valores do espaço reservado em *italico* pelo nome de sua métrica preferida, o valor da métrica, um namespace (desde que não comece com "AWS") e dimensões (opcional) e execute o comando a seguir. [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 console do CloudWatch. Você pode acessá-la fazendo login no AWS Management Console e navegando para a página do CloudWatch. Depois, visualize a métrica navegando até a página de métricas ou procurando-a usando a caixa de pesquisa. Para obter informações sobre a visualização de métricas, consulte [Exibir 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 de rastreamento de metas.

Para criar uma política de escalabilidade de rastreamento de metas () AWS CLI

1. Use o cat comando a seguir para armazenar um valor alvo para sua política de escalabilidade e uma especificação métrica personalizada em um arquivo JSON nomeado config.json em seu diretório inicial. Substitua os valores do espaço reservado em *italico* pelos seus próprios valores. Para o TargetValue, calcule a métrica backlog aceitável por instância e insira-a aqui. Para calcular esse número, escolha 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 com a qual está registrada CloudWatch, que é CloudWatch usada 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 [Enviar mensagens para uma fila \(console\)](#) 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 [Recebimento e exclusão de 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 seus aplicativos no Amazon EC2 Auto Scaling para lidar com o encerramento de instâncias com elegância \(p. 305\)](#).

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 [Receber notificações do Amazon SNS quando o grupo do Auto Scaling escala \(p. 341\)](#).

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 grupos do Auto Scaling 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 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.
4. (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.
5. 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 de 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
```

A seguir está 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: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 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 \(p. 459\)](#).

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 grupos do Auto Scaling 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
```

A seguir está 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  
        }  
    ]  
}
```

Excluir uma política de escalabilidade

Quando você não precisar mais de uma política de escalabilidade, poderá excluí-la. Dependendo do tipo de política de escalabilidade, talvez seja necessário excluir os alarmes do CloudWatch. A exclusão de uma política de escalabilidade de rastreamento de destino também exclui todos os alarmes do CloudWatch associados. A exclusão de uma política de escalabilidade simples ou em etapas excluirá a ação de alarme subjacente, mas não excluirá o alarme do CloudWatch, mesmo se ele não tiver 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 grupos do Auto Scaling 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 Yes, Delete (Sim, excluir).
6. (Opcional) Se você excluiu uma política de escalabilidade em etapas ou uma política de escalabilidade simples, faça o seguinte para excluir o alarme do CloudWatch que foi associado à política. É possível ignorar essas subetapas para manter o alarme para uso futuro.
 - a. Abra o console do CloudWatch 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 Delete (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`]'
```

A seguir está um exemplo de saída.

```
[  
  {  
    "AutoScalingGroupName": "my-asg",  
    "PolicyName": "cpu40-target-tracking-scaling-policy",  
    "PolicyARN": "PolicyARN",  
    "PolicyType": "TargetTrackingScaling",  
    "StepAdjustments": [],  
    "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"  
      }  
    ],  
    "TargetTrackingConfiguration": {  
      "PredefinedMetricSpecification": {  
        "PredefinedMetricType": "ASGAverageCPUUtilization"  
      },  
      "TargetValue": 40.0,  
      "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 cpu40-target-tracking-scaling-policy
```

Para excluir seu alarme do CloudWatch (AWS CLI)

Para políticas de escalonamento simples e por etapas, 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 a AWS Command Line Interface (AWS CLI)

É possível criar políticas de escalabilidade para o Amazon EC2 Auto Scaling por meio do AWS Management Console, da AWS CLI ou de 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](#). Para obter exercícios introdutórios de criação de políticas de escalabilidade da AWS CLI, consulte [Políticas de escalabilidade de rastreamento de destino \(p. 180\)](#) e [Políticas de escalabilidade simples e em etapas \(p. 190\)](#).

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 cpu40-target-tracking-scaling-policy \
--auto-scaling-group-name my-asg --policy-type TargetTrackingScaling \
--target-tracking-configuration file://config.json
{
    "TargetValue": 40.0,
    "PredefinedMetricSpecification": {
        "PredefinedMetricType": "ASGAverageCPUUtilization"
    }
}
```

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"
}
```

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 precisará do ARN ao criar o alarme do CloudWatch.

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 precisará do ARN ao criar o alarme do CloudWatch.

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 precisará do ARN ao criar o alarme do CloudWatch.

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 precisará do ARN ao criar o alarme do CloudWatch.

Escala preditiva para o Amazon EC2 Auto Scaling

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.

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
- 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.

Usar o AWS Management Console, a AWS CLI ou um dos SDKs para adicionar uma política de escalabilidade preditiva a qualquer grupo do Auto Scaling.

Índice

- [Como a escalabilidade preditiva funciona \(p. 223\)](#)
- [Práticas recomendadas \(p. 223\)](#)
- [Criar uma política de escalabilidade preditiva \(console\) \(p. 224\)](#)
- [Criar uma política de escalabilidade preditiva \(AWS CLI\) \(p. 227\)](#)
- [Limitações \(p. 229\)](#)
- [Supported Regions \(Regiões compatíveis\) \(p. 229\)](#)
- [Avaliar as políticas de escalabilidade preditiva \(p. 230\)](#)
- [Substituir valores de previsão usando ações programadas \(p. 236\)](#)

- [Configurações avançadas de política de escalabilidade preditiva usando métricas personalizadas \(p. 239\)](#)

Como a escalabilidade preditiva funciona

A escalabilidade preditiva usa machine learning para prever requisitos de capacidade com base em dados históricos de CloudWatch. O algoritmo de machine learning consome os dados históricos disponíveis e calcula a capacidade que melhor se ajusta ao padrão de carga histórico e, em seguida, aprende continuamente com base em novos dados para tornar as previsões futuras mais precisas.

Para usar a escalabilidade preditiva, crie primeiro uma política de escalabilidade com um par de métricas e uma utilização-alvo. A criação da previsão começará imediatamente após a criação da política se houver pelo menos 24 horas de dados históricos. A escalabilidade preditiva localiza padrões em dados CloudWatch métricos dos 14 dias anteriores para criar uma previsão horária para as próximas 48 horas. Os dados de Forecast são atualizados diariamente com base nos dados CloudWatch métricos mais recentes.

Você pode configurar a escalabilidade preditiva no modo forecast only (somente previsão) para avaliar a previsão antes que a escalabilidade preditiva comece a modificar ativamente a capacidade. Em seguida, você poderá visualizar os dados de previsão e métricas recentes CloudWatch em forma de gráfico no console do Amazon EC2 Auto Scaling. Também é possível acessar dados de previsão usando a AWS CLI ou um dos SDKs.

Quando estiver pronto para começar a escalar com a escalabilidade preditiva, alterne a política do modo somente previsão para o modo previsão e escalabilidade. Depois de mudar para o modo previsão e escalabilidade, o grupo do Auto Scaling começa escalar a capacidade com base na previsão.

Usando a previsão, o Amazon EC2 Auto Scaling escala o número de instâncias no início de cada hora:

- Se a capacidade real for menor que a capacidade prevista, o Amazon EC2 Auto Scaling aumentará a escala do seu grupo do Auto Scaling na horizontal para que sua capacidade desejada seja igual à capacidade prevista.
- Se a capacidade real for maior do que a capacidade prevista, o Amazon EC2 Auto Scaling não terá a capacidade modificada.
- Os valores definidos para a capacidade mínima e máxima do grupo do Auto Scaling serão respeitados se a capacidade prevista estiver fora desse intervalo.

Práticas recomendadas

- 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 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.
- Para começar, use o console do Amazon EC2 Auto Scaling para criar várias políticas de escalabilidade preditiva no modo somente 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.

- Ao escolher uma métrica de carga, verifique se os dados descrevem o carregamento completo da aplicação. Além disso, verifique se é relevante para o aspecto de performance que você deseja escalar.
- Use a escalabilidade preditiva com a escalabilidade dinâmica. A escalabilidade dinâmica é usada para dimensionar automaticamente a capacidade em resposta a alterações em tempo real na utilização de recursos. 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 preditivo, a capacidade desejada do grupo será definida como 10. Recomendamos usar políticas de escalabilidade de rastreamento de metas para a maioria dos clientes que desejam começar a usar a escalabilidade dinâmica. Para obter mais informações, consulte [Políticas de escalabilidade com monitoramento do objetivo para o Amazon EC2 Auto Scaling \(p. 180\)](#).

Criar uma política de escalabilidade preditiva (console)

Você pode criar, visualizar e excluir as políticas de escalação preditiva com o console do Amazon EC2 Auto Scaling.

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 \(métricas personalizadas\) \(p. 225\)](#).

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.

Para criar uma política de escalabilidade preditiva

1. Abra o console do Amazon EC2 em <https://console.aws.amazon.com/ec2/> e escolha grupos do Auto Scaling 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.
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. Métricas personalizadas podem incluir outras métricas fornecidas pela CloudWatch ou as métricas 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 \(métricas predefinidas\) \(p. 224\)](#).

Para criar uma política de escalação preditiva usando métricas personalizadas, você deve fazer o seguinte:

- Você deve fornecer as consultas brutas que permitem que o Amazon EC2 Auto Scaling interaja com as métricas em CloudWatch. Para obter mais informações, consulte [Configurações avançadas de política de escalabilidade preditiva usando métricas personalizadas \(p. 239\)](#). Para ter certeza de que o

Amazon EC2 Auto Scaling possa extrair os dados métricos CloudWatch, confirme se cada consulta está retornando pontos de dados. Confirme isso usando o CloudWatch console ou a operação CloudWatch [GetMetricData](#) API.

Note

Fornecemos exemplos de cargas úteis JSON no editor de JSON no console do Amazon EC2 Auto Scaling. Esses exemplos oferecem uma referência para os pares de chave-valor exigidos para a adição de outras CloudWatch métricas fornecidas pela AWS ou as métricas que 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 \(p. 242\)](#). 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 [GetMetricData](#) 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 [Considerações e solução de problemas \(p. 246\)](#).

Para criar uma política de escalabilidade preditiva

1. Abra o console do Amazon EC2 em <https://console.aws.amazon.com/ec2/> e escolha grupos do Auto Scaling 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, você deve usar a matemática de métricas para agrregar 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 a AWS CLI conforme mostrado a seguir para configurar políticas de escalabilidade preditiva para seu grupo do Auto Scaling. Para obter mais informações sobre as CloudWatch métricas que você pode especificar para uma política de escalabilidade preditiva, consulte [PredictiveScalingMetricSpecification](#)na 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 essa política usando a linha de comando, execute o [put-scaling-policy](#) comando 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-policy](#) comando 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-policy](#) comando 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": []  
}
```

Limitações

- A escalabilidade preditiva requer 24 horas de histórico de métricas antes de começar gerar previsões.
- 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 \(p. 67\)](#), 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 vCPUs 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.

Supported Regions (Regiões compatíveis)

O Amazon EC2 Auto Scaling suporta políticas de escalabilidade preditiva nas seguintes políticas de escalabilidade preditiva nas seguintes áreasRegiões da AWS: Leste dos EUA (Norte da Virgínia), Leste

dos EUA (Oregon), Oeste dos EUA (Norte da Califórnia), África (Cidade do Cabo), Canadá (Central), UE (Frankfurt), UE (Irlanda), UE (Milão), UE (Paris), UE (Estocolmo), Ásia-Pacífico (Hong Kong), UE (Londres), UE (Milão), UE (Paris), UE (Estocolmo), Ásia-Pacífico (Hong Kong), Ásia-Pacífico (Hong Kong), UE (Londres), UE (Milão), UE (Paris), UE (Estocolmo), Ásia-Pacífico (Hong Kong), Ásia-Pacífico (Hong Kong), UE Jacarta), Ásia-Pacífico (Mumbai), Ásia-Pacífico (Osaka), Ásia-Pacífico (Tóquio), Ásia-Pacífico (Singapura), Ásia-Pacífico (Sydney), Oriente Médio (Bahrein), Oriente Médio (Emirados Árabes Unidos), América do Sul (São Paulo), China (Pequim), China (Ningxia), AWS GovCloud (Leste dos EUA) e AWS GovCloud (Oeste dos EUA).

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 escala 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.

Índice

- [Visualizar recomendações de escalabilidade preditiva \(p. 230\)](#)
- [Analisar grafos de monitoramento de escalabilidade preditiva \(p. 231\)](#)
- [Métricas de monitoramento com CloudWatch \(p. 233\)](#)

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 grupos do Auto Scaling 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.

- (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.

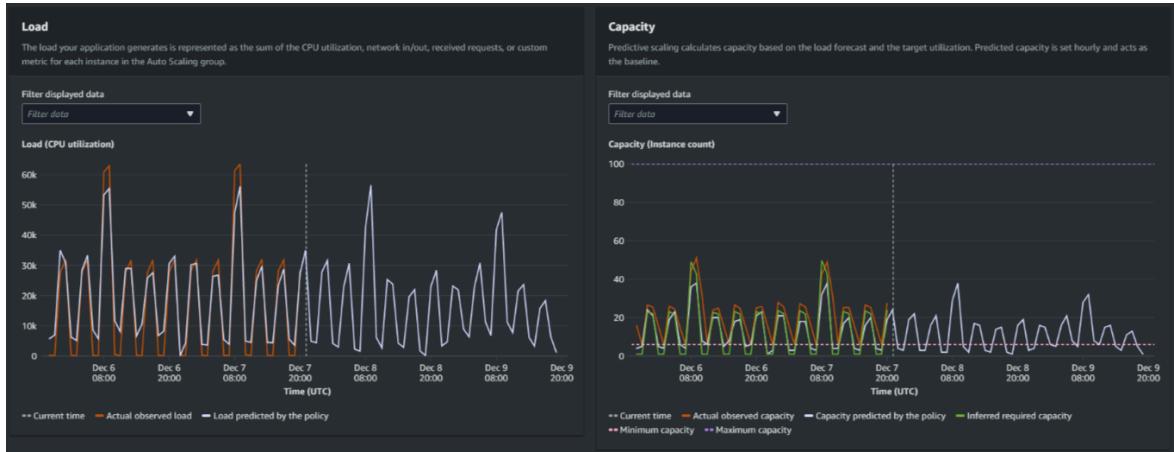
Analisar grafos de monitoramento de escalabilidade preditiva

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 escala a capacidade real.

Para analisar grafos de monitoramento de escalabilidade preditiva no console do Amazon EC2 Auto Scaling

- Escolha uma política na lista Políticas de escalabilidade preditiva.
- Na seção Monitorar, você pode visualizar as previsões passadas e futuras de carga e de capacidade da política em relação aos valores reais. O gráfico Carga exibe a previsão de carga e os valores reais para a métrica de carga escolhida. O gráfico 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 gráficos ficam disponíveis logo após a criação da política.
- (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 gráficos 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 no passado.
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 exibe a capacidade real do grupo do Auto Scaling no passado, o que depende de suas outras políticas de escalabilidade e do tamanho mínimo do grupo em vigor no 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.

Métricas de monitoramento com CloudWatch

Dependendo de suas necessidades, talvez você prefira acessar dados de monitoramento para escalabilidade preditiva da Amazon CloudWatch em vez de usar o console do 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 da coleta desses dados, eles serão armazenados automaticamente em CloudWatch intervalos regulares. Em seguida, você poderá 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 \(p. 233\)](#)
- [Criar métricas de precisão usando matemática métrica \(p. 234\)](#)

Visualizar dados históricos de previsão

Você pode visualizar os dados de previsão de carga e capacidade para uma política de escalabilidade CloudWatch preditiva 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 \(p. 335\)](#).

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 CloudWatch do 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 possa escolher qualquer estatística para cada métrica, nem todas as estatísticas são úteis para PredictiveScalingLoadForecasts PredictiveScalingCapacityForecastmétricas. Por exemplo, as estatísticas Average (Média), Minimum (Mínimo) e Maximum (Máximo) são úteis, mas a estatística Sum (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 métrica CPUUtilization e o grupo específico do Auto Scaling.
 9. (Opcional) Para adicionar o gráfico a um CloudWatch painel, escolha Actions (Ações), Add to dashboard (Adicionar ao painel).

Criar métricas de precisão usando matemática métrica

Com matemática métrica, você pode consultar várias CloudWatch métricas e usar expressões matemáticas para criar novas séries temporais de acordo com essas métricas. Você pode visualizar as séries temporais resultantes no CloudWatch console e adicioná-las aos painéis. Para obter mais informações sobre matemática métrica, consulte [Usar 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 do seu aplicativo 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). m1é 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 chamadamy-predictive-scaling-policy). m2é 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 da métrica de precisão, selecione a guia Métricas no CloudWatch console. Nele, é possível representar graficamente os dados. Para obter mais informações, consulte [Adição de uma expressão matemática a um CloudWatch gráfico](#) no Guia CloudWatch do 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, consulte [Criar um CloudWatch alarme com base em uma expressão matemática métrica](#) no Guia CloudWatch do usuário da Amazon. Para obter mais informações sobre o recebimento de alertas com o Amazon SNS, consulte [Configurar notificações do Amazon SNS](#) no Guia CloudWatch do usuário da Amazon.

Como alternativa, você pode usar [GetMetricData](#) [PutMetricAlarm](#) realizar 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 tempo de execução, 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.

Índice

- [Etapa 1: \(Opcional\) Analisar dados de séries temporais \(p. 236\)](#)
- [Etapa 2: Criar duas ações programadas \(p. 238\)](#)

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
        ]
    }
}
```

```
        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 grupos do Auto Scaling 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 Create (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 programadas, 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 do Auto Scaling especificado às 17h UTC em 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 \(p. 247\)](#).

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 pela AWS ou especificar métricas que você mesmo define e publica. Você também pode usar a matemática de métricas para agregar e transformar métricas existentes em uma nova série temporal que a AWS não rastreia 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.

Índice

- [Práticas recomendadas \(p. 240\)](#)
- [Pré-requisitos \(p. 240\)](#)
- [Estruture o JSON para métricas personalizadas \(p. 241\)](#)
- [Considerações e solução de problemas \(p. 246\)](#)
- [Limitações \(p. 247\)](#)

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 osCapacityForecast objetosLoadForecast e retornados por uma chamada para a [GetPredictiveScalingForecastAPI](#).
- Recomendamos a configuração da escalabilidade preditiva no modo apenas previsão para avaliar a previsão antes que a escalabilidade preditiva comece a modificarativamente a capacidade.

Pré-requisitos

Para adicionar métricas personalizadas à política de escalação preditiva, você deve ter as permissões `cloudwatch:GetMetricData`.

Para especificar suas próprias métricas em vez de usar as métricas que a AWS fornecer, você deve primeiro publicá-las 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 pontos de dados CloudWatch com base na duração do período necessário. Por exemplo, a especificação da métrica de carga usa métricas por hora para medir a carga em sua aplicação. CloudWatch usa seus dados de métrica publicados para

fornecer um único valor de dados para qualquer período de uma hora, agregando todos os pontos de dados com a data/hora que caem dentro de cada período de uma hora.

Estruture o JSON para métricas personalizadas

A seção a seguir contém exemplos de como configurar escalação 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 da métrica que está sendo aplicada.

1. Para criar uma política que obtenha dados diretamente de outras CloudWatch métricas fornecidas pela AWS ou das métricas que você publica CloudWatch, consulte [Exemplo de política de escalação preditiva com métricas personalizadas de carga e de dimensionamento \(AWS CLI\) \(p. 241\)](#).
2. Para criar uma política que possa consultar várias CloudWatch métricas e usar expressões matemáticas para criar novas séries temporais de acordo com essas métricas, consulte [Usar expressões de matemática métrica \(p. 242\)](#).

Exemplo de política de escalação preditiva com métricas personalizadas de carga e de dimensionamento (AWS CLI)

Para criar uma política de escalação preditiva com métricas personalizadas de carga e dimensionamento com a AWS CLI, armazene os argumentos para a `--predictive-scaling-configuration` em um arquivo JSON denominado `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": [  
                    {  
                        "MetricStat": {  
                            "Metric": {  
                                "MetricName": "MyUtilizationMetric",  
                                "Namespace": "MyNameSpace",  
                                "Dimensions": [  
                                    {  
                                        "Name": "MyOptionalMetricDimensionName",  
                                        "Value": "MyOptionalMetricDimensionValue"  
                                    }  
                                ]  
                            },  
                            "Stat": "Average"  
                        }  
                    }  
                ]  
            },  
            "CustomizedLoadMetricSpecification": {  
                "MetricDataQueries": [  
                    {  
                        "MetricStat": {  
                            "Metric": {  
                                "MetricName": "MyLoadMetric",  
                                "Namespace": "MyNameSpace",  
                                "Dimensions": [  
                                    {  
                                        "Name": "MyOptionalLoadMetricDimensionName",  
                                        "Value": "MyOptionalLoadMetricDimensionValue"  
                                    }  
                                ]  
                            },  
                            "Stat": "Sum"  
                        }  
                    }  
                ]  
            }  
        }  
    ]  
}
```

```
        "Name": "MyOptionalMetricDimensionName",
        "Value": "MyOptionalMetricDimensionValue"
    }
],
"Stat": "Sum"
}
]
}
}
```

Para obter mais informações, consulte [MetricDataQuery](#) Referência da API do Auto Scaling do Amazon EC2.

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 os valores exatos de nome da métrica, namespace e dimensões (se aplicável) para uma CloudWatch [métrica](#) com AWS CLI

Para criar essa política, execute o [put-scaling-policy](#) comando usando como entrada o arquivo JSON, como 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.

Índice

- [Noções básicas de matemática métrica \(p. 242\)](#)
- [Exemplo de política de escalação preditiva que combina métricas por meio da matemática de métricas \(AWS CLI\) \(p. 243\)](#)
- [Exemplo de política de escalação preditiva para usar em um cenário de implantação azul/verde \(AWS CLI\) \(p. 245\)](#)

Noções básicas de matemática métrica

Se tudo o que deseja fazer é agregar dados métricos existentes, a matemática CloudWatch métrica poupa o esforço e o custo de publicar outra métrica em CloudWatch. Você pode usar qualquer métrica que a

AWS fornece, e também pode usar métricas definidas como parte de suas aplicações. 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 [Como usar 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 CloudWatch compatível com.
- 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 [Como usar 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 [Escalabilidade baseada no Amazon SQS \(p. 210\)](#).

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": [
                    {
                        "MetricStat": {
                            "MetricName": "ApproximateNumberOfMessagesVisible",
                            "Namespace": "AWS/SQS",
                            "Statistics": [
                                "Sum"
                            ],
                            "Period": 300
                        }
                    },
                    {
                        "MetricStat": {
                            "MetricName": "GroupInServiceInstances",
                            "Namespace": "AWS/EC2",
                            "Statistics": [
                                "Sum"
                            ],
                            "Period": 300
                        }
                    }
                ],
                "Label": "SQSVisibleMessages"
            }
        }
    ]
}
```

```
"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 escalação preditiva para usar em um cenário de implantação azul/verde (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 útil especialmente para implantações azul/verde.

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 do Auto Scaling são criados como parte de uma implantação azul/verde, o histórico de métricas 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 [Using EC2 Auto Scaling predictive scaling policies with Blue/Green deployments](#) (Como usar políticas de escalabilidade preditiva do EC2 Auto Scaling com implantações azul/verde) no AWS Compute Blog (Blog de computação).

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": []
}
```

Considerações e solução de problemas

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 quando você estiver tentando usar uma expressão de pesquisa em um cenário de implantação azul/verde, primeiro certifique-se que você entende como criar uma expressão de pesquisa que procure uma correspondência parcial ao invés 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 [CloudWatch pesquisas](#), [consulte a sintaxe da expressão](#) de pesquisa no Guia CloudWatch do usuário da Amazon.
- Se você não validou uma expressão com antecedência, o [put-scaling-policy](#) comando a valida quando você cria 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 para o [get-metric-data](#) comando. Você também pode solucionar problemas 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 \(console\) \(p. 336\)](#).
- 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.
- Você deve 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.

Escalabilidade programada para o Amazon EC2 Auto Scaling

A escalabilidade programada permite que você defina sua própria programação de escalabilidade de acordo com alterações de carga previsíveis. Por exemplo, vamos supor que toda semana o tráfego para sua aplicação Web comece a aumentar na quarta-feira, permaneça alto na quinta-feira e comece a diminuir na sexta-feira. É possível configurar uma programação para o Amazon EC2 Auto Scaling para aumentar a capacidade na quarta-feira e diminuir a capacidade na sexta-feira.

Para usar a escalabilidade programada, crie ações programadas. As ações programadas são executadas automaticamente como uma função de data e hora. Ao criar uma ação programada, especifique quando a ação de escalabilidade deve ocorrer e os novos tamanhos mínimo e máximo para a ação de escalabilidade. É possível criar ações programadas para escalar uma única vez ou de forma programada.

Índice

- [Considerações \(p. 247\)](#)
- [Programações recorrentes \(p. 248\)](#)
- [Criar e gerenciar ações programadas \(console\) \(p. 248\)](#)
- [Criar e gerenciar ações programadas \(AWS CLI\) \(p. 250\)](#)
- [Limitações \(p. 252\)](#)

Considerações

Ao criar uma ação programada, lembre-se do seguinte:

- Uma ação programada define os tamanhos mínimo, máximo e desejado para o que é especificado pela ação programada no horário especificado. A solicitação pode, opcionalmente, incluir apenas um desses

tamanhos. Por exemplo, você pode criar uma ação programada com apenas a capacidade desejada especificada. Em alguns casos, no entanto, você deve incluir os tamanhos mínimo e máximo para garantir que a nova capacidade desejada especificada na ação não esteja fora desses limites.

- 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 fuso horário que siga o horário de verão, ele se ajustará automaticamente ao horário de verão (DST).
- 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 um processo para um grupo do Auto Scaling \(p. 312\)](#).
- 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 próximasumas das outras podem demorar mais para serem executadas.

Programações recorrentes

É possível criar ações programadas para escalar seu grupo do Auto Scaling com uma programação recorrente.

Para criar uma programação recorrente usando a AWS CLI ou um SDK, especifique uma expressão cron e um fuso horário para descrever quando essa ação agendada deve ocorrer. Opcionalmente, você pode especificar uma data e hora para a hora de início, a hora de término ou ambas.

Para criar uma programação recorrente usando o AWS Management Console, especifique o padrão de recorrência, o fuso horário, a hora de início e a hora de término opcional da ação programada. 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 obter outros exemplos de expressões do 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.
- O horário de início e o horário de término devem ser definidos em UTC quando você usar a AWS CLI ou um SDK.

Criar e gerenciar ações programadas (console)

Use os procedimentos desta seção para criar e gerenciar ações programadas usando o AWS Management Console.

Se você criar uma ação programada usando o console e especificar um fuso horário que observe o horário de verão (DST), tanto a programação recorrente quanto as horas de início e término se ajustam automaticamente ao horário de verão.

Criar uma ação programada

Conclua o procedimento a seguir para criar uma ação programada para escalar o grupo do Auto Scaling.

Para criar uma ação de escalabilidade programada para um grupo do Auto Scaling

1. Abra o console do Amazon EC2 em <https://console.aws.amazon.com/ec2/> e escolha grupos do Auto Scaling 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 Desired capacity (Capacidade desejada), Min. (Mínimo) ,Max (Máximo), escolha o novo tamanho desejado do grupo e a nova capacidade mínima e máxima.
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.

Note

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 Create (Criar). O console exibe as ações programadas para o grupo do Auto Scaling.

Verifique a hora, a data e o fuso horário

Para verificar se a hora, a data e o fuso horário estão configurados corretamente, verifique os valores Start time (Hora de início), End Time (Hora de término) e Time zone (Fuso horário) na tabela Scheduled actions (Ações programadas) na guia Automatic scaling (Escalabilidade automática) do seu grupo do Auto Scaling.

O Amazon EC2 Auto Scaling mostra os valores de Start time (Horário de início) e End time (Horário de término) no seu horário local com o deslocamento de UTC em vigor na data e hora especificadas. O

deslocamento de UTC é a diferença, em horas e minutos, da hora local para a UTC. O valor de Time zone (Fuso horário) mostra seu fuso horário solicitado, por exemplo, America/New_York.

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.

Atualizar uma ação programada

Depois de criar uma ação programada, você pode atualizar qualquer uma de suas configurações, exceto o nome.

Para atualizar uma ação programada

1. Abra o console do Amazon EC2 em <https://console.aws.amazon.com/ec2/> e escolha grupos do Auto Scaling 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), selecione uma ação programada.
4. Selecione Ações, Editar.
5. Faça as alterações necessárias e, em seguida, escolha Save changes (Salvar alterações).

Excluir uma ação programada

Quando você não precisar mais de uma ação programada, poderá excluí-la.

Para excluir uma ação programada

1. Abra o console do Amazon EC2 em <https://console.aws.amazon.com/ec2/> e escolha grupos do Auto Scaling 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. Escolha Actions, Delete.
5. Quando a confirmação for solicitada, escolha Yes, Delete (Sim, excluir).

Criar e gerenciar ações programadas (AWS CLI)

Você pode criar e atualizar ações agendadas que escalam apenas uma vez ou que são escaladas em uma agenda recorrente usando o comando [put-scheduled-update-group-action](#).

Criar uma ação programada que ocorre apenas uma vez

Para escalar automaticamente seu grupo do Auto Scaling apenas uma vez, em uma data e hora especificada, use a opção `--start-time "YYYY-MM-DDThh:mm:ssZ"`.

Exemplo: para escalar apenas uma vez

Para aumentar o número de instâncias em execução em seu grupo do Auto Scaling em um horário específico, use o comando a seguir.

Na data e hora especificadas por `--start-time` (8:00 AM UTC em 31 de março de 2021), se o grupo tiver na ocasião menos de 3 instâncias, sofrerá aumento de escala na horizontal para 3 instâncias.

```
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
```

Exemplo: para escalar apenas uma vez

Para diminuir o número de instâncias em execução em seu grupo do Auto Scaling em um horário específico, use o comando a seguir.

Na data e hora especificadas por `--start-time` (4:00 PM UTC em 31 de março de 2021), se o grupo tiver na ocasião mais de uma instância, sofrerá redução de escala na horizontal para uma instância.

```
aws autoscaling put-scheduled-update-group-action --scheduled-action-name my-one-time-action \
--auto-scaling-group-name my-asg --start-time "2021-03-31T16:00:00Z" --desired-capacity 1
```

Criar uma ação programada que é executada em uma programação recorrente

Para programar a escalabilidade em uma programação recorrente, use a opção `--recurrence` "cron expression".

Veja a seguir um exemplo de uma ação programada que especifica uma expressão do cron.

Na programação especificada (todos os dias às 9:00 AM UTC), se o grupo tiver na ocasião menos de 3 instâncias, sofrerá aumento de escala na horizontal para 3 instâncias. Se o grupo tiver na ocasião mais de 3 instâncias, ele sofrerá redução de escala na horizontal para 3 instâncias.

```
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
```

Criar uma ação programada recorrente que especifica um fuso horário

As ações programadas são definidas para o fuso horário UTC por padrão. Para especificar um fuso horário diferente, inclua a opção `--time-zone` e especifique o nome canônico do fuso horário IANA (America/New_York, por exemplo). Para obter mais informações, consulte https://en.wikipedia.org/wiki/List_of_tz_database_time_zones.

Veja a seguir um exemplo que usa uma opção `--time-zone` ao criar uma ação programada recorrente para escalar capacidade.

Na programação especificada (de segunda-feira a sexta-feira, às 6:00 PM, horário local), se o grupo tiver na ocasião menos de 2 instâncias, sofrerá aumento de escala na horizontal para 2 instâncias. Se o grupo tiver na ocasião mais de 2 instâncias, ele sofrerá redução de escala na horizontal para 2 instâncias.

```
aws autoscaling put-scheduled-update-group-action --scheduled-action-name my-recurring-action \  
    --auto-scaling-group-name my-asg --recurrence "0 18 * * 1-5" --time-zone "America/  
New_York" \  
    --desired-capacity 2
```

Descrever ações programadas

Para descrever as ações agendadas para um grupo de Auto Scaling, use o [describe-scheduled-actions](#) comando a seguir.

```
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  
    }  
  ]  
}
```

Excluir uma ação programada

Para excluir uma ação agendada, use o [delete-scheduled-action](#) comando a seguir.

```
aws autoscaling delete-scheduled-action --auto-scaling-group-name my-asg \  
    --scheduled-action-name my-recurring-action
```

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.

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 ganchos 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 de 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 escalonamento, um gancho de ciclo de vida pausa a instância antes que ela seja encerrada e envia uma notificação usando a AmazonEventBridge. Enquanto a instância está em estado de espera, você pode invocar uma função AWS Lambda ou se conectar à instância para baixar logs ou outros dados antes que a instância seja totalmente terminada.

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.

Índice

- [Disponibilidade de ganchos do ciclo de vida \(p. 253\)](#)
- [Considerações e limitações dos ganchos do ciclo de vida \(p. 254\)](#)
- [Recursos relacionados \(p. 255\)](#)
- [Como os ganchos do ciclo de vida funcionam \(p. 255\)](#)
- [Preparar para adicionar um gancho do ciclo de vida a um grupo do Auto Scaling \(p. 257\)](#)
- [Recuperar o estado de destino do ciclo de vida por meio de metadados de instância \(p. 262\)](#)
- [Adicionar ganchos do ciclo de vida \(p. 263\)](#)
- [Concluir uma ação do ciclo de vida \(p. 266\)](#)
- [Tutorial: configurar dados do usuário para recuperar o estado de destino do ciclo de vida por meio de metadados de instância \(p. 267\)](#)
- [Tutorial: Configurar um gancho do ciclo de vida que invoca uma função do Lambda \(p. 273\)](#)

Disponibilidade de ganchos do ciclo de vida

A tabela a seguir lista os ganchos do ciclo de vida disponíveis para vários cenários.

Evento	Início ou término da instância ¹	Maximum Instance Lifetime (Tempo de vida máximo da instância): instâncias de substituição	Instance Refresh (Atualização das instâncias): 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	✓	✓	✓	✓	✓

Evento	Início ou término da instância ¹	<u>Maximum Instance Lifetime</u> (Tempo de vida máximo da instância): instâncias de substituição	<u>Instance Refresh</u> (Atualização das instâncias): instâncias de substituição	<u>Capacity Rebalancing</u> (Rebalanceamento de capacidade): instâncias de substituição	<u>Warm Pools</u> (Grupos de alta atividade): instâncias entrando e saindo do grupo de alta atividade
Término de instâncias	✓	✓	✓	✓	✓

¹ Aplica-se a instâncias iniciadas ou terminadas quando o grupo é criado ou excluído, quando o grupo é escalado automaticamente ou quando você ajusta manualmente a capacidade desejada do grupo. 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 de ciclo de vida, lembre-se das 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 \(p. 8\)](#). 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 grupo de alta atividade \(p. 285\)](#).
- Ao reduzir a escala, o Amazon EC2 Auto Scaling não conta com uma nova instância como agregada CloudWatch métricas de instância do grupo Auto Scaling (como utilização da CPU, NetworkIn, NetworkOut, e assim por diante) até que o gancho do ciclo de vida de lançamento termine. Quando o aquecimento padrão de instância não estiver ativado ou estiver ativado, mas definido como 0, as instâncias do Auto Scaling começarão a contribuir com dados de uso para as métricas agregadas de instância assim que atingirem o estado InService. Para obter mais informações, consulte [Definir o aquecimento padrão da instância para um grupo do Auto Scaling \(p. 200\)](#).
- 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.
- Quando um grupo do Auto Scaling inicia ou encerra instâncias, as ações de escalabilidade iniciadas por políticas simples de escalabilidade são pausadas. Se os ganchos do ciclo de vida forem invocados, as ações de escalabilidade decorrentes de políticas simples de escalabilidade serão pausadas até que as ações do ciclo de vida tenham sido concluídas e o período de desaquecimento tenha expirado. 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 [Desaquecimento de escalabilidade para o Amazon EC2 Auto Scaling \(p. 205\)](#).
- 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 [Interrupção de instâncias spot](#) no Manual do usuário do Amazon EC2 para instâncias do Linux. 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 [Usar o rebalanceamento de capacidade para lidar com interrupções de spot do Amazon EC2 \(p. 346\)](#).

- 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 abandonado ou continuado. Se uma instância estiver sendo iniciada, continuar indica que suas ações foram bem-sucedidas e que o Amazon EC2 Auto Scaling pode colocar a instância em serviço. Caso contrário, abandono 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 permitindo que a instância seja encerrada. No entanto, abandonar (abandonar) interrompe quaisquer ações restantes, como outros ganchos do ciclo de vida, e continue (continuar) permite que quaisquer outros ganchos de 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 de ciclo de vida.
- O processo de criação e atualização de ganchos do ciclo de vida usando a AWS CLI, o AWS CloudFormation ou um SDK fornece opções não disponíveis ao criar um gancho do ciclo de vida diretamente no AWS Management Console. 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 AmazonEventBridge. É possível filtrar esses eventos e redirecioná-los para serviços da AWS como o Lambda, Amazon SNS e Amazon SQS conforme necessário.
- Você pode adicionar vários ganchos de ciclo de vida a um grupo de Auto Scaling enquanto o cria, chamando o [CreateAutoScalingGroup API](#) usando a AWS CLI, AWS CloudFormation, ou um SDK. No entanto, cada gancho deve ter o mesmo destino de notificação e função 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 de cada vez em chamadas separadas para [PutLifecycleHook API](#).

Recursos relacionados

Para ver um vídeo introdutório, consulte [AWS re:Invent 2018: gerenciamento de capacidade facilitado com o Amazon EC2 Auto Scaling](#) em YouTube.

Também fornecemos alguns trechos de modelo JSON e YAML que você pode usar para entender como declarar ganchos do ciclo de vida em seus modelos de pilha de AWS CloudFormation. Para obter mais informações, consulte o [AWS::AutoScaling::LifecycleHook](#) referência no AWS CloudFormation Guia do usuário.

Você também pode visitar nosso [GitHub repositório](#) para baixar modelos de exemplo e scripts de dados do usuário para ganchos de ciclo de vida.

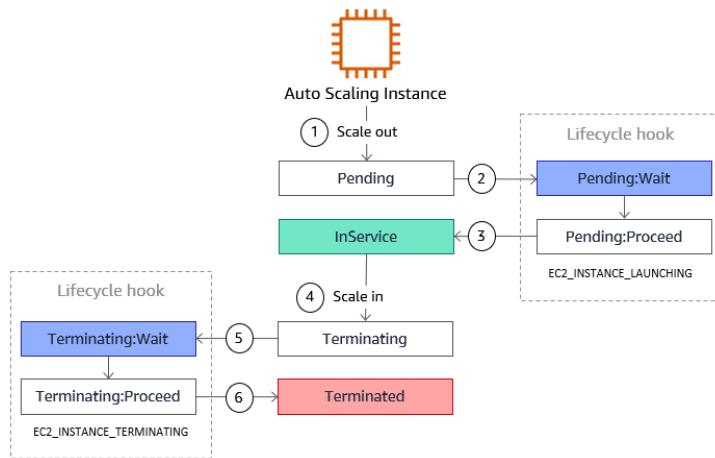
Para ver exemplos do uso de ganchos de ciclo de vida, consulte as seguintes postagens do blog.

- [Criação de um sistema de backup para instâncias escaladas usando o Lambda e o Amazon EC2 Run Command](#)
- [Execute o código antes de encerrar uma instância do EC2 Auto Scaling.](#)

Como os ganchos do ciclo de vida funcionam

Uma instância do Amazon EC2 passa por diferentes estados do momento em que é iniciada até seu término. Você pode criar ganchos do ciclo de vida para agir quando uma instância faz a transição para um estado de espera.

A ilustração a seguir mostra as transições entre estados de instância do Auto Scaling.



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平衡ador 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 grupo de alta atividade 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 grupo de alta atividade \(p. 285\)](#).

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 o Amazon EC2 Auto Scaling quando a ação do ciclo de vida estiver concluída, você deve adicionar o[CompleteLifecycleAction](#)Chame a API para o script e você deve criar manualmente uma função do IAM com uma política que permita que 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 ter mais informações, consulte [Concluir uma ação do ciclo de vida \(p. 266\)](#) e [Funções do IAM para aplicações que são executadas em instâncias do Amazon EC2 \(p. 448\)](#).
- Para usar um serviço como o Lambda para realizar uma ação personalizada, você já deve ter criado umEventBridge e especificou uma função Lambda como seu alvo. Para obter mais informações, consulte [Configurar um destino de notificação para notificações de ciclo de vida \(p. 257\)](#).
- Para permitir que o Lambda sinalize o Amazon EC2 Auto Scaling quando a ação do ciclo de vida estiver concluída, você deve adicionar o[CompleteLifecycleAction](#)Chamada de API para o 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 de ciclo de vida. Para obter mais informações, consulte [Tutorial: Configurar um gancho do ciclo de vida que invoca uma função do Lambda \(p. 273\)](#).
- 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 a função 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 de ciclo de vida \(p. 257\)](#).

Note

Por padrão, quando você adiciona um gancho de ciclo de vida no console, o Amazon EC2 Auto Scaling envia notificações de eventos do ciclo de vida para a AmazonEventBridge. UsandoEventBridgeou um script de dados do usuário é a melhor prática recomendada. Para criar um gancho do ciclo de vida que envie notificações diretamente para o Amazon SNS ou o Amazon SQS, use a AWS CLI, o AWS CloudFormation ou um SDK para adicionar o gancho do ciclo de vida.

Configurar um destino de notificação para notificações de 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.

A primeira abordagem usa a AmazonEventBridgepara invocar uma função Lambda que executa a ação desejada. A segunda abordagem envolve a criação de um tópico do Amazon Simple Notification Service (Amazon SNS) no qual as notificações são publicadas. Os clientes podem se inscrever no tópico do SNS e receber mensagens publicadas usando um protocolo compatível. A última abordagem envolve o uso do Amazon Simple Queue Service (Amazon SQS), um sistema de mensagens usado por aplicações distribuídas para trocar mensagens por meio de um modelo de pesquisa.

Como melhor prática, recomendamos que você useEventBridge. As notificações enviadas ao Amazon SNS e ao Amazon SQS contêm as mesmas informações que as notificações que o Amazon EC2 Auto Scaling envia paraEventBridge. AntesEventBridge, a prática padrão era enviar uma notificação ao SNS ou SQS e integrar outro serviço ao SNS ou SQS para realizar ações programáticas. Hoje,EventBridgeoferece mais opções para quais serviços você pode segmentar e facilita o gerenciamento de eventos usando a arquitetura sem servidor.

Os procedimentos a seguir abordam como configurar seu destino de notificação.

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.

Índice

- [Encaminhe as notificações para o Lambda usandoEventBridge \(p. 258\)](#)
- [Receba notificações usando o Amazon SNS \(p. 260\)](#)
- [Receba notificações usando o Amazon SQS \(p. 261\)](#)
- [Exemplo de mensagem de notificação para o Amazon SNS e o Amazon SQS \(p. 261\)](#)

Important

OEventBridgeA 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 as notificações para o Lambda usandoEventBridge

Você pode configurar umEventBridgeregra para invocar uma função Lambda quando uma instância entra em um estado de espera. O Amazon EC2 Auto Scaling emite uma notificação de evento do ciclo de vida paraEventBridge 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 \(p. 401\)](#).

Note

Quando você usa oAWS Management Consolepara criar uma regra de evento, o console adiciona automaticamente as permissões do IAM necessárias para concederEventBridgepermissão 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 noEventBridgeconsole, veja[Criando a AmazonEventBridgeregras que reagem aos eventosnaAmazôniaEventBridgeGuia do usuário](#).

- 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 \(p. 273\)](#). Este tutorial mostra como criar uma função Lambda simples que escuta eventos de lançamento e os grava em umCloudWatchRegistro de registros.

Para criar umEventBridgeregra que invoca 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 EventBridgealvo. 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 de ciclo de vida de execução de instância. Substitua o texto em **italico** pelo nome do grupo do Auto Scaling.

