aws

Guia do usuário

Amazon EC2 Auto Scaling

Amazon EC2 Auto Scaling: Guia do usuário

Table of Contents

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

Pré-requisitos 30
Etapa 1: Configurar um modelo de execução ou uma configuração de execução 31
Etapa 2: Criar um grupo do Auto Scaling 35
Etapa 3: Verificar se o balanceador de carga está anexado 36
Etapa 4: Próximas etapas 37
Etapa 5: Limpar 38
Recursos relacionados 39
Modelos de inicialização 40
Permissões para trabalhar com modelos de lançamento 41
Operações de API compatíveis com os modelos de execução 41
Criar um modelo de execução para um grupo do Auto Scaling 41
Criar seu modelo de execução (console) 42
Alterar as configurações da interface de rede padrão (console) 45
Modificar a configuração do armazenamento (console) 47
Criar um modelo de execução com base em uma instância existente (console) 51
Recursos relacionados 51
Limitações 52
Criar um modelo de execução usando configurações avançadas 52
Configurações necessárias 52
Configurações avançadas 53
Request Spot Instances 58
Blocos de capacidade para ML 60
Migre seus grupos de Auto Scaling para modelos de lançamento 65
Etapa 1: encontrar grupos do Auto Scaling que usem configurações de execução 65
Etapa 2: copiar uma configuração de execução para um modelo de execução 68
Etapa 3: atualizar um grupo do Auto Scaling para usar um modelo de execução 69
Etapa 4: substituir suas instâncias 70
Mais informações 70
Migre CloudFormation pilhas para modelos de lançamento 71
Encontre grupos do Auto Scaling que usam uma configuração de execução 71
Atualizar uma pilha para usar um modelo de execução 72
Compreender atualização de comportamentos de recursos da pilha 76
Rastrear a migração 77
Referência do mapeamento de configuração de execução 77
AWS CLI exemplos para trabalhar com modelos de lançamento 79
Exemplo de uso 80

Criar um modelo de execução básico ... 80
Especificar etiquetas que marcam instâncias ao iniciar ... 81
Especificar uma função do IAM a ser transmitida às instâncias ... 82
Atribuir um endereço IP público ... 82
Especificar um script de dados do usuário que configura instâncias ao iniciar ... 83
Especificar um mapeamento de dispositivos de blocos ... 83
Especificar hosts dedicados para trazer licenças de software de fornecedores externos ... 83
Especificar uma interface de rede existente ... 84
Criar várias interfaces de rede ... 84
Gerenciar modelos de execução ... 85
Atualizar um grupo do Auto Scaling para usar um modelo de execução ... 88
Use parâmetros do Systems Manager em vez de IDs de AMI ... 89
Crie um modelo de lançamento que especifique um parâmetro para a AMI ... 89
Verifique se um modelo de lançamento obtém a ID de AMI correta ... 94
Recursos relacionados ... 95
Limitações ... 95
Configurações de execução ... 97
Criar uma configuração de execução ... 98
Criar uma configuração de execução ... 98
Configurar IMDS ... 101
Criar uma configuração de execução usando uma instância do EC2 ... 104
Alterar uma configuração de execução ... 109
Grupos do Auto Scaling ... 110
Criar grupos do Auto Scaling usando modelos de execução ... 111
Criar um grupo usando um modelo de execução ... 112
Criar um grupo usando o assistente de execução do EC2 ... 115
Usar vários tipos de instâncias e opções de compra ... 120
Criar grupos do Auto Scaling usando configurações de execução ... 167
Criar um grupo usando uma configuração de execução ... 168
Criar um grupo usando uma instância do EC2 ... 171
Atualizar um grupo do Auto Scaling ... 177
Atualizar instâncias do Auto Scaling ... 179
Marcar grupos e instâncias ... 180
Restrições de nomeação e uso de tags ... 181
Ciclo de vida de marcação de instâncias do EC2 ... 181
Marcar seus grupos do Auto Scaling ... 182

Excluir tags ... 185
Etiquetas para segurança ... 186
Controlar o acesso usando etiquetas ... 187
Usar etiquetas para filtrar grupos do Auto Scaling ... 188
Políticas de manutenção de instância ... 192
Visão geral ... 193
Defina uma política de manutenção de instância no seu grupo ... 201
Ganchos do ciclo de vida ... 205
Disponibilidade de ganchos do ciclo de vida ... 207
Considerações e limitações ... 207
Recursos relacionados ... 209
Como os ganchos do ciclo de vida funcionam ... 210
Preparar para adicionar um gancho de ciclo de vida ... 212
Recuperar o estado de destino do ciclo de vida ... 220
Adicionar ganchos do ciclo de vida ... 222
Concluir uma ação do ciclo de vida ... 226
Tutorial: configurar dados do usuário para recuperar o estado de destino do ciclo de vida por meio de metadados de instância ... 228
Tutorial: Configurar um gancho do ciclo de vida que invoca uma função do Lambda ... 237
Grupos de alta atividade ... 247
Conceitos principais ... 248
Pré-requisitos ... 251
Atualização das instâncias em um grupo de aquecimento ... 252
Recursos relacionados ... 252
Limitações ... 252
Usar ganchos de ciclo de vida ... 254
Criar um grupo de alta atividade para seu grupo do Auto Scaling ... 258
Visualizar status da verificação de integridade ... 260
AWS CLI exemplos para trabalhar com piscinas aquecidas ... 263
Instâncias de desanexação e conexão ... 266
Considerações sobre a separação de instâncias ... 267
Considerações para anexar instâncias ... 268
Mova uma instância para um grupo diferente usando desanexar e anexar ... 268
Remover instâncias temporariamente ... 273
Como o estado de espera funciona ... 274
Considerações ... 275

Status de integridade de uma instância em um estado de espera 276
Remova temporariamente uma instância configurando-a como espera 274
Excluir infraestrutura do Auto Scaling 280
Excluir seu grupo do Auto Scaling 281
(Opcional) Excluir a configuração de execução 282
(Opcional) Excluir o modelo de execução 282
(Opcional) Excluir o balanceador de carga e grupos de destino 283
(Opcional) Excluir CloudWatch alarmes 284
AWS Exemplos de SDK para trabalhar com grupos de Auto Scaling 285
Criar um grupo do Auto Scaling 285
Update an Auto Scaling group 300
Descrever um grupo de Auto Scaling 311
Excluir um grupo do Auto Scaling 326
Recicle suas instâncias 339
Atualização de instância 339
Como funciona uma atualização de instância 340
Entender os valores padrão 346
Iniciar uma atualização de instância 349
Monitore uma atualização de instância 362
Cancelar uma atualização de instância 365
Desfazer alterações com uma reversão 366
Usar a opção de ignorar correspondência 371
Adicionar pontos de verificação 380
Vida útil máxima da instância 387
Considerações 387
Definir o tempo de vida máximo da instância 388
Limitações 389
Escalar o grupo 391
Escolha seu método de escalabilidade 391
Definir limites de escalabilidade 393
Definir o aquecimento de instância padrão 395
Considerações sobre o desempenho de escalabilidade 396
Escolha o tempo padrão de aquecimento da instância 397
Habilitar o aquecimento de instância padrão para um grupo 398
Verificar o aquecimento de instância padrão para um grupo 399