```
{  
    "source": [ "aws.autoscaling" ],  
    "detail-type": [ "EC2 Instance-launch Lifecycle Action" ],  
    "detail": {  
        "AutoScalingGroupName": [ "my-asg" ]  
    }  
}
```

Se o comando for executado com êxito, EventBridge responde 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 [Usar EventBridge para lidar com eventos do Auto Scaling \(p. 400\)](#).

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, enquanto o valor para o parâmetro Arn 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 na entidade principal do serviço do EventBridge (events.amazonaws.com) e nas permissões de escopo 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 do Lambda que deseja que a regra use como destino.
- **my-unique-id** é um identificador exclusivo que você define para descrever a declaração na política de funções do Lambda.
- **source-arn** é o ARN do EventBridge regra.

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 [Adicionar ganchos do ciclo de vida \(p. 263\)](#) 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 Create role (Criar função).
 - 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 \(AWS CLI\) \(p. 265\)](#) 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 Create role (Criar função).
 - 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 \(AWS CLI\) \(p. 265\)](#) como próxima etapa.

Exemplo de mensagem de notificação para o Amazon SNS e o Amazon SQS

Enquanto a instância está em um estado de espera, uma mensagem é publicada no destino de notificação do Amazon SNS ou do Amazon SQS. A mensagem inclui as seguintes informações:

- `LifecycleActionToken` — O token da ação de ciclo de vida.
- `AccountId`: o ID da Conta da AWS.
- `AutoScalingGroupName`: o nome do grupo do Auto Scaling.
- `LifecycleHookName` — O nome do gancho de ciclo de vida.
- `EC2InstanceId` — A ID da instância EC2.
- `LifecycleTransition` — O tipo de gancho de 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
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 de 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 fornecidos pelo Amazon EC2 Auto Scaling paraEventBridge, veja[Referência de eventos do Amazon EC2 Auto Scaling \(p. 401\)](#).

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 do estado do ciclo de vida, você deve recuperar o estado do ciclo de vida de destino por meio dos metadados da instância.

Por exemplo, talvez você precise 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 pesquisa o estado do ciclo de vida de uma instância diretamente da instância. Em seguida, você pode adicionar um gancho de ciclo de vida ao grupo Auto Scaling para manter a instância em execução até que seu código envie ocomplete-lifecycle-action comando 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 grupo de alta atividade, 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 alvo de dentro da instância, você deve 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 Serviço de metadados de instância 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 metadados de instância de uma instância em execução usando um de dois métodos: Instance Metadata Service Version 1 (IMDSv1 – Serviço de metadados de instância versão 1) ou Instance Metadata Service Version 2 (IMDSv2 – Serviço de metadados de instância versão 2). O IMDSv2 usa solicitações orientadas a sessão e mitiga vários tipos de vulnerabilidades que podem ser usadas para tentar ganhar acesso aos metadados de instância. Para obter detalhes sobre esses dois métodos, consulte [Use IMDSv2](#) (Usar o IMDSv2) no Guia do usuário do Amazon EC2 para instâncias Linux.

IMDSv2

```
TOKEN=`curl -X PUT "http://169.254.169.254/latest/api/token" -H "X-aws-ec2-metadata-token-ttl-seconds: 21600"`\n&& curl -H "X-aws-ec2-metadata-token: $TOKEN" -v http://169.254.169.254/latest/meta-data/autoscaling/target-lifecycle-state
```

IMDSv1

```
curl http://169.254.169.254/latest/meta-data/autoscaling/target-lifecycle-state
```

A seguir está 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 de 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 mais informações sobre a recuperação de metadados de instância, consulte [Retrieve instance metadata](#) (Recuperar metadados de instância) no Guia do usuário do Amazon EC2 para instâncias Linux.

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: configurar dados do usuário para recuperar o estado de destino do ciclo de vida por meio de metadados de instância \(p. 267\)](#).

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.

Adicionar ganchos do ciclo de vida

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.

Após criar um grupo do Auto Scaling no AWS Management Console, você pode adicionar um ou mais ganchos do ciclo de vida a ele, até um total de 50 ganchos do ciclo de vida. Também é possível usar a AWS CLI, o AWS CloudFormation ou um SDK para adicionar ganchos do ciclo de vida a um grupo do Auto Scaling conforme você o cria.

Por padrão, quando você adiciona um gancho de ciclo de vida no console, o Amazon EC2 Auto Scaling envia notificações de eventos do ciclo de vida para a AmazonEventBridge. UsandoEventBridgeou um script de dados do usuário é a melhor prática recomendada. Para criar um gancho de ciclo de vida que envia notificações diretamente para o Amazon SNS ou o Amazon SQS, você pode usar o[put-lifecycle-hook](#) comando, conforme mostrado nos exemplos deste tópico.

Índice

- [Adicionar ganchos do ciclo de vida \(console\) \(p. 264\)](#)
- [Adicionar ganchos do ciclo de vida \(AWS CLI\) \(p. 265\)](#)

Adicionar ganchos do ciclo de vida (console)

Siga estas etapas para adicionar ganchos de ciclo de vida ao seu grupo de Auto Scaling. Para adicionar ganchos de ciclo de vida para expansão (inicialização de instâncias) e expansão (instâncias encerrando ou retornando a uma piscina aquecida), você deve 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 \(p. 257\)](#).

Para adicionar um gancho de ciclo de vida para redução de escala

1. Abra o console do Amazon EC2 em <https://console.aws.amazon.com/ec2/> e escolha grupos do Auto Scaling 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 hooks (Ganchos de ciclo de vida), escolha Create lifecycle hook (Criar gancho de ciclo de vida).
4. Para definir um gancho de ciclo de vida para expansão (inicialização 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. ParaTempo limite do batimento cardíaco, especifique a quantidade de tempo, em segundos, para que as instâncias permaneçam em estado de espera durante a expansão, antes que o gancho atinja o tempo limite. O intervalo é de 30 a 7200 segundos. Definir um longo período de tempo limite fornece mais tempo para que sua ação personalizada seja concluída. Em seguida, se você terminar antes que o tempo limite termine, envie o[complete-lifecycle-action](#) comando para permitir que a instância prossiga para o próximo estado.
 - d. Em Default result (Resultado padrão), especifique a ação a ser adotada mediante o término do tempo limite do ganho do ciclo de vida ou quando houver uma falha inesperada. Você pode escolher entreCONTINUARouABANDONO.
 - Se você escolherCONTINUAR, o grupo Auto Scaling pode continuar com qualquer outro ganho do ciclo de vida e, em seguida, colocar a instância em serviço.
 - Se você escolherABANDONO, o grupo Auto Scaling interrompe todas as ações restantes e encerra a instância imediatamente.
 - e. (Opcional) ParaMetadados de notificação, especifique outras informações que você deseja incluir quando o Amazon EC2 Auto Scaling envia uma mensagem para o destino da notificação.
5. Escolha Create (Criar).

Para adicionar um gancho de ciclo de vida para escalar

1. EscolhaCrie um gancho de ciclo de vidapara continuar de onde você parou depois de criar um gancho de ciclo de vida para o scale-out.
2. Para definir um gancho de ciclo de vida para escalar (instâncias que terminam ou retornam a uma piscina aquecida), 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. ParaTempo limite do batimento cardíaco, especifique a quantidade de tempo, em segundos, para que as instâncias permaneçam em estado de espera durante a expansão, antes que o gancho atinja o tempo limite. Recomendamos um curto período de tempo limite de30para120segundos, dependendo de quanto tempo você precisa para realizar qualquer tarefa final, como extrair registros do EC2 doCloudWatch.
 - 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 escolherABANDONO, o grupo Auto Scaling encerra a instância imediatamente.
 - e. (Opcional) ParaMetadados de notificação, especifique outras informações que você deseja incluir quando o Amazon EC2 Auto Scaling envia uma mensagem para o destino da notificação.
3. Escolha Create (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`.

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, [oput-lifecycle-hook](#) comando 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 gancho 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 [orecord-lifecycle-action-heartbeat](#) Comando CLI. Isso estende o tempo limite pelo valor especificado quando você criou o gancho de ciclo de vida. Se você terminar antes que o tempo limite termine, você pode enviar [ocomplete-lifecycle-action](#) Comando CLI para permitir que a instância prossiga para o próximo estado. Para ter mais informações e exemplos, consulte [Concluir uma ação do ciclo de vida \(p. 266\)](#).

Concluir uma ação do ciclo de vida

Quando o grupo do Auto Scaling responde a um evento de 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.

Concluindo a ação do ciclo de vida com um resultado de `CONTINUE` é útil se você terminar antes que o tempo limite tenha expirado. 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 período de tempo limite.

Índice

- [Concluir uma ação do ciclo de vida \(manual\) \(p. 266\)](#)
- [Concluir uma ação do ciclo de vida \(automática\) \(p. 267\)](#)

Concluir uma ação do ciclo de vida (manual)

O procedimento a seguir é para a interface de linha de comando e não tem suporte para o 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 de ciclo de vida recebido com a [notificação \(p. 261\)](#), 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 \(p. 261\)](#), 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 do término do período de tempo limite, use [ocomplete-lifecycle-action](#) comando para que o grupo de Auto Scaling possa continuar iniciando ou encerrando a instância. Você pode especificar o token da ação de 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 concluir manualmente as ações do ciclo de vida. Você pode adicionar [ocomplete-lifecycle-action](#) comando para o 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 [Retrieve instance metadata](#) (Recuperar metadados de instância) no Guia do usuário do Amazon EC2 para instâncias 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 \(p. 273\)](#).

Tutorial: configurar dados do usuário para recuperar o estado de destino do ciclo de vida por meio de metadados de instância

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 o AmazonEventBridge. No entanto, usando um script de dados do usuário para mover o código que configura instâncias e concluir 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 \(p. 262\)](#) 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.

Índice

- [Etapa 1: criar uma função do IAM com permissões para concluir ações de ciclo de vida \(p. 268\)](#)
- [Etapa 2: criar um modelo de execução e incluir a função do IAM e um script de dados de usuário \(p. 269\)](#)
- [Etapa 3: criar um grupo do Auto Scaling \(p. 270\)](#)
- [Etapa 4: Adicionar um gancho do ciclo de vida \(p. 271\)](#)
- [Etapa 5: testar e verificar a funcionalidade \(p. 271\)](#)
- [Etapa 6: Limpar \(p. 272\)](#)
- [Recursos relacionados \(p. 273\)](#)

Etapa 1: criar uma função do IAM com permissões para concluir ações de ciclo de vida

Quando você usa a AWS CLI ou um AWS SDK para enviar um retorno de chamada a fim de concluir ações de ciclo de vida, é necessário usar uma função do IAM com permissões para concluir ações de ciclo de vida.

Para criar a política

1. Abra a página [Policies](#) (Políticas) do console do IAM e escolha Create policy (Criar política).
2. Escolha a guia JSON.
3. Na caixa Policy Document (Documento de política), cole o seguinte documento de política. Substitua **samples text** (texto de amostra) pelo número da sua conta e o nome do grupo do Auto Scaling que deseja criar (**TestAutoScalingEvent-group**).

```
{  
    "Version": "2012-10-17",  
    "Statement": [  
        {  
            "Effect": "Allow",  
            "Action": [  
                "autoscaling:CompleteLifecycleAction"  
            ],  
            "Resource":  
                "arn:aws:autoscaling:*:123456789012:autoScalingGroup:*:autoScalingGroupName/  
                TestAutoScalingEvent-group"  
        }  
    ]  
}
```

4. Escolha Next (próximo).
5. Em Nome da política, insira **TestAutoScalingEvent-policy**. Escolha Create policy (Criar política).

Quando você terminar de criar a política, poderá criar uma função que a utilize.

Para criar a função do

1. No painel de navegação à esquerda, escolha Roles (Funções).
2. Selecione Create role (Criar função).
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. Abaixo Adicionar permissões, escolha a política que você criou (**TestAutoScalingEvent-política**). Em seguida, escolha Next (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 a função do IAM e um script de dados de usuário

Crie um modelo de execução para usar com seu grupo do Auto Scaling. Inclua a função 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 Create launch template (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 Para Imagens de aplicativo 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 Perfil da instância IAM, escolha o nome do perfil da instância do IAM da sua função do IAM (**TestAutoScalingEvent-papel**). Um perfil de instância é um contêiner para uma função do IAM que permite ao Amazon EC2 passar a função do IAM para uma instância quando ela é iniciada.

Se tiver criado uma função 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 de `group_name` pelo nome do grupo do Auto Scaling que deseja criar e `region` pela Região da AWS que deseja que seu grupo do Auto Scaling use.

```
#!/bin/bash

function get_target_state {
    echo $(curl -s http://169.254.169.254/latest/meta-data/autoscaling/target-lifecycle-state)
}

function get_instance_id {
    echo $(curl -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 Create launch template (Criar modelo de execução).
 11. Na página de confirmação, escolha Create Auto Scaling group (Criar grupo do Auto Scaling).

Note

Para ver outros exemplos que você pode usar como referência para desenvolver seu script de dados do usuário, consulte a[GitHubrepositório](#)para o 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 de ciclo de vida

1. Abra o [Página de grupos do Auto Scaling](#)do console 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 hooks (Ganchos do ciclo de vida), escolha Create lifecycle hook (Criar gancho do ciclo de vida).
4. Para definir um gancho de ciclo de vida para expansão (inicialização de instâncias), faça o seguinte:
 - a. Em Lifecycle hook 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 gancho 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 Create (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 Scal

1. Abra o [Página de grupos do Auto Scaling](#)do console 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. Escolha 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 à 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 o [Página de grupos do Auto Scaling](#)do console 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 hooks (Ganchos do ciclo de vida), escolha o gancho do ciclo de vida (TestAutoScalingEvent-hook).
4. Escolha Actions, Delete.
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 a função do IAM e a política de permissões que criou.

Para excluir o grupo do Auto Scaling

1. Abra o [Página de grupos do Auto Scaling](#)do console 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 a função do IAM

1. Abra a página [Roles](#) (Funções) no console do IAM.
2. Selecione o papel da função (TestAutoScalingEvent-role).

3. Escolha Delete (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. Escolha Actions, Delete.
4. Quando for solicitada confirmação, digite o nome da política e escolha Excluir.

Recursos relacionados

Os tópicos relacionados a seguir podem ser úteis ao desenvolver 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 \(p. 262\)](#). 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 \(console\) \(p. 264\)](#). Esse procedimento mostra como adicionar ganchos de ciclo de vida tanto para expansão (inicialização de instâncias) quanto para expansão (instâncias encerrando ou retornando a uma piscina aquecida).
- [Categorias de metadados da instância](#)Guia do usuário do Amazon EC2 para instâncias Linux. Este tópico lista todas as categorias de metadados de instância que você pode usar para invocar ações em instâncias do EC2.

Para um tutorial que mostra como usar a AmazonEventBridge para criar regras que invocam funções do Lambda com base em eventos que acontecem com as instâncias em seu grupo de Auto Scaling, consulte[Tutorial: configurar dados do usuário para recuperar o estado de destino do ciclo de vida por meio de metadados de instância \(p. 267\)](#).

Tutorial: Configurar um gancho do ciclo de vida que invoca uma função do Lambda

Neste exercício, você cria uma AmazonEventBridge regra que inclui um padrão de filtro que, quando combinado, invoca um AWS Lambda que funciona como o alvo da regra. Nós fornecemos o padrão de filtro e o 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 CloudWatchLogs e associa o fluxo de registros 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.

Para obter mais informações sobre o uso de EventBridge, consulte [Usar EventBridge para lidar com eventos do Auto Scaling \(p. 400\)](#).

Índice

- [Pré-requisitos \(p. 274\)](#)
- [Etapa 1: criar uma função do IAM com permissões para concluir ações de ciclo de vida \(p. 274\)](#)
- [Etapa 2: criar uma função do Lambda \(p. 275\)](#)

- [Etapa 3: criar um EventBridge regra \(p. 276\)](#)
- [Etapa 4: Adicionar um gancho do ciclo de vida \(p. 277\)](#)
- [Etapa 5: Testar e verificar o evento \(p. 277\)](#)
- [Etapa 6: Limpar \(p. 278\)](#)
- [Recursos relacionados \(p. 279\)](#)

Pré-requisitos

Antes de iniciar este tutorial, crie um grupo do Auto Scaling, se você ainda não tiver um. Para criar um grupo de Auto Scaling, abra a [Página de grupos do Auto Scaling](#) do console Amazon EC2 e escolha Criar grupo de Auto Scaling.

Etapa 1: criar uma função do IAM com permissões para concluir ações de ciclo de vida

Antes de criar uma função do Lambda, você deve 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. Escolha a guia JSON.
3. Na caixa Policy Document (Documento da política), cole o documento de política a seguir na caixa, substituindo o texto em *íntlico* pelo o número de conta e o nome do grupo do Auto Scaling.

```
{  
    "Version": "2012-10-17",  
    "Statement": [  
        {  
            "Effect": "Allow",  
            "Action": [  
                "autoscaling:CompleteLifecycleAction"  
            ],  
            "Resource":  
                "arn:aws:autoscaling:*:123456789012:autoScalingGroup:*:autoScalingGroupName/my-asg"  
        }  
    ]  
}
```

4. Escolha Next (próximo).
5. Em Nome da política, insira **LogAutoScalingEvent-policy**. Escolha Create policy (Criar política).

Quando você terminar de criar a política, poderá criar uma função que a utilize.

Para criar a função do

1. No painel de navegação à esquerda, escolha Roles (Funções).
2. Selecione Create role (Criar função).
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. Abaixo Adicionar permissões, escolha a política que você criou (LogAutoScalingEvent-política) e a política chamada AWSLambdaBasicExecutionRole. Em seguida, escolha Next (Próximo).

Note

OAWSLambdaBasicExecutionRolea política tem as permissões nas quais a função precisa para gravar registrosCloudWatchRegistros.

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 exemplo, escrita em Node.js, é invocada porEventBridgequando um evento correspondente é emitido pelo Amazon EC2 Auto Scaling.

Como 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. ParaTempo de execução, escolhaNode.js 14.x.
5. Escolha Change default execution role (Alterar a função de execução padrão) e, em Execution role (Função de execução), escolha Use an existing role (Usar uma função existente).
6. ParaFunção existente, escolhaLogAutoScalingEvent-papel.
7. Deixe os outros valores padrão.
8. Escolha Criar função. Você é retornado ao código e configuração da função.
9. Com sua função LogAutoScalingEvent ainda aberta no console, em Function code (Código da função), no editor, cole o seguinte código de exemplo no arquivo chamado index.js.

```
var aws = require("aws-sdk");
exports.handler = (event, context, callback) => {
    console.log('LogAutoScalingEvent');
    console.log('Received event:', JSON.stringify(event, null, 2));
    var autoscaling = new aws.AutoScaling({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;
    autoscaling.completeLifecycleAction(params, function(err, data) {
        if (err) {
            console.log(err, err.stack); // an error occurred
            response = {
                statusCode: 500,
                body: JSON.stringify('ERROR'),
            };
        } else {
            console.log(data);           // successful response
            response = {
                statusCode: 200,
                body: JSON.stringify('SUCCESS'),
            };
        }
    })
}
```

```
});  
return response;  
};
```

Esse código simplesmente registra o evento para que, no final deste tutorial, você possa ver um evento aparecer na CloudWatchLogs e o fluxo de log associado a essa função do Lambda.

10. Escolha Implantar.

Etapa 3: criar umEventBridgeregra

Crie umEventBridgeregra para executar sua função Lambda.

Como criar uma regra usando o console

1. Abra o [console do EventBridge](#).
2. No painel de navegação, escolha Rules (Regras).
3. Escolha Create rule (Criar regra).
4. Em Define rule detail (Definir detalhe da regra), faça o seguinte:
 - a. Em Name (Nome), insira **LogAutoScalingEvent-rule**.
 - b. Em Event Bus (Barramento de eventos), escolha default (padrão). Quando um AWS service (Serviço da AWS) em sua conta gerar um evento, ele sempre irá para o barramento de eventos padrão da sua conta.
 - c. Em Rule type (Tipo de regra), escolha Rule with an event pattern (Regra com um padrão de evento).
 - d. Escolha Next (próximo).
5. Em Build event pattern (Criar padrão de evento), faça o seguinte:
 - a. Para Fonte do evento, escolha AWS events ou EventBridge events parceiros.
 - b. Em Event pattern (Padrão de evento), faça o seguinte:
 - i. Para Event source (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, escolha Specific instance event(s) (Eventos específicos de instância) e selecione EC2 Instance-launch Lifecycle Action (Ação de 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 do Auto Scaling específico, escolha Specific group name(s) (Nomes de grupos específicos) e selecione um ou mais grupos do Auto Scaling.
 - vi. Escolha Next (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. Para Função, escolha LogAutoScalingEvent.
 - d. Escolha Next (Próximo) duas vezes.
7. Na página Review and create (Revisar e criar), escolha Create (Criar).

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 de ciclo de vida

1. Abra o [Página de grupos do Auto Scaling](#)do console 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 hooks (Ganchos do ciclo de vida), escolha Create lifecycle hook (Criar gancho do ciclo de vida).
4. Para definir um gancho de ciclo de vida para expansão (inicialização de instâncias), faça o seguinte:
 - a. Em Lifecycle hook 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 gancho 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 passamosEventBridgecontém todas as informações necessárias para invocar a função Lambda.
5. Escolha Create (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 Scalinh

1. Abra o [Página de grupos do Auto Scaling](#)do console 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. Escolha 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 de grupos de log](#) do console do CloudWatch.
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 à 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 o [Página de grupos do Auto Scaling](#) do console 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 hooks (Ganchos do ciclo de vida), escolha o gancho do ciclo de vida (LogAutoScalingEvent-hook).
4. Escolha Actions, Delete.
5. Para confirmar, escolha Delete (Excluir) novamente.

Para excluir a AmazonEventBridge regra

1. Abra o [Página de regras](#) na AmazôniaEventBridgeconsole.
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 Delete (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. Escolha Actions, Delete.
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 de grupos de log](#) do console do CloudWatch.
2. Selecione o grupo de logs da função (/aws/lambda/LogAutoScalingEvent).
3. Escolha 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 Delete (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. Escolha Actions, Delete.
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 baseadas em eventos que acontecem com as instâncias em seu grupo de Auto Scaling.

- [Usar EventBridge para lidar com eventos do Auto Scaling \(p. 400\)](#). Esta seção mostra exemplos de eventos para outros casos de uso, incluindo eventos para expansão.
- [Adicionar ganchos do ciclo de vida \(console\) \(p. 264\)](#). Esse procedimento mostra como adicionar ganchos de ciclo de vida tanto para expansão (inicialização de instâncias) quanto para expansão (instâncias encerrando ou retornando a uma piscina aquecida).

Para ver um tutorial que mostra como usar o Instance Metadata Service (IMDS) para invocar uma ação de dentro da própria instância, consulte [Tutorial: configurar dados do usuário para recuperar o estado de destino do ciclo de vida por meio de metadados de instância \(p. 267\)](#).

Grupos de alta atividade do Amazon EC2 Auto Scaling

Um grupo de alta atividade 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 grupos de alta atividade, 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 grupos de alta atividade do EC2 Auto Scaling](#).

Important

Criar um grupo de alta atividade 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 grupo de alta atividade.

Índice

- [Conceitos principais \(p. 280\)](#)
- [Pré-requisitos \(p. 282\)](#)
- [Criar um grupo de alta atividade \(p. 283\)](#)
- [Atualizar um grupo de alta atividade \(p. 283\)](#)
- [Excluir um grupo de alta atividade \(p. 284\)](#)
- [Limitações \(p. 284\)](#)
- [Usar ganchos do ciclo de vida com um grupo de alta atividade \(p. 284\)](#)
- [Visualizar o status e o motivo de falhas da verificação de integridade \(p. 287\)](#)
- [Exemplos para criar e gerenciar grupos de alta atividade com a AWS CLI \(p. 289\)](#)

Conceitos principais

Antes de começar a usar, familiarize-se com os seguintes conceitos principais:

Grupo de alta atividade

Um grupo de alta atividade 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 grupo de alta atividade 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 grupo de alta atividade, 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 grupo de alta atividade

Por padrão, o tamanho do grupo de alta atividade 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 grupo de alta atividade será 4 quando você configurar o grupo de alta atividade pela primeira vez e o pool estiver inicializando.

Para especificar a capacidade máxima do grupo de alta atividade separadamente, defina um valor para a capacidade máxima preparada que seja maior que a capacidade atual do grupo. Quando você definir um valor para a capacidade máxima preparada, o tamanho do grupo de alta atividade será calculado como a diferença entre a capacidade máxima preparada e a atual capacidade desejada do grupo. Por exemplo, se a capacidade desejada do grupo do Auto Scaling for 6, a capacidade máxima

for 10 e a capacidade máxima preparada for 8, o tamanho do grupo de alta atividade será 2 quando você configurar o grupo de alta atividade pela primeira vez e o grupo estiver inicializando.

Talvez seja necessário usar apenas a opção de capacidade máxima preparada ao trabalhar com grupos grandes do Auto Scaling para gerenciar os benefícios de custo de ter um grupo de alta atividade. 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 grupo de alta atividade de 100 instâncias seja uma estratégia melhor para ajudar você a atingir seus objetivos do que manter 500 instâncias reservadas para uso futuro no grupo de alta atividade.

Tamanho mínimo do grupo de alta atividade

Considere usar a configuração de tamanho mínimo 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.

Estado da instância do grupo de alta atividade

Você pode manter as instâncias no grupo de alta atividade 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 grupo de alta atividade, 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

[Ganchos do ciclo de vida \(p. 252\)](#) permitem que você coloque 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 grupo de alta atividade, os ganchos do ciclo de vida também podem adiar a interrupção ou a hibernação das instâncias e atrasar sua colocação em serviço durante um evento de aumento de escala na horizontal até que elas tenham concluído a inicialização. Se você adicionar um grupo de alta atividade 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 grupo de alta atividade para substituir as instâncias que foram terminadas.

Se desejar devolver instâncias ao grupo de alta atividade, 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 grupo de alta atividade não seja excessivamente provisionado, o Amazon EC2 Auto Scaling pode terminar instâncias no grupo de alta atividade para reduzir seu tamanho quando for maior do que o necessário, com base em suas configurações. Ao

terminar instâncias no grupo de alta atividade, ele usa a [política de término padrão \(p. 295\)](#) 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 grupo de alta atividade, 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

Decida como você usará ganchos do ciclo de vida para preparar as instâncias para uso. Existem duas formas de executar ações personalizadas em suas instâncias.

- 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 de cloud-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 mais informações, consulte [Retrieve instance metadata](#) (Recuperar metadados de instância) no Guia do usuário do Amazon EC2 para instâncias Linux.

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 o AWS Lambda para fazer algo à medida que as instâncias entram e saem do grupo de alta atividade, você pode criar um gancho do ciclo de vida para o grupo do Auto Scaling e configurar o serviço de destino para executar ações personalizadas com base em notificações de ciclo de vida. Para obter mais informações, consulte [Destinos de notificação compatíveis \(p. 286\)](#).

Para obter mais informações, consulte os exemplos de ganchos de ciclo de vida em nosso repositório. [GitHub](#)

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 [Substituir instâncias do Auto Scaling com base em uma atualização de instância \(p. 108\)](#).

Criar um grupo de alta atividade

Crie um grupo de alta atividade usando o console de acordo com as instruções a seguir.

Antes de começar, confirme se você criou um gancho do ciclo de vida para o grupo do Auto Scaling.

Para criar um grupo de alta atividade (console)

1. Abra o console do Amazon EC2 em <https://console.aws.amazon.com/ec2/> e escolha grupos do Auto Scaling 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. Selecione a guia Instance management (Gerenciamento de instâncias).
4. Em Warm pool (Grupo de alta atividade), escolha Create warm pool (Criar grupo de alta atividade).
5. Para configurar um grupo de alta atividade, 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 grupo de alta atividade. O padrão é Stopped.
- b. Em Minimum warm pool size (Tamanho mínimo do grupo de alta atividade), insira o número mínimo de instâncias que serão mantidas no grupo de alta atividade.
- c. Em Max prepared capacity (Capacidade preparada máxima), é possível especificar a capacidade preparada máxima definindo um certo número de instâncias ou manter a opção padrão para deixar a capacidade preparada máxima indefinida.

Se você mantiver a opção padrão Equal to the Auto Scaling group's maximum capacity (Igual à capacidade máxima do grupo do Auto Scaling), o grupo de alta atividade será dimensionado para corresponder à diferença entre a capacidade máxima do grupo do Auto Scaling e a capacidade desejada. Para facilitar o gerenciamento do tamanho do grupo de alta atividade ajustando a capacidade máxima do grupo, recomendamos usar a opção padrão.

Se escolher a opção Define a set number of instances (Especificar um número definido de instâncias), insira um valor que represente o número máximo de instâncias com permissão para estar no grupo de alta atividade e no grupo do Auto Scaling ao mesmo tempo.

6. Escolha Create (Criar).

Atualizar um grupo de alta atividade

Para alterar o modelo de execução ou a configuração de execução de um grupo de alta atividade, associe um novo modelo de execução ou configuração de execução 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 a inicialização de instâncias substitutas de warm pool que usam o novo modelo de execução ou a configuração de inicialização, você pode iniciar uma atualização de instância para fazer uma atualização contínua do seu grupo. Uma atualização de instância substitui primeiro as instâncias InService. Em seguida, ela substitui as instâncias no grupo de alta atividade. Para obter mais informações, consulte [Substituir instâncias do Auto Scaling com base em uma atualização de instância \(p. 108\)](#).

Excluir um grupo de alta atividade

Quando você não precisar mais do grupo de alta atividade, 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 grupos do Auto Scaling 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. Selecione a guia Instance management (Gerenciamento de instâncias).
4. Em Warm pool (Grupo de alta atividade), escolha Actions (Ações), Delete (Excluir).
5. Quando a confirmação for solicitada, escolha Delete (Excluir).

Limitações

- Não é possível adicionar um grupo de alta atividade a grupos do Auto Scaling que têm uma política de instâncias mistas ou que iniciam instâncias spot.
- 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 grupo de alta atividade 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 você tentar usar grupos de alta atividade 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 grupo de alta atividade com um cluster do Amazon ECS, as instâncias poderão se registrar no cluster antes que a inicialização seja concluída. Para resolver esse problema, você deve 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).
- O suporte à hibernação para piscinas aquecidas está disponível em todos os estabelecimentos comerciais Regiões da AWS em que o Amazon EC2 Auto Scaling está disponível, excluindo a região do Oriente Médio (EAU), as regiões da China e o. AWS GovCloud (US) Regions

Usar ganchos do ciclo de vida com um grupo de alta atividade

As instâncias em um grupo de alta atividade 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 de 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 [Adicionar ganchos do ciclo de vida \(p. 263\)](#). Para obter mais informações sobre a conclusão de uma ação do ciclo de vida, consulte [Concluir uma ação do ciclo de vida \(p. 266\)](#).

Para instâncias que entram no grupo de alta atividade, 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.

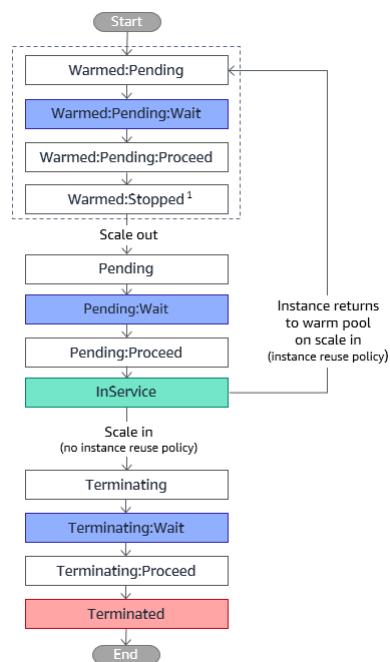
Para instâncias que saem do grupo de alta atividade, talvez você precise de um gancho do ciclo de vida por um dos seguintes motivos:

- 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 grupo de alta atividade

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 grupo de alta atividade:

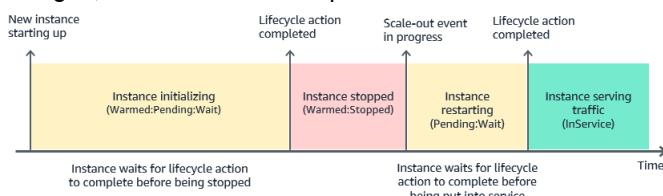


¹ Esse estado varia de acordo com a configuração do estado do grupo de alta atividade. Se o estado do grupo estiver definido como Running, então esse estado será Warmed:Running, em vez disso. Se o estado do grupo estiver definido como Hibernated, então esse estado será Warmed:Hibernated, em vez disso.

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` de ciclo de vida, uma instância recém-iniciada faz uma 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` de 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 grupo de alta atividade na operação de redução da escala horizontalmente em vez de encerrá-las, uma instância que estiver retornando ao grupo de alta atividade fará uma pausa para realizar uma ação personalizada no estado `Warmed:Pending:Wait` para a ação de ciclo de vida `EC2_INSTANCE_TERMINATING`.
- Se a demanda em sua aplicação esgotar o grupo de alta atividade, 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 piscinas aquecidas \(p. 407\)](#).

Destinos de notificação compatíveis

O Amazon EC2 Auto Scaling oferece suporte para definir qualquer um dos seguintes destinos como destinos de notificação para notificações de ciclo de vida:

- Regras do EventBridge
- Tópicos do Amazon SNS
- Filas do Amazon SQS

Important

Lembre-se de que, se você tiver um script de dados do usuário (cloud-init) no modelo de inicialização ou na configuração de inicialização 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: para executar código quando o Amazon EC2 Auto Scaling coloca uma instância em estado de espera, você pode criar uma EventBridge regra e especificar uma função Lambda como seu destino. Para invocar diferentes funções do Lambda com base em notificações de 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 [Criar EventBridge regras para eventos de grupo de alta atividade \(p. 413\)](#).

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 de ciclo de vida de forma diferente com base em um atributo de mensagem. Para obter mais informações, consulte [Receba notificações usando o Amazon SNS \(p. 260\)](#).

Filas do Amazon SQS: para configurar um ponto de entrega para notificações de ciclo de vida em que um consumidor relevante possa buscá-las e processá-las, você pode 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 de 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 \(p. 261\)](#).

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 grupo de alta atividade 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 grupo de alta atividade 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 grupos de alta atividade, 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 [Substituição de instância não íntegra \(p. 323\)](#).

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 [Tarefas personalizadas de detecção de integridade \(p. 322\)](#).

No console do Amazon EC2 Auto Scaling, é possível visualizar o status (íntegra ou não íntegra) das instâncias do grupo de alta atividade. Também é possível visualizar seus status de integridade usando a 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 grupos do Auto Scaling 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 grupo de alta atividade), 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 grupos do Auto Scaling 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 grupo de alta atividade (AWS CLI)

Visualize a piscina aquecida de um grupo de Auto Scaling usando o [describe-warm-pool](#) comando a seguir.

```
aws autoscaling describe-warm-pool --auto-scaling-group-name my-asg
```

Saída de exemplo.

```
{  
    "WarmPoolConfiguration": {  
        "MinSize": 0,  
        "PoolState": "Stopped"  
    },  
    "Instances": [  
        {  
            "InstanceId": "i-0b5e5e7521cf8aa46c",  
            "InstanceType": "t2.micro",  
            "AvailabilityZone": "us-west-2a",  
            "LifecycleState": "Warmed:Stopped",  
            "HealthStatus": "Healthy",  
            "LaunchTemplate": {  
                "LaunchTemplateId": "lt-08c4cd42f320d5dc",  
                "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-08c4cd42f320d5dc",  
                "LaunchTemplateName": "my-template-for-auto-scaling",  
                "Version": "1"  
            }  
        }  
    ]  
}
```

```
        "InstanceType": "t2.micro",
        "AvailabilityZone": "us-west-2a",
        "LifecycleState": "Warmed:Stopped",
        "HealthStatus": "Healthy",
        "LaunchTemplate": {
            "LaunchTemplateId": "lt-08c4cd42f320d5dc",
            "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 grupos de alta atividade com a AWS CLI

Você pode criar e gerenciar grupos de alta atividade usando o AWS Management Console, a AWS Command Line Interface (AWS CLI) ou SDKs.

Os exemplos a seguir mostram como criar e gerenciar grupos de alta atividade usando a AWS CLI.

Índice

- [Exemplo 1: manter instâncias no estado Stopped \(p. 290\)](#)
- [Exemplo 2: manter instâncias no estado Running \(p. 290\)](#)

- [Exemplo 3: manter instâncias no estado Hibernated \(p. 290\)](#)
- [Exemplo 4: retornar instâncias para o grupo de alta atividade ao reduzir a escala na horizontal \(p. 290\)](#)
- [Exemplo 5: especificar o número mínimo de instâncias no grupo de alta atividade \(p. 290\)](#)
- [Exemplo 6: definir separadamente a capacidade máxima do grupo de alta atividade \(p. 291\)](#)
- [Exemplo 7: definir um tamanho de grupo de alta atividade absoluto \(p. 291\)](#)
- [Exemplo 8: exclusão um grupo de alta atividade \(p. 291\)](#)

Exemplo 1: manter instâncias no estado Stopped

O [put-warm-pool](#) exemplo a seguir cria uma piscina aquecida 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-pool](#) exemplo 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-pool](#) exemplo 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 grupo de alta atividade ao reduzir a escala na horizontal

O [put-warm-pool](#) exemplo a seguir cria uma piscina aquecida 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 grupo de alta atividade 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 grupo de alta atividade

O [put-warm-pool](#) exemplo a seguir cria um pool aquecido que mantém no mínimo 4 instâncias, de forma 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 separadamente a capacidade máxima do grupo de alta atividade

Geralmente, você entende o quanto precisa aumentar a capacidade acima da capacidade desejada. Normalmente, não há necessidade de definir um tamanho máximo adicional porque o Amazon EC2 Auto Scaling cria um grupo de alta atividade que é redimensionado dinamicamente com base nas capacidades desejada e máxima do seu grupo. No entanto, você pode usar a opção `--max-group-prepared-capacity` para definir separadamente a capacidade máxima do grupo de alta atividade quando desejado.

O [put-warm-pool](#) exemplo a seguir cria uma piscina aquecida que define sua capacidade máxima separadamente. Suponha que o grupo do Auto Scaling tenha uma capacidade desejada de 800. O tamanho do grupo de alta atividade será 100 quando você executar esse comando e o pool estiver inicializando.

```
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 grupo de alta atividade, 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 grupo de alta atividade absoluto

Se você definir os mesmos valores para as opções `--max-group-prepared-capacity` e `--min-size`, o grupo de alta atividade terá um tamanho absoluto. O [put-warm-pool](#) exemplo a seguir cria uma piscina aquecida que mantém um tamanho constante de piscina aquecida 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 grupo de alta atividade

Use o [delete-warm-pool](#) comando a seguir para excluir uma piscina aquecida.

```
aws autoscaling delete-warm-pool --auto-scaling-group-name my-asg
```

Se houver instâncias na piscina aquecida ou se as atividades de escalonamento estiverem em andamento, use o [delete-warm-pool](#) comando com a `--force-delete` opção. Essa opção também terminará as instâncias do Amazon EC2 e quaisquer ações de ciclo de vida pendentes.

```
aws autoscaling delete-warm-pool --auto-scaling-group-name my-asg --force-delete
```

Controlar quais instâncias do Auto Scaling serão terminadas durante uma redução de escala na horizontal

O Amazon EC2 Auto Scaling usa políticas de término para determinar quais instâncias ele termina primeiro durante eventos de redução de escala na horizontal. As políticas de término definem os critérios de término usados pelo Amazon EC2 Auto Scaling ao escolher quais instâncias serão terminadas.

Seu grupo do Auto Scaling usa uma política padrão de encerramento, mas você também pode escolher ou criar suas próprias políticas de encerramento com seus próprios critérios de encerramento. Isso permite que você garanta que suas instâncias sejam terminadas com base em suas necessidades específicas da aplicação.

O Amazon EC2 Auto Scaling também oferece proteção contra redução de escala na horizontal de instâncias. Quando você habilita esse recurso, ele impede que instâncias sejam terminadas durante eventos de redução de escala na horizontal. É possível habilitar a proteção de redução de escala na horizontal de instâncias ao criar um grupo do Auto Scaling e alterar a configuração em instâncias em execução. Se você habilitar a proteção de redução de escala na horizontal de instâncias em um grupo do Auto Scaling existente, todas as novas instâncias executadas após isso terão a proteção de redução de escala na horizontal de instâncias habilitada.

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 accidental, 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.

Índice

- [Cenários para o uso da política de término \(p. 292\)](#)
- [Trabalhar com políticas de término do Amazon EC2 Auto Scaling \(p. 295\)](#)
- [Criar uma política de término personalizada com o Lambda \(p. 298\)](#)
- [Usar proteção de redução na escala na horizontal de instâncias \(p. 302\)](#)
- [Projete seus aplicativos no Amazon EC2 Auto Scaling para lidar com o encerramento de instâncias com elegância \(p. 305\)](#)

Cenários para o uso da política de término

As seções a seguir descrevem os cenários em que o Amazon EC2 Auto Scaling usa políticas de término.

Índice

- [Eventos de redução de escala na horizontal \(p. 293\)](#)
- [Atualizações de instância \(p. 293\)](#)
- [Rebalanceamento de zona de disponibilidade \(p. 294\)](#)

Eventos de redução de escala na horizontal

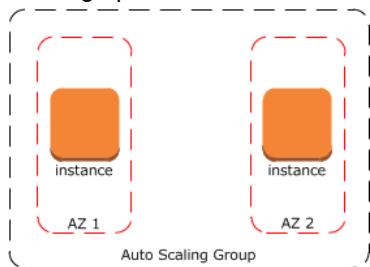
Um evento de redução de escala na horizontal ocorre quando há um novo valor para a capacidade desejada de um grupo do Auto Scaling que é menor do que a capacidade atual do grupo.

Eventos de redução de escala na horizontal ocorrem nos seguintes casos:

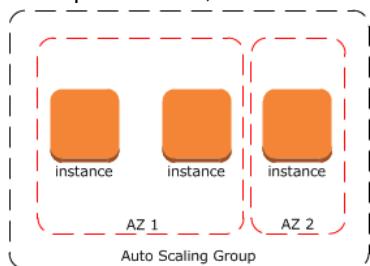
- Ao usar políticas de escalabilidade dinâmica e o tamanho do grupo diminui como resultado de alterações no valor de uma métrica
- Ao usar a escalabilidade programada e o tamanho do grupo diminui como resultado de uma ação programada
- Quando você reduz o tamanho do grupo manualmente

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 iniciadas 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.

Atualizações 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

Este tópico fornece informações detalhadas sobre a política padrão de encerramento e as opções disponíveis para escolher políticas de encerramento diferentes para grupos específicos do Auto Scaling. Usando políticas de término, você pode controlar quais instâncias prefere terminar primeiro quando ocorre um evento de redução de escala na horizontal. Por exemplo, você pode escolher uma política de término diferente para que o Amazon EC2 Auto Scaling priorize o término das instâncias mais antigas 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 que as outras zonas de disponibilidade que são usadas pelo grupo, o Amazon EC2 Auto Scaling aplicará sua política de término especificada nas instâncias da zona de disponibilidade desbalanceada. Se as zonas de disponibilidade usadas pelo grupo estiverem balanceadas, o Amazon EC2 Auto Scaling aplicará a política de término a todas as zonas de disponibilidade do grupo.

Índice

- [Política de término padrão \(p. 295\)](#)
- [Política de término padrão e grupos de instâncias mistas \(p. 296\)](#)
- [Usar políticas de término diferentes \(p. 296\)](#)
 - [Usar políticas de término diferentes \(console\) \(p. 297\)](#)
 - [Usar políticas de término diferentes \(AWS CLI\) \(p. 298\)](#)

Política de término padrão

A política de término padrão aplica vários critérios de terminação antes de selecionar uma instância a ser terminada. 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. Na zona de disponibilidade selecionada, aplica-se o seguinte comportamento de política de término padrão:

1. Determinar se alguma das instâncias elegíveis para encerramento usa a configuração de execução ou o modelo de execução mais antigo:
 - a. [Para grupos do Auto Scaling que usem um modelo de execução]

Determinar se alguma das instâncias usa o modelo de execução mais antigo, a menos que haja instâncias que usem configuração de execução. O Amazon EC2 Auto Scaling termina instâncias que usam uma configuração de execução antes de terminar instâncias que usam um modelo de execução.

- b. [Para grupos do Auto Scaling que usem uma configuração de execução]

Determine se qualquer uma das instâncias usa a configuração de execução mais antiga.

2. Depois de aplicar os critérios anteriores, se houver várias instâncias desprotegidas a serem terminadas, determinar quais instâncias estão mais perto da próxima hora de faturamento. Se houver várias instâncias desprotegidas mais perto da próxima hora de faturamento, encerre uma delas aleatoriamente.

(Observe que encerrar a instância mais perto da próxima hora de faturamento ajuda você a maximizar o uso das suas instâncias que têm uma cobrança por hora.) Se o seu grupo do Auto Scaling usar Amazon Linux, Windows ou o Ubuntu, seu uso do EC2 será cobrado em incrementos de um segundo. Para obter mais informações, consulte [Definição de preço do Amazon EC2](#).

Política de término padrão e grupos de instâncias mistas

Quando um grupo do Auto Scaling com uma [política de instâncias mistas \(p. 67\)](#) reduz a escala na horizontal, o Amazon EC2 Auto Scaling ainda usa políticas de término para priorizar quais instâncias serão terminadas, mas primeiro ele identifica quais dos dois tipos (spot ou sob demanda) devem ser terminados. Em seguida, aplica as políticas de término em cada zona de disponibilidade individualmente. Ele também identifica quais instâncias (dentro da opção de compra identificada) em que zonas de disponibilidade serão terminadas, o que resultará no equilíbrio das zonas de disponibilidade. A mesma lógica se aplica aos grupos do Auto Scaling que usem uma configuração de instâncias mistas com pesos definidos para os tipos de instância.

A política de término padrão muda ligeiramente devido a diferenças em como as [políticas de instâncias mistas \(p. 67\)](#) são implementadas. Aplica-se o seguinte novo comportamento da política de término padrão:

1. Determinar quais instâncias são qualificáveis para término a fim de alinhar as instâncias restantes à [estratégia de alocação \(p. 68\)](#) da instância sob demanda ou spot que está sendo terminada.

Por exemplo, após a execução das suas instâncias, você pode alterar a ordem de prioridade dos tipos de instância preferenciais. Quando ocorre um evento de redução de escala na horizontal, o Amazon EC2 Auto Scaling tenta afastar gradualmente as instâncias sob demanda dos tipos de instância com prioridade mais baixa.

2. Determinar se alguma das instâncias usa o modelo de execução mais antigo, a menos que haja instâncias que usem configuração de execução. O Amazon EC2 Auto Scaling termina instâncias que usem uma configuração de execução antes de terminar instâncias que usem um modelo de execução.
3. Depois de aplicar os critérios anteriores, se houver várias instâncias desprotegidas a serem terminadas, determinar quais instâncias estão mais perto da próxima hora de faturamento. Se houver várias instâncias desprotegidas mais perto da próxima hora de faturamento, encerre uma delas aleatoriamente.

Usar políticas de término diferentes

Para especificar os critérios de término a serem aplicados antes que o Amazon EC2 Auto Scaling escolha uma instância para término, você pode escolher uma das seguintes políticas de término predefinidas:

- **Default.** Terminar instâncias de acordo com a política de término padrão. Essa política é útil quando você deseja que sua estratégia de alocação spot seja avaliada antes de qualquer outra política, de modo que toda vez que suas instâncias spot forem terminadas ou substituídas, você continue fazendo uso de instâncias spot nos grupos ideais. Também é útil, por exemplo, quando você deseja sair das configurações de execução e começar a usar modelos de execução.
- **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. (Apenas instâncias que usam Amazon Linux, Windows ou Ubuntu são cobradas em incrementos de um segundo.)
- `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.

Usar políticas de término diferentes (console)

Depois que o grupo do Auto Scaling tiver sido criado, você poderá atualizar as políticas de término do grupo. A política de término padrão é usada automaticamente. Você tem a opção de substituir a política padrão por uma política de término diferente (como `OldestLaunchTemplate`) ou várias políticas de término listadas na ordem em que elas devem ser aplicadas.

Para escolher diferentes políticas de término

1. Abra o console do Amazon EC2 em <https://console.aws.amazon.com/ec2/> e escolha grupos do Auto Scaling 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 \(p. 298\)](#).

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. Escolha Update (Atualizar).

Usar políticas de término diferentes (AWS CLI)

A política de término padrão é usada automaticamente, a menos que uma política diferente seja especificada.

Para usar uma política de término diferente

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 término personalizada, você deve primeiro criar sua política de término usando o 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 \(p. 298\)](#).

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 \(p. 295\)](#).

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.