Encontre políticas de escalabilidade com um tempo de aquecimento de instância definido anteriormente ... 400
Limpe o aquecimento da instância definido anteriormente para uma política de escalabilidade ... 401
Escalabilidade manual ... 402
Alterar a capacidade desejada de seu grupo do Auto Scaling ... 403
Encerrar uma instância no seu grupo do Auto Scaling (AWS CLI) ... 406
Escalabilidade programada ... 407
Como a escalabilidade programada funciona ... 408
Programações recorrentes ... 409
Fuso horário ... 410
Considerações ... 410
Criar uma ação programada ... 411
Exibir detalhes da ação agendada ... 413
Verificar as atividades de escalabilidade ... 414
Excluir uma ação programada ... 414
Limitações ... 415
Escalabilidade dinâmica ... 415
Como funcionam as políticas de escalabilidade dinâmica ... 416
Várias políticas de escalabilidade dinâmica ... 418
Políticas de escalabilidade de rastreamento de destino ... 419
Políticas de escalabilidade simples e em etapas ... 433
Desaquecimento de escalabilidade ... 451
Escalabilidade baseada no Amazon SQS ... 454
Verificar uma ação de escalabilidade ... 462
Desabilitar uma política de escalabilidade ... 465
Excluir uma política de escalabilidade ... 467
AWS CLI exemplos de políticas de escalabilidade ... 470
Escalabilidade preditiva ... 473
Como a escalabilidade preditiva funciona ... 474
Crie uma política de escalabilidade preditiva ... 477
Avaliar as políticas de escalabilidade preditiva ... 486
Substituir a previsão ... 495
Usar métricas personalizadas ... 500
Controlar o término de instâncias ... 512
Cenários de término ... 512

Configurar políticas de rescisão ... 517
Criar uma política de término personalizada com o Lambda ... 522
Usar proteção de redução na escala na horizontal de instâncias ... 529
Criar para finalização de instância sem problemas ... 534
Suspender-retomar processos ... 538
Tipos de processos ... 539
Considerações ... 540
Suspender processos ... 540
Processos de currículo ... 541
Como os processos suspensos afetam outros processos ... 542
Monitor ... 547
Verificações de integridade ... 549
Sobre verificações de integridade ... 550
Veja o motivo das falhas na verificação de integridade ... 558
Defina o período de carência da verificação de integridade ... 559
Monitor com AWS Health Dashboard ... 562
Monitorar métricas do CloudWatch ... 563
Visualizar grafos de monitoramento no console do Amazon EC2 Auto Scaling ... 564
Métricas do CloudWatch para o Amazon EC2 Auto Scaling ... 569
Configurar monitoramento para instâncias do Auto Scaling ... 577
Registre chamadas de API com AWS CloudTrail ... 580
Informações do Amazon EC2 Auto Scaling em CloudTrail ... 580
Noções básicas sobre entradas do arquivo de log do Amazon EC2 Auto Scaling ... 581
Recursos relacionados ... 583
Opções de notificação do Amazon SNS ... 583
Amazon SNS e Amazon EC2 Auto Scaling ... 584
Trabalhar com outros serviços ... 591
Rebalanceamento de capacidade ... 591
Visão geral ... 592
Comportamento de rebalanceamento de capacidade ... 593
Considerações ... 594
Habilitar o rebalanceamento de capacidade (console) ... 596
Habilitar o rebalanceamento de capacidade (AWS CLI) ... 597
Recursos relacionados ... 602
Limitações ... 602
Reservas de capacidade ... 602

Etapa 1: criar as Reservas de Capacidade ... 603
Etapa 2: criar um grupo de Reserva de Capacidade ... 606
Etapa 3: criar um modelo de execução ... 608
Etapa 4: criar um grupo do Auto Scaling ... 609
Recursos relacionados ... 611
AWS CloudShell ... 612
AWS CloudFormation ... 612
Amazon EC2 Auto Scaling e modelos AWS CloudFormation ... 613
Saiba mais sobre AWS CloudFormation ... 613
Compute Optimizer ... 614
Limitações ... 614
Descobertas ... 615
Exibir recomendações ... 615
Considerações para avaliação das recomendações ... 616
Elastic Load Balancing ... 617
Tipos de Elastic Load Balancing ... 619
Prepare-se para conectar um balanceador de carga ... 620
Anexar um balanceador de carga ... 623
Configurar um balanceador de carga do console do Amazon EC2 Auto Scaling ... 627
Verificar o status do anexo ... 628
Adicionar zonas de disponibilidade ... 629
AWS CLI exemplos para trabalhar com o Elastic Load Balancing ... 633
VPC Lattice ... 641
Preparar para anexar um grupo de destino ... 643
Anexar um grupo de destino do VPC Lattice ... 646
Verificar o status do anexo ... 651
EventBridge ... 652
Referência de eventos do Amazon EC2 Auto Scaling ... 653
Exemplos de eventos e padrões de grupo de aquecimento ... 662
Crie EventBridge regras ... 668
Amazon VPC ... 673
VPC padrão ... 674
VPC não padrão ... 674
Considerações sobre a escolha de sub-redes da VPC ... 675
Endereçamento IP em uma VPC ... 675
Interfaces de rede em uma VPC ... 676

Locação de localização de instância 676
AWS Outposts 677
Mais recursos para saber mais sobre VPCs 677
Segurança 678
Segurança da infraestrutura 679
Recursos relacionados 679
Resiliência 679
Recursos relacionados 681
Proteção de dados 681
Use AWS KMS keys para criptografar volumes do Amazon EBS 682
Recursos relacionados 683
AWS KMS política chave para uso com volumes criptografados 683
Gerenciamento de identidade e acesso 690
Controle de acesso 690
Como o Amazon EC2 Auto Scaling funciona com o IAM 691
Permissões de API 701
Políticas gerenciadas 702
Perfis vinculados ao serviço 707
Exemplos de políticas baseadas em identidade 715
Prevenção contra o ataque “Confused deputy” entre serviços 724
Suporte a modelo de execução 726
Funções do IAM para aplicações que são executadas em instâncias do Amazon EC2 734
Validação de compatibilidade 737
Conformidade do PCI DSS 739
Usar endpoints da VPC para conectividade privada 739
Criar um VPC endpoint de interface 740
Criar uma política de endpoint da VPC 740
Solução de problemas 742
Recuperar uma mensagem de erro 742
Desative as atividades de escalabilidade 744
Recursos adicionais para solução de problemas 745
Falha ao iniciar instância 746
A configuração solicitada não é suportada atualmente. 747
O grupo de segurança <nome do grupo de segurança> não existe. Falha ao ativar a instância EC2. 748

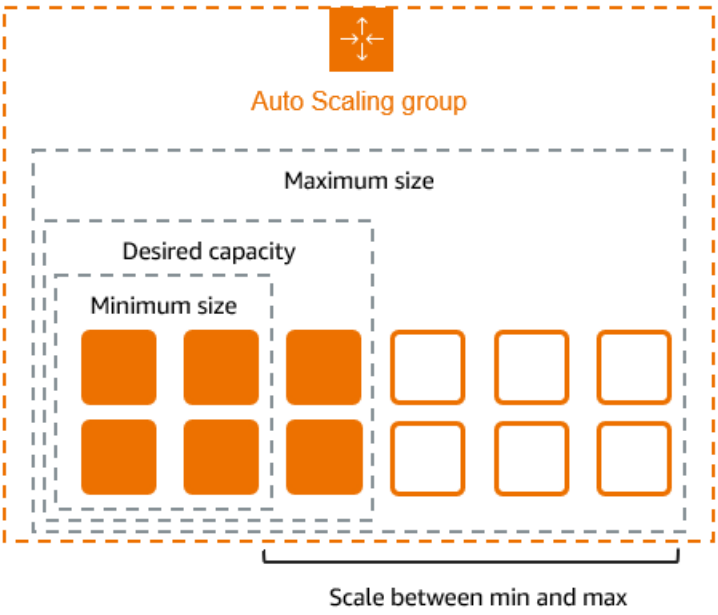
O par de chaves <par de chaves associado à sua instância do EC2> não existe. Falha ao ativar a instância EC2. 748
O tipo de instância solicitado (<tipo de instância>) não tem suporte na Zona de disponibilidade solicitada (<Zona de disponibilidade da instância>)... 749
Seu preço de solicitação spot de 0,015 é inferior ao preço mínimo de atendimento de solicitação spot exigido de 0,0735... 749
Nome de dispositivo inválido <nome do dispositivo> / Carregamento do nome de dispositivo inválido. Falha ao ativar a instância EC2. 749
O valor (<nome associado ao dispositivo de armazenamento de instâncias>) do parâmetro virtualName é inválido... Falha ao ativar a instância EC2. 750
Mapeamentos de dispositivos de blocos do EBS não suportados para AMIs de armazenamento de instância. 751
Os grupos de posicionamento não podem ser usados com instâncias do tipo '<instance type>'. Falha ao ativar a instância EC2. 751
Cliente. InternalError: Erro do cliente na inicialização. 751
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. 753
A reserva solicitada não tem capacidade compatível e disponível suficiente para essa solicitação. Falha ao ativar a instância EC2. 754
Sua reserva do bloco de capacidade <reservation id> ainda não está ativa. Falha ao ativar a instância EC2. 754
Não há capacidade spot disponível que corresponda à sua solicitação. Falha ao ativar a instância EC2. 755
<número de instâncias> instância(s) já estão em execução. Falha ao ativar a instância EC2. 755
Problemas de AMI 756
O ID da AMI <ID de sua AMI> não existe. Falha ao ativar a instância EC2. 756
A AMI <ID da AMI> está pendente e não pode ser executada. Falha ao ativar a instância EC2. 757
Nome do dispositivo inválido <device name>. Falha ao ativar a instância EC2. 757
A arquitetura 'arm64' do tipo de instância especificado não corresponde à arquitetura 'x86_64' da AMI especificada... Falha na execução da instância EC2. 757
A AMI '<AMI ID>' está desabilitada e não pode ser executada. Falha ao ativar a instância EC2. 759
Problemas do balanceador de carga 759