Índice

- [Dados de entrada \(p. 299\)](#)
- [Dados de resposta \(p. 300\)](#)
- [Considerações \(p. 300\)](#)
- [Criar a função do Lambda \(p. 301\)](#)

- [Limitações \(p. 302\)](#)

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 \(p. 67\)](#), 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 serem 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 rescisão personalizada não afeta sua capacidade de também usar proteção escalável para evitar que determinadas instâncias sejam encerradas. Para obter mais informações, consulte [Usar proteção de redução na escala na horizontal de instâncias \(p. 302\)](#).

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 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 a [função vinculada 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. Em Ação, escolha lambda: InvokeFunction.
 - d. Em Statement ID (ID da instrução), insira um ID de instrução exclusivo, como `AllowInvokeByAutoScaling`.
 - e. Escolha Save (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 [Usar políticas de término diferentes \(console\) \(p. 297\)](#).

Note

Para ver exemplos que você pode usar como referência para desenvolver sua função Lambda, consulte o [GitHubrepositório](#) 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, você deve 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.

Usar proteção de redução na escala na horizontal de instâncias

Quando ocorre uma atividade de escalabilidade, o Amazon EC2 Auto Scaling faz o seguinte:

- Aumenta a capacidade do grupo de Auto Scaling (chamado de redução de escala)
- Diminui a capacidade do grupo Auto Scaling (conhecido como escalonamento em)

A configuração de proteção de escalonamento da instância controla se o grupo de Auto Scaling pode encerrar uma instância específica durante o escalonamento. Um caso de uso comum para esse requisito é o dimensionamento de cargas de trabalho baseadas em contêineres.

Você pode proteger as instâncias assim que elas forem iniciadas ativando a configuração de proteção de escalonamento de instâncias em seu grupo de 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, desative a configuração de proteção de escalonamento em instâncias individuais dentro do grupo Auto Scaling. Ao fazer isso, você pode continuar protegendo determinadas instâncias contra encerramentos indesejados.

A proteção contra redução de escala na horizontal de instâncias não protege as instâncias do Auto Scaling contra o seguinte:

- Encerramento manual por meio do `terminate-instance-in-auto-scaling-group` comando ou da `TerminateInstanceInAutoScalingGroup` ação. Para obter mais informações, consulte [TerminateInstanceInAutoScalingGroup](#) Amazon EC2 Auto Scaling API Reference.
- Encerramento manual por meio do console do Amazon EC2, do `terminate-instances` comando do Amazon EC2 ou da ação do Amazon EC2. `TerminateInstances` 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 encerre instâncias.) Para obter mais informações, consulte [Ativação de proteção contra término](#) no Manual do usuário do Amazon EC2 para instâncias do Linux.
- Substituição da verificação se a instância não passar nas verificações de integridade. Para obter mais informações, consulte [Verificações de integridade para instâncias do Auto Scaling \(p. 319\)](#).
- Interrupções de instâncias spot Uma instância spot é encerrada quando a capacidade não está mais disponível ou o preço spot excede seu preço máximo.

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 grupo de alta atividade 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.

Tarefas

- [Ativar a proteção contra redução de escala na horizontal de instâncias para um grupo \(p. 304\)](#)
- [Modificar a configuração de proteção contra redução de escala na horizontal de instâncias para um grupo \(p. 304\)](#)
- [Modificar a configuração de proteção contra redução de escala na horizontal de instâncias para uma instância \(p. 305\)](#)

Note

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 AWS Management Console, o Activity history (Histórico de atividades) para o grupo do Auto Scaling inclui a seguinte mensagem se todas as instâncias em um grupo do Auto Scaling estiverem protegidas contra a redução da escala na horizontal quando ocorrer um evento de

redução de escala na horizontal: Could not scale to desired capacity because all remaining instances are protected from scale-in.

Ativar a proteção contra redução de escala na horizontal de instâncias para um grupo

É possível habilitar a proteção contra redução de escala na horizontal de instâncias ao criar um grupo do Auto Scaling. Por padrão, a proteção de redução de instâncias permanece desabilitada.

Como habilitar a proteção de redução de instâncias (console)

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).

Como habilitar a proteção de redução de instâncias (AWS CLI)

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 ...
```

Modificar a configuração de proteção contra redução de escala na horizontal de instâncias para um grupo

É possível habilitar ou desabilitar a configuração de proteção contra redução de escala na horizontal de instâncias para um grupo do Auto Scaling. Quando a configuração de proteção de redução da instância estiver habilitada, todas as novas instâncias executadas após habilitá-la terão a proteção de redução da instância habilitada. As instâncias executadas anteriormente não são protegidas contra redução de escala na horizontal, a menos que você habilite a configuração de proteção contra redução de escala na horizontal de instâncias para cada instância individualmente.

Como alterar a configuração de proteção de redução de instâncias para um grupo (console)

1. Abra o console do Amazon EC2 em <https://console.aws.amazon.com/ec2/> e escolha grupos do Auto Scaling 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. Em **Proteção contra redução de instâncias**, selecione **Habilitar a proteção contra redução de instâncias**.
5. Escolha **Update (Atualizar)**.

Como alterar a configuração de proteção de redução de instâncias para um grupo (AWS CLI)

Use o [update-auto-scaling-group](#) comando a seguir para ativar a proteção de escalonamento de 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
```

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
```

Modificar a configuração de proteção contra redução de escala na horizontal de instâncias para 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.

Como alterar a configuração de proteção de redução de instâncias para uma instância (console)

1. Abra o console do Amazon EC2 em <https://console.aws.amazon.com/ec2/> e escolha grupos do Auto Scaling 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).

Como alterar a configuração de proteção de redução de instâncias para uma instância (AWS CLI)

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
```

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
```

Projete seus aplicativos no Amazon EC2 Auto Scaling para lidar com o encerramento de instâncias com elegância

Este tópico aborda as diferentes abordagens que você pode adotar se tiver aplicativos em execução em instâncias que, idealmente, não deveriam ser encerradas inesperadamente quando o Amazon EC2 Auto Scaling responde a um evento de escalabilidade.

Por exemplo, suponha que você tenha uma fila do Amazon SQS que coleta mensagens recebidas para trabalhos de longa duração. Quando uma nova mensagem chega, uma instância no grupo Auto Scaling recupera a mensagem e começa a processá-la. Cada mensagem leva 3 horas para ser processada. Conforme o número de mensagens aumenta, novas instâncias são adicionadas automaticamente ao grupo Auto Scaling. À medida que o número de mensagens diminui, as instâncias existentes são

encerradas automaticamente. Nesse caso, o Amazon EC2 Auto Scaling deve decidir qual instância encerrar. Por padrão, é possível que o Amazon EC2 Auto Scaling encerre uma instância com 2,9 horas de processamento de um trabalho de 3 horas, em vez de uma instância que está ociosa no momento. Para evitar problemas com encerramentos inesperados ao usar o Amazon EC2 Auto Scaling, você deve projetar seu aplicativo para responder a esse cenário.

Você pode usar os seguintes recursos para evitar que seu grupo de Auto Scaling encerre instâncias que ainda não estão prontas para serem encerradas ou encerre instâncias muito rapidamente para que elas concluam os trabalhos atribuídos. Todos esses três recursos podem ser usados em combinação ou separadamente.

Índice

- [Proteção de escalabilidade de instâncias \(p. 306\)](#)
- [Política de rescisão personalizada \(p. 306\)](#)
- [Ganchos do ciclo de vida de rescisão \(p. 307\)](#)

Important

Ao projetar seus aplicativos no Amazon EC2 Auto Scaling para lidar com o encerramento de instâncias com elegância, lembre-se desses pontos.

- Se uma instância não estiver íntegra, o Amazon EC2 Auto Scaling a substituirá independentemente do recurso usado (a menos que você suspenda o processo). **ReplaceUnhealthy** Você pode usar um gancho de ciclo de vida para permitir que o aplicativo seja encerrado normalmente ou copie quaisquer dados que você precise recuperar antes que a instância seja encerrada.
- Não é garantido que um gancho do ciclo de vida de encerramento seja executado ou concluído antes que uma instância seja encerrada. Se algo falhar, o Amazon EC2 Auto Scaling ainda encerrará a instância.

Proteção de escalabilidade de instâncias

Você pode usar a proteção de escalabilidade de instâncias em muitas situações em que encerrar instâncias é uma ação crítica que deve ser negada por padrão e permitida apenas explicitamente para instâncias específicas. 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. Serviços como o Amazon ECS incorporaram integrações com proteção de escalabilidade de instâncias em seus produtos.

Você pode ativar a proteção de escalabilidade no grupo Auto Scaling para aplicar a proteção de escalonamento às instâncias quando elas forem criadas e habilitá-la para instâncias existentes. Quando uma instância não tem mais trabalho a fazer, ela pode desativar a proteção. A instância pode continuar pesquisando novos trabalhos e reativar a proteção quando houver novos trabalhos atribuídos.

Os aplicativos podem definir a proteção a partir de um plano de controle centralizado que gerencia se uma instância é encerrável ou não, ou das próprias instâncias. No entanto, uma grande frota pode ter problemas de limitação se um grande número de instâncias estiver continuamente ativando sua proteção de escalabilidade.

Para obter mais informações, consulte [Usar proteção de redução na escala na horizontal de instâncias \(p. 302\)](#).

Política de rescisão personalizada

Assim como a proteção de escalonamento de instâncias, uma política de encerramento personalizada ajuda a impedir que seu grupo de Auto Scaling encerre instâncias específicas.

Por padrão, seu grupo de Auto Scaling usa uma política de encerramento padrão para determinar quais instâncias ele encerra primeiro. Se quiser mais controle sobre quais instâncias são encerradas primeiro, você pode implementar sua própria política de encerramento personalizada usando uma função Lambda. O Amazon EC2 Auto Scaling chama a função sempre que precisa decidir qual instância encerrar. Isso encerrará apenas uma instância retornada pela função. Se a função errar, atingir o tempo limite ou produzir uma lista vazia, o Amazon EC2 Auto Scaling não encerrará as instâncias.

Uma política de encerramento personalizada é útil quando se sabe quando uma instância é suficientemente redundante ou subutilizada para que possa ser encerrada. Para suportar isso, você precisa implementar seu aplicativo com um plano de controle que monitora a carga de trabalho em todo o grupo. Dessa forma, se uma instância ainda estiver processando trabalhos, a função Lambda sabe que não deve incluí-los.

Para obter mais informações, consulte [Criar uma política de término personalizada com o Lambda \(p. 298\)](#).

Ganchos do ciclo de vida de rescisão

Um gancho do ciclo de vida de encerramento prolonga a vida útil de uma instância que já está selecionada para encerramento. Ele fornece tempo extra para concluir todas as mensagens ou solicitações atualmente atribuídas à instância ou para salvar o progresso e transferir o trabalho para outra instância.

Para muitas cargas de trabalho, um gancho de ciclo de vida pode ser suficiente para encerrar normalmente um aplicativo em uma instância selecionada para encerramento. Essa é a abordagem ideal e não pode ser usada para evitar a rescisão em caso de falha.

Para usar um gancho de ciclo de vida, você precisa saber quando uma instância foi selecionada para ser encerrada. Você tem duas maneiras de saber isso:

Opção	Descrição	Melhor usado para	Link para a documentação
Dentro da instância	O Instance Metadata Service (IMDS) é um endpoint seguro que você pode consultar o status de uma instância diretamente da instância. Se os metadados voltarem com <code>Terminated</code> , sua instância está programada para ser encerrada.	Aplicativos nos quais você deve executar uma ação na instância antes que ela seja encerrada.	Recupere o estado do ciclo de vida alvo
Fora da instância	Quando uma instância está sendo encerrada, uma notificação de evento é gerada. Você pode criar regras usando AmazonEventBridge, Amazon SQS ou Amazon SNS para capturar esses eventos e invocar uma resposta, como com uma função Lambda.	Aplicativos que precisam agir fora da instância.	Configurar um alvo de notificação

Para usar um gancho de ciclo de vida, você também precisa saber quando sua instância está pronta para ser totalmente encerrada. O Amazon EC2 Auto Scaling não solicitará que o Amazon EC2 encerre a instância até que receba uma [CompleteLifecycleAction](#) chamada ou o tempo limite termine, o que acontecer 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 realmente em andamento, você pode estender o tempo limite com chamadas de API. [RecordLifecycleActionHeartbeat](#)

Para obter mais informações, consulte [Ganchos do ciclo de vida do Amazon EC2 Auto Scaling \(p. 252\)](#).

Remover temporariamente 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平衡ador 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. Desvincular instâncias pode ser útil se você quiser gerenciar as instâncias como instâncias do EC2 autônomas e possivelmente terminá-las. Para obter mais informações, consulte [Desvincular instâncias do EC2 do seu grupo do Auto Scaling \(p. 174\)](#).

Índice

- [Como o estado de espera funciona \(p. 308\)](#)
- [Considerações \(p. 309\)](#)
- [Status de integridade de uma instância em um estado de espera \(p. 309\)](#)
- [Remoção temporária de uma instância \(console\) \(p. 309\)](#)
- [Remover uma instância temporariamente \(AWS CLI\) \(p. 310\)](#)

Como o estado de espera funciona

O estado de espera funciona da seguinte forma para ajudá-lo 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平衡ador 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 das instâncias em um grupo de Auto Scaling, consulte [Ciclo de vida das instâncias do Amazon EC2 Auto Scaling \(p. 8\)](#)

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 com o mesmo valor.
 - Se você optar por não diminuir a capacidade desejada do grupo de 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 de Auto Scaling, isso impedirá a inicialização 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 de Auto Scaling.
- Para fazer o incremento (e a diminuição), 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, constatar que seu grupo de Auto Scaling não está equilibrado 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 um processo para um grupo do Auto Scaling \(p. 312\)](#).
- 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 do Auto Scaling \(p. 319\)](#).

Remoção temporária de uma instância (console)

Esta seção explica como você pode remover temporariamente uma instância do seu grupo de Auto Scaling usando o console.

Para remover uma instância temporariamente

1. Abra o console do Amazon EC2 em <https://console.aws.amazon.com/ec2/> e escolha grupos do Auto Scaling 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 Definir para espera, mantenha a caixa de seleção Substituir instância marcada 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 Standby estado e, em seguida, escolha Definir como 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.

Remover uma instância temporariamente (AWS CLI)

Esta seção explica como você pode remover temporariamente uma instância do seu grupo de Auto Scaling usando o AWS CLI

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
```

Esta é uma resposta de exemplo.

```
{  
    "AutoScalingInstances": [  
        {  
            "ProtectedFromScaleIn": false,  
            "AvailabilityZone": "us-west-2a",  
            "LaunchTemplate": {  
                "LaunchTemplateName": "my-launch-template",  
                "Version": "1",  
                "LaunchTemplateId": "lt-050555ad16a3f9c7f"  
            },  
            "InstanceId": "i-05b4f7d5be44822a6",  
            "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
```

Esta é uma resposta de exemplo.

```
{  
    "Activities": [  
        {  
            "Description": "Moving EC2 instance to Standby: i-05b4f7d5be44822a6",  
            "AutoScalingGroupName": "my-asg",  
            "ActivityId": "3b1839fe-24b0-40d9-80ae-bcd883c2be32",  
            "Details": "{\"Availability Zone\":\"us-west-2a\"}",  
            "StartTime": "2014-12-15T21:31:26.150Z",  
            "Progress": 50,  
            "Cause": "At 2014-12-15T21:31:26Z instance i-05b4f7d5be44822a6 was moved to  
standby  
in response to a user request, shrinking the capacity from 4 to 3.",  
            "StatusCode": "InProgress"  
        }  
    ]  
}
```

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
```

Esta é uma resposta de exemplo. 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",  
            "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
```

Esta é uma resposta de exemplo.

```
{
```

```
"Activities": [
    {
        "Description": "Moving EC2 instance out of Standby: i-05b4f7d5be44822a6",
        "AutoScalingGroupName": "my-asg",
        "ActivityId": "db12b166-cdcc-4c54-8aac-08c5935f8389",
        "Details": "{\"Availability Zone\":\"us-west-2a\"}",
        "StartTime": "2014-12-15T21:46:14.678Z",
        "Progress": 30,
        "Cause": "At 2014-12-15T21:46:14Z instance i-05b4f7d5be44822a6 was moved
out of standby in
            response to a user request, increasing the capacity from 3 to 4.",
        "StatusCode": "PreInService"
    }
]
```

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
```

Esta é uma resposta de exemplo. 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",
            "AutoScalingGroupName": "my-asg",
            "HealthStatus": "HEALTHY",
            "LifecycleState": "InService"
        },
        ...
    ]
}
```

Suspender e retomar um processo para um grupo do Auto Scaling

Este tópico explica como suspender e, depois, retomar um ou mais dos processos de escalabilidade para o grupo do Auto Scaling. Talvez você queira fazer isso para, por exemplo, investigar um problema de configuração que esteja causando falha no processo ou para impedir que o Amazon EC2 Auto Scaling marque instâncias não íntegras e as substitua enquanto você estiver fazendo alterações em seu grupo do Auto Scaling.

Índice

- [Tipos de processos \(p. 313\)](#)
- [Considerações \(p. 313\)](#)
- [Suspender e retomar processos \(console\) \(p. 316\)](#)
- [Suspender e retomar processos \(AWS CLI\) \(p. 316\)](#)

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 grupo de alta atividade.
- **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 [Usar o Elastic Load Balancing para distribuir tráfego entre as instâncias no grupo do Auto Scaling \(p. 369\)](#).
- **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 \(p. 178\)](#).
- **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 \(p. 7\)](#).
- **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 do Auto Scaling \(p. 319\)](#).
- **InstanceRefresh**: termina e substitui instâncias usando o recurso de atualização de instância. Para obter mais informações, consulte [Substituir instâncias do Auto Scaling com base em uma atualização de instância \(p. 108\)](#).
- **ReplaceUnhealthy**: termina 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 do Auto Scaling \(p. 319\)](#).
- **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 \(p. 247\)](#).

Considerações

Antes de suspender processos, considere o seguinte:

- Você pode suspender e retomar processos individuais ou todos os processos.
- A suspensão de um processo afeta todas as instâncias do grupo do Auto Scaling. Por exemplo, você pode suspender os processos HealthCheck e ReplaceUnhealthy para reiniciar 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 temporariamente instâncias do grupo do Auto Scaling \(p. 308\)](#).

- A suspensão AlarmNotification permite que você interrompa temporariamente as políticas de rastreamento de metas, etapas e escalonamento 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 \(p. 216\)](#).
- 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.
- A suspensão do Terminate processo não impede o encerramento bem-sucedido das instâncias usando a opção forçar exclusão com o [delete-auto-scaling-group](#) comando.
- O processo RemoveFromLoadBalancerLowPriority deverá ser ignorado quando estiver presente em chamadas para descrever grupos do Auto Scaling usando a AWS CLI ou SDKs. Esse processo está defasado e é retido somente para fins de compatibilidade retroativa.

Entender como a suspensão de processos afeta outros processos

As descrições a seguir explicam o que acontece quando tipos de processos individuais são suspensos.

Cenário 1: 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 processoLaunch, o Amazon EC2 Auto Scaling substituirá imediatamente todas as instâncias que ele encerrou enquanto Launch estava suspenso.
- InstanceRefresh não substitui as instâncias.

Cenário 2: 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.

Cenário 3: AddToLoadBalancer está suspenso

- O Amazon EC2 Auto Scaling executa as instâncias, mas não as adiciona ao grupo de destino do平衡ador 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. Você deve registrar essas instâncias manualmente.

Cenário 4: AlarmNotification está suspenso

- O Amazon EC2 Auto Scaling não invoca políticas de escalabilidade quando um limite de CloudWatch 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.

Cenário 5: 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 a expansão, ele executa a instância 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.

Cenário 6: 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.

Cenário 7: 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.

Cenário 8: 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.

Cenário 9: 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 Spot do Amazon EC2 ainda poderá encerrar instâncias caso a capacidade spot não esteja mais disponível. Se Launch estiver suspenso, o Amazon EC2 Auto Scaling não poderá executar 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.

Suspender e retomar processos (console)

Siga o procedimento abaixo para suspender um processo.

Para suspender um processo

1. Abra o console do Amazon EC2 em <https://console.aws.amazon.com/ec2/> e escolha grupos do Auto Scaling 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. Escolha Update (Atualizar).

Quando você estiver pronto, siga o procedimento abaixo para retomar o processo suspenso.

Para retomar um processo

1. Na guia Detalhes, escolha Configurações avançadas, Editar.
2. Em Suspended processes (Processos suspensos), escolha o processo suspenso.
3. Escolha Update (Atualizar).

Suspender e retomar processos (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
```

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
```

Monitore seus grupos de Auto Scaling

O monitoramento é uma parte importante da manutenção da confiabilidade, disponibilidade e desempenho do Amazon EC2 Auto Scaling e do seu Nuvem AWS soluções. AWS fornece as seguintes ferramentas de monitoramento para assistir ao 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 lançar uma nova instância para substituí-la. Para obter mais informações, consulte [Verificações de integridade para instâncias do Auto Scaling \(p. 319\)](#).

AWS Health Dashboard

O AWS Health Dashboard exibe informações e também fornece notificações que são invocadas por mudanças na integridade dos recursos AWS. 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 [Notificações do AWS Health Dashboard para o Amazon EC2 Auto Scaling \(p. 327\)](#).

CloudTrail

Com AWS CloudTrail, você pode rastrear as chamadas feitas para a API do Amazon EC2 Auto Scaling por ou em nome de sua 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 [Registrar chamadas da API do Amazon EC2 Auto Scaling com o AWS CloudTrail \(p. 339\)](#).

Coleta de registros para suas 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 Metrics](#) e [Exibir dados de registro enviados para o CloudWatch Logs](#) na Amazônia CloudWatch Guia do usuário.

Para obter informações sobre outros serviços AWS que podem ajudar você a registrar e coletar dados sobre suas cargas de trabalho, consulte o [Guia de registro e monitoramento para proprietários de aplicativos](#) na AWS Orientação prescritiva.

Amazon CloudWatch

Amazônia CloudWatch ajuda você a analisar registros e, em tempo real, monitorar as métricas do seu AWS recursos e aplicativos hospedados. É possível 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 como usar esse serviço para monitorar as métricas de seus grupos e

instâncias do Auto Scaling, consulte [MonitorCloudWatchmétricas para seus grupos e instâncias do Auto Scaling \(p. 328\)](#).

CloudWatch também rastreia AWS Métricas de uso da API para o Amazon EC2 Auto Scaling. Você pode usar essas métricas para configurar alarmes que alertam quando o volume de chamadas da API viola um limite definido por você. Para obter mais informações, consulte [AWS Métricas de uso na Amazon CloudWatch Guia do usuário](#).

AWS Compute Optimizer

O Compute Optimizer fornece recomendações de instâncias do Amazon EC2 que podem ajudar você a decidir se deve migrar para um novo tipo de instância. Ele analisa se o tipo de instância de um grupo de Auto Scaling é ideal e gera recomendações para reduzir o custo e melhorar o desempenho de suas cargas de trabalho. Para obter mais informações, consulte [Use o AWS Compute Optimizer para obter recomendações para o tipo de instância para um grupo do Auto Scaling \(p. 367\)](#).

Amazon EventBridge

Amazônia EventBridge é um serviço de barramento 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, Software-as-a-Service (SaaS) e AWS serviços e encaminha esses dados para alvos como o Lambda. Isso permite monitorar eventos que acontecem nos serviços e criar arquiteturas orientadas por eventos. Para obter mais informações, consulte [Usar EventBridge para lidar com eventos do Auto Scaling \(p. 400\)](#).

AWS Security Hub

Usar [AWS Security Hub](#) para monitorar seu uso do Amazon EC2 Auto Scaling no que se refere às melhores práticas de segurança. O Security Hub usa detetive-controles de segurança para avaliar as configurações de recursos e padrões de segurança para ajudá-lo a cumprir várias estruturas de conformidade. Para obter mais informações sobre o uso do Security Hub para avaliar os recursos do Amazon EC2 Auto Scaling, consulte [Controles do Amazon EC2 Auto Scaling na AWS Security Hub Guia do usuário](#).

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 [Receber notificações do Amazon SNS quando o grupo do Auto Scaling escala \(p. 341\)](#).

Verificações de integridade para instâncias do Auto Scaling

O status de integridade de uma instância do Auto Scaling indica se ela está íntegra ou não íntegra. Todas as instâncias do grupo do Auto Scaling são iniciadas com um `Healthy` status. Supõe-se que as instâncias estejam íntegras, a menos que o Amazon EC2 Auto Scaling receba uma notificação informando do contrário. Essa notificação pode ser proveniente de fontes como Amazon EC2, Elastic Load Balancing, VPC Lattice ou verificações de integridade personalizadas. Quando o Amazon EC2 Auto Scaling detecta uma instância não íntegra, ele a substituirá.

Índice

- [Tipo de verificação de integridade \(p. 320\)](#)
- [Verificações de integridade do Amazon EC2 \(p. 320\)](#)
- [Verificações de integridade do Elastic Load Balancing \(p. 321\)](#)
- [Verificações de integridade do VPC Lattice \(p. 322\)](#)

- [Tarefas personalizadas de detecção de integridade \(p. 322\)](#)
- [Substituição de instância não íntegra \(p. 323\)](#)
- [Como o Amazon EC2 Auto Scaling minimiza o tempo de inatividade \(p. 324\)](#)
- [Considerações sobre a verificação de integridade \(p. 324\)](#)
- [Informações adicionais \(p. 325\)](#)
- [Definir um período de carência da verificação de integridade para um grupo do Auto Scaling \(p. 325\)](#)

Tipo de verificação de integridade

O Amazon EC2 Auto Scaling pode determinar o status de integridade de uma instância usando uma ou mais das seguintes verificações de integridade:

Health check type (Tipo de verificação de integridade)	O que ele verifica
Verificações de status do Amazon EC2 e eventos programados	<ul style="list-style-type: none">• Verifica se a instância está em execução• Verifica se há problemas subjacentes de hardware ou software capazes de prejudicar a instância <p>Esse é o tipo padrão de verificação de integridade para um grupo do Auto Scaling.</p>
Verificações de integridade do Elastic Load Balancing	<ul style="list-style-type: none">• Verifica se o balanceador de carga relata a instância como íntegra, confirmando se a instância está disponível para processar solicitações. <p>Para executar esse tipo de verificação de integridade, você deve habilitá-lo para seu grupo do Auto Scaling.</p>
Verificações de integridade do VPC Lattice	<ul style="list-style-type: none">• Verifica se o VPC Lattice relata a instância como íntegra, confirmando se a instância está disponível para processar solicitações. <p>Para executar esse tipo de verificação de integridade, você deve habilitá-lo para seu grupo do Auto Scaling.</p>
Verificações de integridade personalizadas	<ul style="list-style-type: none">• 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 a instância for iniciada, ela será anexada ao grupo do Auto Scaling e entra no InService estado. 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 \(p. 8\)](#).

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 ocorre uma falha na verificação de status, o Amazon EC2 Auto Scaling aguarda alguns minutos para a AWS corrigir o problema. Ele não marca imediatamente uma instância `Unhealthy` quando o status das verificações de status se torna `impaired`.

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. Nesse caso, ele 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 integridade 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 que possa habilitar as verificações de integridade do Elastic Load Balancing para seu grupo do Auto Scaling, é necessário fazer o seguinte:

- Configure um balanceador de carga do Elastic Load Balancing e configure uma verificação de integridade que ele use para determinar se suas instâncias estão íntegras.
- Anexe o balanceador de carga ao seu grupo do Auto Scaling.

O seguinte ocorrerá após você realizar as ações acima:

- 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. No entanto, você pode ativar essas verificações de integridade para seu grupo do Auto Scaling. Após fazer isso, quando o Elastic Load Balancing relatar uma instância registrada comoUnhealthy, o Amazon EC2 Auto Scaling marcará a instância Unhealthy em sua próxima verificação de integridade de periódica e a substituirá.

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.

Para saber como habilitar as verificações de integridade do Elastic Load Balancing para seu grupo do Auto Scaling, consulte. [Adicionar verificações de integridade do Elastic Load Balancing a um grupo do Auto Scaling \(p. 376\)](#)

Note

Quando você habilita as verificações de integridade do Elastic Load Balancing para um grupo, o Amazon EC2 Auto Scaling pode encerrar e substituir instâncias relatadas como não íntegras, mas somente depois que o balanceador de carga estiver no estado. InService Para obter mais informações, consulte [Verificar o status do anexo do balanceador de carga \(p. 375\)](#).

Verificações de integridade do VPC Lattice

Por padrão, o Amazon EC2 Auto Scaling ignora os resultados das verificações de integridade do VPC Lattice. Como opção, você pode ativar essas verificações de integridade para seu grupo do Auto Scaling. Após fazer isso, quando o VPC Lattice relatar uma instância registrada comoUnhealthy, o Amazon EC2 Auto Scaling marcará a instância Unhealthy em sua próxima verificação de integridade de periódica e a substituirá. O processo de registrar instâncias e depois verificar sua integridade é o mesmo de como as verificações de integridade do Elastic Load Balancing funcionam.

Para saber como habilitar as verificações de integridade do VPC Lattice para seu grupo do Auto Scaling, consulte. [Anexar um grupo de destino do VPC Lattice \(p. 396\)](#)

Note

Quando você habilita as verificações de integridade do VPC Lattice para um grupo, o Amazon EC2 Auto Scaling pode substituir instâncias relatadas 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 \(p. 399\)](#).

Tarefas personalizadas de detecção de integridade

Também recomendamos executar tarefas personalizadas de detecção de integridade nas instâncias do grupo do Auto Scaling e definir o status de integridade de uma instância como Unhealthy se a tarefa falhar. Isso amplia suas verificações de integridade usando uma combinação de verificações de integridade personalizadas, verificações de status do Amazon EC2 e verificações de integridade do Elastic Load Balancing, se estiverem habilitadas.

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-health](#) comando a seguir para definir o status de integridade da instância especificada comoUnhealthy.

```
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-groups](#) comando 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"  
                },  
                ...  
            ]  
        }  
    ]  
}
```

Substituição de instância não íntegra

Quando o Amazon EC2 Auto Scaling determina que uma instância `InService` não está íntegra, ele encerra a instância enquanto executa uma nova instância substituta. 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 Amazon EC2 Auto Scaling cria uma nova ação de escalabilidade para terminar a instância não íntegra e a encerra. Enquanto a instância estiver sendo encerrada, outra atividade de escalonamento iniciará uma nova instância.

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 grupos do Auto Scaling 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 an ELB system health check failure.

Como o Amazon EC2 Auto Scaling minimiza o tempo de inatividade

As substituições de verificação de integridade exigem que as instâncias sejam encerradas primeiro, o que pode impedir que novas solicitações sejam aceitas até que novas instâncias sejam iniciadas.

Se o Amazon EC2 Auto Scaling determinar que alguma instância não está mais em execução ou foi marcada `Unhealthy` com o [set-instance-health](#) comando, ele imediatamente as 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. Além disso, 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平衡ador de carga estiver no `InService` estado, 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.

Considerações sobre a verificação de integridade

Esta seção contém considerações sobre 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 ganchos 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 \(p. 252\)](#).
- 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 um processo para um grupo do Auto Scaling \(p. 312\)](#).

- Para definir manualmente o status de integridade de uma instância não íntegra de volta para Healthy, você pode tentar usar o [set-instance-health](#) comando. Se você receber um erro, provavelmente a instância já está encerrando. Em geral, definir o status de integridade de uma instância de volta para Healthy com o [set-instance-health](#) comando só é útil nos casos em que o ReplaceUnhealthy processo ou o Terminate processo está suspenso.
- O Amazon EC2 Auto Scaling não executa verificações de integridade em instâncias que estão no estado Standby. Para obter mais informações, consulte [Remover temporariamente instâncias do grupo do Auto Scaling \(p. 308\)](#).
- 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 de IP elásticos](#) no Manual do usuário do Amazon EC2 para instâncias do Linux.
- 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 [Anexar um volume do Amazon EBS a uma instância](#) no Guia do usuário do Amazon EC2 para instâncias do Linux

Informações adicionais

Para mais informações sobre soluções de problemas para as verificações de integridade, consulte [Solucionar problemas com as verificações de integridade do Amazon EC2 Auto Scaling \(p. 473\)](#). Se as verificações de integridade falharem, consulte este tópico para ver as etapas de solução de problemas. Os tópicos a seguir ajudarão a descobrir o que está errado no grupo do Auto Scaling e apresentarão sugestões sobre como corrigir o problema.

O Amazon EC2 Auto Scaling também monitora a integridade das instâncias executadas em um grupo de alta atividade usando o Amazon EC2, o Amazon EBS ou verificações de integridade personalizadas. Para obter mais informações, consulte [Visualizar o status e o motivo de falhas da verificação de integridade \(p. 287\)](#).

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 instância InService não está íntegra, ela termina a instância e inicia uma nova instância substituta. 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平衡ador de carga. O período de carência impede que o Amazon EC2 Auto Scaling marque as instâncias recém-iniciadas Unhealthy e as encerre desnecessariamente se elas não forem aprovadas nessas verificações de integridade imediatamente após entrarem no estado InService.

Por padrão, o período de carência da verificação de integridade é de 300 segundos quando você cria um grupo do Auto Scaling. O valor padrão é de 0 segundos quando você cria um grupo do Auto Scaling usando a AWS CLI ou um SDK.

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 de 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 \(p. 252\)](#).

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 `running` estado Amazon EC2, ele a marcará imediatamente e a substituirá. `Unhealthy` Por exemplo, se você interromper uma instância em um grupo do Auto Scaling, ela será marcada `Unhealthy` e substituirá.

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 (console)

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 (AWS CLI)

Adicione a `--health-check-grace-period` opção ao [`create-auto-scaling-group`](#) comando. 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 (console)

1. Abra o console do Amazon EC2 em <https://console.aws.amazon.com/ec2/> e escolha grupos do Auto Scaling 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 Auto Scaling groups (Grupos do Auto Scaling).

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. Escolha Update (Atualizar).

AWS CLI

Para modificar o período de carência da verificação de integridade para um grupo existente (AWS CLI)

Adicione a `--health-check-grace-period` opção ao [update-auto-scaling-group](#) comando. 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 \(p. 200\)](#).

Notificações do AWS Health Dashboard para o Amazon EC2 Auto Scaling

O AWS Health Dashboard é compatível com as 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.

O exemplo de AWS Health Dashboard é parte do serviço AWS Health. Ele não requer configuração e pode ser visualizado por qualquer usuário autenticado em sua conta. Para obter mais informações, consulte [Conceitos básicos do AWS Health Dashboard](#).

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

MonitorCloudWatchmétricas para seus grupos e instâncias do Auto Scaling

Métricas são o conceito fundamental na AmazonCloudWatch. Uma métrica representa um conjunto ordenado de pontos de dados que são publicados no 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, talvez você prefira acessar os dados de seus grupos e instâncias de Auto Scaling da AmazonCloudWatchem vez do console Amazon EC2 Auto Scaling.

Para obter mais informações, consulte o [AmazôniaCloudWatchGuia do usuário](#).

Índice

- [Visualizar grafos de monitoramento no console do Amazon EC2 Auto Scaling \(p. 328\)](#)
- [AmazôniaCloudWatchmétricas para o Amazon EC2 Auto Scaling \(p. 332\)](#)
- [Configurar monitoramento para instâncias do Auto Scaling \(p. 337\)](#)

Visualizar grafos de monitoramento no console do Amazon EC2 Auto Scaling

Na seção Amazon EC2 Auto Scaling do console do Amazon EC2, você pode monitorar minute-by-minute progresso de grupos individuais de Auto Scaling usando CloudWatchmétricas.

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 \(console\) \(p. 336\)](#). 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. Quando o monitoramento detalhado está ativado, os gráficos de monitoramento mostram dados publicados com granularidade de um minuto para métricas de exemplo. Para obter mais informações, consulte [Configurar monitoramento para instâncias do Auto Scaling \(p. 337\)](#).

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 grupos do Auto Scaling 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 Exhibir registros na ferramenta de menu para visualizar os fluxos de registro associados (se houver) no CloudWatchconsole.
- Para visualizar um gráfico no CloudWatch, escolha Exhibir em métricas na ferramenta de menu. Isso leva você para o CloudWatchpágina para esse 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).

No Escalabilidade automática seção, 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 [Amazon CloudWatch métricas para o Amazon EC2 Auto Scaling \(p. 332\)](#).

Nome de exibição	Nome da métrica do CloudWatch
Tamanho mínimo do grupo	GroupMinSize
Tamanho máximo do grupo	GroupMaxSize
Capacidade desejada	GroupDesiredCapacity
Em instâncias de serviço	GroupInServiceInstances
Instâncias pendentes	GroupPendingInstances
Instâncias em espera	GroupStandbyInstances
Encerrando instâncias	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 [Liste as disponíveis CloudWatch métricas para suas instâncias na Guia do usuário do Amazon EC2 para instâncias Linux](#).

Nome de exibição	Nome da métrica do CloudWatch
Utilização da CPU	CPUUtilization
Leituras de disco	DiskReadBytes
Operações de leitura de disco	DiskReadOps
Gravações em disco	DiskWriteBytes
Operações de gravação em disco	DiskWriteOps
Entrada de rede	NetworkIn
Saída de rede	NetworkOut
Status Check Failed (Any) (Falha na verificação de status (qualquer))	StatusCheckFailed
Status Check Failed (Instance) (Falha na verificação de status (instância))	StatusCheckFailed_Instance
Status Check Failed (System) (Falha na verificação de status (sistema))	StatusCheckFailed_System

Além disso, algumas métricas estão disponíveis para casos de uso específicos no Escalabilidade automática métricas gráficas.

As métricas a seguir são úteis se seu grupo usa pesos que definem quantas unidades cada instância contribui para a capacidade desejada do grupo. Você pode encontrar definições para essas métricas em [Amazônia CloudWatch métricas para o Amazon EC2 Auto Scaling \(p. 332\)](#).

Nome de exibição	Nome da métrica do CloudWatch
Unidades de capacidade em serviço	GroupInServiceCapacity
Unidades de capacidade pendentes	GroupPendingCapacity
Unidades de capacidade de espera	GroupStandbyCapacity
Unidades de capacidade de terminação	GroupTerminatingCapacity
Unidades de capacidade total	GroupTotalCapacity

As métricas a seguir são úteis se o seu grupo usa [a piscina quente \(p. 279\)](#) recurso. Você pode encontrar definições para essas métricas em [AmazôniaCloudWatchmétricas para o Amazon EC2 Auto Scaling \(p. 332\)](#).

Nome de exibição	Nome da métrica do CloudWatch
Tamanho mínimo da piscina quente	WarmPoolMinSize
Capacidade desejada da piscina quente	WarmPoolDesiredCapacity
Unidades de capacidade pendente para piscinas aquecidas	WarmPoolPendingCapacity
Unidades de capacidade de terminação de piscinas quentes	WarmPoolTerminatingCapacity
Unidades de capacidade aquecida para piscinas quentes	WarmPoolWarmedCapacity
Lançadas unidades de capacidade total para piscinas aquecidas	WarmPoolTotalCapacity
Capacidade desejada para grupos e piscinas aquecidas	GroupAndWarmPoolDesiredCapacity
Lançadas unidades de capacidade total para grupos e piscinas aquecidas	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.
- Os painéis do CloudWatch são páginas iniciais personalizáveis no console do CloudWatch. Você pode usar essas páginas para monitorar seus recursos em uma única visualização, incluindo recursos distribuídos por regiões diferentes. Você pode usar os painéis do CloudWatch para criar visualizações

personalizadas das métricas e dos alarmes para seus recursos da AWS. Para obter mais informações, consulte o [AmazôniaCloudWatchGuia do usuário](#).

AmazôniaCloudWatchmétricas para o 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 dependerão de você ter as métricas de grupo ativadas e de quais métricas de grupo você ativou. As métricas de grupo estão disponíveis com granularidade de um minuto sem custo adicional, mas você deve ativá-las.

Quando você ativa as métricas de grupo do Auto Scaling, o Amazon EC2 Auto Scaling envia dados de amostra paraCloudWatcha cada minuto com base no melhor esforço possível. Em casos raros, quandoCloudWatchsofre uma interrupção no serviço, os dados não são preenchidos para preencher lacunas no histórico de métricas do grupo.

Índice

- [Métricas do grupo do Auto Scaling \(p. 332\)](#)
- [Dimensões para métricas do grupo do Auto Scaling \(p. 335\)](#)
- [Métricas e dimensões de escalabilidade preditiva \(p. 335\)](#)
- [Ativar métricas do grupo do Auto Scaling \(console\) \(p. 336\)](#)
- [Habilitar métricas do grupo do Auto Scaling \(AWS CLI\) \(p. 337\)](#)

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	Descrição
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. 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.

Métrica	Descrição
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 estão em serviço ou pendentes. Critérios de relatório: relatado se a coleta de métricas estiver habilitada.
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 a ponderação da instância é usada, as métricas a seguir contam o número de unidades de capacidade usadas pelo seu grupo de Auto Scaling. Se a ponderação da instância não for usada, as métricas a seguir serão preenchidas, mas serão iguais às métricas definidas na tabela anterior. Para obter mais informações sobre ponderação de instâncias, consulte [Configurar ponderação de instâncias para o Amazon EC2 Auto Scaling \(p. 86\)](#) e [Criar um grupo do Auto Scaling usando seleção de tipo de instância baseada em atributos \(p. 92\)](#).

Métrica	Descrição
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	Descrição
	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 grupo de alta atividade. Para obter mais informações, consulte [Grupos de alta atividade do Amazon EC2 Auto Scaling \(p. 279\)](#).

Métrica	Descrição
WarmPoolMinSize	O tamanho mínimo do grupo de alta atividade. 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 grupo de alta atividade. 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 grupo de alta atividade for igual ou maior que a diferença entre o tamanho máximo (ou, se definido, a capacidade máxima preparada) e a capacidade desejada do grupo do Auto Scaling, a capacidade desejada do grupo de alta atividade será equivalente a WarmPoolMinSize. Critérios de relatório: relatado se a coleta de métricas estiver habilitada.
WarmPoolPendingCapacity	A quantidade de capacidade no grupo de alta atividade que está pendente. Essa métrica não inclui instâncias em execução, interrompidas ou sendo terminadas. Critérios de relatório: relatado se a coleta de métricas estiver habilitada.
WarmPoolTerminatingCapacity	A quantidade de capacidade no grupo de alta atividade 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.
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.

Métrica	Descrição
	Critérios de relatório: relatado se a coleta de métricas estiver habilitada.
WarmPoolTotalCapacity	A capacidade total do grupo de alta atividade, 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 grupo de alta atividade combinados. Critérios de relatório: relatado se a coleta de métricas estiver habilitada.
GroupAndWarmPoolTotalCapacity	A capacidade total do grupo do Auto Scaling e o grupo de alta atividade 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	Descrição
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 [Métricas de monitoramento com CloudWatch](#) (p. 233).

Métrica	Descrição	Dimensões
PredictiveScalingLoad	A previsão da quantidade de carga que será gerada por seu aplicativo. 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

Métrica	Descrição	Dimensões
PredictiveScalingCapacityFiducial	<p>A capacidade prevista de capacidade necessária para atender à demanda de aplicativos. Isso se baseia na previsão de carga e no nível de utilização pretendido no qual você deseja manter suas instâncias do Auto Scaling.</p> <p>As estatísticas Average, Minimum e Maximum são úteis, mas a estatística Sum não.</p> <p>Critérios de relatório: reportado após a criação da previsão inicial.</p>	AutoScalingGroupName, PolicyName
PredictiveScalingMetricPairIndex	<p>A métrica contém 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.</p> <p>As estatísticas Average, Minimum e Maximum são úteis, mas a estatística Sum não.</p> <p>Critérios de relatório: reportado após a criação da previsão inicial.</p>	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 grupos do Auto Scaling 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 grupos do Auto Scaling 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 ativar as métricas de grupo do Auto Scaling

Ative uma ou mais métricas de grupo usando o [enable-metrics-collection](#) comando. Por exemplo, o comando a seguir habilita uma única métrica para o grupo especificado de Auto Scaling.

```
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 desativar as métricas de grupo do Auto Scaling

Use [disable-metrics-collection](#) comando 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 [Amazônia CloudWatch preços](#) e [Monitorando suas instâncias usando CloudWatch](#) na Guia do usuário do Amazon EC2 para instâncias Linux.

Para criar um grupo do Auto Scaling, você deve criar uma configuração de execução ou um modelo de execução que permita o tipo de monitoramento adequado ao seu aplicativo. 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.

Índice

- [Habilitar o monitoramento detalhado \(console\) \(p. 337\)](#)
- [Habilitar o monitoramento detalhado \(AWS CLI\) \(p. 338\)](#)
- [Alternar entre monitoramento básico e detalhado \(p. 338\)](#)
- [Colete métricas adicionais usando o CloudWatch agente \(p. 339\)](#)

Habilitar o monitoramento detalhado (console)

Por padrão, o monitoramento básico é habilitado quando você usa o AWS Management Console para criar um modelo ou uma configuração de execução.

Para habilitar o monitoramento detalhado em um modelo de execução

Quando você cria o modelo de lançamento usando o AWS Management Console, na [Detalhes avançados](#) seção, para [Detalhado](#) o CloudWatch monitoramento, escolha Ativar. Caso contrário, o monitoramento básico será habilitado. Para obter mais informações, consulte [Definir configurações avançadas para seu modelo de execução \(p. 29\)](#).

Para habilitar o monitoramento detalhado em uma configuração de execução

Quando você cria a configuração de inicialização usando o AWS Management Console, na [Configuração adicional](#) seção, selecione Habilite o monitoramento detalhado da instância EC2 em CloudWatch. Caso contrário, o monitoramento básico será habilitado. Para obter mais informações, consulte [Criar uma configuração de execução \(p. 49\)](#).

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 [Substituir instâncias do Auto Scaling \(p. 108\)](#).

Ao alternar entre monitoramento básico e detalhado:

Se você tem CloudWatch alarms associados às políticas de escalonamento por etapas ou políticas de escalonamento simples para seu grupo de Auto Scaling, use o [put-metric-alarm](#) comando 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 oCloudWatchagente

Para coletar métricas em nível de sistema operacional, como memória disponível e usada, você deve instalar oCloudWatchagente. Podem ser cobrados taxas adicionais. Você pode usar oCloudWatchagente para coletar métricas do sistema e arquivos de log das instâncias do Amazon EC2. Para obter mais informações, consulte [Métricas coletadas peloCloudWatchagente](#)naAmazôniaCloudWatchGuia do usuário.

Registrar chamadas da API do Amazon EC2 Auto Scaling com o AWS CloudTrail

O Amazon EC2 Auto Scaling é integrado ao AWS CloudTrail, um serviço que fornece um registro das ações realizadas por um usuário, função ou serviço no Amazon EC2 Auto Scaling. CloudTrailcaptura todas as chamadas de API para o Amazon EC2 Auto Scaling como eventos. As chamadas capturadas incluem chamadas do console do Amazon EC2 Auto Scaling e chamadas de código para a API do Amazon EC2 Auto Scaling.

Se você criar uma trilha, poderá habilitar a entrega contínua deCloudTraileventos para um bucket do Amazon S3, incluindo eventos para o Amazon EC2 Auto Scaling. Se não configurar uma trilha, você ainda poderá visualizar os eventos mais recentes no console do CloudTrail em Event history. Usando as informações coletadas porCloudTrail, você pode determinar a solicitação que foi feita ao Amazon EC2 Auto Scaling, o endereço IP do qual a solicitação foi feita, quem fez a solicitação, quando ela foi feita e detalhes adicionais.

Para saber mais sobre CloudTrail, consulte o [Guia do usuário de AWS CloudTrail](#).

Informações do Amazon EC2 Auto Scaling emCloudTrail

CloudTrailestá habilitado em sua conta da Amazon Web Services quando você cria a conta. Quando a atividade ocorre no Amazon EC2 Auto Scaling, essa atividade é registrada em umCloudTrailevento junto com outros eventos da Amazon Web Services emHistórico do evento. Você pode visualizar, pesquisar e baixar os eventos recentes em sua conta da Amazon Web Services. Para obter mais informações, consulte [Visualizar eventos com o histórico de eventos do CloudTrail](#).

Para obter um registro contínuo dos eventos na sua conta da Amazon Web Services, incluindo os eventos do Amazon EC2 Auto Scaling, crie uma trilha. UMAtrilhaativaCloudTrailpara entregar arquivos de log em um bucket do Amazon S3. Por padrão, quando você cria uma trilha no console, ela é aplicada a todas as regiões da . A trilha registra em log eventos de todas as regiões na partição da Amazon Web Services e entrega os arquivos de log para o bucket do Amazon S3 especificado por você. Além disso, você pode configurar outros Amazon Web Services para analisar e agir de acordo com os dados do evento coletados noCloudTrailtroncos. Para obter mais informações, consulte as informações a seguir.

- [Visão geral da criação de uma trilha](#)
- [Serviços e integrações compatíveis com o CloudTrail](#)
- [Configurar notificações do Amazon SNS para o CloudTrail](#)
- [Receber arquivos de log do CloudTrail de várias regiões](#) e [Receber arquivos de log do CloudTrail de várias contas](#)

Todas as ações do Amazon EC2 Auto Scaling são registradas porCloudTraile estão documentados no[Referência da API Amazon EC2 Auto Scaling](#). Por exemplo, as chamadas para as ações CreateLaunchConfiguration, DescribeAutoScalingGroup e UpdateAutoScalingGroup geram entradas nos arquivos de log do CloudTrail.

Cada entrada de log ou evento contém informações sobre quem gerou a solicitação. As informações de identidade ajudam a determinar:

- Se a solicitação foi feita com credenciais de usuário raiz ou do AWS Identity and Access Management (IAM).
- Se a solicitação foi feita com credenciais de segurança temporárias de uma função ou de um usuário federado.
- Se a solicitação foi feita por outro serviço da .

Para obter mais informações, consulte o[CloudTrail userIdentityelemento](#).

Noções básicas sobre entradas do arquivo de log do Amazon EC2 Auto Scaling

Uma trilha é uma configuração que permite a entrega de eventos como arquivos de log a um bucket do Amazon S3 especificado. Os arquivos de log do CloudTrail contêm uma ou mais entradas de log. Um evento representa uma única solicitação de qualquer fonte e inclui informações sobre a ação solicitada, a data e a hora da ação, os parâmetros de solicitação e assim por diante. Os arquivos de log do CloudTrail não são um rastreamento de pilha ordenada de chamadas de API pública. Dessa forma, eles não são exibidos em uma ordem específica.

O exemplo a seguir mostra uma entrada de log do CloudTrail que demonstra a ação CreateLaunchConfiguration.

```
{  
    "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": "standard"  
                }  
            }  
        ]  
    }  
}
```

```
        "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"
}
```

Recursos relacionados

Com CloudWatchLogs, você pode monitorar e receber alertas de eventos específicos capturados pelo CloudTrail. Os eventos enviados para CloudWatchLogs são aqueles configurados para serem registrados pela sua trilha, portanto, certifique-se de ter configurado sua trilha ou trilhas para registrar os tipos de eventos que você está interessado em monitorar. CloudWatchLogs registros podem monitorar as informações nos arquivos de log e notificá-lo quando determinados limites forem atingidos. Você também pode arquivar seus dados de log em armazenamento resiliente. Para obter mais informações, consulte o [AmazonCloudWatchGuia do usuário do Log](#)se o [MonitoramentoCloudTrailarquivos de log com a AmazonCloudWatchRegistros](#)tópico no [AWS CloudTrailGuia do usuário](#).

Receber notificações do Amazon SNS quando o grupo do Auto Scaling escala

Você poderá receber notificações quando o Amazon EC2 Auto Scaling iniciar ou terminar instâncias do EC2 em seu grupo do Auto Scaling. Gerencie notificações usando o Amazon Simple Notification Service (Amazon SNS).

O Amazon SNS coordena e gerencia a entrega ou o envio de notificações a endpoints ou clientes assinantes. O Amazon SNS oferece toda uma variedade de opções de notificação, incluindo a capacidade de entregar notificações como HTTP ou HTTPS POST, e-mail (SMTP, texto sem formatação ou no formato JSON) ou como uma mensagem postada em uma fila do Amazon SQS, o que permite a você manipular essas notificações de forma programática. Para obter mais informações, consulte o [Guia do desenvolvedor do Amazon Simple Notification Service](#).

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.

As notificações são úteis para desenvolver aplicações orientadas por eventos. Se usar notificações para verificar se um recurso entra em um estado desejado, você poderá eliminar sondagens e não encontrar o erro `RequestLimitExceeded` que às vezes resulta da sondagem.

A AWS fornece várias ferramentas que você pode usar para enviar notificações. Como alternativa, você pode usar o EventBridge Amazon SNS para enviar notificações quando seus grupos de Auto Scaling iniciarem ou encerrarem instâncias. No EventBridge, a regra descreve sobre quais eventos você é notificado. No Amazon SNS, o tópico descreve que tipo de notificação você recebe. Com EventBridge, você

pode decidir se determinados eventos devem acionar uma função Lambda em vez disso. Para obter mais informações, consulte [Usar EventBridge para lidar com eventos do Auto Scaling \(p. 400\)](#).

Índice

- [Notificações do SNS \(p. 342\)](#)
- [Configurar notificações do Amazon SNS para o Amazon EC2 Auto Scaling \(p. 343\)](#)
 - [Criar um tópico do Amazon SNS \(p. 343\)](#)
 - [Assinar o tópico do Amazon SNS \(p. 343\)](#)
 - [Confirmar sua assinatura do Amazon SNS \(p. 343\)](#)
 - [Configurar o grupo do Auto Scaling para enviar notificações \(p. 343\)](#)
 - [Testar a notificação \(p. 344\)](#)
 - [Excluir a configuração da notificação \(p. 344\)](#)
- [Política de chaves para um tópico criptografado do Amazon SNS \(p. 345\)](#)

Notificações do SNS

O Amazon EC2 Auto Scaling oferece suporte ao envio de notificações do Amazon SNS quando os seguintes eventos ocorrem.

Evento	Descrição
autoscaling:EC2_INSTANCE_LAUNCH	Ativação de instância bem-sucedida
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 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.

Criar 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 ASCIIIs, 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 da 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 grupos do Auto Scaling 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 Create (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 grupos do Auto Scaling 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. Escolha Update (Atualizar).
 6. Depois de alguns minutos, você receberá uma notificação para o evento. Se não for necessário ter uma 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 grupos do Auto Scaling 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 criptografado do Amazon SNS

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 vinculada ao serviço apropriada que tem acesso permitido à chave. Para obter mais informações, consulte [Configuração de AWS KMS permissões](#) no Guia do desenvolvedor do Amazon Simple Notification Service.

Neste exemplo, a declaração de política fornece à função vinculada ao serviço chamada AWSServiceRoleForAutoScaling permissões para usar a chave gerenciada pelo cliente. Para saber mais sobre a função vinculada ao serviço Amazon EC2 Auto Scaling, consulte [Funções vinculadas ao serviço do Amazon EC2 Auto Scaling \(p. 432\)](#)

```
{  
    "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 aws:SourceAccount condição aws:SourceArn e condição não são suportadas nas principais políticas que permitem que o Amazon EC2 Auto Scaling publique em tópicos criptografados.

Produtos da AWS integrados ao Amazon EC2 Auto Scaling

O Amazon EC2 Auto Scaling pode ser integrado a outros produtos da AWS. Veja as opções de integração a seguir para saber mais sobre como cada serviço funciona com o Amazon EC2 Auto Scaling.

Tópicos

- [Usar o rebalanceamento de capacidade para lidar com interrupções de spot do Amazon EC2 \(p. 346\)](#)
- [Use reservas de capacidade sob demanda para reservar capacidade em zonas de disponibilidade específicas \(p. 354\)](#)
- [Crie um grupo do Auto Scaling na linha de comando usando o AWS CloudShell. \(p. 360\)](#)
- [Criar um grupo do Auto Scaling com AWS CloudFormation \(p. 360\)](#)
- [Use o AWS Compute Optimizer para obter recomendações para o tipo de instância para um grupo do Auto Scaling \(p. 367\)](#)
- [Usar o Elastic Load Balancing para distribuir tráfego entre as instâncias no grupo do Auto Scaling \(p. 369\)](#)
- [roteie o tráfego para o grupo do Auto Scaling \(p. 393\)](#)
- [Usar EventBridge para lidar com eventos do Auto Scaling \(p. 400\)](#)
- [Fornecer conectividade de rede para suas instâncias do Auto Scaling usando a Amazon VPC \(p. 414\)](#)

Usar o rebalanceamento de capacidade para lidar com interrupções de spot do Amazon EC2

Você pode configurar o Amazon EC2 Auto Scaling para monitorar e responder automaticamente a alterações que afetam a disponibilidade de suas instâncias spot. O rebalanceamento de capacidade ajuda a manter a disponibilidade da workload aumentando proativamente sua frota com uma nova instância spot antes que uma instância em execução seja interrompida por Amazon EC2.

O objetivo do rebalanceamento de capacidade é continuar processando sua workload sem interrupção. Quando as instâncias spot apresentam risco elevado de interrupção, o Amazon EC2 Spot Service notifica o Amazon EC2 Auto Scaling com uma recomendação de rebalanceamento de instância do EC2.

Quando você habilita o rebalanceamento de capacidade do grupo do Auto Scaling, o Amazon EC2 Auto Scaling tenta substituir proativamente as instâncias spot do grupo que receberam uma recomendação de rebalanceamento. Você pode decidir rebalancear sua workload em instâncias spot novas ou existentes que não tenham risco elevado de interrupção. A workload pode continuar processando o trabalho enquanto o Amazon EC2 Auto Scaling inicia novas instâncias spot antes que instâncias existentes sejam interrompidas.

Quando você não usa o rebalanceamento de capacidade, o Amazon EC2 Auto Scaling não substitui 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.

Índice

- [Visão geral \(p. 347\)](#)

- [Comportamento de reequilíbrio de capacidade \(p. 347\)](#)
- [Considerações \(p. 348\)](#)
- [Habilitar o rebalanceamento de capacidade \(console\) \(p. 349\)](#)
- [Habilitar o rebalanceamento de capacidade \(AWS CLI\) \(p. 350\)](#)
- [Recursos relacionados \(p. 353\)](#)
- [Limitações \(p. 353\)](#)

Visão geral

Para usar o Rebalanceamento de Capacidade com seu grupo de Auto Scaling, as etapas básicas são:

1. Configure seu grupo de 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.
2. Adicione ganchos de ciclo de vida conforme necessário para realizar um desligamento normal do seu aplicativo dentro das instâncias que recebem a notificação de rebalanceamento. Para obter mais informações, consulte [Ganchos do ciclo de vida do Amazon EC2 Auto Scaling \(p. 252\)](#).

A seguir estão alguns motivos pelos quais você pode usar um gancho de 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 registros do sistema ou do aplicativo e enviá-los para o Amazon Simple Storage Service (Amazon S3)
3. Desenvolva uma ação personalizada para o gancho do ciclo de vida. Para fazer com que o gancho do ciclo de vida invoque sua ação personalizada, 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 um EventBridge e automatize a ação a ser tomada quando um padrão de evento coincide com a 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, você pode começar a usar o rebalanceamento de capacidade.

Comportamento de reequilíbrio 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 espera até que a nova instância passe pela 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 de ciclo de vida ao seu grupo de Auto Scaling, o Amazon EC2 Auto Scaling começará a encerrar as instâncias anteriores assim que as novas instâncias passarem pela verificação de saúde.

- Se você adicionou um gancho de ciclo de vida, isso aumenta o tempo necessário até começarmos a encerrar as instâncias anteriores pelo valor de tempo limite que você especificou para o gancho do ciclo de vida.
- Se você estiver usando políticas de escalabilidade ou escalabilidade programada, 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 obter sucesso.

Na pior das hipóteses, se novas instâncias falharem na inicialização ou suas verificações de integridade falharem, o Amazon EC2 Auto Scaling continuará tentando reiniciá-las. Enquanto estiver tentando iniciar novas instâncias, as anteriores acabarão sendo interrompidas e encerradas à força com um aviso de interrupção de dois minutos.

Considerações

Considere o seguinte ao usar o rebalanceamento de capacidade:

Projete seu aplicativo para ser tolerante a interrupções pontuais

Seu aplicativo deve ser capaz de lidar com mudanças dinâmicas no número de instâncias e na possibilidade de uma instância spot ser interrompida precocemente. Por exemplo, se seu grupo de Auto Scaling estiver por trás de um balanceador de carga do Elastic Load Balancing, o Amazon EC2 Auto Scaling espera que a instância cancele o registro do balanceador de carga antes de chamar seu gancho de 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 que o Amazon EC2 envie o sinal de recomendação de rebalanceamento antes do aviso de interrupção da instância spot de dois minutos. Às vezes, o sinal de recomendação de rebalanceamento 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 substitutas podem ter um risco elevado de interrupção se você usar o `lowest-price` estratégia de alocação. Isso ocorre porque lançamos instâncias no pool de menor preço que tem capacidade disponível naquele momento, mesmo que suas instâncias spot substitutas provavelmente sejam interrompidas logo após o lançamento. Para evitar um risco elevado de interrupção, é altamente recomendável que você não use `lowest-price` estratégia de alocação. Em vez disso, recomendamos o `price-capacity-optimized` estratégia de alocação. Essa estratégia lança instâncias spot substitutas em pools spot que têm menor probabilidade de serem interrompidas e têm o menor preço 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 a mesma 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 podem ser substituídas do que se você esperasse que o Amazon EC2 Auto Scaling iniciasse uma nova instância depois que a instância em risco fosse interrompida.

Embora você possa substituir mais instâncias com o rebalanceamento de capacidade ativado, você se beneficia de 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 de Capacidade lançando uma nova instância com antecedência, você oferece aos processos existentes uma melhor chance de serem concluídos em sua instância em risco. Você também pode iniciar os procedimentos de encerramento da instância, impedir que novos trabalhos sejam agendados em sua instância em risco e preparar a instância recém-lançada para assumir o controle do aplicativo. Com a substituição proativa no rebalanceamento de capacidade, você se beneficia de uma continuidade elegante.

O exemplo teórico a seguir demonstra os riscos e benefícios do uso do rebalanceamento de capacidade:

- 14h — Uma recomendação de rebalanceamento é recebida, por exemplo A. O Amazon EC2 Auto Scaling tenta imediatamente iniciar a instância B de substituição, dando a você tempo para iniciar seus procedimentos de desligamento.
- 14h30 — Uma recomendação de rebalanceamento é recebida para a instância B, que é substituída pela instância C. Isso lhe dá tempo para iniciar os procedimentos de desligamento.
- 14h32 — Se o rebalanceamento de capacidade não estiver ativado e se um aviso de interrupção da instância spot tivesse sido recebido às 14h32, por exemplo A, você teria apenas dois minutos para agir. No entanto, a Instância A teria continuado em execução até esse momento.

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 grupos do Auto Scaling no painel de navegação.
2. Selecione Criar grupo do Auto Scaling.
3. ParaEtapa 1: Escolha o modelo ou a configuração de inicialização, insira um nome para o grupo Auto Scaling, escolha um modelo de lançamento e, em seguida, escolhaPróximopara prosseguir para a próxima etapa.
4. Em Step 2: Choose instance launch options (Etapa 2: escolher as opções para iniciar a instância), em Network (Rede), escolha as opções, conforme desejado. Verifique se as sub-redes que você deseja utilizar se encontram em diferentes zonas de disponibilidade.
5. ParaRequisitos de tipo de instância, escolha as configurações para criar um grupo de instâncias mistas. Isso inclui os tipos de instância que ele pode iniciar, as opções de compra da instância e as estratégias de alocação para instâncias spot e sob demanda. 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 \(p. 67\)](#).

6. Na seção Allocation strategies (Estratégias de alocação), na parte inferior da página, escolha uma estratégia de alocação spot. Selecione ou desmarque a caixa de seleção Capacity rebalance (Rebalanceamento de capacidade) para habilitar ou desabilitar o rebalanceamento de capacidade. Você só vê essa opção quando solicita que uma porcentagem do grupo de Auto Scaling seja lançada como instâncias spot noOpções de compra de instânciasseção.
7. Crie o grupo do Auto Scaling.
8. (Opcional) Adicione ganchos de ciclo de vida conforme necessário. Para obter mais informações, consulte [Ganchos do ciclo de vida do Amazon EC2 Auto Scaling \(p. 252\)](#).

Para ativar ou desativar o rebalanceamento de capacidade para um grupo existente de Auto Scaling

1. Abra o console do Amazon EC2 em <https://console.aws.amazon.com/ec2/> e escolha grupos do Auto Scaling 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. Sob o Estratégias de alocação seção, ative ou desative o Rebalanceamento de Capacidade selecionando ou desmarcando o Reequilíbrio de capacidade caixa de seleção.
5. Escolha Atualizar.

Habilitar o rebalanceamento de capacidade (AWS CLI)

Os exemplos a seguir mostram como usar a AWS CLI para habilitar e desabilitar o rebalanceamento de Capacidade.

Use o [create-auto-scaling-group](#) ou [update-auto-scaling-group](#) comando com o seguinte parâmetro:

- `--capacity-rebalance / --no-capacity-rebalance`: valor booleano que indica se o rebalanceamento de capacidade está habilitado.

Antes de ligar para o [create-auto-scaling-group](#) comando, 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 \(p. 23\)](#).

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, será necessário 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 seguinte [create-auto-scaling-group](#) comando para criar um novo grupo de Auto Scaling e ativar o rebalanceamento de capacidade. Esse comando faz referência a um arquivo JSON como o 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 \(p. 67\)](#), 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 seguinte [create-auto-scaling-group](#) comando para criar um novo grupo de Auto Scaling e ativar o rebalanceamento de capacidade. Esse comando faz referência a um arquivo YAML como o único parâmetro para seu grupo de 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 seguinte [update-auto-scaling-group](#) comando 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 seguinte [describe-auto-scaling-groups](#) comando para verificar se o rebalanceamento de capacidade está ativado e para visualizar os detalhes.

```
aws autoscaling describe-auto-scaling-groups --auto-scaling-group-name my-asg
```

Esta é uma resposta de exemplo.

```
{  
    "AutoScalingGroups": [  
        {  
            "AutoScalingGroupName": "my-asg",  
            "AutoScalingGroupARN": "arn",  
            ...  
            "CapacityRebalance": true  
        }  
    ]  
}
```

Para desabilitar o rebalanceamento de capacidade

Use o [update-auto-scaling-group](#) comando com o `--no-capacity-rebalance` opção para 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 [Gerencie proativamente o ciclo de vida da instância spot usando o novo recurso de rebalanceamento de capacidade para EC2 Auto Scaling](#) no AWSBlog de computação.

Para obter mais informações sobre as recomendações de rebalanceamento de instâncias do EC2, consulte [Recomendações de rebalanceamento de instâncias do EC2](#) na Guia do usuário do Amazon EC2 para instâncias Linux.

Para saber mais sobre ganchos de ciclo de vida, consulte os seguintes recursos.

- [Tutorial: Configurar um gancho do ciclo de vida que invoca uma função do Lambda \(p. 273\)](#) (usando EventBridge)
- [Tutorial: configurar dados do usuário para recuperar o estado de destino do ciclo de vida por meio de metadados de instância \(p. 267\)](#)

Limitações

- O Amazon EC2 Auto Scaling pode encerrar a instância que recebe a notificação de rebalanceamento somente se a instância não estiver protegida da expansão. No entanto, essa proteção não impede o encerramento devido a uma interrupção do Spot.
- O suporte para rebalanceamento de capacidade está disponível em todos os estabelecimentos comerciais Regiões da AWS onde o Amazon EC2 Auto Scaling está disponível, exceto na região do Oriente Médio (EAU).

Use reservas de capacidade sob demanda para reservar capacidade em zonas de disponibilidade específicas

As reservas de capacidade sob demanda do Amazon EC2 ajudam você a reservar a capacidade computacional em zonas de disponibilidade específicas. Para começar a usar reservas de capacidade, você cria uma reserva de capacidade em uma zona de disponibilidade específica. Depois, é possível executar instâncias na capacidade reservada, visualizar a utilização da capacidade em tempo real e aumentar ou diminuir a capacidade conforme necessário.

As Reservas de Capacidade são configuradas como open ou targeted. Se a reserva de capacidade for open, todas as instâncias novas e existentes que têm atributos correspondentes são executadas automaticamente na capacidade da reserva de capacidade. Se a Reserva de capacidade for targeted, as instâncias deverão usá-la como destino especificamente para executar na capacidade reservada.

Este tópico mostra como criar um grupo de Auto Scaling que inicia instâncias sob demanda em targetedReservas de capacidade. Isso lhe dá mais controle sobre quando usar reservas de capacidade específicas.

As etapas básicas são:

1. Crie reservas de capacidade em várias zonas de disponibilidade que tenham o mesmo tipo de instância, plataforma e contagem de instâncias.
2. Reservas de capacidade de grupo usando AWSGrupos de recursos.
3. Crie um grupo de Auto Scaling com um modelo de lançamento direcionado ao grupo de recursos, usando as mesmas zonas de disponibilidade das reservas de capacidade.

Índice

- [Etapa 1: Criar as reservas de capacidade \(p. 354\)](#)
- [Etapa 2: Criar um grupo de reserva de capacidade \(p. 356\)](#)
- [Etapa 3: criar um modelo de lançamento \(p. 357\)](#)
- [Etapa 4: criar um grupo de Auto Scaling \(p. 358\)](#)
- [Recursos relacionados \(p. 360\)](#)

Etapa 1: Criar as reservas de capacidade

A primeira etapa é criar uma reserva de capacidade em cada zona de disponibilidade em que seu grupo de Auto Scaling será implantado.

Note

Você só pode criartargetedreservas quando você cria as reservas de capacidade pela primeira vez.

Console

Para criar suas reservas de capacidade

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

2. Selecione Reservas de Capacidade e Create Reserva de capacidade (Criar Reserva de capacidade).
3. Sobre o Criar uma reserva de capacidade página, preste atenção às seguintes configurações na Detalhes da instância seção. O tipo de instância, a plataforma e a zona de disponibilidade das instâncias iniciadas devem corresponder ao tipo de instância, à plataforma e à zona de disponibilidade especificadas aqui ou a Reserva de capacidade não será aplicada.
 - a. Para Tipo de instância, escolha o tipo de instância a ser executada na capacidade reservada.
 - b. Para Plataforma, escolha o sistema operacional para suas instâncias.
 - c. Para Zona de disponibilidade, escolha a primeira zona de disponibilidade na qual você deseja reservar capacidade.
 - d. Para Quantidade, escolha o número de instâncias que você precisa. Calcule o número total de instâncias necessárias para seu grupo de Auto Scaling dividido pelo número de zonas de disponibilidade que você planeja usar.
4. Abaixo Detalhes da reserva de capacidade, para A reserva de capacidade termina, escolha uma das seguintes opções:
 - Em um horário específico—Cancele a reserva de capacidade automaticamente na data e hora especificadas.
 - Manually (Manualmente) — reserve a capacidade até que você a cancele explicitamente.
5. Para Elegibilidade da instância, escolha Destinado: somente instâncias que visam a reserva de capacidade.
6. (Opcional) Para Etiquetas, especifique quaisquer etiquetas a serem associadas à reserva de capacidade.
7. Escolha Create (Criar).
8. Anote o ID da reserva de capacidade recém-criada. Você precisa dele para configurar o grupo de reserva de capacidade.

Repita esse procedimento para cada zona de disponibilidade que você deseja ativar para seu grupo de Auto Scaling, alterando somente o valor do Zona de disponibilidade opção.

AWS CLI

Para criar suas reservas de capacidade

Use o seguinte [create-capacity-reservation](#) comando para criar as reservas de capacidade. Substitua os valores de amostra por --availability-zone, --instance-type, --instance-platform, e --instance-count.

```
aws ec2 create-capacity-reservation \
--availability-zone us-east-1a \
--instance-type c5.xlarge \
--instance-platform Linux/UNIX \
--instance-count 3 \
--instance-match-criteria targeted
```

Exemplo de ID de reserva de capacidade resultante

```
{  
    "CapacityReservation": {  
        "CapacityReservationId": "cr-1234567890abcdef1",  
        "OwnerId": "123456789012",  
        "CapacityReservationArn": "arn:aws:ec2:us-east-1:123456789012:capacity-  
reservation/cr-1234567890abcdef1",  
        "InstanceType": "c5.xlarge",  
        "InstancePlatform": "Linux/UNIX",  
    }  
}
```

```
        "AvailabilityZone": "us-east-1a",
        "Tenancy": "default",
        "TotalInstanceCount": 3,
        "AvailableInstanceCount": 3,
        "EbsOptimized": false,
        "EphemeralStorage": false,
        "State": "active",
        "StartDate": "2023-07-26T21:36:14+00:00",
        "EndDateType": "unlimited",
        "InstanceMatchCriteria": "targeted",
        "CreateDate": "2023-07-26T21:36:14+00:00"
    }
}
```

Anote o ID da reserva de capacidade recém-criada. Você precisa dele para configurar o grupo de reserva de capacidade.

Repita esse comando para cada zona de disponibilidade que você deseja ativar para seu grupo de Auto Scaling, alterando somente o valor do --availability-zoneopção.

Etapa 2: Criar um grupo de reserva de capacidade

Quando terminar de criar as reservas de capacidade, você poderá agrupá-las usando o AWS Serviço de grupos de recursos. Os Grupos de Recursos oferecem suporte a vários tipos diferentes de grupos para diferentes usos. O Amazon EC2 usa um grupo para fins especiais, conhecido como grupo de recursos vinculados a serviços, para atingir um grupo de reservas de capacidade. Para interagir com esse grupo de recursos vinculado ao serviço, você pode usar o AWS CLI ou um SDK, mas não o console. Para obter mais informações sobre grupos de recursos vinculados a serviços, consulte [Configurações de serviço para grupos de recursos](#) na AWS Guia do usuário de grupos de recursos.

Para criar um grupo de reserva de capacidade usando o AWS CLI

Use o [criar grupo](#) comando para criar um grupo de recursos que pode conter somente reservas de capacidade. Neste exemplo, o grupo de recursos é chamado de **my-cr-group**.

```
aws resource-groups create-group \
  --name my-cr-group \
  --configuration '[{"Type": "AWS::EC2::CapacityReservationPool"}, \
  {"Type": "AWS::ResourceGroups::Generic", "Parameters": [{"Name": "allowed-resource-types", \
  "Values": ["AWS::EC2::CapacityReservation"]}]}'
```

Esta é uma resposta de exemplo.

```
{
  "Group": {
    "GroupArn": "arn:aws:resource-groups:us-east-1:123456789012:group/my-cr-group",
    "Name": "my-cr-group"
  },
  "GroupConfiguration": {
    "Configuration": [
      {
        "Type": "AWS::EC2::CapacityReservationPool"
      },
      {
        "Type": "AWS::ResourceGroups::Generic",
        "Parameters": [
          {
            "Name": "allowed-resource-types",
            "Values": [
              "AWS::EC2::CapacityReservation"
            ]
          }
        ]
      }
    ]
  }
}
```

```
        "AWS::EC2::CapacityReservation"
    ]
}
],
"Status": "UPDATE_COMPLETE"
}
```

Anote o ARN do grupo de recursos. Você precisa dele para configurar o modelo de lançamento para seu grupo de Auto Scaling.

Para associar suas reservas de capacidade ao grupo recém-criado usando o AWS CLI

Use o seguinte comando para associar as reservas de capacidade ao grupo de reserva de capacidade recém-criado. Para o `--resource-arns` opção, especifique as reservas de capacidade usando seus ARNs. Crie os ARNs usando a região relevante, o ID da sua conta e os IDs de reserva que você anotou anteriormente. Neste exemplo, as reservas com IDs `cr-1234567890abcdef1` e `cr-54321abcdef567890` serão agrupados no grupo chamado `my-cr-group`.

```
aws resource-groups group-resources \
--group my-cr-group \
--resource-arns \
arn:aws:ec2:region:account-id:capacity-reservation/cr-1234567890abcdef1 \
arn:aws:ec2:region:account-id:capacity-reservation/cr-54321abcdef567890
```

Esta é uma resposta de exemplo.

```
{
  "Succeeded": [
    "arn:aws:ec2:us-east-1:123456789012:capacity-reservation/cr-1234567890abcdef1",
    "arn:aws:ec2:us-east-1:123456789012:capacity-reservation/cr-54321abcdef567890"
  ],
  "Failed": [],
  "Pending": []
}
```

Para obter informações sobre como modificar ou excluir o grupo de recursos, consulte o [AWS Referência da API de grupos de recursos](#).

Etapa 3: criar um modelo de lançamento

Console

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 Create launch template (Criar modelo de execução). Insira um nome e forneça uma descrição para a versão inicial do modelo de execução.
4. Em Auto Scaling guidance (Guia do Auto Scaling), marque a caixa de seleção.
5. Crie o modelo de lançamento. Escolha uma AMI e um tipo de instância que correspondam às reservas de capacidade que você planeja usar e, opcionalmente, um par de chaves, um ou mais grupos de segurança e quaisquer volumes adicionais do EBS ou volumes de armazenamento de instâncias para suas instâncias.

6. ExpandirDetalhes avançadose faça o seguinte:
 - a. ParaReserva de capacidade, escolhaAlvo por grupo.
 - b. ParaReserva de capacidade - Alvo por grupo, escolha o grupo de Reservas de Capacidade que você criou na seção anterior e, em seguida, escolhaSalvar.
7. Escolha Create launch template (Criar modelo de execução).
8. Na página de confirmação, escolha Create Auto Scaling group (Criar grupo do Auto Scaling).

AWS CLI

Para criar um modelo de execução

Use o seguinte [create-launch-template](#) comando para criar um modelo de lançamento que especifica que a reserva de capacidade é direcionada a um grupo de recursos específico. Substitua o valor da amostra por --launch-template-name. Substituir `c5.xlarge` com o tipo de instância que você usou na reserva de capacidade `eami-0123456789EXAMPLE` com o ID da AMI que você deseja usar. Substituir `arn:aws:resource-groups:region:account-id:group/my-cr-group` com o ARN do grupo de recursos que você criou no início da seção anterior.

```
aws ec2 create-launch-template \
--launch-template-name my-launch-template \
--launch-template-data \
'{"InstanceType": "c5.xlarge", \
 "ImageId": "ami-0123456789EXAMPLE", \
 "CapacityReservationSpecification": \
 {"CapacityReservationTarget": \
 { "CapacityReservationResourceGroupArn": "arn:aws:resource- \
 groups:region:account-id:group/my-cr-group" } \
 } \
}'
```

Esta é uma resposta de exemplo.

```
{ \
    "LaunchTemplate": { \
        "LaunchTemplateId": "lt-0dd77bd41dEXAMPLE", \
        "LaunchTemplateName": "my-launch-template", \
        "CreateTime": "2023-07-26T21:42:48+00:00", \
        "CreatedBy": "arn:aws:iam::123456789012:user/Bob", \
        "DefaultVersionNumber": 1, \
        "LatestVersionNumber": 1 \
    } \
}
```

Etapa 4: criar um grupo de Auto Scaling

Console

Crie seu grupo de Auto Scaling como você costuma fazer, mas ao escolher suas sub-redes VPC, escolha uma sub-rede de cada zona de disponibilidade que corresponda à `targetedReserves` de capacidade que você criou. Então, quando seu grupo de Auto Scaling iniciar uma instância sob demanda em uma dessas zonas de disponibilidade, a instância será executada na capacidade reservada para essa zona de disponibilidade. Se o grupo de recursos ficar sem reservas de capacidade antes que a capacidade desejada seja atendida, lançaremos qualquer coisa além da capacidade reservada como capacidade normal sob demanda.

Para criar um grupo simples de Auto Scaling

1. Abra o console do Amazon EC2 em <https://console.aws.amazon.com/ec2/> e escolha grupos do Auto Scaling 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. Em Launch template (Modelo de execução), escolha um modelo de execução existente.
6. 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.
7. Na página Choose instance launch options (Escolher as opções de execução da instância) em Network (Rede), para VPC, selecione uma VPC. O grupo do Auto Scaling deve ser criado na mesma VPC do grupo de segurança especificado no modelo de execução. Se você não especificou um grupo de segurança em seu modelo de lançamento, pode escolher qualquer VPC que tenha sub-redes nas mesmas zonas de disponibilidade de suas reservas de capacidade.
8. Para Zonas de disponibilidade e sub-redes, escolha sub-redes de cada zona de disponibilidade que você deseja incluir, com base nas zonas de disponibilidade em que suas reservas de capacidade estão.
9. Escolha Next (Próximo) duas vezes.
10. Sobre o Configurar o tamanho do grupo e as políticas de escalabilidade, para Capacidade desejada, insira o número inicial de instâncias a serem iniciadas. Quando esse número é alterado para um valor fora dos limites de capacidade mínima ou máxima, é necessário atualizar os valores de Minimum capacity (Capacidade mínima) ou Maximum capacity (Capacidade máxima). Para obter mais informações, consulte [Definir limites de capacidade no grupo do Auto Scaling \(p. 166\)](#).
11. Escolha Skip to review (Ir para revisão).
12. Na página Review (Revisão), escolha Create Auto Scaling group (Criar grupo do Auto Scaling).

AWS CLI

Para criar um grupo simples de Auto Scaling

Use o seguinte `create-auto-scaling-group` comando e especifique o nome e a versão do seu modelo de lançamento como o valor do `--launch-template` opção. Substitua os valores de amostra por `--auto-scaling-group-name`, `--min-size`, `--max-size`, e `--vpc-zone-identifier`.

Para o `--availability-zones` opção, especifique as zonas de disponibilidade para as quais você criou reservas de capacidade. Por exemplo, se suas reservas de capacidade especificarem `us-east-1a` e `us-east-1b` Zonas de disponibilidade, então você deve criar seu grupo de Auto Scaling nas mesmas zonas. Então, quando seu grupo de Auto Scaling iniciar uma instância sob demanda em uma dessas zonas de disponibilidade, a instância será executada na capacidade reservada para essa zona de disponibilidade. Se o grupo de recursos ficar sem reservas de capacidade antes que a capacidade desejada seja atendida, lançaremos qualquer coisa além da capacidade reservada como capacidade normal sob demanda.

```
aws autoscaling create-auto-scaling-group \
    --auto-scaling-group-name my-asg \
    --launch-template LaunchTemplateName=my-launch-template,Version='1' \
    --min-size 6 \
    --max-size 6 \
    --vpc-zone-identifier "subnet-5f46ec3b,subnet-0ecac448" \
```

```
--availability-zones us-east-1a us-east-1b
```

Recursos relacionados

Para ver um exemplo de implementação, consulte AWS CloudFormation modelo a seguir AWS amostras GitHub repositório: <https://github.com/aws-samples/aws-auto-scaling-backed-by-on-demand-capacity-reserves/>.

Para obter mais informações sobre reservas de capacidade, consulte os seguintes recursos.

- [Criar uma reserva de capacidade](#) na Guia do usuário do Amazon EC2 para instâncias Linux
- [Reservas de capacidade sob demanda](#) na Guia do usuário do Amazon EC2 para instâncias Linux
- [Segmentar um grupo de reservas de capacidade sob demanda do Amazon EC2](#) no AWS Blog sobre operações e migrações na nuvem

Crie um grupo do Auto Scaling na linha de comando usando o AWS CloudShell.

Em [Regiões da AWS compatíveis](#), você pode executar comandos da AWS CLI usando o AWS CloudShell para um shell pré-autenticado baseado em navegador que é iniciado diretamente do AWS Management Console. Você pode correr AWS CLI comandos contra serviços usando seu shell preferido (Bash, PowerShell, ou concha Z).

Você pode iniciar o AWS CloudShell pelo AWS Management Console usando um dos seguintes dois métodos:

- Clique no ícone AWS CloudShell 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 CloudShell, em seguida, escolha o CloudShell opção.

Quando o AWS CloudShell for iniciado em uma nova janela do navegador pela primeira vez, um painel de boas-vindas vai exibir e listar 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](#).

Criar um grupo do Auto Scaling com AWS CloudFormation

O Amazon EC2 Auto Scaling é integrado ao AWS CloudFormation, um serviço que ajuda você a modelar e configurar os recursos da AWS, para passar menos tempo criando e gerenciando os recursos e a infraestrutura. Você cria um modelo que descreve todos os recursos da AWS desejados (como os grupos do Auto Scaling), e o AWS CloudFormation provisiona e configura esses recursos para você.

Quando você usa o AWS CloudFormation, é possível reutilizar seu modelo para configurar seus recursos do Amazon EC2 Auto Scaling repetidamente e de forma consistente. Descreva seus recursos uma vez e depois provisione os mesmos recursos repetidamente em várias regiões e contas da AWS.

Amazon EC2 Auto Scaling e modelos AWS CloudFormation

Para provisionar e configurar recursos para Amazon EC2 Auto Scaling e serviços relacionados, você deve entender [AWS CloudFormation modelos](#). Os modelos são arquivos de texto formatados em JSON ou YAML. Esses modelos descrevem os recursos que você deseja provisionar nas suas pilhas do AWS CloudFormation. Se você não estiver familiarizado com JSON ou YAML, poderá usar o AWS CloudFormation Designer para ajudá-lo a começar a usar os modelos do AWS CloudFormation. Para obter mais informações, consulte [O que é o Designer?](#) (O que é o AWS CloudFormation Designer) no Manual do usuário do AWS CloudFormation.

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::LaunchTemplate](#).
- Crie um grupo de Auto Scaling usando Group [AWS::AutoScaling::AutoScaling](#).

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平衡amento de carga](#) no Guia do usuário do AWS CloudFormation.

Você encontra outros exemplos úteis de modelos que criam grupos do Auto Scaling e recursos relacionados na seção [Treichos de modelos do Auto Scaling](#) do Guia do usuário do AWS CloudFormation. Para obter mais informações e trechos de exemplo, consulte a [Referência do tipo de recurso do Amazon EC2 Auto Scaling](#) no Guia do usuário do AWS CloudFormation.

Saiba mais sobre o AWS CloudFormation

Para saber mais sobre o AWS CloudFormation, consulte os seguintes recursos:

- [AWS CloudFormation](#)
- [Manual do usuário do AWS CloudFormation](#)
- [AWS CloudFormation Referência da API](#)
- [Guia do usuário da interface de linha de comando do AWS CloudFormation](#)

Migrar AWS CloudFormation pilhas de configurações de execução

Você pode migrar seus modelos de AWS 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 Auto Scaling no modelo de pilha. Em seguida, use seu modelo modificado para atualizar sua pilha.

Ao migrar para modelos de inicialização, este tópico economiza seu tempo ao fornecer instruções para reescrever as configurações de inicialização em seus modelos de CloudFormation pilha como modelos de execução. Para obter mais informações sobre como migrar uma configurações de execução para modelos de execução, consulte [Migre para lançar modelos \(p. 32\)](#)

Tópicos

- [Localizar grupos do Auto Scaling que usem uma configuração de execução \(p. 362\)](#)
- [Atualizar uma pilha para usar um modelo de execução \(p. 362\)](#)

- [Compreender o comportamento de atualização \(p. 365\)](#)
- [Acompanhar a migração \(p. 365\)](#)
- [Referência do mapeamento de configuração \(p. 366\)](#)

Localizar grupos do Auto Scaling que usem uma configuração de execução

Para encontrar grupos do Auto Scaling que usem uma configuração de execução

- Use o [describe-auto-scaling-groups](#) comando a seguir para listar os nomes dos grupos de 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 AWS CLI comandos adicionais úteis para encontrar grupos de Auto Scaling para migrar e filtrar a saída. [Migre para lançar modelos \(p. 32\)](#)

Important

Se os recursos da pilha tiverem AWSEB em seu nome, isso significa que eles foram criados por meio de AWS Elastic Beanstalk. Nesse caso, você deve atualizar o ambiente Beanstalk para orientar o Elastic Beanstalk a remover a configuração de inicialização e substituí-la por um modelo de execução.

Atualizar uma pilha para usar um modelo de execução

Siga as etapas nesta seção para fazer o seguinte:

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

Para modificar o modelo de pilha e atualizar a pilha

1. Siga os mesmos procedimentos gerais para modificar o modelo de pilha descrito em [Modificar um modelo de pilha no Guia do usuário](#). AWS CloudFormation
2. Reescreva a configuração de lançamento como um modelo de lançamento. Veja o exemplo a seguir:

Exemplo: Uma configuração de inicialização 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 lançamento 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 [AWS::EC2::LaunchTemplate](#) Guia do AWS CloudFormation usuário.

Observe como o modelo de lançamento inclui a `LaunchTemplateName` propriedade 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 `IamInstanceProfile` propriedade estiver presente na sua configuração de execução, você deverá 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 `PlacementTenancy` propriedades `AssociatePublicIpAddress``InstanceMonitoring`, ou estiverem presentes em sua configuração de inicialização, você deverá convertê-las em uma estrutura. Para ver exemplos, consulte [AWS::EC2::LaunchTemplate](#).

Uma exceção é quando o valor da `MapPublicIpOnLaunch` propriedade nas sub-redes que você usou para seu grupo de Auto Scaling corresponde ao valor da `AssociatePublicIpAddress` propriedade em sua configuração de inicialização. Nesse caso, você pode ignorar a `AssociatePublicIpAddress` propriedade. A `AssociatePublicIpAddress` propriedade só é usada para substituir a `MapPublicIpOnLaunch` propriedade para alterar se as instâncias recebem um endereço IPv4 público na inicialização.

5. Você pode copiar grupos de segurança da **SecurityGroups** propriedade para um dos dois lugares em seu modelo de lançamento. Normalmente, você copia os grupos de segurança para a `SecurityGroupIds` propriedade. No entanto, se você criar uma `NetworkInterfaces` estrutura em seu modelo de execução para especificar a `AssociatePublicIpAddress` propriedade, deverá copiar os grupos de segurança para a `Groups` propriedade da interface de rede.
6. Se alguma `BlockDeviceMapping` estrutura estiver presente em sua configuração de execução com `NoDevice` set to `true`, você deverá especificar uma string vazia para `NoDevice` em seu modelo de execução para que o Amazon EC2 omita o dispositivo.
7. Se a **SpotPrice** propriedade estiver presente em sua configuração de lançamento, recomendamos que você a omita do seu modelo de lançamento. Suas instâncias spot serão executadas pelo preço spot atual atual. Esse preço nunca excederá o preço sob demanda.

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

- A primeira é usar a `InstanceMarketOptions` estrutura em seu modelo de lançamento (não recomendado). Para obter mais informações, consulte [AWS::EC2::LaunchTemplateInstanceMarketOptions](#) Guia AWS CloudFormation do usuário.
 - A outra é adicionar uma `MixedInstancesPolicy` estrutura ao seu grupo do Auto Scaling. Isso fornece mais opções de como fazer a solicitação. Uma solicitação de instância spot em seu modelo de execução não suporta mais de uma seleção de tipo de instância por grupo de Auto Scaling. No entanto, uma política de instâncias mistas oferece suporte à seleção de mais de um tipo de instância por grupo de Auto Scaling. As solicitações de instâncias spot se beneficiam de ter mais de um tipo de instância para escolher. Para obter mais informações, consulte [AWS::AutoScaling::AutoScalingMixedInstancesPolicy](#) e [AWS::AutoScaling::AutoScalingGroup](#) no Guia AWS CloudFormation do usuário.
8. Remova a **LaunchConfigurationName** propriedade do recurso [AWS::AutoScaling::AutoScalingAWS::AutoScalingGroup](#). Adicione o modelo de lançamento em seu lugar.