Um ou mais grupos de destino não encontrados. Falha na validação da configuração do balanceador de carga. 760
Não é possível encontrar o Load Balancer <seu load balancer>. Falha na validação da configuração do balanceador de carga. 761
Não há nenhum balanceador de carga ATIVO chamado <nome do balanceador de carga>. Falha ao atualizar a configuração do balanceador de carga. 761
A instância do EC2 <ID da instância> não está na VPC. Falha ao atualizar a configuração do balanceador de carga. 762
Problemas em modelos de execução 762
Você deve usar um modelo de inicialização totalmente formado válido (valor inválido) 762
Você não está autorizado a usar o modelo de execução (permissões insuficientes) 763
Verificações de integridade 764
Uma instância foi retirada de serviço em resposta a uma falha de verificação de status de instância do EC2 765
Uma instância foi retirada de serviço em resposta a uma reinicialização programada do EC2 766
Uma instância foi retirada de serviço em resposta a uma verificação de integridade do EC2 que indicou que ela tinha sido terminada ou interrompida 767
Uma instância foi retirada de serviço em resposta a uma falha na verificação de integridade do sistema ELB 768
Informações relacionadas 770
Histórico do documento 773
..... dcccxvii

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



Características do Amazon EC2 Auto Scaling

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

gerenciamento. Os grupos do Auto Scaling usam modelos de execução (ou configurações de execução) como modelos de configuração para suas instâncias do EC2.

A seguir estão os principais recursos do Amazon EC2 Auto Scaling:

Monitorando a integridade das instâncias em execução

O Amazon EC2 Auto Scaling monitora automaticamente a saúde e a disponibilidade de suas instâncias usando verificações de saúde do EC2 e substitui instâncias encerradas ou prejudicadas para manter a capacidade desejada.

Verificações de integridade personalizadas

Além das verificações de saúde integradas, você pode definir verificações de saúde personalizadas específicas do seu aplicativo para verificar se ele está respondendo conforme o esperado. Se uma instância falhar na verificação de integridade personalizada, ela será substituída automaticamente para manter a capacidade desejada.

Equilibrando a capacidade em todas as zonas de disponibilidade

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

Vários tipos de instâncias e várias opções de compra

Em um único grupo de Auto Scaling, você pode lançar vários tipos de instância e opções de compra (instâncias spot e sob demanda), permitindo otimizar os custos por meio do uso da instância spot. Você também pode aproveitar os descontos da Instância Reservada e do Savings Plan usando-os em conjunto com as Instâncias sob demanda do grupo.

Substituição automatizada de instâncias spot

Se o seu grupo inclui instâncias spot, o Amazon EC2 Auto Scaling pode solicitar automaticamente a substituição da capacidade spot se suas instâncias spot forem interrompidas. Por meio do rebalanceamento de capacidade, o Amazon EC2 Auto Scaling também pode monitorar e substituir proativamente suas instâncias spot que correm um risco elevado de interrupção.

Balanceamento de carga

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

íntegras. Sempre que as instâncias são iniciadas ou encerradas, o Amazon EC2 Auto Scaling registra e cancela automaticamente o registro das instâncias do balanceador de carga.

Escalabilidade

O Amazon EC2 Auto Scaling também fornece várias maneiras de escalar seus grupos de Auto Scaling. O uso do escalonamento automático permite que você mantenha a disponibilidade dos aplicativos e reduza os custos adicionando capacidade para lidar com picos de carga e removendo a capacidade quando a demanda é menor. Você também pode ajustar manualmente o tamanho do seu grupo de Auto Scaling conforme necessário.

Atualização de instância

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

Ganchos do ciclo de vida

Ganchos de ciclo de vida são úteis para definir ações personalizadas que são invocadas quando novas instâncias são iniciadas ou antes que as instâncias sejam encerradas. Esse recurso é particularmente útil para criar arquiteturas orientadas por eventos, mas também ajuda a gerenciar instâncias em seu ciclo de vida.

Support para cargas de trabalho com estado

Os ganchos de ciclo de vida também oferecem um mecanismo para persistir o estado no desligamento. Para garantir a continuidade de aplicativos com estado, você também pode usar proteção escalável ou políticas de encerramento personalizadas para evitar que instâncias com processos de longa execução sejam encerradas mais cedo.

Para obter mais informações sobre os benefícios do Amazon EC2 Auto Scaling consulte Benefícios do Amazon EC2 Auto Scaling.

Preços do Amazon EC2 Auto Scaling

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

Conceitos básicos

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

Trabalhar com grupos do Auto Scaling

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

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

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

Benefícios do Amazon EC2 Auto Scaling

Adicionar o Amazon EC2 Auto Scaling à sua arquitetura de aplicativos é uma forma 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.