Nos exemplos a seguir, a função intrínseca `Ref` obtém o ID do [AWS::EC2::LaunchTemplate](#) recurso com o ID lógico `myLaunchTemplate`. A `GetAtt` função 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 myLaunchTemplateLatestVersionNumber  
      ...
```

Exemplo: Com uma política de instâncias mistas

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

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

Para obter informações de referência sobre todas as propriedades que o Amazon EC2 Auto Scaling suporta, consulte [AWS::AutoScaling::AutoScaling](#) de grupos no Guia do AWS CloudFormationusuário.

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

Compreender o comportamento de atualização

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

CloudFormationsuporta o [UpdatePolicy](#)atributo para grupos de Auto Scaling. Durante uma atualização, se UpdatePolicy definido comoAutoScalingRollingUpdate, CloudFormation substitui InService as instâncias após você executar as etapas deste procedimento. Se UpdatePolicy estiver definido comoAutoScalingReplacingUpdate, CloudFormation substitui o grupo Auto Scaling e sua piscina aquecida (se houver).

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

Acompanhar a migração

Para rastrear a migração

1. No [console do AWS 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 mais recentes, selecione o botão "atualizar" no console do CloudFormation.
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 lançamento, consulte [Solucionar problemas do Amazon EC2 Auto Scaling: modelos de execução \(p. 471\)](#) as possíveis causas.
4. (Opcional) Dependendo do uso do UpdatePolicy atributo, você pode monitorar o progresso do seu grupo de 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 AWS CloudFormation usuário.
6. Depois que a atualização da pilha estiver concluída, abra a página de [modelos do Launch e a página de configurações do Launch](#) do console do Amazon EC2. Você notará que um novo modelo de lançamento foi criado e a configuração de inicialização foi excluída.

Referência do mapeamento de configuração

Para fins de referência, a tabela a seguir lista todas as propriedades de nível superior no [AWS::AutoScaling::LaunchConfiguration](#) recurso com suas propriedades correspondentes no [AWS::EC2::LaunchTemplate](#) recurso.

Iniciar a propriedade da fonte de configuração	Propriedade de destino do modelo
AssociatePublicIpAddress	NetworkInterfaces.AssociatePublicIpAddress
BlockDeviceMappings	BlockDeviceMappings
ClassicLinkVPCId	Não disponível
ClassicLinkVPCSecurityGroups	Não disponível
EbsOptimized	EbsOptimized
IamInstanceProfile	IamInstanceProfile.ArnOuIamInstanceProfile.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	SecurityGroupIdsOuNetworkInterfaces.Groups, mas não ambos
SpotPrice	InstanceMarketOptions.SpotOptions.MaxPrice
UserData	UserData

¹ As ClassicLinkVPCSecurityGroups propriedades ClassicLinkVPCId e não estão disponíveis para uso em um modelo de lançamento. Se sua configuração de execução tiver essas propriedades, você poderá usar essa oportunidade para migrar do EC2-Classic para a VPC. Para obter mais informações,

consulte [Migrar do EC2-Classic para uma VPC](#) no Guia do usuário do Amazon EC2 para instâncias do Linux.

Use o AWS Compute Optimizer para obter recomendações para o tipo de instância para um grupo do Auto Scaling

A AWS fornece recomendações de instâncias do Amazon EC2 para ajudar a melhorar a performance, economizar dinheiro ou ambos, usando recursos oferecidos pelo AWS Compute Optimizer. É possível usar essas recomendações para decidir se deseja passar para um novo tipo de instância.

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 [Getting Started with AWS Compute Optimizer](#) (Conceitos básicos do AWS Compute Optimizer) no AWS Compute Optimizer User Guide (Manual do usuário do AWS Compute Optimizer).

Índice

- [Limitações \(p. 367\)](#)
- [Descobertas \(p. 367\)](#)
- [Exibir recomendações \(p. 368\)](#)
- [Considerações para avaliação das recomendações \(p. 368\)](#)

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 oferecidas pelo processadores AWS Graviton2 (por exemplo, C6g) e para tipos de instância -n que têm maior performance 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.

Descobertas

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 \(p. 367\)](#).

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 informações recentesCloudWatchdados métricos (utilização média da CPU, média de entrada de rede e média de saída de rede) para o grupo de 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 \(p. 292\)](#).

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 ter mais informações, consulte [Substituir instâncias do Auto Scaling com base na vida útil máxima da instância \(p. 135\)](#) e [Substituir instâncias do Auto Scaling com base em uma atualização de instância \(p. 108\)](#).

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. Também é possível exibir dados de métricas (média, pico, percentil) no CloudWatch

para aprofundar a avaliação de suas recomendações de instâncias 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 na Amazônia CloudWatch Guia do usuário](#).

- O Compute Optimizer pode fornecer recomendações para instâncias expansíveis, que são as instâncias T3, T3a e T2. Se você ultrapassa periodicamente a linha de base, verifique se poderá continuar a fazer isso com base nas vCPUs do novo tipo de instância. Para obter mais informações, consulte [Créditos de CPU e performance de linha de base para instâncias expansíveis](#) no Manual do usuário do Amazon EC2 para instâncias do Linux.
- 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 Manual do usuário do Amazon EC2 para instâncias do Linux.
- 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

Além dos tópicos desta página, consulte os seguintes recursos:

- [Tipos de instância do Amazon EC2](#)
- [Manual do usuário do AWS Compute Optimizer](#)

Usar o Elastic Load Balancing para distribuir tráfego entre as instâncias no 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 \(p. 372\)](#). 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 para notificá-lo se a contagem de hosts íntegros do grupo de destino for menor do que o permitido.

Índice

- [Tipos de Elastic Load Balancing \(p. 370\)](#)
- [Pré-requisitos para começar a usar o Elastic Load Balancing \(p. 371\)](#)
- [Anexar um balanceador de carga ao grupo do Auto Scaling \(p. 372\)](#)
- [Configurar um Application Load Balancer ou Network Load Balancer pelo console do Amazon EC2 Auto Scaling \(p. 374\)](#)
- [Verificar o status do anexo do balanceador de carga \(p. 375\)](#)
- [Adicionar verificações de integridade do Elastic Load Balancing a um grupo do Auto Scaling \(p. 376\)](#)
- [Adicionar e remover zonas de disponibilidade \(p. 377\)](#)
- [Exemplos para trabalhar com Elastic Load Balancing com a AWS Command Line Interface \(AWS CLI\) \(p. 379\)](#)
- [Tutorial: Configurar uma aplicação escalonada e com平衡amento de carga \(p. 385\)](#)

Tipos de Elastic Load Balancing

O Elastic Load Balancing oferece quatro tipos de平衡adores de carga que podem ser usados com seu grupo do Auto Scaling:平衡adores de carga de aplicação,平衡adores de carga de rede,平衡adores de carga de gateway e平衡adores de carga clássicos

Há uma diferença fundamental em como os tipos de平衡adores de carga são configurados. Com os平衡adores de carga de aplicação,平衡adores de carga de rede e平衡adores 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平衡adores de carga clássicos, as instâncias são registradas diretamente no平衡ador de carga.

Application Load Balancer

Roteia e faz平衡amento 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平衡amento de carga na camada de transporte (camada 4 do TCP/UDP) com base nas informações de endereço extraídas do cabeçalho da camada 4. Os平衡adores 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平衡ador 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平衡adores 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平衡ador de carga de gateway.

Classic Load Balancer

Roteia e faz平衡amento de carga na camada de transporte (TCP/SSL) ou na camada da aplicação (HTTP/HTTPS).

Para saber mais sobre Elastic Load Balancing, consulte os seguintes tópicos:

- [O que é Elastic Load Balancing?](#)
- [O que é um Application Load Balancer?](#)
- [O que é um Network Load Balancer?](#)
- [O que é um平衡ador de carga de gateway?](#)
- [O que é um Classic Load Balancer?](#)

Pré-requisitos para começar a usar o Elastic Load Balancing

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.

Você pode criar o balanceador de carga antes de criar o grupo Auto Scaling ou durante a criação do grupo Auto Scaling.

- Siga os procedimentos na documentação do Elastic Load Balancing para criar o balanceador de carga antes de criar o grupo do Auto Scaling. Ignore a etapa para registrar suas instâncias do Amazon EC2. O Amazon EC2 Auto Scaling cuida automaticamente do registro (e cancelamento de registro) de instâncias. Para obter mais informações, consulte [Conceitos básicos do Elastic Load Balancing](#) no Manual do usuário do Elastic Load Balancing.
- Como alternativa, você pode criar e anexar um Application Load Balancer ou Network Load Balancer com uma configuração básica do console do Amazon EC2 Auto Scaling. Para obter mais informações, consulte [Configurar um Application Load Balancer ou Network Load Balancer pelo console do Amazon EC2 Auto Scaling \(p. 374\)](#).

Para anexar um balanceador de carga ao seu grupo do Auto Scaling, primeiro verifique se você atendeu aos seguintes pré-requisitos:

- O balanceador de carga e seu grupo de destino devem estar na mesma Conta da AWS, VPC e região que o grupo do 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 o modelo de execução não contiver um grupo de segurança que permita que o tráfego do balanceador de carga alcance o grupo do Auto Scaling, você deverá 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.

Note

As instâncias existentes não são atualizadas com as novas configurações quando o modelo de execução é modificado. Para atualizar instâncias existentes, você pode encerrar instâncias existentes no grupo do Auto Scaling. O Amazon EC2 Auto Scaling inicia imediatamente novas instâncias para substituir as instâncias que você terminou. Como alternativa, você pode iniciar uma atualização de instância para substituir as instâncias. Para obter mais informações, consulte [Substituir instâncias do Auto Scaling com base em uma atualização de instância \(p. 108\)](#).

- 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.

Antes de permitir que seu grupo do Auto Scaling use os resultados das verificações de integridade do Elastic Load Balancing para determinar a integridade de uma instância, certifique-se de ter cumprido esse pré-requisito adicional:

- Os grupos de segurança que você especifica no modelo de execução devem permitir o acesso do balanceador de carga na porta correta para que o Elastic Load Balancing execute suas verificações de integridade. Para obter mais informações, consulte [Adicionar verificações de integridade do Elastic Load Balancing a um grupo do Auto Scaling \(p. 376\)](#).

Tip

Se você tiver scripts de bootstrap que demoram um pouco 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 criar o grupo. Para obter mais informações, consulte [Ganchos do ciclo de vida do Amazon EC2 Auto Scaling \(p. 252\)](#).

Anexar um balanceador de carga ao grupo do Auto Scaling

Este tópico descreve como anexar um balanceador de carga Elastic Load Balancing ao grupo do Auto Scaling. O Amazon EC2 Auto Scaling integra-se ao Elastic Load Balancing para ajudar você a inserir um Application Load Balancer, Network Load Balancer, Classic Load Balancer ou balanceador de carga de gateway na frente do seu grupo do Auto Scaling. Para saber mais sobre os diferentes tipos de平衡adores de carga, consulte [Tipos de Elastic Load Balancing \(p. 370\)](#).

Ao anexar um Application Load Balancer, Network Load Balancer ou balanceador de carga de gateway, você anexa um grupo de destino. O Amazon EC2 Auto Scaling adiciona instâncias ao grupo de destino anexado quando elas são iniciadas. Você pode anexar um ou vários grupos de destino e configurar verificações de integridade por grupo de destino.

Para obter um guia introdutório para anexar um grupo-alvo ao seu grupo do Auto Scaling, consulte [Tutorial: Configurar uma aplicação escalonada e com平衡amento de carga \(p. 385\)](#).

Important

Antes de continuar, preencha todos os [pré-requisitos \(p. 371\)](#) na seção anterior.

Índice

- [Anexar um balanceador de carga \(p. 372\)](#)
- [Desvincular um balanceador de carga \(p. 373\)](#)

Anexar um balanceador de carga

Use os procedimentos a seguir para associar um balanceador de carga ao grupo do Auto Scaling. Você pode especificar o balancer ao escolher um grupo de destino que você criou para um Application Load Balancer, Network Load Balancer ou balancer ou balancer Classic Load Balancer.

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 grupos do Auto Scaling 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 existing load balancer (Anexar a um balanceador de carga existente).
5. Em Attach to an existing load balancer (Anexar a um balanceador de carga existente), siga um destes procedimentos:
 - a. Para平衡adores de carga de aplicação,平衡adores de carga de rede e平衡adores 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平衡adores de carga clássicos:

Escolha Choose from Classic Load Balancers (Escolher entre平衡adores de carga clássicos) e, em seguida, escolha seu balanceador de carga no campo Classic Load Balancers (Balanceadores de carga clássicos).
6. 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 grupo do Auto Scaling após sua criação

1. Abra o console do Amazon EC2 em <https://console.aws.amazon.com/ec2/> e escolha grupos do Auto Scaling 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 Detalhes, 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平衡adores 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. Escolha Update (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 do balanceador de carga \(p. 375\)](#).

Desvincular um balanceador de carga

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 grupos do Auto Scaling 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. Escolha Update (Atualizar).

Configurar um Application Load Balancer ou Network Load Balancer pelo console do Amazon EC2 Auto Scaling

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 grupos do Auto Scaling 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. 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平衡adores de carga existentes, desde que use portas distintas.

Verificar o status do anexo do balanceador de carga

Depois de anexar um balanceador de carga, ele entra no `noAdding` estado ao registrar as instâncias no grupo. Quando todas as instâncias do grupo são registradas, ele entra no `noAdded` estado. 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 após o cancelamento do registro. 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 os SDKs. Você não pode verificar o status do anexo no console.

Para usar o AWS CLI para verificar o status do anexo

O [comando `describe-traffic-sources`](#) comando a seguir retorna o status do anexo de todas as fontes de tráfego para o 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 anexado ao grupo Auto Scaling, junto com o status da anexação do grupo-alvo no `State` elemento.

```
{  
    "TrafficSources": [  
        {  
            "Identifier": "arn:aws:elasticloadbalancing:region:account-id:targetgroup/my-targets/1234567890123456",  
            "State": "InService",  
            "Type": "elbv2"  
        }  
    ]  
}
```

Adicionar verificações de integridade do Elastic Load Balancing a um grupo do Auto Scaling

As verificações de integridade padrão para um grupo do Auto Scaling são somente verificações de integridade do EC2. Se uma instância falhar nessas verificações de integridade, ela será marcada como não íntegra e terminada. Ao mesmo tempo, o Amazon EC2 Auto Scaling iniciará uma nova instância de substituição. Para obter mais informações, consulte [Verificações de integridade para instâncias do Auto Scaling \(p. 319\)](#).

É possível anexar um ou mais grupos de destino de balanceador de carga, um ou mais平衡adores de carga clássicos ou ambos ao seu grupo do Auto Scaling. No entanto, por padrão, o grupo do Auto Scaling não considerará uma instância não íntegra e a substituirá se ela apresentar falha nas verificações de integridade do Elastic Load Balancing.

Para garantir que o grupo do Auto Scaling possa determinar a integridade da instância com base em testes adicionais do balanceador de carga, é possível configurar o grupo do Auto Scaling para usar verificações de integridade do Elastic Load Balancing (ELB). O balanceador de carga envia periodicamente pings, realiza tentativas de conexão ou envia solicitações para testar as instâncias do EC2 e determinar se uma instância não está íntegra. Se você configurar o grupo do Auto Scaling para usar as verificações de integridade do Elastic Load Balancing, ele considerará a instância não íntegra se ela não passar 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平衡adores 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.

Adicionar verificações de integridade do Elastic Load Balancing

Para adicionar verificações de integridade do Elastic Load Balancing usando o console do Amazon EC2 Auto Scaling, execute as seguintes etapas.

Para adicionar verificações de integridade do Elastic Load Balancing a um novo grupo

Ao criar o grupo Auto Scaling, na página Configurar opções avançadas, para Verificações de integridade, Tipos adicionais de verificação de Health, selecione Ativar verificações de saúde do Elastic Load Balancing. Em seguida, para o período de carência da verificação de Health, 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 \(p. 325\)](#).

Para adicionar verificações de integridade do Elastic Load Balancing a um grupo existente

1. Abra o console do Amazon EC2 em <https://console.aws.amazon.com/ec2/> e escolha grupos do Auto Scaling 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. Para Verificações de integridade, Tipos adicionais de verificação de Health, selecione Ativar verificações de integridade do Elastic Load Balancing.
6. Para o período de carência da verificação de Health, insira a quantia 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 \(p. 325\)](#).

7. Escolha Update (Atualizar).
8. Na guia Instance management (Gerenciamento de instâncias), em Instances (Instâncias), é possível visualizar o status de integridade de instâncias. A coluna Health Status (Status de integridade) exibe os resultados das verificações de integridade recém-adicionadas.

Consulte também

- Para configurar verificações de integridade para seu Application Load Balancer, consulte [Verificações de integridade de grupos de destino](#) no Manual do usuário de平衡adores de carga de aplicação.
- Para configurar verificações de integridade para seu Network Load Balancer, consulte [Verificações de integridade de grupos de destino](#) no Manual do usuário de Network Load Balancers.
- Para configurar verificações de integridade para seu balanceador de carga de gateway, consulte [Verificações de integridade de grupos de destino](#) no Manual do usuário de balanceadores de carga de gateway.
- Para configurar verificações de integridade para seu Classic Load Balancer, consulte [Configurar as verificações de integridade do seu Classic Load Balancer](#) no Manual do usuário de Classic Load Balancers.

Adicionar e remover zonas 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.

Você deve 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).

Índice

- [Adicione uma Zona de disponibilidade \(p. 378\)](#)
- [Remover uma Zona de disponibilidade \(p. 378\)](#)
- [Limitações \(p. 379\)](#)

Adicione uma Zona de disponibilidade

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 grupos do Auto Scaling 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. Escolha Update (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. Faça um dos seguintes procedimentos:
 - Para平衡adores de carga de aplicação e平衡adores 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平衡adores 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 Save (Salvar).

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 grupos do Auto Scaling 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. Escolha Update (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. Faça um dos seguintes procedimentos:
 - Para平衡adores de carga de aplicação e平衡adores 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), desmarque a caixa de seleção para remover a sub-rede da zona de disponibilidade selecionada.
 - Para平衡adores de carga clássicos em uma VPC:
 1. Na guia Instances, 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 Save (Salvar).

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平衡adores de carga voltados para a Internet, as sub-redes especificadas para o balanceador de carga devem ter pelo menos oito endereços IP disponíveis.
- Para平衡adores de carga de aplicação, é necessário habilitar pelo menos duas zonas de disponibilidade.
- Para平衡adores de carga de rede, você não pode desabilitar as zonas de disponibilidade habilitadas, mas pode habilitar zonas adicionais.
- Para os平衡adores de carga de gateway, não é possível alterar as zonas de disponibilidade ou sub-redes que foram adicionadas quando o balanceador de carga foi criado.

Exemplos para trabalhar com Elastic Load Balancing com a AWS Command Line Interface (AWS CLI)

Use oAWS CLI para conectar, separar e descrever平衡adores de carga e grupos-alvo, adicionar e remover verificações de integridade do Elastic Load Balancing e alterar quais zonas de disponibilidade estão habilitadas.

Este tópico mostra exemplos de AWS CLI comandos da que executam tarefas comuns para o Amazon EC2 Auto Scaling.

Important

Para obter exemplos adicionais de comandos, consulte [aws elbv2](#) e [aws elbna](#) Referência de AWS CLI Comandos.

Índice

- [Conecte seu grupo-alvo ou o Classic Load Balancer \(p. 380\)](#)
- [Descreva seus grupos-alvo ou平衡adores de carga clássicos \(p. 380\)](#)
- [Adicionar verificações de integridade do Elastic Load Balancing \(p. 381\)](#)
- [Alterar suas zonas de disponibilidade \(p. 381\)](#)
- [Separar seu grupo-alvo ou o Classic Load Balancer \(p. 383\)](#)
- [Remover as verificações de integridade do Elastic Load Balancing \(p. 383\)](#)
- [Comandos herdados \(p. 383\)](#)

Conecte seu grupo-alvo ou o Classic Load Balancer

Use o [create-auto-scaling-group](#) comando a seguir para criar um grupo de Auto Scaling e simultaneamente anexar um grupo-alvo especificando seu nome de recurso da Amazon (ARN). O grupo de destino pode ser associado a um Application Load Balancer, um Network Load Balancer ou um Gateway Load Balancer.

Substitua os valores da amostra por `--auto-scaling-group-name`, `--vpc-zone-identifier`, `--min-size`, `--max-size` e `--launch-template`. Como `--launch-template` opção, substitua `my-launch-template` pelo nome e versão de um modelo de execução para seu grupo de Auto Scaling. Como `--traffic-sources` opção, substitua o ARN de exemplo pelo ARN de um grupo de destino para um Application Load Balancer, Network Load Balancer ou Load Balancer 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-sources](#) comando para anexar grupos-alvo adicionais 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 `--type` opções `--traffic-sources` e 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 `--type` opção, especifique um valor de `elb`.

```
--traffic-sources "Identifier=my-classic-load-balancer" --type elb
```

Descreva seus grupos-alvo ou balanceadores de carga clássicos

Para descrever os平衡adores de carga ou grupos-alvo vinculados ao seu grupo de Auto Scaling, use o [describe-traffic-sources](#) comando a seguir. 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 do balanceador de carga \(p. 375\)](#).

Adicionar verificações de integridade do Elastic Load Balancing

Para adicionar verificações de integridade do Elastic Load Balancing às verificações de integridade que seu grupo do Auto Scaling executa nas instâncias, use o [update-auto-scaling-group](#) comando a seguir e especifique **ELB** o valor para a **--health-check-type** opção. **my-asg** Substitua 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 **--health-check-grace-period** opção ao usar [update-auto-scaling-group](#), como no exemplo a seguir. Substitua **300** pelo número de segundos para manter as novas instâncias em serviço antes de encerrá-las se elas não estiverem íntegras.

```
--health-check-grace-period 300
```

Para obter mais informações, consulte [Verificações de integridade para instâncias do Auto Scaling \(p. 319\)](#).

Alterar suas zonas de disponibilidade

Alterar suas zonas de disponibilidade apresenta algumas limitações das quais você deve estar ciente. Para obter mais informações, consulte [Limitações \(p. 379\)](#).

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 Auto Scaling para verificar se há disponibilidade para seus tipos de instância nas zonas especificadas.

Para atualizar as zonas de disponibilidade do seu grupo do Auto Scaling, use o [update-auto-scaling-group](#) comando a seguir. Substitua os IDs de sub-rede de destino pelos IDs das sub-redes nas zonas de disponibilidade para ativar. As sub-redes especificadas substituem as sub-redes habilitadas anteriormente. *my-asg* Substitua 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-groups](#) comando 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. *my-asg* Substitua 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 os IDs de sub-rede de destino pelos IDs das sub-redes nas zonas de disponibilidade para ativar. Você pode especificar somente uma sub-rede por Zona de disponibilidade. As sub-redes especificadas substituem as sub-redes habilitadas anteriormente. *my-lb-arn* Substitua 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 Auto Scaling para verificar se há disponibilidade para seus tipos de instância nas zonas especificadas.

Para atualizar as zonas de disponibilidade do seu grupo do Auto Scaling, use o [update-auto-scaling-group](#) comando a seguir. Substitua os IDs de sub-rede de destino pelos IDs das sub-redes nas zonas de disponibilidade para ativar. As sub-redes especificadas substituem as sub-redes habilitadas anteriormente. *my-asg* Substitua 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-groups](#) comando 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. *my-asg* Substitua 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 ativar uma nova zona de disponibilidade para o Classic Load Balancer. Substitua a ID da amostra pela ID da sub-rede para que a zona de disponibilidade seja ativada. *my-lb* Substitua pelo nome do balanceador de carga.

```
aws elb attach-load-balancer-to-subnets --load-balancer-name my-lb \  
--subnets subnet-41767929 subnet-cb663da2 subnet-8360a9e7
```

```
--subnets subnet-cb663da2
```

Para desativar uma zona de disponibilidade, use o seguinte comando [detach-load-balancer-from-subnets](#). Substitua o ID da sub-rede de amostra pelo ID da sub-rede para que a zona de disponibilidade seja desativada. *my-lb* Substitua pelo nome do balanceador de carga.

```
aws elb detach-load-balancer-from-subnets --load-balancer-name my-lb \
--subnets subnet-8360a9e7
```

Separe seu grupo-alvo ou o Classic Load Balancer

O [detach-traffic-sources](#) comando a seguir desvincula um grupo de destino do seu grupo do Auto Scaling quando ele não é mais necessário.

Para a `--auto-scaling-group-name` opção, *my-asg* substitua pelo nome do seu grupo. Como `--traffic-sources` opção, substitua o ARN de exemplo pelo ARN de um grupo de destino para um Application Load Balancer, Network Load Balancer ou Load Balancer 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 separar um Classic Load Balancer do seu grupo, especifique `--type` opções `--traffic-sources` e, como no exemplo a seguir. *my-classic-load-balancer* Substitua pelo nome de um Classic Load Balancer. Para a `--type` opção, 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 integridade do Elastic Load Balancing do seu grupo de Auto Scaling, use o [update-auto-scaling-group](#) comando a seguir e especifique `EC2` como o valor da `--health-check-type` opção. *my-asg* Substitua 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 do Auto Scaling \(p. 319\)](#).

Comandos herdados

Os exemplos a seguir mostram como você pode usar comandos da CLI herdados para anexar, desvincular e descrever平衡adores 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 CLI de “fontes de tráfego”, que podem anexar e desanexar vários tipos de fontes de tráfego. Você pode usar os comandos da CLI legados e os comandos CLI “fontes de tráfego” no mesmo grupo do Auto Scaling.

Conecte seu grupo-alvo ou o Classic Load Balancer (legado)

Para anexar seu grupo de destino

O [create-auto-scaling-group](#) comando a seguir cria um grupo do Auto Scaling com um grupo de destino 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 de destino a um grupo do Auto Scaling existente.

```
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-group](#) comando a seguir cria um grupo do 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-balancers](#) comando a seguir anexa o Classic Load Balancer especificado a um grupo do Auto Scaling existente.

```
aws autoscaling attach-load-balancers --auto-scaling-group-name my-asg \
--load-balancer-names my-lb
```

Descreva seu grupo-alvo ou o Classic Load Balancer (legado)

Para descrever grupos de destino

Para descrever os grupos de destino associados a um grupo do Auto Scaling, use o comando [describe-load-balancer-target-groups](#). O exemplo a seguir lista os grupos de destino para *my-asg*.

```
aws autoscaling describe-load-balancer-target-groups --auto-scaling-group-name my-asg
```

Para descrever Classic Load Balancers

Para descrever os平衡adores de carga clássicos associados a um grupo do Auto Scaling, use o [describe-load-balancers](#) comando. O exemplo a seguir lista os平衡adores de carga clássicos para *my-asg*.

```
aws autoscaling describe-load-balancers --auto-scaling-group-name my-asg
```

Separe seu grupo-alvo ou o Classic Load Balancer (legado)

Para desanexar um grupo de destino

O comando [detach-load-balancer-target-groups](#) a seguir desvincula um grupo de destino do seu grupo do Auto Scaling quando ele não é mais necessário.

```
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"
```

Para desvincular um Classic Load Balancer

O [detach-load-balancers](#) comando a seguir desvincula um Classic Load Balancer do seu grupo do Auto Scaling quando ele não é mais necessário.

```
aws autoscaling detach-load-balancers --auto-scaling-group-name my-asg \  
--load-balancer-names my-lb
```

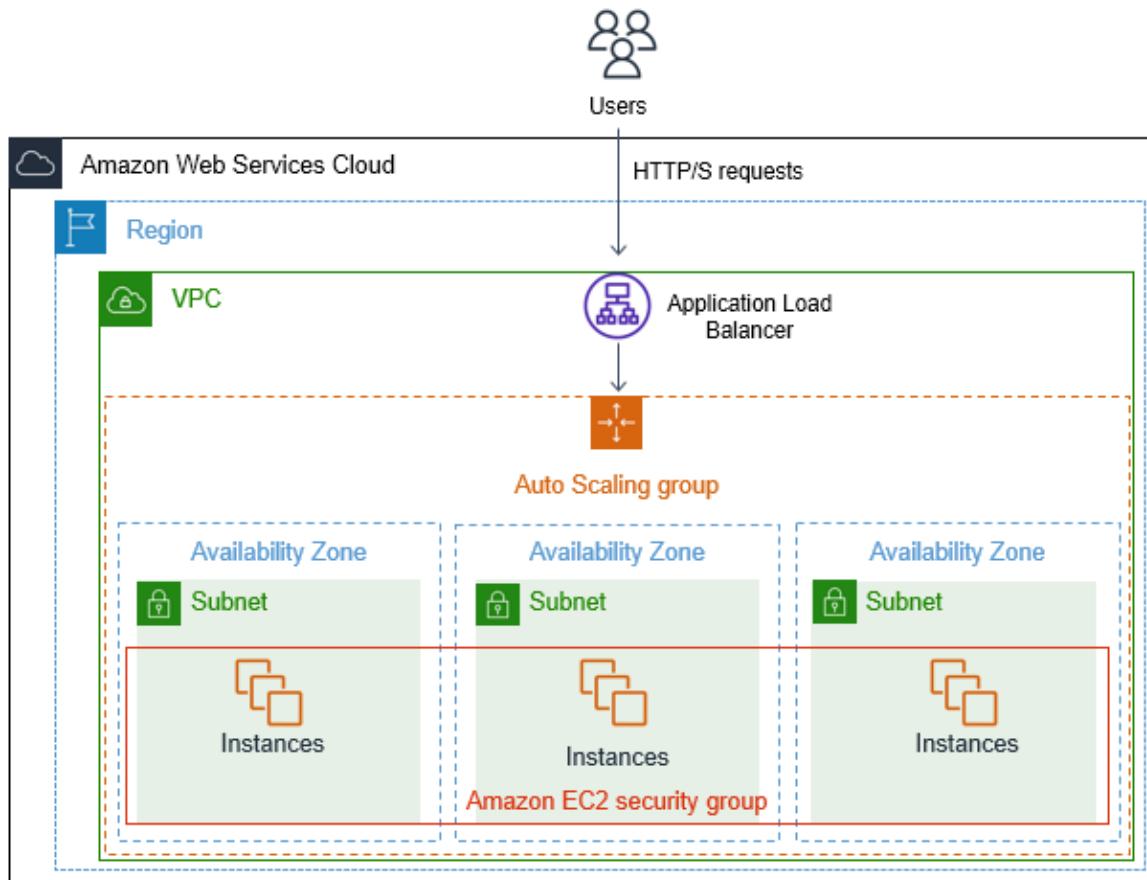
Tutorial: Configurar uma aplicação escalonada e com balanceamento de carga

Important

Antes de explorar este tutorial, recomendamos que você primeiramente examine o tutorial introdutório: [Conceitos básicos do Amazon EC2 Auto Scaling \(p. 15\)](#).

O registro do seu grupo do Auto Scaling em um平衡eador 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平衡adores de carga. Recomendamos que você use um Application Load Balancer para este tutorial.

Para obter mais informações sobre como introduzir um平衡ador de carga em sua arquitetura, consulte [Usar o Elastic Load Balancing para distribuir tráfego entre as instâncias no grupo do Auto Scaling \(p. 369\)](#).

Tarefas

- [Pré-requisitos \(p. 386\)](#)
- [Etapa 1: Configurar um modelo de execução ou uma configuração de execução \(p. 387\)](#)
- [Etapa 2: Criar um grupo do Auto Scaling \(p. 389\)](#)
- [Etapa 3: Verificar se o平衡ador de carga está anexado \(p. 390\)](#)
- [Etapa 4: Próximas etapas \(p. 391\)](#)
- [Etapa 5: Limpar \(p. 391\)](#)
- [Recursos relacionados \(p. 392\)](#)

Pré-requisitos

- Um平衡ador de carga e grupo de destino. Certifique-se de escolher as mesmas zonas de disponibilidade para o平衡ador 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 conceda à sua aplicação acesso a AWS.
- (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 a função do IAM (se tiver criado uma) 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.

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

Note

Como alternativa, é possível usar uma configuração de execução em vez de um modelo de execução. Para obter as instruções de configuração de execução, consulte [Criar ou selecionar uma configuração de execução \(p. 388\)](#).

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 Create launch template (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 AMIs disponíveis ou selecionar uma AMI na lista Recents (Recentes) ou Quick Start (Início rápido). Caso não veja a AMI de que precisa, escolha Browser more AMIs (Pesquisar mais AMIs) para navegar pelo catálogo completo de AMIs.
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. (Opcional) Em Atribuir IP público automaticamente, mantenha o valor padrão, Não incluir no modelo de execução. Ao criar o grupo do Auto Scaling, é possível atribuir um endereço IPv4 público a instâncias no grupo do Auto Scaling usando sub-redes que têm o atributo de endereçamento IPv4 público habilitado, como as sub-redes padrão na VPC padrão. 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), digite o nome de recurso da Amazon (ARN) da sua função 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 Create launch template (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

Se você já tiver uma configuração de ativação que gostaria de usar, selecione-a usando o procedimento a seguir.

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 em que 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 solicitada a confirmação, escolha **Exibir configurações de lançamento** para confirmar que você deseja visualizar o [Configurações de lançamento](#) na página.
2. Na barra de navegação superior, escolha a região em que o balanceador de carga foi criado.
3. Selecione [Create launch configuration](#) (Criar uma configuração de execução), e insira um nome para sua configuração de execução.
4. Em [Amazon machine image \(AMI\)](#) (Imagen 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 EC2, em [IAM instance profile](#) (Perfil da instância do IAM), escolha sua função do IAM. Para obter mais informações, consulte [Funções do IAM para aplicações que são executadas em instâncias do Amazon EC2 \(p. 448\)](#).
 - 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 Next (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 \(p. 417\)](#).
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 a 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平衡adores 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 usar as verificações de integridade do Elastic Load Balancing, em Health checks (Verificações de integridade), escolha ELB em Health check type (Tipo de verificação de integridade).
10. Quando terminar de configurar o grupo do Auto Scaling, escolha Skip to review (Pular para revisão).
11. 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. Parastep-by-stepinstruções, consulte [Escalabilidade manual \(p. 168\)](#).

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 \(p. 459\)](#).
4. Na guia Instance management (Gerenciamento de instâncias), em Instances (Instâncias), é possível verificar se 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:

- Você pode configurar seu grupo do Auto Scaling para usar as verificações de integridade do Elastic Load Balancing. 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 [Adicionar verificações de integridade do Elastic Load Balancing \(p. 376\)](#).
- 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 [Adicionar zonas de disponibilidade \(p. 377\)](#).
- 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 \(p. 180\)](#).

Etapa 5: Limpar

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

Para excluir seu grupo do Auto Scaling

1. Abra o [Página de grupos do Auto Scaling](#) do console Amazon EC2.
2. Marque a caixa de seleção ao lado do seu grupo do Auto Scaling.
3. Escolha Delete (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 Desired (Desejado), Min (Mínimo) e Max

(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 Delete (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 Yes, Delete (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 Yes, Delete (Sim, excluir).

Recursos relacionados

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

Para ver um passo a passo que usa um modelo de pilha para provisionar um grupo de Auto Scaling e um Application Load Balancer, consulte [Passo a passo: criar um aplicativo dimensionado e com balanceamento de carga na AWS CloudFormation](#). Use o passo a passo e o modelo de amostra como ponto de partida para criar modelos semelhantes que atendam às suas necessidades.

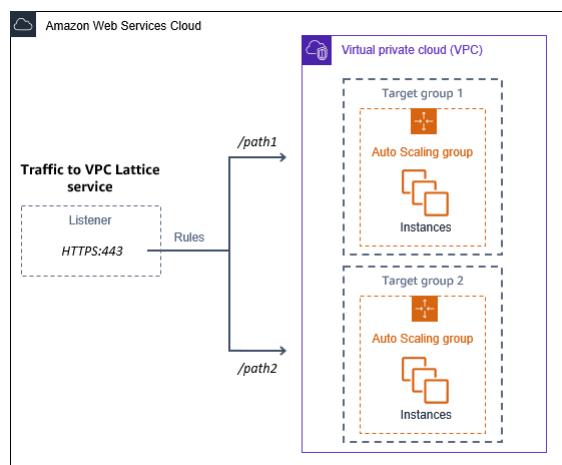
roteie o tráfego para o grupo do Auto Scaling

Você pode usar o Amazon VPC Lattice para gerenciar o fluxo de tráfego e chamadas de API entre seus aplicativos e serviços executados em recursos separados, como grupos de 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 uns aos outros. Esses recursos incluem os serviços, os listeners, as regras de listener e os grupos de destino.

Para associar um grupo do Auto Scaling a um serviço VPC Lattice, crie um grupo de destino para o serviço que roteie as solicitações para instâncias registradas por ID da instância e adicione um listener ao serviço que envie solicitações para o grupo de destino. Em seguida, vincular o grupo de destino ao Auto Scaling. O Amazon EC2 Auto Scaling registra automaticamente as instâncias do EC2 como destino com o grupo de destino. Posteriormente, quando o Amazon EC2 Auto Scaling precisa encerrar uma instância, ele automaticamente cancela o registro da instância do grupo-alvo antes do encerramento.

Depois de anexar o grupo-alvo, ele é o ponto de entrada para todas as solicitações recebidas ao seu grupo de Auto Scaling. Como mostra o exemplo no diagrama a seguir, as solicitações recebidas podem então ser roteadas para o grupo-alvo apropriado usando as regras de escuta especificadas para um serviço VPC Lattice.



Quando o tráfego é roteado pelo VPC Lattice para seu grupo de Auto Scaling, o VPC Lattice equilibra as solicitações entre as instâncias do grupo usando o balanceamento de carga round robin. O VPC Lattice também pode monitorar a integridade de suas instâncias registradas e direcionar o tráfego somente para instâncias saudáveis.

Para manter suas instâncias disponíveis para solicitações recebidas, opcionalmente, você pode adicionar verificações de integridade do VPC Lattice ao seu grupo de Auto Scaling. Dessa forma, se uma das instâncias do EC2 falhar, seu 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 conectividade, a segurança e o monitoramento entre serviços com o Amazon VPC Lattice — agora disponível](#) ao público no AWS blog.

Índice

- [Preparar para anexar um grupo de destino do VPC Lattice \(p. 394\)](#)

- [Anexar um grupo de destino do VPC Lattice \(p. 396\)](#)
- [Verifique o status do anexo do grupo de destino do VPC Lattice \(p. 399\)](#)

Preparar para anexar um grupo de destino do VPC Lattice

Antes de anexar um grupo de destino do VPC Lattice ao Auto Scaling

- Você já deve ter criado uma rede de serviços, serviço, ouvinte e grupo-alvo do VPC Lattice. Para obter mais informações, consulte um dos tópicos a seguir no Guia do usuário do VPC Lattice.
 - [Redes de serviços](#)
 - [Serviços](#)
 - [Listeners](#)
 - [Grupos-alvo](#)
- O grupo de destino deve estar na mesma conta da AWS, VPC e região que o grupo do 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 de 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-alvo.

```
{  
    "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 de Auto Scaling não contiver as configurações corretas para o VPC Lattice, como um grupo de segurança compatível, você deverá atualizar o modelo de execução. As instâncias existentes não são atualizadas com as novas configurações quando o modelo de execução é modificado. Para atualizar instâncias existentes, você pode encerrar as instâncias existentes no grupo do Auto Scaling. O Amazon EC2 Auto Scaling inicia imediatamente novas instâncias para substituir as instâncias que você terminou. Como alternativa, você pode iniciar uma atualização de instância para substituir as instâncias. Para obter mais informações, consulte [Substituir instâncias do Auto Scaling com base em uma atualização de instância \(p. 108\)](#).
- Antes de ativar as verificações de integridade do VPC Lattice em seu grupo de Auto Scaling, você pode configurar uma verificação de integridade baseada em aplicativos para verificar se seu aplicativo está respondendo conforme o esperado. Para obter mais informações, consulte [Verificações de Health para seus grupos-alvo](#) no Guia do usuário do VPC Lattice.

Grupos de segurança: regras de entrada e saída

Os security groups atuam como firewall para instâncias associadas do EC2, controlando o tráfego de entrada e de saída no nível da instância.

Note

A configuração da rede é suficientemente complexa para que seja altamente recomendável que você crie um novo grupo de segurança para uso com o VPC Lattice. Também torna mais fácil ajudá-loAWS Support se você precisar entrar em contato com eles. As seções a seguir são baseadas 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 de 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 informações sobre como criar um grupo de segurança, consulte [Criar um grupo de segurança](#) no Guia do Usuário do Amazon EC2 para instâncias Linux e use a tabela a seguir para determinar quais opções selecionar.

Opção	Value (Valor)
Name (Nome)	Um nome fácil de lembrar.
Descrição	Uma descrição para ajudar 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 dentro de uma rede de serviço VPC Lattice é permitido até que você adicione regras de entrada ao security group.

Para permitir que clientes em uma rede de serviços VPC Lattice se conectem a instâncias em seu grupo de Auto Scaling, o grupo de segurança do seu grupo de Auto Scaling deve estar configurado corretamente. Nesse caso, defina uma regra de entrada para permitir tráfego a partir do nome da lista de prefixosAWS gerenciados para o VPC Lattice, em vez de um endereço IP específico. A lista de prefixos do VPC Lattice é um intervalo de endereços IP usados pelo VPC Lattice na notação CIDR. Para obter mais informações, consulte [Trabalhar com listasAWS de prefixos gerenciados](#) da no Guia do usuário da Amazon VPC.

Para obter informações sobre como adicionar regras a um grupo de segurança, consulte [Adicionar regras ao grupo de segurança](#) no Guia do usuário da Amazon VPC e use a tabela a seguir para determinar quais opções selecionar.

Opção	Value (Valor)
Regra HTTP	Type (Tipo): HTTP Fonte: com.amazonaws. <i>região</i> .vpc-lattice
Regra HTTPS	Tipo: HTTPS Fonte: com.amazonaws. <i>região</i> .vpc-lattice

O grupo de segurança é do tipo com estado: ele permite o tráfego de clientes da rede de serviço VPC Lattice para instâncias em seu grupo do Auto Scaling e envia a resposta de volta para o cliente que ela deixou 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 para [grupos de instâncias mistas \(p. 67\)](#). Se você tentar anexar um grupo-alvo do VPC Lattice a um grupo de Auto Scaling que tenha uma política de instâncias mistas, receberá a mensagem de erro Atualmente, os grupos de Auto Scaling com instâncias mistas não podem ser integrados a um serviço do VPC Lattice. Isso ocorre porque o algoritmo de平衡amento de carga distribui uniformemente a carga em todos os recursos disponíveis e assume que as instâncias são semelhantes o suficiente para lidar com cargas iguais.

Anexar um grupo de destino do VPC Lattice

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 as instâncias que o VPC Lattice relata como não saudáveis.

Por padrão, o Amazon EC2 Auto Scaling substitui somente instâncias que não são íntegras ou não podem ser acessadas 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-alvo do VPC Lattice que você anexar ao grupo Auto Scaling denunciar que ela não está íntegra. Para obter mais informações, consulte [Verificações de integridade para instâncias do Auto Scaling \(p. 319\)](#).

Important

Antes de continuar, preencha todos os [pré-requisitos \(p. 394\)](#) na seção anterior.

Anexar um grupo-alvo 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
- Ative as verificações de integridade do VPC Lattice

Para anexar um grupo de destino do VPC Lattice

1. Abra o console do Amazon EC2 em <https://console.aws.amazon.com/ec2/> e escolha grupos do Auto Scaling no painel de navegação.
2. Na barra de navegação, na parte superior da tela, escolha a mesma naRegião da AWS qual você criou o grupo de destino.

3. Selecione Criar grupo do Auto Scaling.
4. Nas etapas 1 e 2, escolha as opções desejadas e prossiga para Etapa 3: Configurar opções avançadas.
5. Para opções de integração do VPC Lattice, escolha Attach to VPC Lattice service.
6. Em Escolher grupo alvo do VPC Lattice, escolha seu grupo-alvo.
7. (Opcional) Para verificações de integridade, tipos adicionais de verificação de Health, selecione Ativar verificações de integridade do VPC Lattice.
8. (Opcional) Para o período de carência da verificação de Health, 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 \(p. 325\)](#).
9. Prossiga para criar o grupo do Auto Scaling. Suas instâncias serão registradas automaticamente no grupo de destino do VPC Lattice após a criação do grupo do Auto Scaling.

Para anexar um grupo de destino do VPC Lattice

Siga o procedimento a seguir para associar um grupo de destino de um serviço a um grupo do Auto Scaling

1. Abra o console do Amazon EC2 em <https://console.aws.amazon.com/ec2/> e escolha grupos do Auto Scaling 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. Em Opções de integração do VPC Lattice, escolha Attach to VPC Lattice service.
5. Em Escolher grupo alvo do VPC Lattice, escolha seu grupo-alvo.
6. Escolha Update (Atualizar).

Ao terminar de anexar o grupo-alvo, você pode, opcionalmente, ativar as verificações de saúde que o usam.

Para ativar as verificações de integridade do VPC Lattice

1. Na guia Detalhes, escolha Verificações de integridade, Editar.
2. Para Verificações de Health, Tipos adicionais de verificação de integridade, selecione Ativar verificações de integridade do VPC Lattice.
3. Em Período de carência da verificação de Health, 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 \(p. 325\)](#).
4. Escolha Update (Atualizar).

AWS CLI

Siga as etapas nesta seção para usar a AWS CLI para:

- Anexar um grupo de destino do VPC Lattice
- Ative as verificações de integridade do VPC Lattice

Para anexar um grupo de destino do VPC Lattice

Use o [create-auto-scaling-group](#) comando 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 da amostra por --auto-scaling-group-name --vpc-zone-identifier--min-size, --max-size e. Como --launch-template opção, substitua *my-launch-template* e *1* pelo nome e versão do modelo de execução que você criou para instâncias registradas em um grupo-alvo do VPC Lattice. Como --traffic-sources opção, substitua o ARN de exemplo pelo ARN do 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-sources](#) comando a seguir para anexar um grupo-alvo do VPC Lattice a um grupo de Auto Scaling depois que ele já estiver 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 aplicativos para seu grupo-alvo do VPC Lattice, você pode ativar essas verificações de integridade. Use o [update-auto-scaling-group](#) comando [create-auto-scaling-group](#) ou com a --health-check-type opção e um valor de **VPC_LATTICE**. Para especificar o período de carência para as verificações de saúde realizadas pelo seu grupo do Auto Scaling, inclua a --health-check-grace-period opção e forneça seu valor em segundos.

```
--health-check-type "VPC_LATTICE" --health-check-grace-period 60
```

Desanexar um grupo-alvo do VPC Lattice

Se você não precisar mais usar o VPC Lattice, use o procedimento a seguir para desanexar o grupo de destino do Auto Scaling.

Console

Siga as etapas nesta seção para usar o console para:

- Desvincular um grupo de destino do VPC Lattice
- Desative as verificações de integridade do VPC Lattice

Para desanexar um grupo de destino do VPC Lattice

1. Abra o console do Amazon EC2 em <https://console.aws.amazon.com/ec2/> e escolha grupos do Auto Scaling 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. Escolha Update (Atualizar).

Ao terminar de separar o grupo-alvo, 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 Health, tipos adicionais de verificação de integridade, desmarque Ativar verificações de integridade do VPC Lattice.
3. Escolha Update (Atualizar).

AWS CLI

Siga as etapas nesta seção para usar aAWS CLI para:

- Desvincular um grupo de destino do VPC Lattice
- Desative as verificações de integridade do VPC Lattice

Use o [detach-traffic-sources](#) comando para desanexar um grupo de destino do Auto Scaling quando ele não é mais necessário.