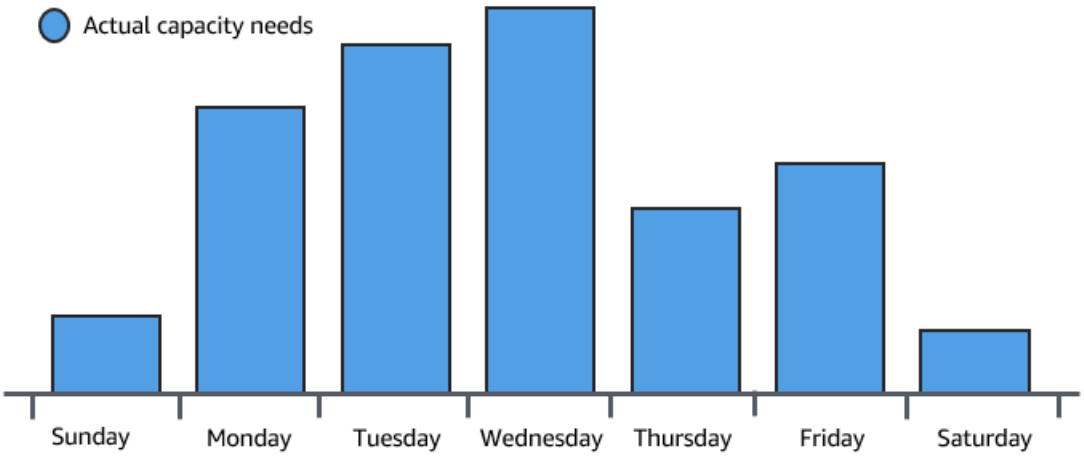
Conteúdo

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

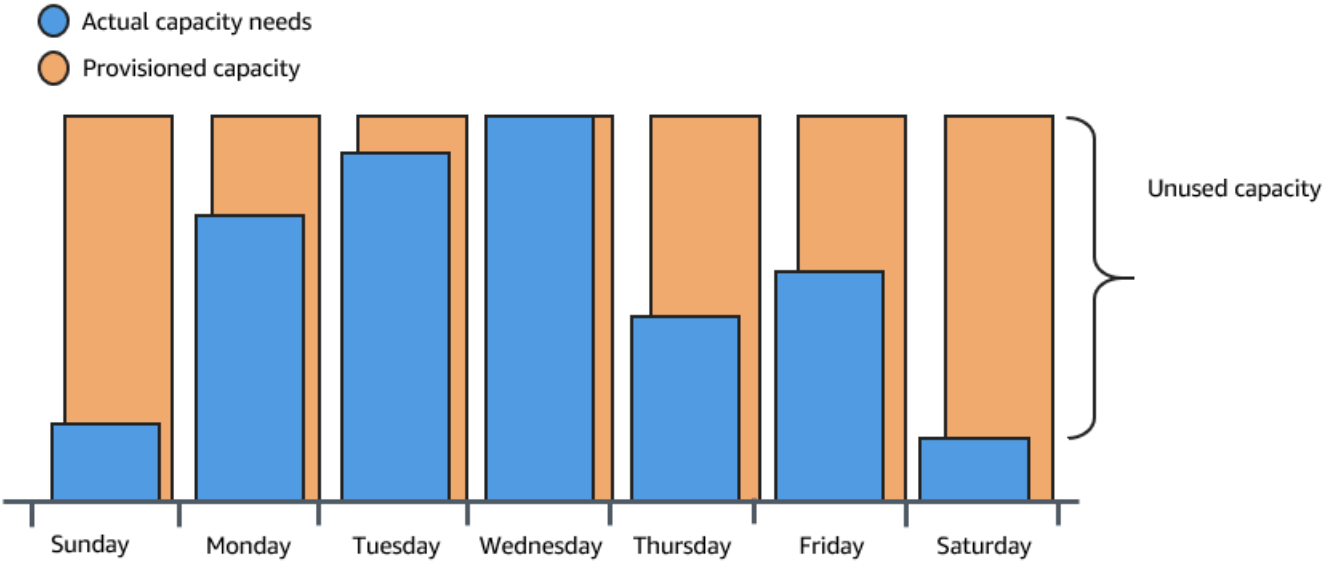
Exemplo: atender a demanda variável

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

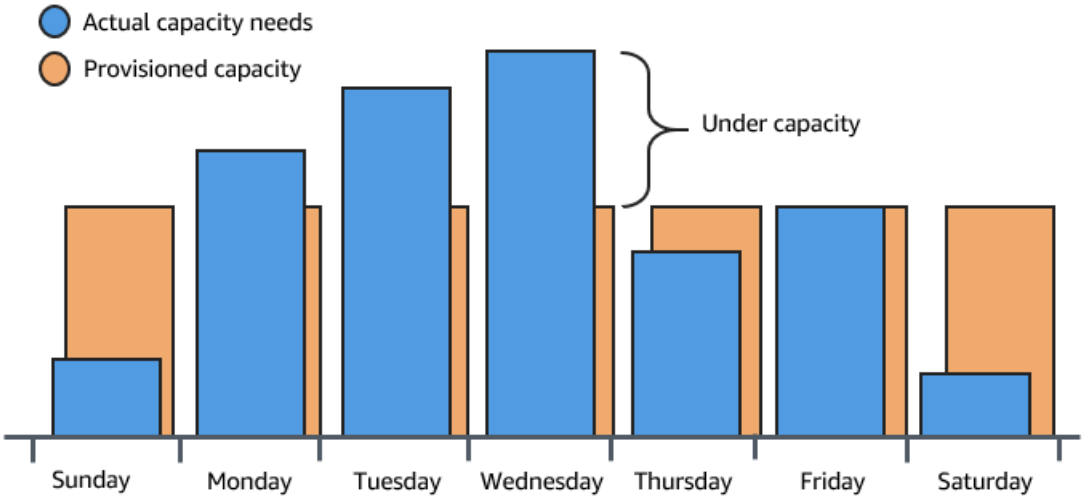
O gráfico a seguir mostra quanto da capacidade da aplicação é usado durante o período de uma semana.



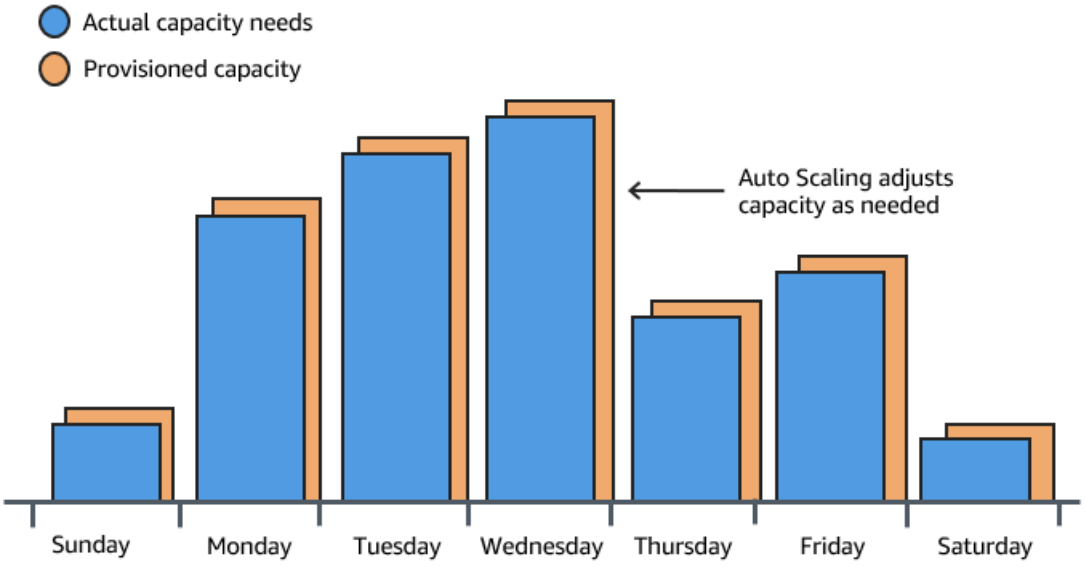
Tradicionalmente, há duas maneiras de planejar essas alterações na capacidade. A primeira opção é adicionar servidores suficientes para que a aplicação sempre tenha capacidade suficiente para atender à demanda. A desvantagem dessa opção, no entanto, é que há dias em que a aplicação não precisa de toda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



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



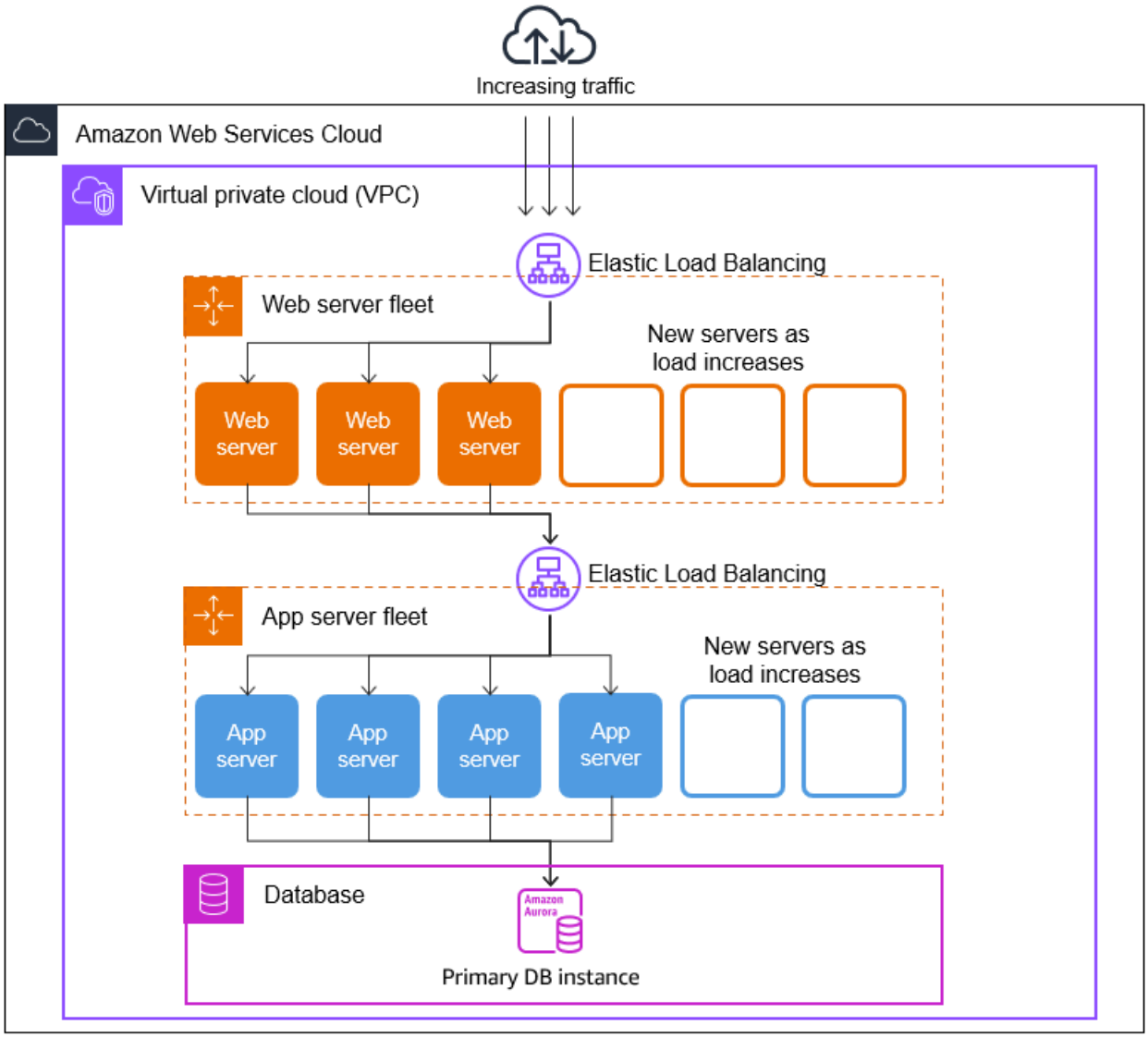
Ao adicionar o Amazon EC2 Auto Scaling a essa aplicação, você passa a ter uma terceira opção disponível. Você pode adicionar novas instâncias à aplicação somente quando necessário e encerrá-las quando não forem mais necessárias. Como o Amazon EC2 Auto Scaling usa instâncias do EC2, você só precisa pagar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



Exemplo: arquitetura de aplicação Web

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

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



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

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

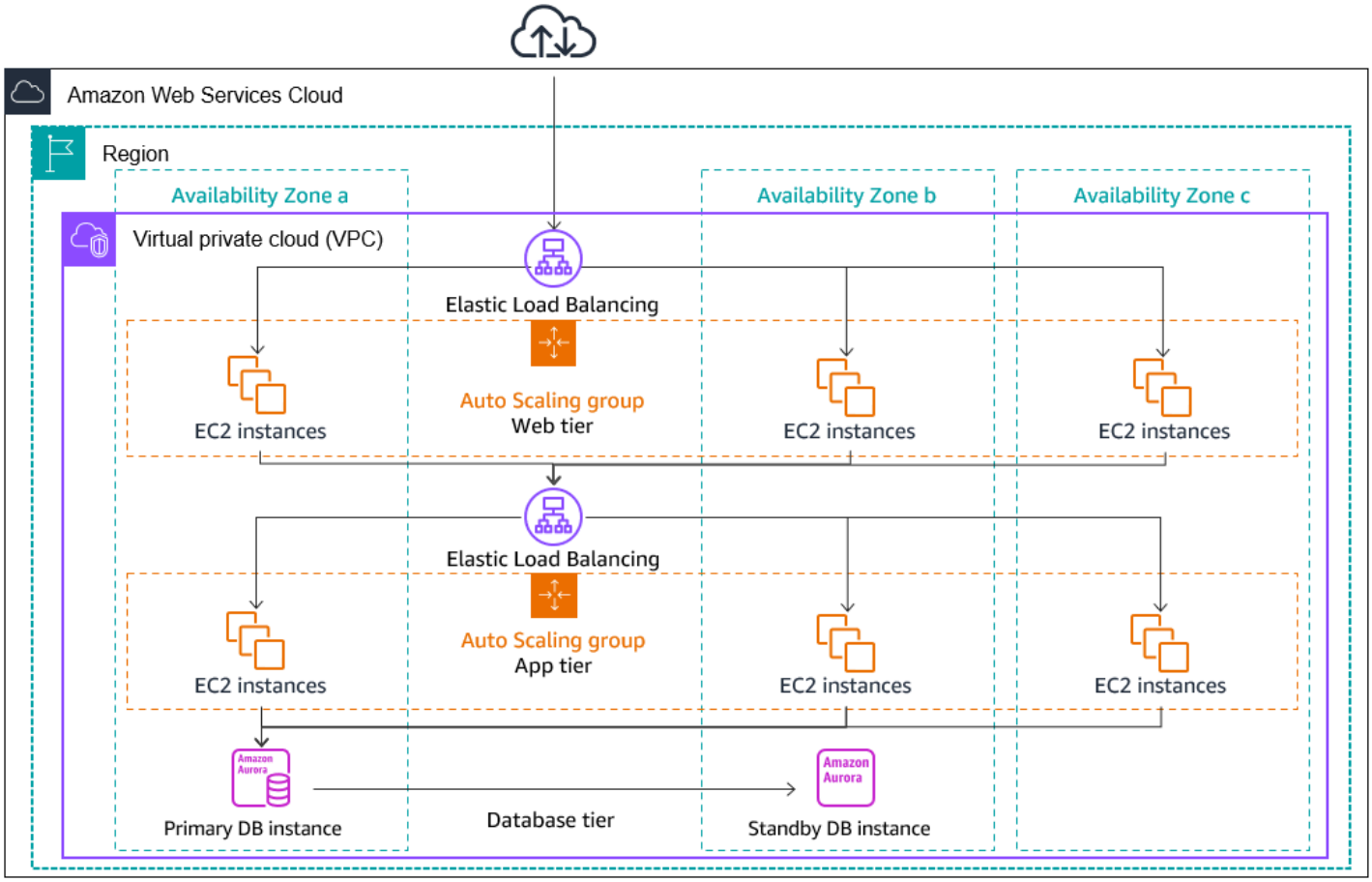
Exemplo: distribuir instâncias entre zonas de disponibilidade

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

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

Ao criar um grupo do Auto Scaling, 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. Por padrão, isso pode ser feito com uma margem de 10% ou uma instância, o que for maior. A margem só é estendida se o grupo estiver usando toda ou quase toda a capacidade máxima e precisar ser rebalanceado. A extensão dura somente o tempo necessário para rebalancear o grupo (em geral, alguns minutos).