```
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 integridade em um grupo do Auto Scaling para que ele não use mais as verificações de integridade do VPC Lattice, use o [update-auto-scaling-group](#) comando. Inclua a `--health-check-type` opção 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

Depois que você associa um grupo de destino do VPC Lattice a um grupo do Auto Scaling, ele entra no `Adding` estado ao registrar as instâncias no grupo. Quando todas as instâncias do grupo são registradas, ele entra no `Added` estado. Depois que pelo menos uma instância registrada passa nas verificações de integridade, ele entra no estado `InService`. Quando o grupo de destino está no `InService` estado, 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 `InService` estado. O Amazon EC2 Auto Scaling não termina e substitui as instâncias.

Quando você desvincula um grupo de destino de um serviço, ele entra no `Removing` estado ao cancelar o registro das instâncias no grupo. As instâncias permanecem em execução após o cancelamento do

registro. Por padrão, a descarga da conexão (atraso de cancelamento de registro) é habilitada. Se a descarga da conexão estiver habilitada, a VPC Lattice 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 os 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-sources](#) comando a seguir retorna o status do anexo de todas as fontes de tráfego para o 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 vinculado ao grupo Auto Scaling, junto com o status do anexo do grupo-alvo no `State` elemento.

```
{  
    "TrafficSources": [  
        {  
            "Identifier": "arn:aws:vpc-lattice:region:account-id:targetgroup/tg-0e2f2665eEXAMPLE",  
            "State": "InService",  
            "Type": "vpc-lattice"  
        }  
    ]  
}
```

Usar EventBridge para lidar com eventos do Auto Scaling

A AmazonEventBridge, anteriormente chamada de CloudWatch Eventos, ajuda a configurar regras orientadas por eventos que monitoram recursos e iniciam ações-alvo que utilizam outros serviços da AWS.

Os eventos do Amazon EC2 Auto Scaling são entregues quase EventBridge 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 no processo de iniciar ou terminar, você pode invocar uma função do AWS Lambda para executar uma tarefa pré-configurada.

Os destinos das EventBridge regras podem incluir AWS Lambda funções do, tópicos do Amazon SNS, destinos de API, barramentos de eventos em outras Contas da AWS e muito mais. Para obter informações sobre alvos suportados, consulte os [EventBridgealvos](#) da Amazon no Guia EventBridge do usuário da Amazon.

Comece criando EventBridge regras com um exemplo usando um tópico do Amazon SNS e uma EventBridge regra. 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 [Criar EventBridge regras para eventos de atualização de instância \(p. 411\)](#).

Índice

- [Referência de eventos do Amazon EC2 Auto Scaling \(p. 401\)](#)
- [Exemplos de eventos e padrões de piscinas aquecidas \(p. 407\)](#)

- [Crie EventBridge regras \(p. 411\)](#)

Referência de eventos do Amazon EC2 Auto Scaling

Usando a AmazonEventBridge, você pode criar regras que correspondam aos eventos recebidos e encaminhá-los aos alvos para processamento.

Índice

- [Eventos de ação do ciclo de vida \(p. 401\)](#)
- [Eventos de aumento \(p. 402\)](#)
- [Eventos de aumento \(p. 404\)](#)
- [Eventos de atualização de instância \(p. 405\)](#)

Eventos de ação do ciclo de vida

Quando você adiciona ganchos do ciclo de vida ao seu grupo do Auto Scaling, o Amazon EC2 Auto Scaling envia eventos para EventBridge quando uma instância faz a transição para um estado de espera. Os eventos são emitidos com base no melhor esforço.

Event types (Tipos de evento)

- [Ação de expansão do ciclo de vida \(p. 401\)](#)
- [Ampliar a ação do ciclo de vida \(p. 401\)](#)

Ação de expansão 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 execução.

```
{  
    "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"  
    }  
}
```

Ampliar a 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 gancho do ciclo de vida de encerramento.

Important

Quando um grupo de Auto Scaling retorna instâncias para um pool aquecido em escala, o retorno de instâncias para o pool aquecido também pode gerar EC2 Instance-terminate Lifecycle Action eventos. Os eventos que são entregues quando uma instância passa para o estado de espera em escala em têm WarmPool como valor paraDestination. 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 aumento

Os exemplos a seguir mostram os tipos de eventos de escalabilidade bem-sucedidos. Os eventos são emitidos com base no melhor esforço.

Event types (Tipos de evento)

- [Evento de aumento \(p. 402\)](#)
- [Evento de expansão bem-sucedido \(p. 403\)](#)

Evento de aumento

O exemplo de evento a seguir mostra que o Amazon EC2 Auto Scaling iniciou com êxito uma instância.

```
{  
    "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",  
        "Status": "In Progress",  
        "Reason": "The instance was successfully launched.",  
        "ReasonCode": "Success",  
        "ImageId": "ami-00000000",  
        "InstanceId": "i-1234567890abcdef0",  
        "ImageType": "Amazon Linux 2",  
        "LaunchTime": "2023-01-01T12:00:00Z",  
        "PrivateDns": "ip-10-0-0-1.us-west-2.compute.internal",  
        "PrivateIpAddress": "10.0.0.1",  
        "PublicDns": "ip-10-0-0-1.us-west-2.compute.amazonaws.com",  
        "PublicIpAddress": "10.0.0.1",  
        "Region": "us-west-2",  
        "SecurityGroupIds": ["sg-00000000"],  
        "SubnetId": "subnet-00000000",  
        "VpcId": "vpc-00000000",  
        "KernelId": null,  
        "RamdiskId": null,  
        "BlockDeviceMappings": [{"DeviceName": "/dev/sda1", "Ebs": {"VolumeSize": 20, "VolumeType": "Standard"}, "VirtualName": null}, {"DeviceName": "/dev/xvda", "Ebs": {"VolumeSize": 20, "VolumeType": "Standard"}, "VirtualName": null}],  
        "NetworkInterfaces": [{"AssociatePublicIpAddress": true, "Description": "Primary network interface", "MacAddress": "54-0c-27-00-00-01", "PrivateDns": "ip-10-0-0-1.us-west-2.compute.internal", "PrivateIpAddress": "10.0.0.1", "PublicDns": "ip-10-0-0-1.us-west-2.compute.amazonaws.com", "PublicIpAddress": "10.0.0.1", "SubnetId": "subnet-00000000", "VpcId": "vpc-00000000"}]  
    }  
}
```

```
        "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 expansão bem-sucedido

O evento de exemplo a seguir mostra que o Amazon EC2 Auto Scaling encerrou com êxito uma instância.

Important

Quando um grupo de Auto Scaling retorna instâncias para um pool aquecido em escala, o retorno de instâncias para o pool aquecido também pode gerar EC2 Instance Terminate Successful eventos. Os eventos que são entregues quando uma instância retorna com êxito à piscina aquecida têm WarmPool como valor o valor deDestination. 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": [  

```

Eventos de aumento

Os exemplos a seguir mostram os tipos de eventos de escalabilidade malsucedidos. Os eventos são emitidos com base no melhor esforço.

Event types (Tipos de evento)

- [Evento de aumento \(p. 404\)](#)
- [Evento de escalonamento malsucedido \(p. 404\)](#)

Evento de aumento

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": {  
        "Status": "Failed",  
        "AutoScalingGroupName": "my-asg",  
        "ActivityId": "87654321-4321-4321-210987654321",  
        "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 escalonamento malsucedido

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 grupo de alta atividade ao reduzir a escala na horizontal, a falha ao retornar as instâncias ao grupo de alta atividade também pode gerar EC2 Instance Terminate Unsuccessful eventos. Os eventos que são entregues quando uma instância não retorna à piscina aquecida têm WarmPool como valor o valor deDestination. 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": {  
        "Status": "Unsuccessful",  
        "AutoScalingGroupName": "my-asg",  
        "ActivityId": "87654321-4321-4321-210987654321",  
        "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"  
    }  
}
```

```
"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"
}
}
```

Eventos de atualização de instância

Os exemplos a seguir mostram eventos do recurso de atualização de instância. Os eventos são emitidos com base no melhor esforço.

Event types (Tipos de evento)

- [Ponto de controle alcançado \(p. 405\)](#)
- [Atualização de instância iniciada \(p. 406\)](#)
- [A atualização de instância foi bem-sucedida \(p. 406\)](#)
- [Failed na atualização de instância \(p. 406\)](#)
- [Atualização de instância cancelada \(p. 407\)](#)

Ponto de controle alcançado

Quando o número de instâncias que foram 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 é alterado para `InProgress`, 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 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 atualização de instância foi bem-sucedida

Quando o status de uma atualização de instância é alterado para `Succeeded`, 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 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"  
    }  
}
```

Failed na atualização de instância

Quando o status de uma atualização de instância é alterado para `Failed`, 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 Failed",  
    "source": "aws.autoscaling",  
    "account": "123456789012",  
    "time": "yyyy-mm-ddThh:mm:ssZ",  
    "region": "us-west-2",  
    "resources": [  
        "auto-scaling-group-arn"  
    ]  
}
```

```
        "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 é alterado para `Cancelled`, 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 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"
    }
}
```

Exemplos de eventos e padrões de piscinas aquecidas

O Amazon EC2 Auto Scaling oferece suporte a 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 grupo de alta atividade. Você deve inserir o padrão como um objeto JSON. Esta seção e o tópico [Criar EventBridge regras para eventos de grupo de alta atividade \(p. 413\)](#) mostram como usar um padrão de evento para selecionar eventos e enviá-los para destinos.

Para criar EventBridge regras que filtrem eventos relacionados ao pool quente para os quais o Amazon EC2 Auto Scaling envia `EventBridge`, inclua os `Destination` campos `Origin` e da seção do `detail` evento.

Os valores de `Origin` e `Destination` podem ser:

`EC2 | AutoScalingGroup | WarmPool`

Índice

- [Eventos de exemplo do \(p. 407\)](#)
- [Exemplo de padrões de eventos \(p. 409\)](#)

Eventos de exemplo do

Quando você adiciona ganchos do ciclo de vida ao seu grupo do Auto Scaling, o Amazon EC2 Auto Scaling envia eventos para EventBridge quando uma instância faz a transição para um estado de

espera. Para obter mais informações, consulte [Usar ganchos do ciclo de vida com um grupo de alta atividade \(p. 284\)](#).

Esta seção inclui exemplos desses eventos quando seu grupo de Auto Scaling tem uma piscina aquecida. Os eventos são emitidos com base no melhor esforço.

Note

Para eventos aos quais o Amazon EC2 Auto Scaling envia EventBridge quando o aumento é bem-sucedido, consulte. [Eventos de aumento \(p. 402\)](#) Para eventos em que o dimensionamento não é bem-sucedido, consulte. [Eventos de aumento \(p. 404\)](#)

Exemplos de evento

- [Ação de expansão do ciclo de vida \(p. 408\)](#)
- [Ampliar a ação do ciclo de vida \(p. 409\)](#)

Ação de expansão do ciclo de vida

Os eventos que são entregues quando uma instância faz a transição para um estado de espera por eventos de aumento da escala na horizontal têm o EC2 Instance-launch Lifecycle Action valor de detail-type No detail objeto, os valores dos Destination atributos Origin e mostram de onde a instância está vindo 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 foi adicionada ao pool aquecido. Para obter mais informações, consulte [Transições de estado do ciclo de vida para instâncias em um grupo de alta atividade \(p. 285\)](#).

```
{  
    "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 aumento da escala na horizontal, o estado da instância muda para Pending:Wait porque ela é adicionada ao grupo do Auto Scaling a partir do grupo de alta atividade. Para obter mais informações, consulte [Transições de estado do ciclo de vida para instâncias em um grupo de alta atividade \(p. 285\)](#).

```
{  
    "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": "HighActivityPool"  
    }  
}
```

```
"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"
}
}
```

Ampliar a ação do ciclo de vida

Os eventos que são entregues quando uma instância faz a transição para um estado de espera por eventos de aumento de escala na horizontal têm o EC2 Instance-terminate Lifecycle Action valor de. detail-type No detail objeto, os valores dos Destination atributos Origin e mostram de onde a instância está vindo e para onde está indo.

Neste exemplo de evento de aumento da escala na horizontal, o estado de uma instância muda para Warmed:Pending:Wait porque ela é retornada ao grupo de alta atividade. Para obter mais informações, consulte [Transições de estado do ciclo de vida para instâncias em um grupo de alta atividade \(p. 285\)](#).

```
{
    "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.

EventBridgeos padrões de eventos têm a mesma estrutura que os 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 grupo de alta atividade.

```
{  
    "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 grupo de alta atividade devido a um evento de aumento de escala na horizontal.

```
{  
    "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 grupo de alta atividade ao reduzir a escala na horizontal.

```
{  
    "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" ]  
}
```

Crie EventBridge regras

Quando um evento é emitido pelo Amazon EC2 Auto Scaling, uma notificação de evento é enviada para a Amazon EventBridge como um arquivo JSON. Você pode escrever uma EventBridge regra de e automatizar quais ações tomar quando o padrão de evento corresponder à regra. Se o EventBridge detectar um padrão de evento que corresponda a um padrão definido em uma regra, o EventBridge invocará um destino (ou destinos) especificado (s) 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 \(p. 273\)](#).
- Para invocar uma função do Lambda em chamadas de API registradas com CloudTrail, consulte [Tutorial: Registre chamadas de AWS API usando EventBridge no Guia do usuário da Amazon EventBridge](#).
- Para obter mais informações sobre como criar regras de eventos da Amazon, consulte [Criar EventBridge regras da Amazon que reajam a eventos](#) no Guia EventBridge do usuário da Amazon.

Tópicos

- [Criar EventBridge regras para eventos de atualização de instância \(p. 411\)](#)
- [Criar EventBridge regras para eventos de grupo de alta atividade \(p. 413\)](#)

Criar EventBridge regras para eventos de atualização de instância

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 [Substituir instâncias do Auto Scaling com base em uma atualização de instância \(p. 108\)](#).

Criar 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 ASCIIIs, 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 da 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 Rules (Regras).
3. Escolha Create rule (Criar regra).
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 Event Bus (Barramento de eventos), escolha default (padrão). Quando um serviço da AWS em sua conta gerar um evento, ele sempre irá para o barramento de eventos padrão da sua conta.
 - c. Em Rule type (Tipo de regra), escolha Rule with an event pattern (Regra com um padrão de evento).
 - d. Escolha Próximo.
5. Em Build event pattern (Criar padrão de evento), faça o seguinte:
 - a. Em Origem do evento, escolha AWSEventos ou eventos de EventBridge parceiros.
 - b. Em Event pattern (Padrão de evento), faça o seguinte:
 - i. Para Event source (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 [Criar EventBridge regras da Amazon que reajam a eventos](#) (etapa 16) 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).

Criar EventBridge regras para eventos de grupo de alta atividade

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 grupo de alta atividade.

Antes de criar a regra, crie a função do AWS Lambda que deseja que a regra use como destino. Você deve especificar essa função como o destino da regra. O procedimento a seguir fornece apenas as etapas para criar a EventBridge regra que atua quando novas instâncias entrarem no grupo de alta atividade. 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 \(p. 273\)](#).

Para obter mais informações sobre como criar e trabalhar com grupos de alta atividade, consulte [Grupos de alta atividade do Amazon EC2 Auto Scaling \(p. 279\)](#).

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 Rules (Regras).
3. Escolha Create rule (Criar regra).
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 Event Bus (Barramento de eventos), escolha default (padrão). Quando um AWS service (Serviço da AWS) em sua conta gerar um evento, ele sempre irá para o barramento de eventos padrão da sua conta.

- c. Em Rule type (Tipo de regra), escolha Rule with an event pattern (Regra com um padrão de evento).
 - d. Escolha Próximo.
5. Em Build event pattern (Criar padrão de evento), faça o seguinte:
 - a. Em Origem do evento, escolha AWSeventos ou eventos de EventBridge parceiros.
 - b. Em Event pattern (Padrão de evento), escolha Custom pattern (JSON editor) (Padrão personalizado [editor JSON]) e cole o padrão a seguir na caixa Event pattern (Padrão de evento), substituindo o texto em *itálico* pelo nome do seu grupo do Auto Scaling.

```
{  
    "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 \(p. 409\)](#).

- 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 [Criar EventBridge regras da Amazon que reajam a eventos](#) (etapa 16) 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

Estamos aposentando o EC2-Classic. Recomendamos que você migre do EC2-Classic para uma VPC. Para mais informações, consulte a postagem de blog [EC2-Classic Networking is Retiring - Here's How to Prepare](#).

A Amazon Virtual Private Cloud (Amazon VPC) permite que você defina um ambiente de redes virtuais em uma seção privada e isolada na Nuvem AWS. Você tem controle total sobre seu ambiente de rede virtual.

Em uma Virtual Private Cloud (VPC), você pode iniciar recursos da AWS como grupos do Auto Scaling. Um grupo do Auto Scaling em uma VPC funciona basicamente da mesma maneira como no EC2-Classic e oferece suporte ao mesmo conjunto de recursos.

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, você deve 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 Manual do usuário da Amazon VPC.

Índice

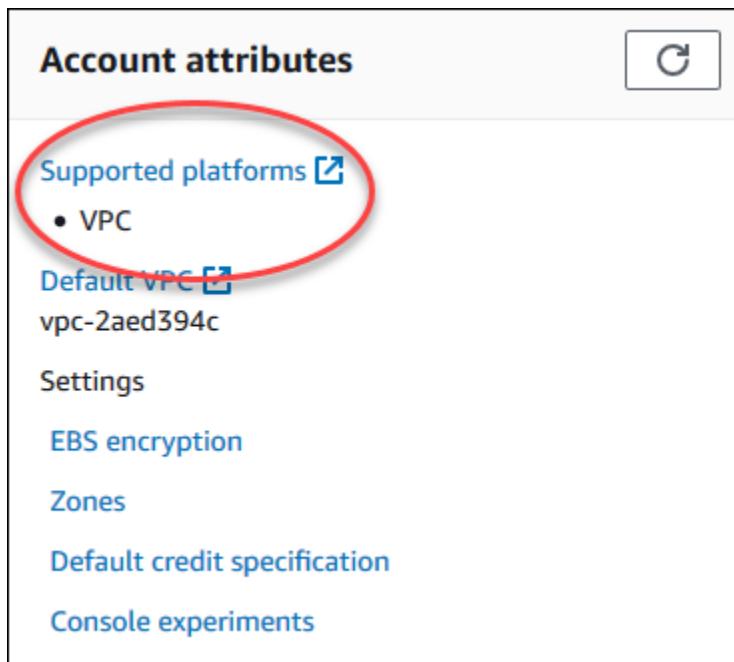
- [EC2-Classic \(p. 415\)](#)
- [VPC padrão \(p. 416\)](#)
- [VPC não padrão \(p. 416\)](#)
- [Considerações sobre a escolha de sub-redes da VPC \(p. 417\)](#)
- [Endereçamento IP em uma VPC \(p. 417\)](#)
- [Interfaces de rede em uma VPC \(p. 417\)](#)
- [Locação de localização de instância \(p. 418\)](#)
- [AWS Outposts \(p. 418\)](#)
- [Mais recursos para saber mais sobre VPCs \(p. 418\)](#)

EC2-Classic

Se você criou sua Conta da AWS antes de 4 de dezembro de 2013, ela pode permitir que você escolha entre a Amazon VPC e o EC2-Classic em determinadas regiões. Se você tiver uma dessas contas mais antigas, talvez tenha grupos do Auto Scaling no EC2-Classic em algumas regiões do em vez da Amazon VPC.

Para obter mais informações sobre a migração do EC2 Classic, consulte a postagem no blog [EC2-Classic Networking is Retiring - Here's How to Prepare](#) (O EC2-Classic está sendo descontinuado - Veja como se preparar). Para obter mais informações sobre as diferenças entre instâncias no EC2-Classic e uma VPC, consulte [EC2-Classic](#) no Manual do usuário do Amazon EC2 para instâncias do Linux.

Para determinar se alguma das Regiões da AWS que você usa ainda está usando o EC2-Classic, abra o console do Amazon EC2. Se o campo Supported platforms (Plataformas compatíveis) indicar apenas VCP, como mostrado no exemplo a seguir, a sua Conta da AWS na Região da AWS atual usa a plataforma VPC e usa uma VPC padrão. O nome da VPC padrão é mostrado abaixo da plataforma com suporte.



Tip

Qualquer grupo do Auto Scaling criado em uma região que tenha uma VPC padrão pode iniciar instâncias na VPC padrão ou em uma VPC não padrão, mas não no EC2-Classic.

VPC padrão

Se você tiver criado sua Conta da AWS depois de 4 de dezembro de 2013, ou se estiver criando seu grupo do Auto Scaling em uma nova 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.

É possível ver suas VPCs na página [Your VPCs](#) (Suas VPCs) do console do Amazon VPC.

Para obter mais informações sobre a VPC padrão, consulte [VPC padrão e sub-redes padrão](#) no Guia do usuário do Amazon VPC.

VPC não padrão

Você pode optar por criar VPCs adicionais acessando a página [VPC Dashboard](#) (Painel da VPC) no AWS Management Console e selecionando Create VPC (Criar VPC).

Para obter mais informações, 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 os aplicativos hospedados 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 os aplicativos 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平衡ador 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 em sub-redes públicas para oferecer suporte à 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 sem ingresso do Auto Scaling usando o gerenciador de sessão do AWS Systems Manager, as instâncias poderão ser iniciadas 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 endereços IPv4 públicos às instâncias. A atribuição de endereços IP públicos às suas instâncias permite que elas se comuniquem com a Internet ou com outros produtos da AWS.

Quando você inicia instâncias em uma sub-rede configurada para atribuir automaticamente endereços IPv6, elas recebem endereços IPv4 e IPv6. Caso contrário, elas recebem apenas endereços IPv4. Para obter mais informações, consulte [Endereços IPv6](#) no Manual do usuário do Amazon EC2 para instâncias do Linux.

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 poderá mais atribuir um endereço IPv4 público à interface de rede principal. Você não poderá se conectar às suas instâncias por meio de um endereço IPv4 público, a menos que associe endereços de IP elástico 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?](#) na Central de conhecimento da AWS.

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. No entanto, o suporte para hosts dedicados só está disponível para grupos do Auto Scaling que usem um modelo de execução. Para obter mais informações, consulte [Configurar a locação de instância com uma configuração de execução \(p. 58\)](#).

AWS Outposts

O AWS Outposts estende um Amazon VPC de uma região da AWS para um Outpost com os componentes da VPC que estão acessíveis na região, incluindo gateways da Internet, gateways privados virtuais, Amazon VPC Transit Gateways e endpoints da VPC. 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 saber mais sobre VPCs

Use os tópicos a seguir para saber mais sobre VPCs e sub-redes.

- Sub-redes privadas em uma VPC
 - [VPC com sub-redes públicas e privadas \(NAT\)](#)
 - [Gateways NAT](#)
- Sub-redes públicas em uma VPC
 - [VPC com uma única sub-rede pública](#)
- Sub-redes para seu Application Load Balancer
 - [Sub-redes para seu平衡ador de carga](#)
- Informações gerais da VPC
 - [Guia do usuário da Amazon VPC](#)
 - [Emparelhamento de VPC](#)
 - [Interfaces de rede elástica](#)
 - [Usar endpoints da VPC para conectividade privada \(p. 457\)](#)
 - [Migrar do EC2-Classic para uma VPC](#)

Segurança no Amazon EC2 Auto Scaling

A segurança para com a nuvem na AWS é a nossa maior prioridade. Como cliente da AWS, você se contará com um datacenter e uma arquitetura de rede criados para atender aos requisitos das organizações com as maiores exigências de segurança.

A segurança é uma responsabilidade compartilhada entre a AWS e você. O [modelo de responsabilidade compartilhada](#) descreve isso como segurança da nuvem e segurança na nuvem:

- Segurança da nuvem: a AWS é responsável pela proteção da infraestrutura que executa produtos da AWS na Nuvem AWS. A AWS também fornece serviços que podem ser usados com segurança. Auditores externos testam e verificam regularmente a eficácia da nossa segurança como parte dos [Programas de conformidade da AWS](#). Para saber mais sobre os programas de compatibilidade que se aplicam ao Amazon EC2 Auto Scaling, consulte [Serviços da AWS em escopo por programa de compatibilidade](#).
- Segurança na nuvem: sua responsabilidade é determinada pelo serviço da AWS 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 regulamentos 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. Saiba também como usar outros serviços da AWS que ajudam você a monitorar e proteger os recursos do Amazon EC2 Auto Scaling.

Tópicos

- [Amazon EC2 Auto Scaling e proteção de dados \(p. 419\)](#)
- [Gerenciamento de identidade e acesso para o Amazon EC2 Auto Scaling \(p. 421\)](#)
- [Validação de compatibilidade do Amazon EC2 Auto Scaling \(p. 455\)](#)
- [Resiliência no Amazon EC2 Auto Scaling \(p. 456\)](#)
- [Segurança da infraestrutura no Amazon EC2 Auto Scaling \(p. 457\)](#)
- [Amazon EC2 Auto Scaling e endpoints da VPC da interface \(p. 457\)](#)

Amazon EC2 Auto Scaling e proteção de dados

O [modelo de responsabilidade compartilhada](#) da AWS se aplica à proteção de dados no Amazon EC2 Auto Scaling. Conforme descrito nesse modelo, a AWS é responsável por proteger a infraestrutura global que executa toda a Nuvem AWS. Você é responsável por manter o controle sobre seu conteúdo hospedado nessa infraestrutura. Esse conteúdo inclui as tarefas de configuração e gerenciamento de segurança dos Serviços da AWS que você usa. Para ter mais informações sobre a privacidade de dados, consulte as [Perguntas frequentes sobre privacidade de dados](#). Para ter mais informações sobre a proteção de dados na Europa, consulte a postagem do blog [AWS Shared Responsibility Model and GDPR](#) no Blog de segurança da AWS.

Para fins de proteção de dados, recomendamos que você proteja as credenciais da Conta da AWS e configure as contas de usuário individuais com o AWS IAM Identity Center (successor to AWS Single Sign-

On) ou o 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 os recursos da AWS. Exigimos TLS 1.2 e recomendamos TLS 1.3.
- Configure o registro em log das atividades da API e do usuário com o AWS CloudTrail.
- Use as soluções de criptografia da AWS, juntamente com todos os controles de segurança padrão dos Serviços da AWS.
- Use serviços gerenciados de segurança avançada, como o Amazon Macie, que ajuda a localizar e proteger dados sigilosos armazenados no Amazon S3.
- Se você precisar de módulos criptográficos validados pelo FIPS 140-2 ao acessar a AWS por meio de uma interface de linha de comando ou uma API, use um endpoint do FIPS. Para ter mais informações sobre endpoints do FIPS, consulte [Federal Information Processing Standard \(FIPS\) 140-2](#).

É altamente recomendável que nunca sejam colocadas informações de identificação confidenciais, como endereços de email dos seus clientes, em marcações ou campos de formato livre, como um campo Name (Nome). Isso também vale para o uso do Amazon EC2 Auto Scaling ou de outros Serviços da AWS com o console, a API, a AWS CLI ou os SDKs da AWS. Quaisquer dados inseridos em tags ou 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, recomendamos fortemente que não sejam incluídas informações de credenciais no URL para validar a solicitação a esse servidor.

Ao executar uma instância no Amazon EC2, você tem a opção de passar dados de usuário para a instância para realizar configurações adicionais na inicialização da instância. Também recomendamos que você nunca coloque informações confidenciais ou confidenciais nos dados do usuário que serão passadas para uma instância.

Usar 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 Scaling oferece suporte a chaves gerenciadas pela AWS e chaves 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 \(p. 23\)](#).

Também é possível configurar uma chave gerenciada pelo cliente na AMI baseada no EBS antes de configurar o modelo ou a configuração de execução, ou usar a criptografia por padrão para impor a criptografia dos novos volumes do EBS e cópias de snapshots que você criar.

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 chaves do AWS KMS obrigatórias para uso com volumes criptografados \(p. 450\)](#). Para obter informações sobre como criar, armazenar e gerenciar suas chaves de criptografia do AWS KMS, consulte [O que é o AWS Key Management Service?](#)

Tópicos relacionados

- [Proteção de dados no Amazon EC2](#) no Manual do usuário do Amazon EC2 para instâncias do Linux
- [Usar criptografia com AMIs baseadas no EBS](#) no Manual do usuário do Amazon EC2 para instâncias do Linux

- [Criptografia por padrão](#) no Manual do usuário do Amazon EC2 para instâncias do Linux

Gerenciamento de identidade e acesso para o Amazon EC2 Auto Scaling

O AWS Identity and Access Management (IAM) é um serviço da AWS service (Serviço da AWS) que ajuda a controlar o acesso aos recursos da AWS de forma segura. 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 pode ser usado sem custo adicional.

Para usar o Amazon EC2 Auto Scaling, você precisa de um Conta da AWS e suas credenciais de segurança para entrar em sua conta. Para obter mais informações, consulte [AWS credenciais de segurança](#) Referência geral da AWS.

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, você deve ter permissões para criar grupos de 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 \(p. 421\)](#)
- [Permissões da API Amazon EC2 Auto Scaling \(p. 428\)](#)
- [Políticas gerenciadas pela AWS para o Amazon EC2 Auto Scaling \(p. 429\)](#)
- [Funções vinculadas ao serviço do Amazon EC2 Auto Scaling \(p. 432\)](#)
- [Exemplos de políticas baseadas em identidade do Amazon EC2 Auto Scaling \(p. 437\)](#)
- [Prevenção do problema do substituto confuso entre serviços \(p. 442\)](#)
- [Suporte a modelo de execução \(p. 443\)](#)
- [Funções do IAM para aplicações que são executadas em instâncias do Amazon EC2 \(p. 448\)](#)
- [Política de chaves do AWS KMS obrigatórias para uso com volumes criptografados \(p. 450\)](#)

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	Supporte a Amazon EC2 Auto Scaling
Políticas baseadas em identidade (p. 422)	Sim
Políticas baseadas em recursos (p. 422)	Não
Ações de políticas (p. 423)	Sim
Recursos de políticas (p. 424)	Sim
Chaves de condição de política (específicas do serviço) (p. 425)	Sim
ACLs (p. 426)	Não
ABAC (etiquetas em políticas) (p. 427)	Parcial
Credenciais temporárias (p. 427)	Sim
Funções de serviço (p. 428)	Sim
Funções vinculadas ao serviço (p. 428)	Sim

Para obter uma visualização de alto nível de como o Amazon EC2 Auto Scaling e outros Serviços da AWS funcionam com a maioria dos recursos do IAM, consulte [Serviços da AWS que funcionam com o IAM](#) no Guia do usuário do IAM.

Políticas baseadas em identidade para o Amazon EC2 Auto Scaling

É compatível com políticas baseadas em identidade	Sim
---	-----

As políticas baseadas em identidade são documentos de políticas de permissões JSON que você pode anexar a uma identidade, como usuário, grupo de usuários ou função do IAM. Essas políticas controlam quais ações os usuários e funções podem realizar, em quais recursos e em que condições. Para saber como criar uma política baseada em identidade, consulte [Criar políticas do IAM](#) no Guia do usuário do IAM.

Com as políticas baseadas em identidade do IAM, é possível especificar ações ou recursos permitidos ou negados, bem como as condições sob as quais as ações são permitidas ou negadas. Você não pode especificar a entidade principal em uma política baseada em identidade porque ela se aplica ao usuário ou função à qual ela está anexado. Para saber mais sobre todos os elementos que podem ser usados em uma política JSON, consulte [Referência de elementos de política JSON do IAM](#) no Guia do usuário do IAM.

Exemplos de políticas baseadas em identidade para o Amazon EC2 Auto Scaling

Para visualizar exemplos de políticas baseadas em identidade do Amazon EC2 Auto Scaling, consulte [Exemplos de políticas baseadas em identidade do Amazon EC2 Auto Scaling \(p. 437\)](#).

Políticas baseadas em recursos no Amazon EC2 Auto Scaling

Oferece suporte a políticas baseadas em recursos	Não
--	-----

Políticas baseadas em recurso são documentos de políticas JSON que você anexa a um recurso. São exemplos de políticas baseadas em recursos as políticas de confiança de função 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 recurso ao qual a política está anexada, a política define quais ações um principal especificado pode executar nesse recurso e em que condições. Você deve especificar um principal em uma política baseada em recursos. As entidades principais podem incluir contas, usuários, funções, usuários federados ou Serviços da AWS.

Para permitir o acesso entre contas, você pode especificar uma conta inteira ou as entidades do IAM em outra conta como a entidade principal em uma política baseada em recurso. Adicionar um principal entre contas à política baseada em recurso é apenas metade da tarefa de estabelecimento da relação de confiança. Quando a entidade principal e o recurso estão em diferentes Contas da AWS, um administrador do IAM da conta confiável também deve conceder à entidade principal (usuário ou função) permissão para acessar o recurso. Eles concedem permissão ao anexar uma política baseada em identidade para a entidade. No entanto, se uma política baseada em recurso conceder acesso a um principal na mesma conta, nenhuma política baseada em identidade adicional será necessária. Para obter mais informações, consulte Como as funções do IAM diferem de políticas baseadas em recursos no Guia do usuário do IAM.

Ações de políticas para o Amazon EC2 Auto Scaling

Oferece suporte a ações de políticas	Sim
--------------------------------------	-----

Os administradores podem usar AWS as políticas JSON da para especificar quem tem acesso a quê. Ou seja, qual principal pode executar ações em quais recursos, e em que condições.

O elemento Action de uma política JSON descreve as ações que você pode usar para permitir ou negar acesso em uma política. As ações de política geralmente têm o mesmo nome que a operação de API da AWS associada. Existem algumas exceções, como ações somente de permissão, que não têm uma operação de API correspondente. Há também algumas operações que exigem várias ações em uma política. Essas ações adicionais são chamadas de ações dependentes.

Inclua 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*"
```

Para visualizar exemplos de políticas baseadas em identidade do Amazon EC2 Auto Scaling, consulte Exemplos de políticas baseadas em identidade do Amazon EC2 Auto Scaling (p. 437).

Recursos de política para o Amazon EC2 Auto Scaling

Oferece suporte a recursos de políticas	Sim
---	-----

Os administradores podem usar AWS as políticas JSON da para especificar quem tem acesso a quê. Ou seja, qual principal pode executar ações em quais recursos, e em que condições.

O elemento Resource de política JSON especifica o objeto ou os objetos aos quais a ação se aplica. As instruções devem incluir um elemento Resource ou um elemento NotResource. Como prática recomendada, especifique um recurso usando seu [Nome do recurso da Amazon \(ARN\)](#). Isso pode ser feito para ações que oferecem suporte a um tipo de recurso específico, conhecido como permissões em nível de recurso.

Para ações que não oferecem suporte a permissões em nível de recurso, como operações de listagem, use um curinga (*) para indicar que a instrução se aplica a todos os recursos.

```
"Resource": "*"
```

É possível usar ARNs para identificar os grupos do Auto Scaling e as configurações de execução 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: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 e seus ARNs, consulte [Recursos definidos pelo Amazon EC2 Auto Scaling](#) na Referência de autorização do serviço. Para saber com quais ações você pode especificar o ARN de cada recurso, consulte [Ações definidas pelo Amazon EC2 Auto Scaling](#).

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 do serviço	Sim
--	-----

Os administradores podem usar AWS as políticas JSON da para especificar quem tem acesso a quê. Ou seja, qual entidade principal pode executar ações em quais recursos e em que condições.

O elemento `Condition` (ou bloco de `Condition`) permite que você especifique condições nas quais uma instrução está em vigor. O elemento `Condition` é opcional. É possível criar expressões condicionais que usam [agentes de condição](#), como “igual a” ou “menor que”, para fazer a condição da política corresponder aos valores na solicitação.

Se você especificar vários elementos `Condition` em uma instrução ou várias chaves em um único elemento `Condition`, a AWS os avaliará usando uma operação lógica AND. Se você especificar vários valores para uma única chave de condição, a AWS avaliará a condição usando uma operação lógica OR. Todas as condições devem ser atendidas para que as permissões da instrução sejam concedidas.

Você também pode usar variáveis de espaço reservado ao especificar as condições. Por exemplo, é possível conceder a um usuário do IAM permissão para acessar um recurso somente se ele estiver marcado com seu nome de usuário do IAM. Para obter mais informações, consulte [Elementos de política do IAM: variáveis e tags](#) no Guia do usuário do IAM.

A AWS oferece suporte a chaves de condição globais e chaves de condição específicas do serviço. Para ver todas as chaves de condição globais da AWS, consulte [Chaves de contexto de condição globais da AWS](#) no Manual do usuário do IAM.

O Amazon EC2 Auto Scaling define seu próprio conjunto de chaves de condição e também oferece suporte ao uso de algumas chaves de condição globais.

O Amazon EC2 Auto Scaling oferece suporte às seguintes chaves de condição que você pode usar em políticas de permissão para determinar quem pode acessar o Amazon EC2 Auto Scaling:

- `autoscaling:InstanceTypes`
- `autoscaling:LaunchConfigurationName`
- `autoscaling:LaunchTemplateVersionSpecified`
- `autoscaling:LoadBalancerNames`
- `autoscaling:MaxSize`
- `autoscaling:MinSize`
- `autoscaling:ResourceTag/key-name: tag-value`
- `autoscaling:TargetGroupARNs`
- `autoscaling:VPCZoneIdentifiers`

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 \(p. 138\)](#).

- `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 como usar chaves de condição do Amazon EC2 Auto Scaling, consulte [Chaves de condição para o Amazon EC2 Auto Scaling](#).

Para ver exemplos de políticas do IAM que você pode usar para controlar o acesso, consulte os seguintes tópicos:

- Para exemplos que usam chaves de condição para controlar o acesso a ações em grupos do Auto Scaling e configurações de inicialização, consulte [Etiquetar grupos e instâncias do Auto Scaling \(p. 138\)](#) e [Exemplos de políticas baseadas em identidade do Amazon EC2 Auto Scaling \(p. 437\)](#).
- Para exemplos adicionais, incluindo um exemplo que nega acesso a grupos do Auto Scaling se uma configuração de inicialização for especificada na solicitação, consulte [Suporte a modelo de execução \(p. 443\)](#).

ACLs no Amazon EC2 Auto Scaling

Oferece suporte a ACLs	Não
------------------------	-----

As listas de controle de acesso (ACLs) controlam quais entidades principais (membros, usuários ou funções da conta) têm permissões para acessar um recurso. As 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

Oferece suporte a ABAC (tags em políticas)	Parcial
--	---------

O controle de acesso baseado em atributo (ABAC) é uma estratégia de autorização que define permissões com base em atributos. Na AWS, esses atributos são chamados de tags. É possível anexar tags a entidades do IAM (usuários ou funções) e a muitos recursos da AWS. A marcação de entidades e recursos é a primeira etapa do ABAC. Em seguida, você cria políticas de ABAC para permitir operações quando a tag da entidade principal corresponder à tag do recurso que ela está tentando acessar.

O ABAC é útil em ambientes que estão crescendo rapidamente e ajuda em situações em que o gerenciamento de políticas se torna um problema.

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 `aws:TagKeys` chaves de condição.

Se um serviço oferecer suporte às três chaves de condição para cada tipo de recurso, o valor será Yes (Sim) para o serviço. Se um serviço oferecer suporte às três chaves de condição somente para alguns tipos de recursos, o valor será Partial (Parcial).

Para obter mais informações sobre o ABAC, consulte [O que é ABAC?](#) no Guia do usuário do IAM. Para visualizar um tutorial com etapas para configurar o ABAC, consulte [Usar controle de acesso baseado em atributos \(ABAC\)](#) (Use attribute-based access control [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 sobre a marcação de grupos do Auto Scaling, consulte [Etiquetar grupos e instâncias do Auto Scaling \(p. 138\)](#).

Para visualizar um exemplo de política baseada em identidade para limitar o acesso a um grupo do Auto Scaling com base nas tags desse grupo, consulte [Etiquetas para segurança \(p. 142\)](#).

Uso de credenciais temporárias com o Amazon EC2 Auto Scaling

Oferece suporte a credenciais temporárias	Sim
---	-----

Alguns Serviços da AWS não funcionam quando você faz login usando credenciais temporárias. Para obter informações adicionais, incluindo quais Serviços da AWS funcionam com credenciais temporárias, consulte [Serviços da AWS que funcionam com o IAM](#) no Guia do usuário do IAM.

Você está usando credenciais temporárias se faz login no AWS Management Console usando qualquer método, exceto um nome de usuário e uma senha. Por exemplo, quando você acessa a AWS usando o link de autenticação única (SSO) da sua empresa, esse processo cria automaticamente credenciais temporárias. Você também cria automaticamente credenciais temporárias quando faz login no console como usuário e, em seguida, alterna funções. Para obter mais informações sobre como alternar funções, consulte [Alternar para uma função \(console\)](#) no Guia do usuário do IAM.

Você pode criar credenciais temporárias manualmente usando a AWS CLI ou a API da AWS. Em seguida, você pode usar essas credenciais temporárias para acessar a AWS. A AWS recomenda que você gere credenciais temporárias dinamicamente em vez de usar chaves de acesso de longo prazo. Para obter mais informações, consulte [Credenciais de segurança temporárias no IAM](#).

Perfis de serviço para o Amazon EC2 Auto Scaling

Oferece suporte a funções de serviço	Sim
--------------------------------------	-----

A função de serviço é uma [função do IAM](#) que um serviço assume para realizar ações em seu nome. Um administrador do IAM pode criar, modificar e excluir um perfil de serviço do IAM. Para obter mais informações, 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, você deve especificar uma função 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 uma função com um conjunto suficiente de permissões usando uma política gerenciada. Para obter mais informações, consulte [Receba notificações usando o Amazon SNS \(p. 260\)](#) e [Receba notificações usando o Amazon SQS \(p. 261\)](#).

Ao criar um grupo do Auto Scaling, você pode opcionalmente transmitir um perfil de serviço para permitir que instâncias do Amazon EC2 acessem outros Serviços da AWS em seu nome. 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 a AWS CLI para criar ou editar esse perfil de serviço. Para obter mais informações, consulte [Funções do IAM para aplicações que são executadas em instâncias do Amazon EC2 \(p. 448\)](#).

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.

Funções vinculadas ao serviço do Amazon EC2 Auto Scaling

Oferece suporte a funções vinculadas ao serviço	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 a função de executar uma ação em seu nome. Os perfis vinculados ao serviço aparecem em sua Conta da AWS e são de propriedade do serviço. Um administrador do IAM pode visualizar, mas não pode editar as permissões para funções vinculadas ao serviço.

Para obter detalhes sobre como criar ou gerenciar funções vinculadas ao serviço do Amazon EC2 Auto Scaling, consulte [Funções vinculadas ao serviço do Amazon EC2 Auto Scaling \(p. 432\)](#).

Permissões da API Amazon EC2 Auto Scaling

Você deve conceder aos usuários permissão para chamar as ações da API do Amazon EC2 Auto Scaling de que precisam, conforme descrito em [Ações de políticas para o Amazon EC2 Auto Scaling \(p. 423\)](#). 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 outrosAWSAPIs

Além das permissões da API Amazon EC2 Auto Scaling, os usuários devem ter as seguintes permissões de outrosAWSAPIs para executar com êxito a ação associada.

Criar um grupo do Auto Scaling (`autoscaling:CreateAutoScalingGroup`)

- `iam:CreateServiceLinkedRole`— Para criar a função padrão vinculada ao serviço, caso essa função ainda não exista.
- `iam:PassRole`— Transmitir uma função do IAM para o serviço ou para as instâncias do EC2 no lançamento. Necessário quando uma função não padrão vinculada ao serviço, uma função do IAM para um gancho do ciclo de vida ou um modelo de execução que especifica um perfil de instância (um contêiner para uma função do IAM) é fornecido.
- `ec2:RunInstances`— Para iniciar instâncias quando um modelo de execução é fornecido.
- `ec2:CreateTags`— Marcar instâncias e volumes na inicialização 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 uma função do IAM para o serviço. Necessário quando uma função do IAM é fornecida.

Anexar um grupo-alvo do VPC Lattice (`autoscaling:AttachTrafficSources`)

- `vpc-lattice:RegisterTargets`— Para registrar automaticamente as instâncias com o grupo-alvo.

Separe um grupo-alvo do VPC Lattice (`autoscaling:DetachTrafficSources`)

- `vpc-lattice:DeregisterTargets`— Para cancelar automaticamente o registro de instâncias com o grupo-alvo.

Criar uma configuração de execução (`autoscaling:CreateLaunchConfiguration`)

- `ec2:DescribeImages`
- `ec2:DescribeInstances`
- `ec2:DescribeInstanceAttribute`
- `ec2:DescribeKeyPairs`
- `ec2:DescribeSecurityGroups`
- `ec2:DescribeSpotInstanceRequests`
- `ec2:DescribeVpcClassicLink`
- `iam:PassRole`— Passar uma função do IAM para instâncias do EC2 no lançamento. Necessário quando uma configuração de execução especifica um perfil de instância (um contêiner para uma função do IAM).

Políticas gerenciadas pela AWS para o Amazon EC2 Auto Scaling

Uma política gerenciada pela AWS é uma política independente criada e administrada pela AWS. As políticas gerenciadas pela AWS são criadas para fornecer permissões a vários casos de uso comuns a fim de que você possa começar a atribuir permissões a usuários, grupos e perfis.

Lembre-se de que as políticas gerenciadas pela AWS podem não conceder permissões de privilégio mínimo para seus casos de uso específicos porque estão disponíveis para todos os clientes da AWS usarem. Recomendamos que você reduza ainda mais as permissões definindo [políticas gerenciadas pelo cliente](#) específicas para seus casos de uso.

Você não pode alterar as permissões definidas em políticas gerenciadas pela AWS. Se a AWS atualiza as permissões definidas em um política gerenciada pela AWS, a atualização afeta todas as identidades de entidades principais (usuários, grupos e perfis) às quais a política está vinculada. É mais provável que a AWS atualize uma política gerenciada pela AWS quando um novo AWS service (Serviço da AWS) é lançado ou novas operações de API são disponibilizadas para os serviços existentes.

Para obter mais informações, consulte [Políticas gerenciadas pela AWS](#) no Manual do usuário do IAM.

Políticas gerenciadas do Amazon EC2 Auto Scaling

Você pode anexar as políticas gerenciadas às suas identidades do AWS Identity and Access Management (IAM) (usuários ou perfis). Cada política fornece acesso a todas ou a algumas das ações de API para o Amazon EC2 Auto Scaling.

- [AutoScalingFullAccess](#)— Concede acesso total ao Amazon EC2 Auto Scaling para identidades IAM que precisam de acesso total ao Amazon EC2 Auto Scaling a partir do AWS CLI ou SDKs, mas não AWS Management Console acesso.
- [AutoScalingReadOnlyAccess](#)— Concede acesso somente de leitura ao Amazon EC2 Auto Scaling para identidades IAM que estão fazendo chamadas somente para o AWS CLI ou SDKs.
- [AutoScalingConsoleFullAccess](#)— Concede acesso total ao Amazon EC2 Auto Scaling usando o AWS Management Console. 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 AWS Management Console. Esta política funciona quando você usa configurações de execução, mas não quando usa modelos de execução.

Ao usar modelos de execução via console, você precisa conceder permissões adicionais específicas para os modelos de execução, o que é debatido em [Suporte a modelo de execução \(p. 443\)](#). 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 AWSPolítica gerenciada

Você não pode anexar [AutoScalingServiceRolePolicy](#) às suas identidades do IAM. Essa política é anexada a uma função vinculada ao serviço que permite que o Amazon EC2 Auto Scaling inicie e termine instâncias. Para obter mais informações, consulte [Funções vinculadas ao serviço do Amazon EC2 Auto Scaling \(p. 432\)](#).

Atualizações do Amazon EC2 Auto Scaling a políticas gerenciadas pela AWS

Visualize detalhes sobre atualizações em políticas gerenciadas pela AWS para o 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 adiciona permissões à respectiva função vinculada ao serviço	A política <u>AutoScalingServiceRolePolicy</u> agora concede permissões ao serviço para acessar as ações de API necessárias para uma integração com o VPC Lattice.	6 de dezembro de 2022

Alteração	Descrição	Data
	<ul style="list-style-type: none"> Ações <code>GetTargetGroup</code> e <code>ListTargetGroup</code>. Obrigatório para recuperar informações sobre grupos de destino VPC Lattice. Ações <code>RegisterTargets</code> e <code>DeregisterTargets</code>. Obrigatório para registrar e cancelar o registro de instâncias de grupos de destino VPC Lattice. <code>ListTargets</code>. Permite que o Amazon EC2 Auto Scaling recupere informações de integridade para instâncias registradas em grupos de destino VPC Lattice. <p>Para obter mais informações, consulte Funções vinculadas ao serviço do Amazon EC2 Auto Scaling (p. 432).</p>	
O Amazon EC2 Auto Scaling adiciona permissões à respectiva função vinculada ao serviço	Para apoiar o uso de um AWS Systems Manager Parâmetro como alias para uma ID de AMI ao criar um modelo de execução, a <code>oAutoScalingServiceRolePolicy</code> política agora concede permissão para ligar para o AWS Systems Manager <code>GetParameters</code> Ação da API. Para obter mais informações, consulte Funções vinculadas ao serviço do Amazon EC2 Auto Scaling (p. 432) .	28 de março de 2022
O Amazon EC2 Auto Scaling adiciona permissões à respectiva função vinculada ao serviço	Para apoiar a escalabilidade preditiva, a <code>oAutoScalingServiceRolePolicy</code> política agora inclui permissão para chamar o CloudWatch <code>GetMetricData</code> Ação da API. Para obter mais informações, consulte Funções vinculadas ao serviço do Amazon EC2 Auto Scaling (p. 432) .	19 de maio de 2021
O Amazon EC2 Auto Scaling começou a monitorar alterações	O Amazon EC2 Auto Scaling começou a monitorar alterações nas políticas gerenciadas pela AWS.	19 de maio de 2021

Funções vinculadas 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. O perfil vinculado ao serviço é um tipo exclusivo de perfil do IAM vinculado diretamente ao 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 uma função vinculada ao serviço. Para obter mais informações, consulte [Usar funções vinculadas ao serviço](#) no Manual do usuário do IAM. As funções vinculadas ao serviço também permitem que todas as chamadas de API fiquem visíveis por meio do 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 [Registrar chamadas da API do Amazon EC2 Auto Scaling com o AWS CloudTrail \(p. 339\)](#).

As seções a seguir descrevem como criar e gerenciar funções vinculadas 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. Para ter mais informações, consulte [Usar funções vinculadas a serviço](#) no Guia do usuário do IAM.

Índice

- [Visão geral \(p. 432\)](#)
- [Permissões concedidas pela função vinculada ao serviço \(p. 433\)](#)
- [Criar uma função vinculada ao serviço \(automática\) \(p. 435\)](#)
- [Criar uma função vinculada ao serviço \(manual\) \(p. 436\)](#)
- [Editar a função vinculada ao serviço \(p. 437\)](#)
- [Excluir a função vinculada ao serviço \(p. 437\)](#)
- [Regiões compatíveis com funções vinculadas ao serviço do Amazon EC2 Auto Scaling \(p. 437\)](#)

Visão geral

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_**meu sufixo**.

As permissões de uma função vinculada ao serviço com sufixo personalizado são idênticas às da função vinculada 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 da função.

Você pode especificar uma dessas funções ao editar suas políticas de chaves do AWS Key Management Service para permitir que as instâncias iniciadas pelo Amazon EC2 Auto Scaling sejam criptografadas com sua CMK gerenciada pelo cliente. No entanto, se você planeja conceder acesso granular a uma determinada CMK gerenciada pelo cliente, você deverá usar uma função vinculada ao serviço com sufixo personalizado. O uso de uma função vinculada 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 seu 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 uma função vinculada ao serviço com sufixo personalizado:

1. Crie uma função vinculada ao serviço com um sufixo personalizado. Para obter mais informações, consulte [Criar uma função vinculada ao serviço \(manual\) \(p. 436\)](#).
2. Conceda à função vinculada 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 uma função vinculada ao serviço, consulte [Política de chaves do AWS KMS obrigatórias para uso com volumes criptografados \(p. 450\)](#).
3. Dê aos usuários acesso à função vinculada 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 \(usandoPassRole\) \(p. 441\)](#). Se os usuários tentarem especificar uma função vinculada ao serviço sem permissão para passar essa função para o serviço, eles receberão um erro.

Permissões concedidas pela função vinculada ao serviço

O Amazon EC2 Auto Scaling usa a função vinculada ao serviço chamada `AWSServiceRoleForAutoScaling` ou sua função vinculada ao serviço com sufixo personalizado.

A função vinculada ao serviço confia no seguinte serviço para assumir a função:

- `autoscaling.amazonaws.com`

A função usa a política [AutoScalingServiceRolePolicy \(p. 429\)](#), que inclui as seguintes permissões:

```
{  
    "Version": "2012-10-17",  
    "Statement": [  
        {  
            "Sid": "EC2InstanceManagement",  
            "Effect": "Allow",  
            "Action": [  
                "ec2:AttachClassicLinkVpc",  
                "ec2:CancelSpotInstanceRequests",  
                "ec2>CreateFleet",  
                "ec2>CreateTags",  
                "ec2>DeleteTags",  
                "ec2:Describe*",  
                "ec2:DetachClassicLinkVpc",  
                "ec2:ModifyInstanceAttribute",  
                "ec2:RequestSpotInstances",  
                "ec2:RunInstances",  
                "ec2:StartInstances",  
                "ec2:StopInstances",  
                "ec2:TerminateInstances"  
            ],  
            "Resource": "*"  
        },  
        {  
            "Sid": "EC2InstanceProfileManagement",  
            "Effect": "Allow",  
            "Action": [  
                "iam:PassRole"  
            ],  
            "Resource": "*"  
        }  
    ]  
}
```

```
"Condition":{  
    "StringLike":{  
        "iam:PassedToService":"ec2.amazonaws.com*"  
    }  
}  
,  
{  
    "Sid":"EC2SpotManagement",  
    "Effect":"Allow",  
    "Action": [  
        "iam:CreateServiceLinkedRole"  
    ],  
    "Resource": "*",  
    "Condition": {  
        "StringEquals": {  
            "iam:AWSServiceName": "spot.amazonaws.com"  
        }  
    }  
,  
{  
    "Sid":"ELBManagement",  
    "Effect":"Allow",  
    "Action": [  
        "elasticloadbalancing:Register*",  
        "elasticloadbalancing:Deregister*",  
        "elasticloadbalancing:Describe"  
    ],  
    "Resource": "*"  
},  
{  
    "Sid":"CWManagement",  
    "Effect":"Allow",  
    "Action": [  
        "cloudwatch:DeleteAlarms",  
        "cloudwatch:DescribeAlarms",  
        "cloudwatch:GetMetricData",  
        "cloudwatch:PutMetricAlarm"  
    ],  
    "Resource": "*"  
},  
{  
    "Sid":"SNSManagement",  
    "Effect":"Allow",  
    "Action": [  
        "sns:Publish"  
    ],  
    "Resource": "*"  
},  
{  
    "Sid":"EventBridgeRuleManagement",  
    "Effect":"Allow",  
    "Action": [  
        "events:PutRule",  
        "events:PutTargets",  
        "events:RemoveTargets",  
        "events:DeleteRule",  
        "events:DescribeRule"  
    ],  
    "Resource": "*",  
    "Condition": {  
        "StringEquals": {  
            "events:ManagedBy": "autoscaling.amazonaws.com"  
        }  
    }  
,  
{
```

```
"Sid":"SystemsManagerParameterManagement",
"Effect":"Allow",
"Action":[
    "ssm:GetParameters"
],
"Resource": "*"
},
{
    "Sid":"VpcLatticeManagement",
    "Effect":"Allow",
    "Action":[
        "vpc-lattice:DeregisterTargets",
        "vpc-lattice:GetTargetGroup",
        "vpc-lattice>ListTargets",
        "vpc-lattice>ListTargetGroups",
        "vpc-lattice:RegisterTargets"
    ],
    "Resource": "*"
}
]
```

A função tem permissões para fazer o seguinte:

- **ec2**— Crie, descreva, modifique, inicie/interrompa e encerre instâncias do EC2.
- **iam**—[Passe funções do IAM \(p. 448\)](#) às instâncias do EC2 para que os aplicativos em execução nas instâncias possam acessar credenciais temporárias para a função.
- **iam**— Crie o AWSServiceRoleForEC2Spot função vinculada ao serviço para permitir que o Amazon EC2 Auto Scaling lance instâncias spot em seu nome.
- **elasticloadbalancing**— Registre e cancele o registro de instâncias com o Elastic Load Balancing e verifique a integridade dos alvos registrados.
- **cloudwatch**— Criar, descrever, modificar e excluir CloudWatch Alarms para políticas de escalonamento e métricas de recuperação usadas para escalabilidade preditiva.
- **sns**— Publique notificações no Amazon SNS quando as instâncias são iniciadas ou encerradas.
- **events**— Criar, descrever, atualizar e excluir EventBridge regras em seu nome.
- **ssm**— Leia os parâmetros do Parameter Store ao usar um parâmetro do Systems Manager como alias para uma ID de AMI em um modelo de execução.
- **vpc-lattice**— Registre e cancele o registro de instâncias com o VPC Lattice e verifique a integridade dos alvos registrados.

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.

Important

Você deve ter permissões do IAM para criar a função vinculada ao serviço. Caso contrário, a criação automática falhará. Para obter mais informações, consulte [Permissões de função vinculada ao serviço](#) no Manual do usuário do IAM e [Criar uma função vinculada a serviço \(p. 441\)](#) neste guia.

O Amazon EC2 Auto Scaling começou a oferecer suporte a funções vinculadas ao serviço em março de 2018. Se você criou um grupo do Auto Scaling antes disso, o Amazon EC2 Auto Scaling criou a função AWSServiceRoleForAutoScaling em sua conta. Para obter mais informações, consulte [Uma nova função surgiu em minha Conta da AWS](#) no Manual do usuário do IAM.

Criar uma função vinculada ao serviço (manual)

Para criar uma função vinculada ao serviço (console)

1. Abra o console do IAM em <https://console.aws.amazon.com/iam/>.
2. No painel de navegação, escolha Roles e depois Create Role.
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 essa função), 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 funções vinculadas ao serviço durante a criação.
6. Na página Review (Revisão), deixe em branco Role name (Nome da função) para criar uma função vinculada ao serviço com o nome AWSServiceRoleForAutoScaling ou insira um sufixo para criar uma função vinculada ao serviço com o nome AWSServiceRoleForAutoScaling_**sufixo**.
7. (Opcional) Em Role description (Descrição da função), edite a descrição para a função vinculada ao serviço.
8. Selecione Create role (Criar função).

Para criar uma função vinculada a serviço (AWS CLI)

Use o seguinte [create-service-linked-role](#) Comando CLI para criar uma função vinculada ao serviço para o Amazon EC2 Auto Scaling com o nome AWSServiceRoleForAutoScaling_**sufixo**.

```
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_sufix",  
        "Arn": "arn:aws:iam::123456789012:role/aws-service-role/autoscaling.amazonaws.com/AWSServiceRoleForAutoScaling_sufix",  
        "Path": "/aws-service-role/autoscaling.amazonaws.com/",  
        "AssumeRolePolicyDocument": {  
            "Version": "2012-10-17",  
            "Statement": [  
                {  
                    "Action": [  
                        "sts:AssumeRole"  
                    ],  
                    "Principal": {  
                        "Service": [  
                            "autoscaling.amazonaws.com"  
                        ]  
                    },  
                    "Effect": "Allow"  
                }  
            ]  
        }  
    }  
}
```

Para obter mais informações, consulte [Criação de uma função vinculada ao serviço](#) no Manual do usuário do IAM.

Editar a função vinculada ao serviço

Você não pode editar as funções vinculadas ao serviço criadas para o Amazon EC2 Auto Scaling. Depois de criar uma função vinculada ao serviço, você não pode alterar o nome da função ou suas permissões. No entanto, você poderá editar a descrição da função. Para obter mais informações, consulte [Editar uma função vinculada a serviço](#) no Manual do usuário do IAM.

Note

Na AWS, não se recomenda modificar funções vinculadas ao serviço porque isso pode causar problemas de segurança, como [confused deputy](#) entre serviços.

Excluir a função vinculada ao serviço

Se você não estiver usando um grupo do Auto Scaling, recomendamos excluir a função vinculada ao serviço. Excluir a função evita que você tenha uma entidade que não é usada ou mantida e monitorada ativamente.

Você poderá excluir uma função vinculada 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 uma função vinculada ao serviço é usada com vários grupos do Auto Scaling, você deve excluir todos os grupos do Auto Scaling que usam a função vinculada ao serviço antes de excluí-la. Para obter mais informações, consulte [Excluir infraestrutura do Auto Scaling](#) (p. 146).

É possível usar o IAM para excluir uma função vinculada ao serviço. Para ter 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.

Regiões compatíveis com funções vinculadas ao serviço do Amazon EC2 Auto Scaling

O Amazon EC2 Auto Scaling é compatível com perfis vinculados ao serviço em todas as regiões da Regiões da AWS em que o serviço está disponível.

Exemplos de políticas baseadas em identidade do Amazon EC2 Auto Scaling

Por padrão, um usuário totalmente novo na sua Conta da AWS não 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.

```
{  
    "Version": "2012-10-17",  
    "Statement": [{  
        "Effect": "Allow",  
        "Action": [  
            "autoscaling>CreateAutoScalingGroup",
```

```
        "autoscaling:UpdateAutoScalingGroup",
        "autoscaling>DeleteAutoScalingGroup"
    ],
    "Resource": "*",
    "Condition": {
        "StringEquals": { "autoscaling:ResourceTag/purpose": "test" }
    }
},
{
    "Effect": "Allow",
    "Action": "autoscaling:Describe*",
    "Resource": "*"
}]
}
```

Esse exemplo de política concede permissões para criar, modificar e excluir grupos do Auto Scaling, mas somente se o grupo usar a etiqueta **purpose=test**. Como as ações `Describe` não oferecem suporte a permissões em nível de recurso, é necessário especificá-las em uma declaração separada sem condições.

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. É possível restringir os recursos de destino para essas ações especificando ARNs de grupos do Auto Scaling individuais. 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.

Tópicos

- [Controle o tamanho dos grupos do Auto Scaling que podem ser criados \(p. 438\)](#)
- [Controlar quais chaves de tag e valores de tag podem ser usados \(p. 439\)](#)
- [Controle quais grupos do Auto Scaling podem ser excluídos \(p. 440\)](#)
- [Controlar quais políticas de escalabilidade podem ser excluídas \(p. 440\)](#)
- [Criar uma função vinculada a serviço \(p. 441\)](#)
- [Controle qual função vinculada ao serviço pode ser passada \(usandoPassRole\) \(p. 441\)](#)

Controle o tamanho dos 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.

```
{
    "Version": "2012-10-17",
    "Statement": [
        {
            "Effect": "Allow",
            "Action": [
                "autoscaling>CreateAutoScalingGroup",
                "autoscaling:UpdateAutoScalingGroup"
            ],
            "Resource": "*",
            "Condition": {
                "StringEquals": { "autoscaling:ResourceTag/environment": "development" },
                "NumericLessThan": { "aws:MaxSize": 10 },
                "NumericGreaterThan": { "aws:MinSize": 1 }
            }
        }
    ]
}
```

```
        "NumericGreaterThanOrEqualIfExists": { "autoscaling:MinSize": 1 },
        "NumericLessThanOrEqualIfExists": { "autoscaling:MaxSize": 10 }
    }]
}
```

Controlar quais chaves de tag e valores de tag podem ser usados

Você também pode usar condições em suas políticas do IAM para controlar as chaves e os valores das tags que podem ser aplicados aos grupos de 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`.

```
{
    "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).

```
{
    "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**.

```
{
```

```
"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.

Controle quais grupos do Auto Scaling podem ser excluídos

Se você não estiver usando as teclas de condição para controlar o acesso aos grupos do Auto Scaling, poderá especificar os ARNs dos recursos na `Resource` elemento para controlar o acesso em vez disso.

A política a seguir concede aos usuários permissões para usar a `DeleteAutoScalingGroup` ação de API, mas somente para grupos de Auto Scaling cujo nome começa com `devteam-`.

```
{ "Version": "2012-10-17",
  "Statement": [{
      "Effect": "Allow",
      "Action": "autoscaling:DeleteAutoScalingGroup",
      "Resource": "arn:aws:autoscaling:region:account-
id:autoScalingGroup:*:autoScalingGroupName/devteam-*"
  }]
}
```

Também possível especificar vários ARNs incluindo-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"
]
```

Para obter mais informações sobre a especificação dos ARNs dos recursos do Amazon EC2 Auto Scaling no `Resource` elemento, veja [Recursos de política para o Amazon EC2 Auto Scaling \(p. 424\)](#).

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`.

```
{  
    "Version": "2012-10-17",  
    "Statement": [{  
        "Effect": "Allow",  
        "Action": "autoscaling:DeletePolicy",  
        "Resource": "*"  
    },  
    {  
        "Effect": "Deny",  
        "Action": "autoscaling:DeletePolicy",  
        "Resource": "*",  
        "Condition": {  
            "StringEquals": { "autoscaling:ResourceTag/environment": "production" }  
        }  
    }]  
}
```

Criar uma função vinculada a serviço

O Amazon EC2 Auto Scaling requer permissões para criar uma função vinculada ao serviço na primeira vez que qualquer usuário em sua Conta da AWS chama as ações de 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. O perfil vinculado ao serviço concede permissões para que o Amazon EC2 Auto Scaling possa chamar outros Serviços da AWS em seu nome.

Para que a criação automática da função 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 uma função vinculada ao serviço do Amazon EC2 Auto Scaling para o Amazon EC2 Auto Scaling.

```
{  
    "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 (usandoPassRole)

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 permissão `iam:PassRole` para proteger a segurança de suas chaves gerenciadas pelo cliente do AWS KMS se fornecer acesso a diferentes chaves para diferentes perfis vinculados ao serviço. 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 aprovar oAWSServiceRoleForAutoScaling_devteamkeyaccess função para qualquer grupo de 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.

```
{  
    "Version": "2012-10-17",  
    "Statement": [{  
        "Effect": "Allow",  
        "Action": "iam:PassRole",  
        "Resource": "arn:aws:iam::account-id:role/aws-service-role/  
autoscaling.amazonaws.com/AWSServiceRoleForAutoScaling_devteamkeyaccess",  
        "Condition": {  
            "StringEquals": { "iam:PassedToService": [ "autoscaling.amazonaws.com" ] },  
            "StringLike": { "iam:AssociatedResourceARN":  
[ "arn:aws:autoscaling:region:account-  
id:autoScalingGroup:*:autoScalingGroupName/devteam-*" ] }  
        }  
    }]  
}
```

Para obter mais informações sobre funções vinculadas ao serviço com sufixo personalizado, consulte [Funções vinculadas ao serviço do Amazon EC2 Auto Scaling \(p. 432\)](#).

Prevenção do problema do substituto confuso entre serviços

O problema de "confused deputy" é uma questão de segurança em que uma entidade que não tem permissão para executar uma ação pode coagir uma entidade mais privilegiada a executá-la.

Em AWS, a personificação entre serviços pode resultar no problema do 'confused deputy'. 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, o AWS fornece ferramentas que ajudam você a proteger seus dados para todos os serviços com entidades principais de serviço 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 para oSourceArneSourceAccountos campos são definidos quando o Amazon EC2 Auto Scaling usaAWS Security Token Service(AWS STS) 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 \(p. 428\)](#) é uma função que um serviço assume para realizar ações em seu nome. Nos casos em que você deseja criar ganchos de ciclo de vida que enviam notificações para qualquer lugar que não seja a AmazonEventBridge, 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.

```
{  
    "Version": "2012-10-17",  
    "Statement": {  
        "Sid": "ConfusedDeputyPreventionExamplePolicy",  
        "Effect": "Allow",  
        "Principal": {  
            "Service": "autoscaling.amazonaws.com"  
        },  
        "Action": "sts:AssumeRole",  
        "Condition": {  
            "ArnLike": {  
                "aws:SourceArn":  
                    "arn:aws:autoscaling:region:account_id:autoScalingGroup:uuid:autoScalingGroupName/my-asg"  
            },  
            "StringEquals": {  
                "aws:SourceAccount": "account_id"  
            }  
        }  
    }  
}
```

No exemplo anterior:

- O Principal elemento especifica o principal de serviço do serviço (autoscaling.amazonaws.com).
- O Action elemento especifica sts:AssumeRole.
- O Condition elemento especifica oaws:SourceArn e aws:SourceAccount chaves de condição global. O ARN da fonte inclui o ID da conta, portanto, não é necessário usar aws:SourceAccount com aws:SourceArn.

Informações adicionais

Para obter mais informações, consulte [Chaves de contexto de condição globais da AWS](#), [O problema de confused deputy](#) e [Modificar a política de confiança de uma função \(console\)](#) no Manual do usuário do IAM.

Suporte a modelo de execução

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, modificar e excluir todos os grupos do Auto Scaling, mas somente se o grupo usa a etiqueta `purpose=test`. 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`.

```
{  
    "Version": "2012-10-17",  
    "Statement": [  
        {  
            "Effect": "Allow",  
            "Action": [  
                "autoscaling>CreateAutoScalingGroup",  
                "autoscaling:UpdateAutoScalingGroup",  
                "autoscaling>DeleteAutoScalingGroup"  
            ],  
            "Resource": "*",  
            "Condition": {  
                "StringEquals": { "autoscaling:ResourceTag/purpose": "test" }  
            }  
        },  
        {  
            "Effect": "Allow",  
            "Action": [  
                "autoscaling:Describe*",  
                "ec2:RunInstances"  
            ],  
            "Resource": "*"  
        }  
    ]  
}
```

`ec2:RunInstances` é verificado quando um grupo do Auto Scaling é criado ou atualizado usando um modelo de execução. Use este exemplo de política para criar sua própria política que limite as permissões `ec2:RunInstances`, por exemplo, exigindo modelos de execução com etiquetas específicas.

Os exemplos a seguir mostram declarações de políticas que você pode usar para controlar as permissões que os usuários têm ao usar políticas de lançamento.

Tópicos

- [Exigir modelos de execução que têm uma tag específica \(p. 445\)](#)
- [Exigir um modelo de execução e um número de versão \(p. 445\)](#)
- [Exigir o uso do Instance Metadata Service Version 2 \(IMDSv2\) \(p. 446\)](#)
- [Restringir o acesso aos recursos do Amazon EC2 \(p. 446\)](#)

- [Permissões necessárias para marcar instâncias e volumes \(p. 447\)](#)
- [Permissões adicionais \(p. 447\)](#)
- [Validação de permissões para ec2:RunInstances com o IAM:PassRole \(p. 448\)](#)

Exigir modelos de execução que têm uma tag específica

O exemplo a seguir restringe o acesso à chamada da ação `ec2:RunInstances` com modelos de execução que estão localizados na região especificada e que têm a tag `purpose=test`.

```
{  
    "Version": "2012-10-17",  
    "Statement": [  
        {  
            "Effect": "Allow",  
            "Action": "ec2:RunInstances",  
            "Resource": "arn:aws:ec2:region:account-id:launch-template/*",  
            "Condition": {  
                "StringEquals": { "ec2:ResourceTag/purpose": "test" }  
            }  
        }  
    ]  
}
```

Exigir um modelo de execução e um número de versão

O exemplo a seguir permite que os usuários criem e modifiquem grupos do Auto Scaling se especificarem o número da versão do modelo de execução e, em seguida, nega permissão para criar ou modificar grupos do Auto Scaling usando uma configuração de execução. Se usuários com essa política omitirem o número da versão para especificar a versão `$Latest` ou `$Default` do modelo de execução ou especificarem uma configuração de execução, a ação falhará.

```
{  
    "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 Instance Metadata Service Version 2 (IMDSv2)

Para segurança adicional, é possível definir as permissões dos usuários para exigir o uso de um modelo de execução que exige IMDSv2. Para obter mais informações, consulte [Configuração do serviço de metadados de instância](#) no Manual do usuário do Amazon EC2 para instâncias do Linux.

O exemplo de política a seguir especifica que os usuários não poderão chamar a ação ec2:RunInstances a menos que a instância também esteja configurada para exigir o uso de IMDSv2 (indicado por "ec2:MetadataHttpTokens": "required").

```
{  
    "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 Auto Scaling que iniciam um novo modelo de execução ou uma nova versão de um modelo de execução com as opções de metadados de instância configuradas, é possível terminar instâncias existentes no grupo. O Amazon EC2 Auto Scaling inicia imediatamente novas instâncias para substituir as instâncias que você terminou. Como alternativa, você pode iniciar uma atualização de instância para fazer uma atualização contínua do seu grupo. Para obter mais informações, consulte [Substituir instâncias do Auto Scaling com base em uma atualização de instância \(p. 108\)](#).

Restringir o acesso aos recursos do Amazon EC2

O exemplo a seguir controla a configuração das instâncias que um usuário pode iniciar restringindo o acesso aos recursos do Amazon EC2.

```
{  
    "Version": "2012-10-17",  
    "Statement": [  
        {  
            "Effect": "Allow",  
            "Action": "ec2:RunInstances",  
            "Resource": [  
                "arn:aws:ec2:region:account-id:subnet/subnet-1a2b3c4d",  
                "arn:aws:ec2:region:account-id:security-group/sg-903004f88example",  
                "arn:aws:ec2:region:account-id:network-interface/*",  
                "arn:aws:ec2:region:account-id:volume/*",  
                "arn:aws:ec2:region::image/ami-04d5cc9b88example"  
            ]  
        },  
        {  
            "Effect": "Allow",  
            "Action": "ec2:RunInstances",  
            "Resource": "arn:aws:ec2:region:account-id:instance/*",  
            "Condition": {  
                "StringEquals": { "ec2:InstanceType": "t2.micro" }  
            }  
        }  
    ]  
}
```

```
        }
    ]
}
```

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. Isso também concede aos usuários acesso a recursos adicionais necessários para executar instâncias: interfaces de rede e volumes.
- A segunda declaração permite que os usuários executem instâncias somente de um tipo de instância específico (**t2.micro**).

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.

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": "ec2:CreateTags",
      "Resource": "arn:aws:ec2:region:account-id:/*/*",
      "Condition": {
        "StringEquals": { "ec2:CreateAction": "RunInstances" }
      }
    }
  ]
}
```

Permissões adicionais

Dependendo de a quais cenários você deseja oferecer suporte, é possível especificar essas ações adicionais no elemento Action de uma declaração de política do IAM.

Você deve 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.

Você também deve decidir quem deve ter permissões para criar, modificar, descrever e excluir modelos de execução e versões de modelos de execução. Para obter mais informações, consulte [Controlar o uso de modelos de execução](#) e [Exemplo: trabalhar com modelos de execução](#) no Guia do usuário do Amazon EC2 para instâncias do Linux.

Além das ações do Amazon EC2, os usuários que criam ou atualizam grupos do Auto Scaling com um modelo de execução que especifica um perfil de instância (um contêiner para um perfil do IAM) precisam da permissão `iam:PassRole`. Para obter mais informações e um exemplo de política do IAM, consulte [Funções do IAM para aplicações que são executadas em instâncias do Amazon EC2 \(p. 448\)](#).

Os usuários que criam ou atualizam grupos do Auto Scaling com um modelo de execução que usa um parâmetro do AWS Systems Manager precisam da permissão `ssm:GetParameters`. Para obter mais

informações, consulte [Usar parâmetros do AWS Systems Manager em vez de IDs de AMI em modelos de execução \(p. 44\)](#).

Validação de permissões para `aec2:RunInstances` e `iam:PassRole`

Os usuários podem especificar qual versão de um modelo de lançamento seu grupo de Auto Scaling usa. Dependendo de suas permissões, essa pode ser uma versão numerada específica ou a `$Latest` ou `$Default` versão do modelo de lançamento. Se for o último, tome cuidado especial. Isso pode anular as permissões de `ec2:RunInstances` que você pretendia restringir.

Esta seção explica o cenário de usar a versão mais recente ou padrão do modelo de lançamento com um grupo de Auto Scaling.

Quando um usuário liga para o `CreateAutoScalingGroup`, `UpdateAutoScalingGroup`, `StartInstanceRefresh` APIs, o Amazon EC2 Auto Scaling verifica suas permissões em relação à versão do modelo de execução que é a versão mais recente ou padrão naquele momento antes de prosseguir com a solicitação. Isso valida as permissões para que as ações sejam concluídas ao iniciar instâncias, como `aec2:RunInstances`. Para fazer isso, emitimos um Amazon EC2 [RunInstances](#) chamada de execução seca 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, o Amazon EC2 Auto Scaling falhará na solicitação e retornará um erro ao usuário contendo informações sobre a permissão ausente.

Depois que a verificação e a solicitação iniciais são concluídas, sempre que as instâncias são iniciadas, o Amazon EC2 Auto Scaling as inicia com a versão mais recente ou padrão, mesmo que ela tenha sido alterada, usando as permissões de seu [função vinculada ao serviço \(p. 433\)](#). Isso significa que um usuário que está usando o modelo de execução pode potencialmente atualizá-lo para passar uma função do IAM para uma instância, mesmo que não tenha a `iam:PassRole` permissão.

Use o `autoScaling:LaunchTemplateVersionSpecified` chave de condição se você quiser limitar quem tem acesso à configuração de grupos para usar o `$Latest` ou `$Default` versão. Isso garante que o grupo Auto Scaling só aceite uma versão numerada específica quando um usuário chama `CreateAutoScalingGroup` ou `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 \(p. 445\)](#).

Para grupos de Auto Scaling configurados para usar o `$Latest` ou `$Default` versão do modelo de lançamento, considere limitar quem pode criar e gerenciar versões do modelo de lançamento, incluindo `ec2:ModifyLaunchTemplate` ação que permite que um usuário especifique a versão padrão do modelo de inicialização. Para obter mais informações, consulte [Controlar o uso de modelos de execução e Exemplo: trabalhar com modelos de execução](#) no Guia do usuário do Amazon EC2 para instâncias do Linux.

Funções do IAM para aplicações que são executadas em instâncias do Amazon EC2

Aplicativos executados em instâncias do Amazon EC2 precisam de credenciais para acessar outros Serviços da AWS. Para fornecer essas credenciais de uma maneira segura, use uma função 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 uma função do IAM que permite ao Amazon EC2 passar a função do IAM para

uma instância quando ela é iniciada. Primeiro, crie uma função do IAM que tenha todas as permissões necessárias para acessar os recursos da AWS. Depois, crie o perfil da instância e atribua a função a ele.

Note

Como melhor prática, é altamente recomendável criar a função para que ela tenha as permissões mínimas para outros Serviços da AWS que sua aplicação exige.

Índice

- [Pré-requisitos \(p. 449\)](#)
- [Criar um modelo de execução \(p. 450\)](#)
- [Consulte também \(p. 450\)](#)

Pré-requisitos

Crie a função 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.

Quando você usa o console do IAM em vez da AWS CLI ou um dos AWS SDKs, o console cria automaticamente um perfil de instância e atribui a ele o mesmo nome da função correspondente.

Para criar uma função 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 Create role (Criar função).
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 Next (próximo).
9. Na página Name, review, and create (Nomear, revisar e criar), em Role name (Nome da função), insira um nome de função para ajudar você a identificar a finalidade dessa função. Esse nome deve ser exclusivo em sua Conta da AWS. Como outros recursos da AWS podem fazer referência à função, não é possível editar o nome da função depois que ela é criada.
10. Reveja a função e escolha Create role (Criar função).

Permissões do IAM

Use uma política baseada em identidade do IAM para controlar o acesso ao novo perfil do IAM. Oiam:PassRoleé necessária permissão na identidade do IAM (usuário ou função) que cria ou atualiza um grupo de 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-**.

{

```
"Version": "2012-10-17",
"Statement": [
    {
        "Effect": "Allow",
        "Action": "iam:PassRole",
        "Resource": "arn:aws:iam::account-id: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 as permissões para `iam:PassRole`ação para um grupo de Auto Scaling que usa um modelo de lançamento, consulte [Validação de permissões para `ec2:RunInstances` e `iam:PassRole` \(p. 448\)](#).

Criar um modelo de execução

Ao criar o modelo de execução usando o AWS Management Console, na seção Advanced Details (Detalhes avançados), selecione a função no IAM instance profile (Perfil de instância do IAM). Para obter mais informações, consulte [Definir configurações avançadas para seu modelo de execução \(p. 29\)](#).

Quando você cria o modelo de lançamento usando o `create-launch-template` comando 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 funções do IAM para Amazon EC2, consulte:

- [Funções do IAM para Amazon EC2](#) no Manual do usuário do Amazon EC2 para instâncias do Linux
- [Uso de perfis de instância](#) e [Uso de funções do IAM para conceder permissões a aplicações em execução nas instâncias do Amazon EC2](#) no Manual do usuário do IAM

Política de chaves do AWS KMS obrigatorias para uso com volumes criptografados

O Amazon EC2 Auto Scaling usa [funções vinculadas ao serviço \(p. 432\)](#) para delegar permissões para outros Serviços da AWS. As funções vinculadas ao serviço do Amazon EC2 Auto Scaling são predefinidas e incluem as permissões que o Amazon EC2 Auto Scaling requer para chamar outros Serviços da AWS em seu nome. As permissões predefinidas também incluem acesso a suas 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.

Índice

- [Visão geral \(p. 451\)](#)
- [Configurar políticas de chave \(p. 451\)](#)
- [Exemplo 1: seções da política de chaves que permitem acesso à chave gerenciada pelo cliente \(p. 452\)](#)
- [Exemplo 2: seções da política de chaves que permitem acesso entre contas à chave gerenciada pelo cliente \(p. 453\)](#)
- [Editar políticas de chaves no console do AWS KMS \(p. 454\)](#)

Visão geral

A seguinte AWS KMS keys pode ser usada para 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. A Chave gerenciada pela AWS será usada 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. Ela inclui permissões para executar as operações Encrypt, Decrypt, ReEncrypt*, GenerateDataKey* e DescribeKey do AWS KMS na chave.
- A segunda declaração permite que a identidade do IAM especificada no elemento Principal use concessões para delegar um subconjunto de suas próprias permissões aos Serviços da AWS integrados com o AWS KMS ou outra entidade 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.

Em cada um dos exemplos a seguir, os argumentos que devem ser substituídos, como um ID de chave ou o nome de uma função vinculada ao serviço, são mostrados como *texto substituível em itálico*. Na maioria dos casos, você pode substituir o nome da função vinculada ao serviço pelo nome de uma função vinculada ao serviço do Amazon EC2 Auto Scaling.

Para mais informações, consulte os seguintes recursos do :

- Para criar uma chave com a AWS CLI, consulte [create-key](#).
- Para atualizar uma política de chave com oAWS CLI, veja[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 as funções vinculadas ao serviço do Amazon EC2 Auto Scaling, consulte [Funções vinculadas ao serviço do Amazon EC2 Auto Scaling \(p. 432\)](#).
- Para obter mais informações sobre o Amazon EBS e o KMS, consulte [Criptografia do Amazon EBS](#) no Manual do usuário do Amazon EC2 para instâncias do Linux e o [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 instruções de política a seguir à política de chave da chave gerenciada pelo cliente, substituindo o ARN de exemplo pelo ARN da função vinculada ao serviço apropriada que tem acesso permitido 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, substituindo **111122223333** pelo ID da conta efetiva da Conta da AWS na qual você deseja criar o grupo do Auto Scaling. Isso permite que você conceda permissão para que um usuário ou uma função 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 da função vinculada a serviço apropriada. O key-id é o ARN da chave.

Veja a seguir um exemplo de comando [create-grant](#) da CLI que concede à função vinculada a serviço chamada AWSServiceRoleForAutoScaling na conta [111122223333](#) 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 a seguir de política do IAM permite que uma identidade do IAM (usuário ou perfil) na conta [111122223333](#) crie uma concessão para a chave gerenciada pelo cliente na conta [444455556666](#).

```
{
    "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 da função vinculada 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 a função vinculada ao serviço.

Editar políticas de chaves no console do AWS KMS

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 chave é usar a visualização padrão do console do AWS KMS para políticas de chave e tornar uma identidade do IAM (usuário ou perfil) um dos usuários da chave para política de chave apropriada. Para obter mais informações, consulte [Uso da visualização padrão do AWS Management Console](#) no Guia do desenvolvedor do AWS Key Management Service.

Important

Tenha cuidado. As declarações de política da visualização padrão do console incluem permissões para executar operações Revoke do AWS KMS na chave gerenciada pelo cliente. Se você conceder acesso a uma chave gerenciada pelo cliente para uma conta da Conta da AWS em sua

conta e accidentalmente revogar a concessão que concedia essa permissão à conta, os usuários externos não poderão mais acessar os dados criptografados ou a chave que foi usada para criptografar os dados.

Validação de compatibilidade do Amazon EC2 Auto Scaling

Para saber se um AWS service (Serviço da AWS) está no escopo de programas de conformidade específicos, consulte [Serviços da AWS no escopo por programa de conformidade](#) e selecione o programa de conformidade em que você está interessado. Para obter informações gerais, consulte [Programas de conformidade da AWS](#).

É possível fazer download de relatórios de auditoria de terceiros usando o AWS Artifact. Para obter mais informações, consulte [Downloading Reports in AWS Artifact](#).

Sua responsabilidade de conformidade ao usar o Serviços da AWS é determinada pela confidencialidade dos seus dados, pelos objetivos de conformidade da sua empresa e pelos regulamentos e leis aplicáveis. A AWS fornece os seguintes recursos para ajudar com a conformidade:

- [Guias de início rápido de segurança e conformidade](#): estes guias de implantação discutem considerações sobre arquitetura e fornecem as etapas para a implantação de ambientes de linha de base focados em segurança e conformidade na AWS.
- [Architecting for HIPAA Security and Compliance on Amazon Web Services](#) (Arquitetura para segurança e conformidade com HIPAA no Amazon Web Services): esse whitepaper descreve como as empresas podem usar a AWS para criar aplicações adequadas aos padrões HIPAA.

Note

Nem todos os Serviços da AWS estão qualificados pela HIPAA. Para mais informações, consulte a [Referência dos serviços qualificados pela HIPAA](#).

- [Recursos de conformidade da AWS](#): essa coleção de manuais e guias pode ser aplicada a seu setor e local.
- [Avaliar recursos com regras](#) no AWS Config Developer Guide (Guia do desenvolvedor do CCI): o serviço AWS Config avalia como as configurações de recursos estão em conformidade com práticas internas, diretrizes do setor e regulamentos.
- [AWS Security Hub](#): este AWS service (Serviço da AWS) fornece uma visão abrangente do seu estado de segurança na AWS. O Security Hub usa controles de segurança para avaliar os recursos da AWS e verificar a conformidade com os padrões e as práticas recomendadas do setor de segurança. Para obter uma lista dos serviços e controles aceitos, consulte a [Referência de controles do Security Hub](#).
- [AWS Audit Manager](#): esse AWS service (Serviço da AWS) ajuda a auditar continuamente seu uso da AWS para simplificar a forma como você gerencia os riscos e a conformidade com regulamentos e padrões do setor.

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 pacote de conformidade com o PCI da AWS, consulte [Nível 1 do PCI DSS](#).

Para obter informações sobre como alcançar a compatibilidade com o PCI DSS para suas workloads da AWS, consulte o seguinte guia de compatibilidade:

- [Payment Card Industry Data Security Standard \(PCI DSS\) 3.2.1 na AWS](#)

Resiliência no Amazon EC2 Auto Scaling

A infraestrutura global da AWS se baseia em Regiões da AWS e zonas de disponibilidade. A Regiões da AWS oferece várias zonas de disponibilidade separadas e isoladas fisicamente que são conectadas com baixa latência, throughputs elevados e em redes altamente redundantes. 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 mais altamente disponíveis, tolerantes a falhas e escaláveis que uma ou várias infraestruturas de data center tradicionais.

Para obter mais informações sobre Regiões da AWS e zonas de disponibilidade, consulte [Infraestrutura globalAWS](#).

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 aplicativos:

- 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 oferecem suporte a políticas de escalabilidade e pausa as atividades de escalabilidade quando métricas confiáveis não estão disponíveis, como quando pontos de dados estão ausentes.
- 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 uma zona de disponibilidade diferente que está habilitada 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 executar 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平衡adores de carga associados ao seu grupo do Auto Scaling. Dessa forma, você não precisa registrar e cancelar o registro de instâncias separadamente.

Para obter informações sobre os recursos para ajudar a suportar suas necessidades de resiliência e backup de dados do Amazon EC2, consulte [Resiliência no Amazon EC2](#) no Guia do usuário do Amazon EC2 para instâncias do Linux.

Segurança da infraestrutura no Amazon EC2 Auto Scaling

Como um serviço gerenciado, o Amazon EC2 Auto Scaling é protegido por AWS Segurança de rede global. Para obter informações sobre serviços de segurança da AWS e como a AWS protege a infraestrutura, consulte [Segurança na Nuvem AWS](#). Para projetar seu ambiente da AWS usando as práticas recomendadas de segurança de infraestrutura, consulte [Proteção de infraestrutura](#) em Pilar segurança: AWS Well-Architected Framework.

Você usa chamadas à API publicadas pela AWS para acessar o Amazon EC2 Auto Scaling via rede. Os clientes devem oferecer suporte para:

- 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.

Além disso, as solicitações devem ser assinadas usando um ID da chave de acesso e uma chave de acesso secreta associada a uma entidade principal do IAM. Ou você pode usar o [AWS Security Token Service](#) (AWS STS) para gerar credenciais de segurança temporárias para assinar solicitações.

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 ter mais informações, consulte [Amazon EC2 Auto Scaling e endpoints da VPC da interface \(p. 457\)](#).

Tópicos relacionados

- [Segurança da infraestrutura no Amazon EC2](#) no Manual do usuário do Amazon EC2 para instâncias do Linux

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 habilitados pela AWS PrivateLink, uma tecnologia que permite acessar APIs do Amazon EC2 Auto Scaling de maneira privada ao restringir todo o tráfego de rede entre sua VPC e o Amazon EC2 Auto Scaling à rede da 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 o AWS PrivateLink, mas é recomendável. Para obter mais informações sobre o AWS PrivateLink e endpoints da VPC, consulte [What is AWS PrivateLink?](#) (O que é o?) no AWS PrivateLink Guia.

Tópicos

- [Criar um VPC endpoint de interface \(p. 458\)](#)

- [Criar uma política de endpoint da VPC \(p. 458\)](#)

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 [Access an AWS using an interface VPC endpoint](#) (Acessar um por meio de um endpoint da VPC de interface) no AWS PrivateLink Guia.

Você não precisa alterar nenhuma configuração do Amazon EC2 Auto Scaling. O Amazon EC2 Auto Scaling chama outros serviços da AWS usando endpoints de serviço ou endpoints da VPC da 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 [VPC endpoint policies](#) (Políticas de endpoint da VPC) no AWS PrivateLink Guide (Guia do).

Solucionar problemas do Amazon EC2 Auto Scaling

O Amazon EC2 Auto Scaling fornece erros específicos e descriptivos 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 \(p. 459\)](#)
- [Recursos adicionais para solução de problemas \(p. 460\)](#)
- [Solucionar problemas do Amazon EC2 Auto Scaling: falhas ao iniciar instâncias do EC2 \(p. 461\)](#)
- [Solucionar problemas do Amazon EC2 Auto Scaling: problemas de AMI \(p. 467\)](#)
- [Solucionar problemas do Amazon EC2 Auto Scaling: problemas do平衡ador de carga \(p. 469\)](#)
- [Solucionar problemas do Amazon EC2 Auto Scaling: modelos de execução \(p. 471\)](#)
- [Solucionar problemas com as verificações de integridade do Amazon EC2 Auto Scaling \(p. 473\)](#)

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 dura 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-a6c42d8abaf 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 ações de dimensionamento para o grupo do Auto Scaling excluído, adicione a --include-deleted-groups opção ao [describe-scaling-activities](#) comando como descrito a seguir.