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

Rebalanceamento de capacidade

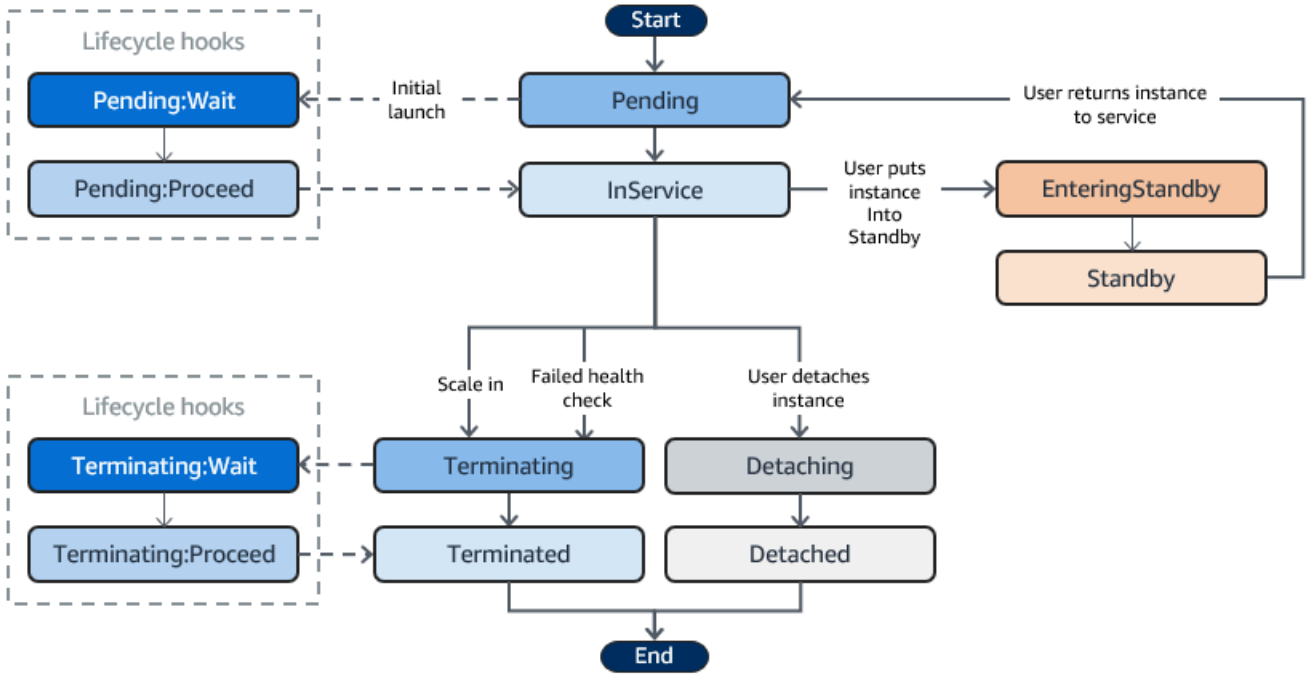
Você pode habilitar o rebalanceamento de capacidade nos grupos do Auto Scaling usando instâncias spot. O Amazon EC2 Auto Scaling tenta iniciar uma instância spot sempre que o Amazon EC2 informa que uma instância spot está em alto risco de ser interrompida. Após iniciar uma nova instância, ele encerra uma instância mais antiga. Para ter mais informações, consulte Usar o rebalanceamento de capacidade para lidar com interrupções de spot do Amazon EC2.

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:

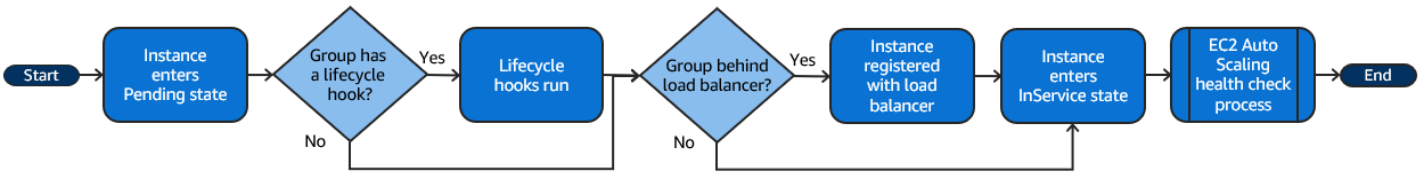
- Você aumenta o tamanho do grupo manualmente. Para ter mais informações, consulte Alterar a capacidade desejada de um grupo do Auto Scaling existente.
- Você cria uma política de escalabilidade para aumentar automaticamente o tamanho do grupo com base em um aumento especificado na demanda. Para ter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

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

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

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

Veja a seguir um resumo das etapas para registrar uma instância com um balanceador de carga para um evento de expansão.



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 ter mais informações, consulte Controlar quais instâncias do Auto Scaling serão terminadas durante uma redução de escala na horizontal.
- Você coloca a instância em um estado `Standby`. Para ter mais informações, consulte Entrar e sair de espera.
- Você desvincula a instância do grupo do Auto Scaling. Para ter mais informações, consulte Desanexar ou anexar instâncias.
- 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 ter mais informações, consulte Verificações de integridade para instâncias em um grupo do Auto Scaling.

Reduzir a escala na horizontal

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

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

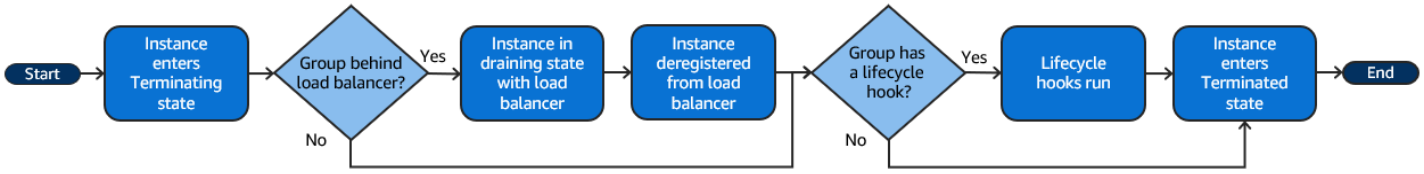
É importante criar um evento de redução correspondente para cada evento de expansão criado. Isso ajuda a garantir que os recursos atribuídos à aplicação correspondam à demanda por esses recursos da maneira mais próxima possível.

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

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

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

Veja a seguir um resumo das etapas para cancelar o registro de uma instância com um balanceador de carga para um evento de escalabilidade.



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 ter mais informações, consulte Desanexar ou anexar instâncias.

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 ter mais informações, consulte Desanexar ou anexar instâncias.

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

Entrar e sair de espera

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

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

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

Cotas do Amazon EC2 Auto Scaling

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

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

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

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

Recursos do Amazon EC2 Auto Scaling

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

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

Configuração do grupo do Auto Scaling

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

Recurso	Cota
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 de destino do Elastic Load Balancing por grupo do Auto Scaling	50

Recurso	Cota
Grupos de destino do VPC Lattice por grupo do Auto Scaling	5

Operações da API do grupo do Auto Scaling

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

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

Limitação de solicitações para a API Amazon EC2 Auto Scaling

As solicitações da API do Amazon EC2 Auto Scaling são limitadas usando um esquema de token bucket para manter a largura de banda do serviço. Para obter mais informações, consulte a taxa de solicitação de API na Referência de API do Amazon EC2 Auto Scaling.

Taxas de encerramento do EC2

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

Outros produtos da

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

Para obter cotas específicas para modelos de execução, consulte Restrições de modelos de execução no Guia do usuário do Amazon EC2 para instâncias do 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
- Preparar-se para usar a AWS CLI

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 grupos de Auto Scaling a partir da linha de comando usando AWS CloudShell.

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

Conceitos básicos do Amazon EC2 Auto Scaling

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

Tópicos

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

Para ver tutoriais adicionais que se concentram em ferramentas específicas para gerenciar o ciclo de vida de instâncias em um grupo de Auto Scaling, consulte os tópicos a seguir:

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

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

- Quantas zonas de disponibilidade o grupo do Auto Scaling deve abranger.
- Quais recursos existentes podem ser usados, como grupos de segurança ou 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 provisionar um servidor.

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

Tutorial: Crie seu primeiro grupo de Auto Scaling

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

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

Tarefas

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

Preparar para a demonstração

Este passo a passo pressupõe que você esteja familiarizado com a execução de instâncias do EC2 e que já criou um par de chaves e um grupo de segurança. Para obter mais informações, consulte Configuração do Amazon EC2 no Manual do usuário do Amazon EC2 para instâncias do Linux.

Para começar a usar o Amazon EC2 Auto Scaling, você pode usar a VPC padrão para seu. Conta da AWS A VPC padrão inclui uma sub-rede pública padrão em cada zona de disponibilidade e um gateway de Internet conectado à VPC. Você pode 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 cria um modelo de execução que especifica o tipo de instância do EC2 que o Amazon EC2 Auto Scaling cria para você. Inclua informações, como o ID da imagem de máquina da Amazon (AMI) a ser usada, o tipo de instância, o par de chaves e os grupos de segurança.

Para criar um modelo de execução

1. Abra o console do Amazon EC2 e acesse a página de modelos do Launch.
2. Na barra de navegação superior, selecione um Região da AWS. O modelo de execução e os recursos do grupo do Auto Scaling que você cria são vinculados à região que você especifica.
3. Escolha Criar modelo de execução.
4. Para o Launch template name (Nome do modelo de execução), insira **`my-template-for-auto-scaling`**.
5. Em Auto Scaling guidance (Guia do Auto Scaling), marque a caixa de seleção.
6. Em Application and OS Images (Amazon Machine Image) (Imagens de aplicações e sistemas operacionais [imagem de máquina da Amazon]), escolha uma versão do Amazon Linux 2 (HVM) na lista Quick Start (Início rápido). A AMI serve como modelo de configuração básico para suas instâncias.
7. Em Instance type (Tipo de instância), selecione uma configuração de hardware que seja compatível com a AMI que você especificou.
8. (Opcional) Em Key pair (login) (Par de chaves [login]), escolha um par de chaves existente. Você usa pares de chaves para se conectar a uma instância do Amazon EC2 co o SSH. A conexão a uma instância não está incluída como parte deste tutorial. Portanto, não é necessário especificar um par de chaves, a menos que pretenda se conectar à instância usando SSH.
9. Em Network settings (Configurações de rede), expanda Advanced network configuration (Configuração de rede avançada) e execute estas ações:

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

    b. Para atribuir automaticamente IP público, especifique se sua instância recebe um endereço IPv4 público. Por padrão, o Amazon EC2 atribui um endereço IPv4 público se a instância do EC2 for iniciada em uma sub-rede padrão ou se a instância for iniciada em uma sub-rede configurada para atribuir automaticamente um endereço IPv4 público. Se você não precisar se conectar à sua instância, escolha Desativar.

c. Para ID do grupo de segurança, escolha um grupo de segurança na mesma VPC que você planeja usar como VPC para seu grupo de Auto Scaling. Se você não especificar um grupo de segurança, sua instância será automaticamente associada ao grupo de segurança padrão da VPC.

d. Em Excluir ao encerrar, escolha Sim para excluir a interface de rede quando a instância for excluída.

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

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

Use o procedimento a seguir para continuar de onde você parou depois de criar um modelo de lançamento.

Para criar um grupo do Auto Scaling

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

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

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

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

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

Etapa 3: Verificar seu grupo do Auto Scaling

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

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

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

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

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

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

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

   - A guia Lifecycle (Ciclo de vida) mostra o estado de sua instância. Inicialmente, sua instância está no estado `Pending`. Quando uma instância está pronta para receber tráfego, seu estado é `InService`.
   - A coluna Health status mostra o resultado das verificações de saúde do Amazon EC2 Auto Scaling 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 a página de grupos do Auto Scaling do console do Amazon EC2.
2. Marque a caixa de seleção ao lado do seu grupo do Auto Scaling.
3. Na guia Instance management (Gerenciamento de instâncias), em Instances (Instâncias), selecione o ID da instância.

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

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

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

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

Etapa 5: Próximas etapas

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

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

Você também pode começar a se familiarizar com os conceitos de escalonamento lendo sobre Políticas de escalabilidade de rastreamento de destino. Se a carga 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 escalabilidade para seu grupo do Auto Scaling.

Etapa 6: limpar

Você pode excluir sua infraestrutura de escalabilidade ou excluir apenas seu grupo de Auto Scaling e manter seu modelo de lançamento para uso posterior.

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

Para excluir seu grupo do Auto Scaling

1. Abra a página de grupos do Auto Scaling do console do Amazon EC2.
2. Marque a caixa de seleção ao lado do seu grupo do Auto Scaling (`my-first-asg`).
3. Escolha Delete.
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

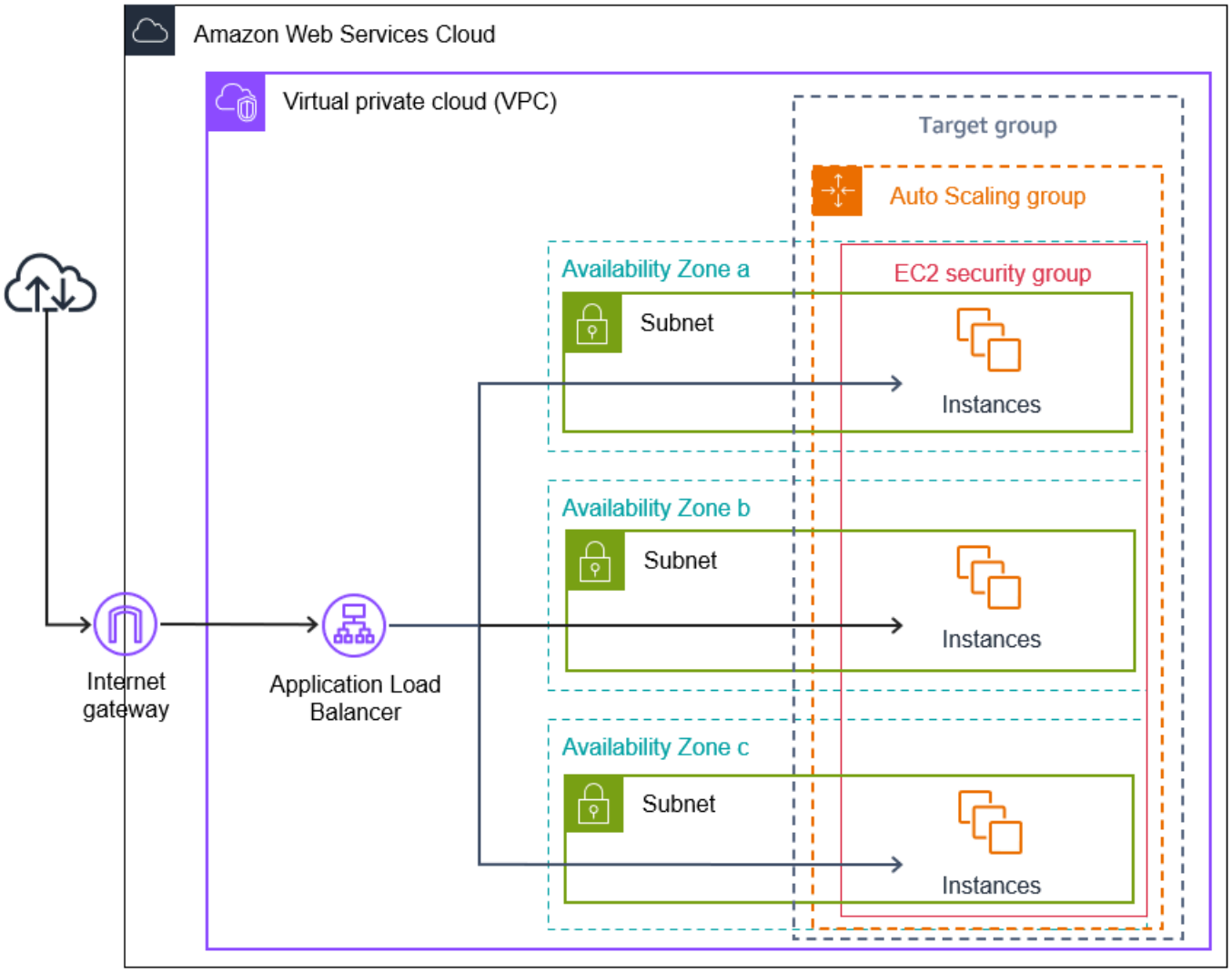
1. Abra a página Launch templates (Modelos de execução) do console do Amazon EC2.
2. Selecione o modelo de execução (`my-template-for-auto-scaling`).
3. Escolha Actions (Ações), Delete template (Excluir modelo).
4. Quando a confirmação for solicitada, digite **`Delete`** para confirmar a exclusão do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

Tutorial: Configurar uma aplicação escalonada e com balanceamento de carga

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

O registro do seu grupo do Auto Scaling em um balanceador de carga Elastic Load Balancing ajuda você a configurar uma aplicação com balanceamento de carga. O Elastic Load Balancing funciona com o Amazon EC2 Auto Scaling para distribuir o tráfego de entrada entre suas instâncias íntegras do Amazon EC2. Isso aumenta a escalabilidade e a disponibilidade da sua aplicação. É possível habilitar o Elastic Load Balancing em várias zonas de disponibilidade para aumentar a tolerância a falhas das aplicações.