```
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"  
            },  
            ...  
        }  
    ]  
}
```

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 \(p. 215\)](#)
- [Visualizar grafos de monitoramento no console do Amazon EC2 Auto Scaling \(p. 328\)](#)
- [Verificações de integridade para instâncias do Auto Scaling \(p. 319\)](#)
- [Considerações e limitações dos ganchos do ciclo de vida \(p. 254\)](#)
- [Concluir uma ação do ciclo de vida \(p. 266\)](#)
- [Fornecer conectividade de rede para suas instâncias do Auto Scaling usando a Amazon VPC \(p. 414\)](#)
- [Remover temporariamente instâncias do grupo do Auto Scaling \(p. 308\)](#)
- [Desabilitar uma política de escalabilidade para um grupo do Auto Scaling \(p. 216\)](#)
- [Suspender e retomar um processo para um grupo do Auto Scaling \(p. 312\)](#)
- [Controlar quais instâncias do Auto Scaling serão terminadas durante uma redução de escala na horizontal \(p. 292\)](#)
- [Excluir infraestrutura do Auto Scaling \(p. 146\)](#)
- [Cotas do Amazon EC2 Auto Scaling \(p. 11\)](#)

Os seguintes recursos da AWS também podem ajudar:

- [Tópicos do Amazon EC2 Auto Scaling na Central de conhecimento da AWS](#)
- [Perguntas sobre o Amazon EC2 Auto Scaling no AWS re:Post](#)
- [Publicações sobre o Amazon EC2 Auto Scaling no AWS Compute Blog \(Blog da computação\)](#)
- [Solução de problemas CloudFormation no GuiaAWS CloudFormation do usuário](#)

Geralmente, a solução de problemas requer consulta e descoberta iterativas por um especialista ou de uma comunidade de ajudantes. Se continuar enfrentando problemas após tentar aplicar as sugestões desta seção, entre em contato com o AWS Support (no AWS Management Console, clique em Support [Suporte], Support Center [Central de suporte]) ou faça uma pergunta no [AWS re:Post](#) usando a etiqueta 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 \(p. 459\)](#).

Quando suas instâncias 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. \(p. 462\)](#)
- [O grupo de segurança <nome do grupo de segurança> não existe. Falha ao ativar a instância EC2. \(p. 462\)](#)
- [O par de chaves <par de chaves associado à sua instância do EC2> não existe. Falha ao ativar a instância EC2. \(p. 463\)](#)
- [A Zona de disponibilidade solicitada não é mais suportada. Tente sua solicitação novamente... \(p. 463\)](#)
- [O tipo de instância solicitado \(<tipo de instância>\) não tem suporte na Zona de disponibilidade solicitada \(<Zona de disponibilidade da instância>\)... \(p. 463\)](#)

- Seu preço de solicitação spot de 0,015 é inferior ao preço mínimo de atendimento de solicitação spot exigido de 0,0735... (p. 464)
- Nome de dispositivo inválido <nome do dispositivo> / Carregamento do nome de dispositivo inválido. Falha ao ativar a instância EC2. (p. 464)
- O valor (<nome associado ao dispositivo de armazenamento de instâncias>) do parâmetro virtualName é inválido... (p. 464)
- Mapeamentos de dispositivos de blocos do EBS não suportados para AMIs de armazenamento de instância. (p. 465)
- Os placement groups não podem ser usados com instâncias do tipo 'm1.large'. Falha ao ativar a instância EC2. (p. 465)
- de de de de InternalError: Erro do cliente na inicialização. (p. 465)
- No momento, não temos capacidade de <tipo de instância> suficiente para tipo de instância na zona de disponibilidade solicitada. Falha ao ativar a instância EC2. (p. 466)
- Não há capacidade spot disponível que corresponda à sua solicitação. Falha ao ativar a instância EC2. (p. 467)
- <número de instâncias> instância(s) já estão em execução. Falha ao ativar a instância EC2. (p. 467)

A configuração solicitada não é suportada atualmente.

- Causa: algumas opções no modelo ou na configuração de execução podem não ser compatíveis com o tipo de instância, ou a configuração de instância pode não ser compatível com a região da AWS ou 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 instâncias do Linux.

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 Graviton AWS baseado em Arm em vez de um processador Intel Xeon, você precisará de uma AMI compatível com ARM.
- 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 [Como encontrar tipos de instância do Amazon EC2](#) no Manual do usuário do Amazon EC2 para instâncias do Linux.
- 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 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 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](#).

A Zona de disponibilidade solicitada não é mais suportada. Tente sua solicitação novamente...

- Mensagem de erro: a zona de disponibilidade solicitada não tem mais suporte. Tente sua solicitação novamente sem especificar uma Zona de disponibilidade ou escolher <lista de Zonas de disponibilidade disponíveis>. Falha ao ativar a instância EC2.
- Causa: a zona de disponibilidade associada a seu grupo do Auto Scaling pode não estar disponível no momento.
- Solução: atualize seu grupo do Auto Scaling com as recomendações na mensagem de erro.

O tipo de instância solicitado (<tipo de instância>) não tem suporte na Zona de disponibilidade solicitada (<Zona de disponibilidade da instância>)...

- Mensagem de erro: Your requested instance type (<instance type>) is not supported in your requested Availability Zone (<instance Availability Zone>). (O tipo de instância solicitado (<tipo de instância>) não tem suporte na zona de disponibilidade solicitada (<zona de disponibilidade da instância>).) Tente a solicitação novamente sem especificar uma Zona de disponibilidade ou escolher <lista de Zonas de disponibilidade que oferecem suporte ao tipo de instância>. Falha ao ativar a instância EC2.
- Causa: o tipo de instância escolhido pode não estar disponível atualmente nas zonas de disponibilidade especificadas em seu grupo do Auto Scaling.
- Solução: atualize seu grupo do Auto Scaling com as recomendações na mensagem de erro.

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 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 para instâncias do Linux.
 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 para instâncias do Linux.
 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...

- Mensagem de erro: Value (<name associated with the instance storage device>) for parameter virtualName is invalid. (O valor (<nome associado ao dispositivo de armazenamento de instâncias>) do parâmetro virtualName é inválido.) Formato esperado: 'ephemeralNUMBER'. Falha ao ativar a instância 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 para instâncias do Linux.
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](#).

Mapeamentos de dispositivos de blocos do EBS não suportados para AMIs de armazenamento de instância.

- 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 obter mais informações, consulte [Mapeamento de dispositivos de blocos](#) no Manual do usuário do Amazon EC2 para instâncias do Linux.
 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 placement groups não podem ser usados com instâncias do tipo 'm1.large'. Falha ao ativar a instância 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 para instâncias do Linux.
 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](#).

de de de de de 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 do AWS KMS usada para criptografá-la. Para obter mais informações, consulte [Política de chaves do AWS KMS obrigatórias para uso com volumes criptografados \(p. 450\)](#).
- Causa 1: você precisa de uma política de chaves que conceda permissão para usar a chave gerenciada pelo cliente para a função vinculada ao serviço adequada.

No momento, não temos capacidade de <tipo de instância> suficiente para tipo de instância na zona de disponibilidade solicitada. Falha ao ativar a instância EC2.

-
- Solução 1: permita que a função vinculada ao serviço use a chave gerenciada pelo cliente da seguinte forma:

1. Determine que função vinculada ao serviço deve ser usada para esse grupo do Auto Scaling.
2. Atualize a política de chaves na chave gerenciada pelo cliente e permita que a função vinculada ao serviço use a chave gerenciada pelo cliente.
3. Atualize o grupo do Auto Scaling para usar a função vinculada ao serviço.

Para obter um exemplo de uma política de chave que permita que a função vinculada ao serviço use a chave gerenciada pelo cliente, consulte [Exemplo 1: seções da política de chaves que permitem acesso à chave gerenciada pelo cliente \(p. 452\)](#).

- Causa 2: se a chave gerenciada pelo cliente e o grupo do Auto Scaling estiverem em contas da AWS diferentes, será necessário configurar o acesso entre contas à chave gerenciada pelo cliente a fim de conceder permissão para usar a chave gerenciada pelo cliente para a função vinculada ao serviço adequada.
- Solução 2: permita que a função vinculada 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 accesse a chave gerenciada pelo cliente.
 2. Defina um usuário ou uma função do IAM na conta do grupo do Auto Scaling que possa criar uma concessão.
 3. Determine que função vinculada ao serviço deve ser usada para esse grupo do Auto Scaling.
 4. Crie uma concessão para a chave gerenciada pelo cliente com a função vinculada ao serviço como o principal favorecido.
 5. Atualize o grupo do Auto Scaling para usar a função vinculada ao serviço.

Para obter mais informações, consulte [Exemplo 2: seções da política de chaves que permitem acesso entre contas à chave gerenciada pelo cliente \(p. 453\)](#).

- 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 a função vinculada 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 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 EC2.
- Causa: no momento, o Amazon EC2 não pode oferecer suporte a seu tipo de instância na zona de disponibilidade solicitada.
- Solução:

Para resolver esse problema, experimente o seguinte:

- Espere alguns minutos e envie uma solicitação novamente; a capacidade pode mudar com frequência.
- Envie uma nova solicitação de acordo com as recomendações na mensagem de erro.
- Envie uma nova solicitação com um número de instâncias reduzido (que pode ser aumentado posteriormente).

Não há capacidade spot disponível que corresponda à sua solicitação. Falha ao ativar a instância 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. Se a capacidade não estiver disponível, a solicitação spot continuará a fazer a solicitação de lançamento automaticamente até que a capacidade seja disponibilizada. Quando a capacidade se tornar disponível, o serviço spot do Amazon EC2 atenderá à solicitação spot.
- Siga a prática recomendada de usar um conjunto diversificado de tipos de instância para que você não dependa de um tipo de instância específico. Para obter mais informações, incluindo uma lista de práticas recomendadas para usar instâncias spot com êxito, consulte [Grupos do Auto Scaling com vários tipos de instâncias e opções de compra \(p. 67\)](#).
- Envie uma nova solicitação com um número de instâncias reduzido (que pode ser aumentado posteriormente).

<número de instâncias> instância(s) já estão em execução. Falha ao ativar a instância EC2.

- Causa: você atingiu o limite do número total de instâncias que pode iniciar em uma região. Ao criar uma conta da AWS, 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 [Cotas de serviço do Amazon EC2](#) no Manual do usuário do Amazon EC2 para instâncias do Linux.
- 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 a suas AMIs, 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 \(p. 459\)](#).

Quando suas instâncias 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 EC2. \(p. 468\)](#)
- [A AMI <ID da AMI> está pendente e não pode ser executada. Falha ao ativar a instância EC2. \(p. 468\)](#)
- [O valor do \(<ID da ami>\) para o parâmetro virtualName é inválido. \(p. 468\)](#)
- [A arquitetura do tipo de instância solicitado \(i386\) não corresponde à arquitetura no manifesto da ami-6622f00f \(x86_64\). Falha ao ativar a instância EC2. \(p. 469\)](#)

O ID da AMI <ID de sua AMI> não existe. Falha ao ativar a instância 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 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: você deve aguardar até que sua AMI esteja disponível e, em seguida, criar um modelo de execução ou uma configuração de execução.

O valor do (<ID da ami>) para o parâmetro virtualName é inválido.

- Causa: Valor incorreto. O parâmetro virtualName se refere ao nome virtual associado ao dispositivo.
- Solução:
 1. Crie um novo modelo de execução ou uma nova configuração de execução especificando o nome do dispositivo virtual de sua instância para o parâmetro virtualName.
 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 do tipo de instância solicitado (i386) não corresponde à arquitetura no manifesto da ami-6622f00f (x86_64). Falha ao ativar a instância EC2.

- Causa: a arquitetura do InstanceType mencionado no modelo de execução ou na configuração de execução não corresponde à arquitetura da imagem.
- Solução:
 1. Crie um novo modelo de execução ou uma nova configuração de execução usando a arquitetura da AMI que corresponde à arquitetura do tipo de instância solicitado.
 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平衡ador 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 \(p. 459\)](#).

Quando houver falha ao iniciar suas instâncias 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. \(p. 470\)](#)
- [Não é possível encontrar o Load Balancer <seu load balancer>. Falha na validação da configuração do balanceador de carga. \(p. 470\)](#)
- [Não há nenhum balanceador de carga ATIVO chamado <nome do balanceador de carga>. Falha ao atualizar a configuração do balanceador de carga. \(p. 470\)](#)
- [A instância do EC2 <ID da instância> não está na VPC. Falha ao atualizar a configuração do balanceador de carga. \(p. 471\)](#)
- [A instância do EC2 <ID da instância> está na VPC. Falha ao atualizar a configuração do balanceador de carga. \(p. 471\)](#)

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 do Auto Scaling sem o grupo alvo usando o console do Amazon EC2 Auto Scaling ou o comando [detach-load-balancer-target-groups](#).

Causa 2: O grupo-alvo existe, mas houve um problema ao tentar especificar o ARN do grupo-alvo ao criar o 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 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 do Auto Scaling sem o balanceador de carga ou remover o balanceador de carga não usado do grupo do Auto Scaling usando o console do Amazon EC2 Auto Scaling ou o [detach-load-balancers](#) comando.

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平衡器 de carga associado à instância ou criar um novo grupo do Auto Scaling.

A instância do EC2 <ID da instância> está na VPC. Falha ao atualizar a configuração do balanceador de carga.

Causa: o平衡器 de carga está no EC2-Classic, mas o grupo do Auto Scaling está em uma VPC.

Solução: verifique se o平衡器 de carga e o grupo do Auto Scaling estão na mesma rede (EC2-Classic ou VPC).

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 \(p. 461\)](#).

Você deve 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 (Você deve 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 (Você deve 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 \(p. 23\)](#).

Você não está autorizado a usar o modelo de execução (permisões 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: 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`.

Solução 2: verifique se a política `AmazonEC2FullAccess` está atribuída às credenciais do IAM que você está usando para fazer a solicitação. Esta política AWS gerenciada pela concede acesso total a todos os recursos do Amazon EC2 e serviços relacionados, incluindo Amazon EC2 Auto Scaling e Elastic Load Balancing, CloudWatch

Causa 2: se estiver tentando usar um modelo de execução que especifica um perfil da instância, você deverá ter permissão do IAM para transmitir a função do IAM que está associada ao perfil da instância.

Solução 3: verifique se as credenciais do IAM que você está usando para fazer a solicitação têm as permissões `iam:PassRole` corretas para passar a função especificada ao serviço Amazon EC2 Auto Scaling. Para obter mais informações e um exemplo de política do IAM, consulte [Funções do IAM para aplicações que são executadas em instâncias do Amazon EC2 \(p. 448\)](#). 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 estiver tentando usar um modelo de execução que especifica uma AMI em outro Conta da AWS e a AMI for privada e não compartilhada com o Conta da AWS que você está usando, você receberá um erro informando que não está autorizado a usar o modelo de execução.

Solução 4: 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 para instâncias do Linux.

Important

Para obter mais informações sobre permissões para modelos de execução, incluindo exemplos de políticas do IAM, consulte [Suporte a modelo de execução \(p. 443\)](#).

Solucionar problemas com as verificações de integridade do Amazon EC2 Auto Scaling

Esta página fornece informações sobre suas instâncias do EC2 que são terminadas devido a uma verificação de integridade. Ela descreve as possíveis causas e as etapas que podem ser adotadas para resolver os problemas.

Para recuperar uma mensagem de erro, consulte [Recuperar uma mensagem de erro de ações de escalabilidade \(p. 459\)](#).

Problemas de verificação de integridade

- [Uma instância foi retirada de serviço em resposta a uma falha de verificação de status de instância do EC2 \(p. 473\)](#)
- [Uma instância foi retirada de serviço em resposta a uma reinicialização programada do EC2 \(p. 474\)](#)
- [Uma instância foi retirada de serviço em resposta a uma verificação de integridade do EC2 que indicou que ela tinha sido terminada ou interrompida \(p. 474\)](#)
- [Uma instância foi retirada de serviço em resposta a uma falha na verificação de integridade do sistema ELB \(p. 475\)](#)

Note

Você pode ser notificado quando o Amazon EC2 Auto Scaling termina as instâncias no grupo do Auto Scaling, inclusive quando a causa do término da instância não é o resultado de uma atividade de escalabilidade. Para obter mais informações, consulte [Receber notificações do Amazon SNS quando o grupo do Auto Scaling escala \(p. 341\)](#).

As seções a seguir descrevem os erros e causas mais comuns de verificação de integridade que você encontrará. Se um problema diferente surgir, consulte os seguintes artigos da Central de Conhecimento da AWS para obter ajuda adicional para solucioná-lo:

- [Por que o Amazon EC2 Auto Scaling falhou ao terminar uma instância?](#)
- [Por que o Amazon EC2 Auto Scaling não terminou uma instância não íntegra?](#)

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. As verificações de status são integradas ao Amazon EC2, portanto elas não podem ser desabilitadas ou excluídas. 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.

Solução 1: 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 com a investigação de instâncias prejudicadas, consulte [Solução de problemas em instâncias com falha nas verificações de status](#) no Manual do usuário do Amazon EC2 para instâncias do Linux e [Solução de problemas de instâncias do Windows](#) no Manual do usuário do Amazon EC2 para instâncias do Windows.
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.

Causa 2: há uma incompatibilidade entre o período de carência da verificação de integridade e o tempo de inicialização da instância.

Solução 2: edite o período de carência da verificação de integridade do grupo do Auto Scaling para um período de tempo apropriado para a aplicação. As instâncias executadas em um grupo do Auto Scaling exigem tempo de aquecimento suficiente (período de carência) para evitar o encerramento antecipado devido a uma substituição de verificação de integridade. Para obter mais informações, consulte [Definir um período de carência da verificação de integridade para um grupo do Auto Scaling \(p. 325\)](#).

Uma instância foi retirada de serviço em resposta a uma reinicialização programada do EC2

Problema: instâncias do Auto Scaling são substituídas quando um evento programado indica um problema com a instância.

Causa: o Amazon EC2 Auto Scaling substitui instâncias por um evento futuro programado de manutenção ou desativação.

Solução: esses eventos não ocorrem com frequência. 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 ganchos permitem que você execute uma ação personalizada à medida que o Amazon EC2 Auto Scaling inicia ou termina instâncias. Para obter mais informações, consulte [Ganchos do ciclo de vida do Amazon EC2 Auto Scaling \(p. 252\)](#).

Se não desejar que as instâncias sejam substituídas devido a um evento programado, você poderá suspender o processo de verificação de integridade para qualquer grupo individual do Auto Scaling. Para obter mais informações, consulte [Suspender e retomar um processo para um grupo do Auto Scaling \(p. 312\)](#).

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 uma verificação de integridade falhar porque um usuário interrompeu, reinicializou ou terminou manualmente a instância, isso se deve ao funcionamento das verificações de integridade do Amazon EC2 Auto Scaling. A instância deve ser íntegra e acessível. Se precisar reiniciar as instâncias no seu grupo do Auto Scaling, recomendamos colocar as instâncias em espera primeiro. Para obter mais informações, consulte [Remover temporariamente instâncias do grupo do Auto Scaling \(p. 308\)](#).

Observe que, quando instâncias são terminadas manualmente, os ganchos do ciclo de vida de término e o cancelamento do registro do Elastic Load Balancing (e a descarga da conexão) devem ser concluídos antes que a instância seja realmente terminada.

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 \(p. 67\)](#).
- 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 [Usar o rebalanceamento de capacidade para lidar com interrupções de spot do Amazon EC2 \(p. 346\)](#).

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: 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. Para obter mais informações, consulte [Adicionar verificações de integridade do Elastic Load Balancing a um grupo do Auto Scaling \(p. 376\)](#).

Solução 1: para passar nas verificações de integridade do Elastic Load Balancing:

- 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 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.
- Considere iniciar um gancho do ciclo de vida de inicialização ao grupo do Auto Scaling para garantir que as aplicações nas instâncias estejam prontas para aceitar tráfego antes de serem registradas no balanceador de carga no final do ganho do ciclo de vida.
- 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.

- Verifique se o balanceador de carga está configurado nas mesmas zonas de disponibilidade do grupo do Auto Scaling.

Para obter mais informações, consulte os tópicos a seguir:

- [Verificações de integridade para seus grupos de destino](#) no Manual do usuário de Application Load Balancers
- [Verificações de integridade para seus grupos de destino](#) no Manual do usuário de Network Load Balancers
- [Verificações de integridade para seus grupos de destino](#) no Manual do usuário de平衡adores de carga de gateway
- [Configurar verificações de integridade para seu Classic Load Balancer](#) no Manual do usuário de Classic Load Balancers

Solução 2: atualizar o grupo do Auto Scaling para desativar as verificações de integridade do Elastic Load Balancing.

Informações relacionadas

Os recursos relacionados a seguir podem ajudar você à medida que trabalha com este serviço.

Recurso	Descrição
Referência da API Amazon EC2 Auto Scaling	A documentação de cada operação de API mostra os parâmetros da 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 do AWS CLI comandos que você pode usar para trabalhar com grupos de Auto Scaling.
Referência do cmdlet do AWS Tools for PowerShell	O AWS PowerShell ferramentas para permitem que você crie scripts de operações em seu AWS recursos do PowerShell linha de comando.
Criar um grupo do Auto Scaling com AWS CloudFormation (p. 360)	O AWS::AutoScaling::AutoScalingGroup O recurso permite criar, modelar e gerenciar seus grupos de Auto Scaling sem ações manuais.
Endpoints e cotas do Amazon EC2 Auto Scaling Referência geral da AWS	Informações sobre regiões e endpoints do Amazon EC2 Auto Scaling.
Páginas do produtos	A página principal da web para obter informações 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.
Crie uma AMI Guia do usuário do Amazon EC2 para instâncias Linux	Saiba como criar uma Amazon Machine Image (AMI) a partir de uma instância personalizada.
Como se conectar à sua instância Linux Guia do usuário do Amazon EC2 para instâncias Linux	Saiba como se conectar às instâncias do Linux que você executa.
Como se conectar à sua instância do Windows Guia do usuário do Amazon EC2 para instâncias do Windows	Saiba como se conectar às instâncias do Windows que você executa.
Criação de um alarme de cobrança para monitorar sua estimativa AWS cobranças Guia do usuário da Amazônia CloudWatch	Saiba como monitorar suas cobranças estimadas usando CloudWatch.

Os seguintes recursos gerais estão disponíveis para ajudar você a aprender mais sobre AWS.

- [Aulas e workshops](#) — Links para cursos de especialidades e baseados em perfil, bem como laboratórios autoguiados para ajudar a aperfeiçoar suas habilidades na AWS e a obter experiência prática.
- [Centro dos desenvolvedores da AWS](#) — Explore tutoriais, baixe ferramentas e informe-se sobre eventos para desenvolvedores da AWS.

- [Ferramentas do desenvolvedor da AWS](#) — Links para ferramentas de desenvolvedor, SDKs, toolkits de IDE e ferramentas da linha de comando para desenvolver e gerenciar aplicativos da AWS.
- [Centro de recursos de conceitos básicos](#) — Saiba como configurar a Conta da AWS, participar da comunidade da AWS e lançar seu primeiro aplicativo.
- [Tutoriais práticos](#) — Sigastep-by-steptutoriais para lançar seu primeiro aplicativo emAWS.
- [Whitepapers da AWS](#) — Links para uma lista abrangente de whitepapers técnicos da AWS que abrangem tópicos como arquitetura, segurança e economia, elaborados pelos arquitetos de soluções da AWS ou por outros especialistas técnicos.
- [AWS Support Center](#): a central para criar e gerenciar seus casos do AWS Support. Também inclui links para outros recursos úteis, como fóruns, perguntas frequentes técnicas, status de integridade do serviço e AWS Trusted Advisor.
- [AWS Support](#)— A página principal da web para obter informações sobreAWS Support1, ane-on-one, canal de suporte de resposta rápida para ajudar você a criar e executar aplicativos na nuvem.
- [Entrar em contato](#) – Um ponto central de contato para consultas relativas a faturas da AWS, contas, eventos, uso abusivo e outros problemas.
- [Termos do site da AWS](#): informações detalhadas sobre nossos direitos autorais e marca registrada; sua conta, licença e acesso ao site, entre outros tópicos.

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
Novos recursos de atualização de instância (p. 479)	Agora você pode configurar a atualização de uma instância para definir seu status como falha e, opcionalmente, reverter quando detectar que uma determinada CloudWatch alarme entrou no ALARM estado. Para obter mais informações, consulte Desfazer alterações com uma reversão na Guia do usuário do Amazon EC2 Auto Scaling.	31 de julho de 2023
Alterações do guia (p. 479)	Um novo tópico sobre o lançamento de instâncias sob demanda em reservas de capacidade foi adicionado ao guia. Para obter mais informações, consulte Use reservas de capacidade sob demanda para reservar capacidade em zonas de disponibilidade específicas na Guia do usuário do Amazon EC2 Auto Scaling.	28 de julho de 2023
Alterações do guia (p. 479)	Um novo tópico sobre como migrar seu AWS CloudFormation pilhas de configurações de inicialização a modelos de inicialização foram adicionadas ao guia. Para obter mais informações, consulte Migrar AWS CloudFormation pilhas desde configurações de lançamento até modelos de lançamento na Guia do usuário do Amazon EC2 Auto Scaling.	18 de abril de 2023
Suporte para novas operações de API (p. 479)	Esta versão adiciona três novas operações de API, <code>AttachTrafficSources</code> , <code>DetachTrafficSources</code> , e <code>eDescribeTrafficSources</code> . Além disso, um novo campo, <code>TrafficSources</code> , foi adicionado aos resultados de <code>DescribeAutoScalingGroups</code> operações.	31 de março de 2023

Um novo status de atividade, `WaitingForConnectionDraining`, foi adicionado aos resultados `deDescribeScalingActivities` operações. O Amazon EC2 Auto Scaling também suporta um novo valor, `VPC_LATTICE`, para o `HealthCheckType` campo em `CreateAutoScalingGroup`, `UpdateAutoScalingGroup`, e `DescribeAutoScalingGroups` operações. Para obter mais informações, consulte a [Referência da API do Amazon EC2 Auto Scaling](#).

[Suporte para Amazon VPC Lattice \(p. 479\)](#)

Esta é a versão de disponibilidade geral do VPC Lattice para Amazon EC2 Auto Scaling. Para obter mais informações, consulte [Encaminhe o tráfego para seu grupo de Auto Scaling com um grupo-alvo do VPC Lattice](#) na Guia do usuário do Amazon EC2 Auto Scaling.

31 de março de 2023

[Alterações do guia \(p. 479\)](#)

A seção com AWS CLI exemplos para trabalhar com o Elastic Load Balancing agora incluem exemplos novos e atualizados. Para obter mais informações, consulte [Exemplos de como trabalhar com o Elastic Load Balancing com a AWS Command Line Interface \(AWS CLI\)](#) na Guia do usuário do Amazon EC2 Auto Scaling.

31 de março de 2023

[Suporte para escalonamento preditivo adicional Regiões da AWS \(p. 479\)](#)

Agora você pode criar políticas de escalabilidade preditiva no Oriente Médio (EAU) e AWS GovCloud Regiões (Leste dos EUA). Para obter mais informações, consulte [Escalabilidade preditiva no Amazon EC2 Auto Scaling](#) no Manual do usuário do Amazon EC2 Auto Scaling.

16 de março de 2023

<u>Novos recursos de atualização de instância (p. 479)</u>	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 <u>Substituir instâncias do Auto Scaling com base na atualização de uma instância</u> no Manual do usuário do Amazon EC2 Auto Scaling.	10 de fevereiro de 2023
<u>Suporte ao uso de um parâmetro do AWS Systems Manager para um ID de AMI (p. 479)</u>	Agora você pode usar um parâmetro do Systems Manager em vez de um ID de AMI no modelo de execução. Para obter mais informações, consulte <u>Usar parâmetros do AWS Systems Manager em vez de IDs de AMI em modelos de execução</u> no Guia do usuário do Amazon EC2 Auto Scaling.	19 de janeiro de 2023
<u>Recomendações de escalabilidade preditiva (p. 479)</u>	Já é possível obter recomendações para avaliar e escolher a política de escalabilidade preditiva correta no console do Amazon EC2 Auto Scaling. Para obter mais informações, consulte <u>Avaliar políticas de escalabilidade preditiva</u> no Guia do usuário do Amazon EC2 Auto Scaling.	18 de janeiro de 2023
<u>Previsões de escalabilidade preditiva (p. 479)</u>	As previsões geradas pelo dimensionamento preditivo agora são atualizadas a cada seis horas, em vez de diariamente. Para obter mais informações, consulte <u>Escalabilidade preditiva o Amazon EC2 Auto Scaling</u> no Manual do usuário do Amazon EC2 Auto Scaling.	6 de janeiro de 2023

Suporte para CloudWatch metrics (p. 479)	Agora você pode usar a matemática métrica ao criar políticas de dimensionamento de rastreamento de destino. Com a matemática métrica, você pode consultar vários CloudWatch métricas e use 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 matemática de escalabilidade para o Amazon EC2 Auto Scaling usando matemática de escalabilidade.	8 de dezembro de 2022
Atualizar permissões de função vinculada a serviços do IAM (p. 479)	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 (p. 479)	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) (p. 479)	Agora você pode criar políticas de escalação preditiva na Região Ásia-Pacífico (Jacarta). Para obter mais informações, consulte Escalabilidade preditiva o Amazon EC2 Auto Scaling no Manual do usuário do Amazon EC2 Auto Scaling.	13 de outubro de 2022

<u>Compatibilidade com métricas personalizadas para escalação preditiva no console (p. 479)</u>	Agora você pode usar métricas personalizadas ao criar políticas de escalação preditiva no console do Amazon EC2 Auto Scaling. Para obter mais informações, consulte <u>Escalabilidade preditiva o Amazon EC2 Auto Scaling</u> no Manual do usuário do Amazon EC2 Auto Scaling.	13 de outubro de 2022
<u>CloudWatchmonitoramento de métricas de escalabilidade preditiva (p. 479)</u>	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 <u>Monitore métricas de escalabilidade preditiva com CloudWatch</u> na Guia do usuário do Amazon EC2 Auto Scaling.	7 de julho de 2022
<u>Compatibilidade com escalação preditiva na região Ásia-Pacífico (Osaka) (p. 479)</u>	Agora você pode criar políticas de escalação preditiva na Região Ásia-Pacífico (Osaka). Para obter mais informações, consulte <u>Escalabilidade preditiva o Amazon EC2 Auto Scaling</u> no Manual do usuário do Amazon EC2 Auto Scaling.	6 de julho de 2022
<u>Compatibilidade com hibernação de grupo de alta atividade passa a ser oferecida em regiões adicionais (p. 479)</u>	Agora você pode hibernar instâncias em um grupo de alta atividade em mais quatro regiões: África (Cidade do Cabo), Ásia-Pacífico (Jacarta), Ásia-Pacífico (Osaka) e Europa (Milão). Para obter mais informações sobre grupos de alta atividade, consulte <u>Grupos de alta atividade para o Amazon EC2 Auto Scaling</u> no Guia do usuário do Amazon EC2 Auto Scaling.	5 de julho de 2022

<u>Atualização das verificações de integridade (p. 479)</u>	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 informações, consulte <u>Como o Amazon EC2 Auto Scaling minimiza o tempo de inatividade</u> no Guia do usuário do Amazon EC2 Auto Scaling.	21 de maio de 2022
<u>Aquecimento de instância padrão (p. 479)</u>	Agora é possível unificar todas as configurações de aquecimento e desaquecimento de um grupo do Auto Scaling e otimizar a performance de políticas de escalabilidade que escalam continuamente, habilitando o aquecimento de instâncias padrão. Para obter mais informações, consulte <u>Set the default instance warmup for an Auto Scaling group</u> (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
<u>Alterações do guia (p. 479)</u>	Um novo capítulo sobre integração com outros produtos da AWS foi adicionado ao guia. Para obter mais informações, consulte <u>AWS serviços integrados ao Amazon EC2 Auto Scaling</u> no Guia do usuário do Amazon EC2 Auto Scaling.	29 de março de 2022
<u>Atualizar permissões de função vinculada a serviços do IAM (p. 479)</u>	A política <code>AutoScalingServiceRolePolicy</code> concede permissões de leitura adicionais ao Amazon EC2 Auto Scaling. Para obter mais informações, consulte <u>Políticas gerenciadas pela AWS para o Amazon EC2 Auto Scaling</u> no Guia do usuário do Amazon EC2 Auto Scaling.	28 de março de 2022

<u>Os metadados da instância fornecem o estado de destino do ciclo de vida (p. 479)</u>	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 <u>Retrieve the target lifecycle state through instance metadata</u> (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
<u>Suporte à nova funcionalidade de grupo de alta atividade (p. 479)</u>	Agora você pode hibernar instâncias em um grupo de alta atividade para interromper instâncias sem excluir o conteúdo da memória (RAM). Agora você também pode devolver instâncias ao grupo de alta atividade em redução de escala na horizontal, em vez de sempre terminar a capacidade da instância que você precisará posteriormente. Para obter mais informações, consulte <u>Grupos de alta atividade para o Amazon EC2 Auto Scaling</u> no Manual do usuário do Amazon EC2 Auto Scaling.	24 de fevereiro de 2022
<u>Alterações do guia (p. 479)</u>	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 <u>Iniciar ou cancelar uma atualização de instância (console)</u> no Guia do usuário do Amazon EC2 Auto Scaling.	3 de fevereiro de 2022

<u>Métricas personalizadas para políticas de escalabilidade preditiva (p. 479)</u>	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 <u>Advanced predictive scaling policy configurations using custom metrics</u> (Configurações avançadas de política de escalabilidade preditiva usando métricas personalizadas).	24 de novembro de 2021
<u>Nova estratégia de alocação sob demanda (p. 479)</u>	Agora, é possível escolher se deseja executar 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 <u>Allocation strategies</u> (Estratégias de alocação) no Guia do usuário do Amazon EC2 Auto Scaling.	27 de outubro de 2021
<u>Seleção de tipo de instância baseada em atributos (p. 479)</u>	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 <u>Creating an Auto Scaling group using attribute-based instance type selection</u> (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

<u>Suporte para filtragem de grupos por etiquetas (p. 479)</u>	Agora você pode filtrar seus grupos do Auto Scaling usando filtros de etiquetas ao recuperar informações sobre grupos do Auto Scaling usando o comando <code>describe-auto-scaling-groups</code> . 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
<u>Alterações do guia (p. 479)</u>	O console do Amazon EC2 Auto Scaling foi atualizado para ajudar você a criar políticas de encerramento personalizadas com o 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
<u>Suporte para cópia de configurações de execução para modelos de execução (p. 479)</u>	Agora você pode copiar todas as configurações de execução em uma região da AWS para novos modelos de execução a partir do console do Amazon EC2 Auto Scaling. Para obter mais informações, consulte Copiar configurações de execução para modelos de execução no Manual do usuário do Amazon EC2 Auto Scaling.	9 de agosto de 2021
<u>Expande a funcionalidade de atualização de instância (p. 479)</u>	Agora você pode incluir atualizações, como uma nova versão de um modelo de execução, ao substituir instâncias, adicionando a configuração desejada ao comando <code>start-instance-refresh</code> . Você também pode ignorar a substituição de instâncias que já têm a configuração desejada ativando a correspondência de ignorar. 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.	5 de agosto de 2021

<u>Suporte para políticas de término personalizadas (p. 479)</u>	Agora é possível criar políticas de término personalizadas com o AWS Lambda. Para obter mais informações, consulte <u>Criação de uma política de término personalizada com o Lambda</u> . A documentação para especificar políticas de término foi atualizada de maneira adequada.	29 de julho de 2021
<u>Alterações do guia (p. 479)</u>	O console do Amazon EC2 Auto Scaling foi atualizado e aprimorado com outros recursos para ajudá-lo a criar ações programadas com um fuso horário especificado. A documentação para <u>Escalabilidade programada</u> foi revisada em conformidade.	3 de junho de 2021
<u>volumes gp3 em configurações de execução (p. 479)</u>	Agora você pode especificar volumes gp3 nos mapeamentos de dispositivos de bloco para configurações de execução.	2 de junho de 2021
<u>Suporte para escalabilidade preditiva (p. 479)</u>	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 <u>Escalabilidade preditiva no Amazon EC2 Auto Scaling</u> no Manual do usuário do Amazon EC2 Auto Scaling. Com essa atualização, a <u>AutoScalingServiceRolePolicy</u> política gerenciada agora inclui permissão para chamar a <code>ocloudwatch:GetMetricData</code> Ação da API.	19 de maio de 2021
<u>Alterações do guia (p. 479)</u>	Agora você pode acessar modelos de exemplo para ganchos de ciclo de vida em GitHub. Para obter mais informações, consulte <u>Ganchos do ciclo de vida do Amazon EC2 Auto Scaling</u> no Manual do usuário do Amazon EC2 Auto Scaling.	9 de abril de 2021

<u>Suporte a grupos de alta atividade (p. 479)</u>	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 grupos de alta atividade aos grupos do Auto Scaling. Para obter mais informações, consulte <u>Grupos de alta atividade para o Amazon EC2 Auto Scaling</u> no Manual do usuário do Amazon EC2 Auto Scaling.	8 de abril de 2021
<u>Suporte para pontos de verificação (p. 479)</u>	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 <u>Adição de pontos de verificação a uma atualização de instância</u> no Manual do usuário do Amazon EC2 Auto Scaling.	18 de março de 2021
<u>Alterações do guia (p. 479)</u>	Documentação aprimorada para uso EventBridge com eventos e ganchos de ciclo de vida do Amazon EC2 Auto Scaling. Para obter mais informações, consulte <u>Usando o Amazon EC2 Auto Scaling com EventBridge</u> no <u>Tutorial: Configurar um gancho de ciclo de vida que invoque uma função Lambda</u> no Guia do usuário do Amazon EC2 Auto Scaling.	18 de março de 2021
<u>Suporte para fusos horários locais (p. 479)</u>	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 <u>Escalabilidade programada</u> no Manual do usuário do Amazon EC2 Auto Scaling.	9 de março de 2021

<u>Expande a funcionalidade para políticas de instâncias mistas (p. 479)</u>	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 optimiza a capacidade. Para obter mais informações, consulte <u>Grupos de Auto Scaling com vários tipos de instância e opções de compra</u> no Manual do usuário do Amazon EC2 Auto Scaling.	8 de março de 2021
<u>Escalabilidade de atividades para grupos excluídos (p. 479)</u>	Agora você pode visualizar atividades de escalabilidade para grupos do Auto Scaling excluídos adicionando a opção <code>--include-deleted-groups</code> ao comando <code>describe-scaling-activities</code> . Para obter mais informações, consulte <u>Solução de problemas do Amazon EC2 Auto Scaling</u> no Manual do usuário do Amazon EC2 Auto Scaling.	23 de fevereiro de 2021
<u>Melhorias no console (p. 479)</u>	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 <u>Criar e anexar um novo Application Load Balancer ou Network Load Balancer (console)</u> no Guia do usuário do Amazon EC2 Auto Scaling.	24 de novembro de 2020
<u>Várias interfaces de rede (p. 479)</u>	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 <u>Interfaces de rede em uma VPC</u> .	23 de novembro de 2020
<u>Vários modelos de execução (p. 479)</u>	Vários modelos de execução podem agora ser usados com grupos do Auto Scaling. Para obter mais informações, consulte <u>Especificar um modelo de execução diferente para um tipo de instância</u> no Manual do usuário do Amazon EC2 Auto Scaling.	19 de novembro de 2020

<u>Balanceadores de carga de gateway (p. 479)</u>	Guia atualizado para mostrar como anexar um balanceador de carga de gateway a um grupo do Auto Scaling para garantir que as instâncias de dispositivo executadas pelo Amazon EC2 Auto Scaling sejam registradas automaticamente e canceladas 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
<u>Vida útil máxima da instância (p. 479)</u>	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
<u>Rebalanceamento de capacidade (p. 479)</u>	Você pode configurar seu grupo do Auto Scaling para iniciar uma instância spot de substituição quando o Amazon EC2 emitir uma recomendação 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
<u>Instance Metadata Service Version 2 (p. 479)</u>	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 (p. 479)	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 (p. 479)	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 (p. 479)	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 (p. 479)	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

<u>Desabilitar políticas de escalabilidade (p. 479)</u>	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 <u>Desativação de uma política de escalabilidade para um grupo do Auto Scaling</u> no Manual do usuário do Amazon EC2 Auto Scaling.	18 de fevereiro de 2020
<u>Adicionar funcionalidade de notificação (p. 479)</u>	Agora, o Amazon EC2 Auto Scaling envia eventos para o seu AWS Health Dashboard quando os grupos do Auto Scaling não podem ser expandidos devido a um grupo de segurança ou modelo de execução ausente. Para obter mais informações, consulte <u>Notificações do AWS Health Dashboard para o Amazon EC2 Auto Scaling</u> no Manual do usuário do Amazon EC2 Auto Scaling.	12 de fevereiro de 2020
<u>Alterações do guia (p. 479)</u>	Várias melhorias e correções nas seções <u>Como o Amazon EC2 Auto Scaling funciona com o IAM</u> , <u>Exemplo de políticas baseadas em identidade do Amazon EC2 Auto Scaling</u> , <u>Política de chave da CMK necessária para uso com volumes criptografados</u> e <u>Monitoramento das suas instâncias e grupos do Auto Scaling</u> no Manual do usuário do Amazon EC2 Auto Scaling.	10 de fevereiro de 2020
<u>Alterações do guia (p. 479)</u>	Melhoria na documentação para grupos do Auto Scaling que usam peso de instâncias. Saiba como usar políticas de escalabilidade ao usar “unidades de capacidade” para medir a capacidade desejada. Para obter mais informações, consulte <u>Como funcionam as políticas de escalabilidade</u> e <u>Tipos de ajuste da escalabilidade</u> no Manual do usuário do Amazon EC2 Auto Scaling.	6 de fevereiro de 2020

<u>Novo capítulo "Segurança" (p. 479)</u>	Um novo capítulo sobre <u>Segurança</u> no Guia do usuário do Amazon EC2 Auto Scaling ajuda você a entender como aplicar o <u>modelo de responsabilidade compartilhada</u> 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, <u>Gerenciamento de identidades e acesso para o Amazon EC2 Auto Scaling</u> .	4 de fevereiro de 2020
<u>Recomendações para tipos de instância (p. 479)</u>	O AWS Compute Optimizer fornece recomendações para instâncias do Amazon EC2 para ajudar a melhorar a performance, economizar dinheiro ou ambos. Para obter mais informações, consulte <u>Obtenção de recomendações de um tipo de instância</u> no Manual do usuário do Amazon EC2 Auto Scaling.	3 de dezembro de 2019
<u>Hosts dedicados e grupos de recursos de host (p. 479)</u>	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 especifica uma AMI de BYOL a ser usada em hosts dedicados. Para obter mais informações, consulte <u>Criação de um modelo de execução para um grupo do Auto Scaling</u> no Manual do usuário do Amazon EC2 Auto Scaling.	3 de dezembro de 2019
<u>Suporte a endpoints da Amazon VPC (p. 479)</u>	Agora você pode estabelecer uma conexão privada entre sua VPC e o Amazon EC2 Auto Scaling. Para obter mais informações, consulte <u>Amazon EC2 Auto Scaling e endpoints da VPC da interface</u> no Manual do usuário do Amazon EC2 Auto Scaling.	22 de novembro de 2019

<u>Vida útil máxima da instância (p. 479)</u>	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 <u>Substituir instâncias do Auto Scaling com base na vida útil máxima da instância</u> no Manual do usuário do Amazon EC2 Auto Scaling.	19 de novembro de 2019
<u>Ponderação de instâncias (p. 479)</u>	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 <u>Ponderação de instâncias do Auto Scaling do Amazon EC2</u> no Manual do usuário do Amazon EC2 Auto Scaling.	19 de novembro de 2019
<u>Número mínimo de tipos de instância (p. 479)</u>	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 <u>Grupos de Auto Scaling com vários tipos de instância e opções de compra</u> no Manual do usuário do Amazon EC2 Auto Scaling.	16 de setembro de 2019
<u>Suporte para a nova estratégia de alocação spot (p. 479)</u>	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 <u>Grupos de Auto Scaling com vários tipos de instância e opções de compra</u> no Manual do usuário do Amazon EC2 Auto Scaling.	12 de agosto de 2019

Alterações do guia (p. 479)	Documentação do Amazon EC2 Auto Scaling melhorada nos tópicos Funções vinculadas ao serviço e Política de chaves de CMK necessária para uso com volumes criptografados .	1 de agosto de 2019
Suporte para aprimoramento de marcação (p. 479)	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 (p. 479)	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 funções vinculadas ao serviço com sufixo personalizado específicas para o Amazon EC2 Auto Scaling.	13 de junho de 2019
Suporte para novos recursos do Amazon EBS (p. 479)	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 (p. 479)	Melhoria na documentação do Amazon EC2 Auto Scaling nas seguintes seções: Controle de quais instâncias do Auto Scaling são terminadas 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 (p. 479)	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 (p. 479)	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	Descrição	Data de lançamento
Suporte para as políticas de escalabilidade de rastreamento de destino	Configure a escalabilidade dinâmica da sua aplicação em apenas algumas etapas. Para obter mais informações, consulte Políticas de escalabilidade com monitoramento do objetivo do Amazon EC2 Auto Scaling .	12 de julho de 2017
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 Monitorando seus grupos e instâncias de Auto Scaling usando a AmazonCloudWatch .	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 informações, 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 paraEventBridge quando chama ganchos de ciclo de vida. Para obter mais informações,	24 de fevereiro de 2016

Recurso	Descrição	Data de lançamento
	consulte Obtendo EventBridge quando seu grupo de Auto Scaling é escalado.	
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 escalar com base no tamanho da violação do alarme. Para obter mais informações, consulte Tipos de política de escalabilidade .	06 de julho de 2015
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
Suporte para 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
Gerenciar tags	Gerencie seus grupos do Auto Scaling usando o AWS Management Console. 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
Criar um grupo ou configuração de ativação a partir de uma instância 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, consulte Criação de um grupo do Auto Scaling usando uma instância do EC2 .	02 de janeiro de 2014

Recurso	Descrição	Data de lançamento
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 Limites do Auto Scaling .	02 de janeiro de 2014
Suporte para console do Amazon EC2 Auto Scaling	Acesse o Amazon EC2 Auto Scaling usando o AWS Management Console. 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
Suporte para funções do IAM	Iniciar instâncias do EC2 com um perfil de instância do IAM. Você pode usar esse recurso para atribuir funções 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 uma função 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 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	Descrição	Data de lançamento
Suporte para o Amazon SNS	<p>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.</p> <p>O Amazon EC2 Auto Scaling também adicionou os seguintes novos recursos:</p> <ul style="list-style-type: none"> • 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 iniciada ao平衡ador 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	Supor te 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
O suporte a clusters HPC	Supor te adicional para clusters de computação de alta performance (HPC).	2 de dezembro de 2010
Suporte a verificações de integridade	Supor te adicional 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 Uso de verificações de integridade do ELB com o Auto Scaling .	2 de dezembro de 2010
Suporte paraCloudWatchalarms	Removeu o mecanismo de gatilho antigo e redesenhou o Amazon EC2 Auto Scaling para usar o CloudWatch recurso de alarme. Para obter mais informações, consulte Escalabilidade dinâmica do Amazon EC2 Auto Scaling .	2 de dezembro de 2010
Suspender e retomar a escalabilidade	Supor te adicional para suspender e retomar processos de escalabilidade.	2 de dezembro de 2010
Supor te ao IAM	Adicionado suporte ao IAM. Para obter mais informações, 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.