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



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

Para obter mais informações sobre como introduzir um balanceador de carga em sua arquitetura, consulte Usar o Elastic Load Balancing para distribuir tráfego entre as instâncias no grupo do Auto Scaling.

Tarefas

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

Pré-requisitos

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

  Opcionalmente, se as instâncias tiverem endereços IP públicos, também será possível permitir tráfego de SSH para conexão com as instâncias.
- (Opcional) Uma função do IAM que concede acesso ao seu aplicativo AWS a.
- (Opcional) Uma imagem de máquina da Amazon (AMI) definida como sendo o modelo de origem para suas instâncias do Amazon EC2. Para criar um agora, execute uma instância. Especifique 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.

Tópicos

- Selecione ou crie um modelo de lançamento
- Criar ou selecionar uma configuração de execução

Selecione ou crie um modelo de lançamento

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

Para selecionar um modelo de execução existente

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

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

Para criar um modelo de execução

1. Abra a página Launch templates (Modelos de execução) do console do Amazon EC2.
2. Na barra de navegação, na parte superior da tela, escolha a região onde o balanceador de carga foi criado.
3. Escolha Criar modelo de execução.
4. Insira um nome e forneça uma descrição para a versão inicial do modelo de execução.
5. Em Application and OS Images (Amazon Machine Image) (Imagens de aplicações e sistemas operacionais [imagem de máquina da Amazon]), escolha o ID da AMI de suas instâncias. Você pode pesquisar todas as 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. Para atribuir automaticamente IP público, especifique se suas instâncias recebem endereços IPv4 públicos. Por padrão, o Amazon EC2 atribui um endereço IPv4 público se a instância do EC2 for iniciada em uma sub-rede padrão ou se a instância for iniciada em uma sub-rede configurada para atribuir automaticamente um endereço IPv4 público. Se você não precisar se conectar às suas instâncias, escolha Desativar para evitar que as instâncias do seu grupo recebam tráfego diretamente da Internet. Nesse caso, elas receberão tráfego somente do load balancer.

    c. Em Security group ID (ID do grupo de segurança), especifique um grupo de segurança para suas instâncias a partir da mesma VPC que o balanceador de carga.

    d. Em Delete on termination (Excluir ao término), escolha Yes (Sim). Isso excluirá a interface de rede quando o grupo do Auto Scaling reduzir a escala na horizontal e terminará a instância na qual a interface de rede está anexada.

9. (Opcional) Para distribuir as credenciais de forma segura para as suas instâncias, em Advanced details (Detalhes avançados), IAM instance profile (Perfil de instância do IAM), digita o nome de recurso da Amazon (ARN) 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 Criar modelo de execução.
12. Na página de confirmação, escolha Create Auto Scaling group (Criar grupo do Auto Scaling).

Criar ou selecionar uma configuração de execução

> **Note**
>
> Nós desencorajamos fortemente o uso de configurações de lançamento em novos aplicativos porque é um recurso antigo sem investimento planejado. Além disso, 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. Para ter mais informações, consulte Configurações de execução.

Para selecionar uma configuração de ativação existente

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

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

Para criar uma configuração de execução

1. Abra a página Launch configurations (Configurações de execução) do console do Amazon EC2. Quando a confirmação for solicitada, escolha Exibir configurações de inicialização para confirmar que deseja visualizar a página Configurações de inicialização.
2. Na barra de navegação superior, escolha a região onde o balanceador de carga foi criado.
3. SelecioneCreate launch configuration (Criar uma configuração de execução), e insira um nome para sua configuração de execução.

4. Em Amazon machine image (AMI) (Imagem de máquina da Amazon), insira o ID da AMI para suas instâncias como critério de pesquisa.
5. Em Instance type (Tipo de instância), selecione uma configuração de hardware para sua instância.
6. Em Additional configuration (Configuração adicional), preste atenção aos seguintes campos:

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

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

   c. (Opcional) Em Advanced details (Detalhes avançados), IP address type (Tipo de endereço IP), mantenha o valor padrão. Ao criar seu grupo do Auto Scaling, é possível atribuir um endereço IP público a instâncias no seu grupo do Auto Scaling usando sub-redes que têm o atributo de endereçamento IP público habilitado, como as sub-redes padrão na VPC padrão. Como alternativa, se você não precisar se conectar às suas instâncias, escolha Do not assign a public IP address to any instances (Não atribuir um endereço IP público a nenhuma instância) para impedir que as instâncias no seu grupo recebam tráfego diretamente da Internet. Nesse caso, elas receberão tráfego somente do load balancer.
7. Em Security groups (Grupos de segurança), escolha um grupo de segurança existente na mesma VPC que o balanceador de carga. Se você mantiver Create a new security group (Criar um novo grupo de segurança) selecionado, uma regra SSH padrão será configurada para instâncias do Amazon EC2 que executem Linux. Uma função do RDP padrão é configurada para instâncias do Amazon EC2 que executem o Windows.
8. Em Key pair (login) (Par de chaves - login), escolha uma opção em Key pair options (Opções de par de chaves).

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

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

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

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

Etapa 2: Criar um grupo do Auto Scaling

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

Para criar um grupo do Auto Scaling

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

   A página Choose instance launch options (Escolher as opções de execução da instância) será exibida, permitindo a você escolher as configurações de rede VPC que você deseja que o grupo do Auto Scaling use e oferecendo opções de execução para instâncias spot e sob demanda (se você escolher um modelo de execução).
4. Na seção Network (Rede), para VPC, selecione a VPC usada para o balanceador de carga. Se você escolher a VPC padrão, ela será configurada automaticamente para fornecer conectividade com a Internet às instâncias. Essa VPC inclui uma sub-rede pública em cada zona de disponibilidade na região.
5. Para Availability Zones and subnets (Zonas de disponibilidade e sub-redes), selecione uma ou mais sub-redes de cada zona de disponibilidade que você deseja incluir, baseando-se em quais zonas de disponibilidade o balanceador de carga se encontra. Para ter mais informações, consulte Considerações sobre a escolha de sub-redes da VPC.

6. [Apenas modelo de execução] Na seção Instance type requirements (Requisitos de tipo de instância), use a configuração padrão para simplificar esta etapa. (Não substitua o modelo de execução.) Neste tutorial, você fará o execução apenas das Instâncias sob demanda usando o tipo de instância especificado no modelo de execução.
7. Selecione Next (Próximo) para ir até a página Configure advanced options (Configurar opções avançadas).
8. Para anexar o grupo a um balanceador de carga existente, na seção Load balancing (Balanceamento de carga), selecione Attach to an existing load balancer (Anexar a um balanceador de carga existente). É possível selecionar Choose from your load balancer target groups (Escolher entre seus grupos de destino do balanceador de carga) ou Choose from Classic Load Balancers (Escolher entre balanceadores de carga clássicos). Em seguida, você pode escolher o nome de um grupo de destino para o Application Load Balancer ou Network Load Balancer criado ou escolher o nome de um Classic Load Balancer.
9. (Opcional) Para 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. Para step-by-step obter instruções, consulteAlterar a capacidade desejada de seu grupo do Auto Scaling.

Etapa 3: Verificar se o balanceador de carga está anexado

Como verificar se o balanceador de carga está associado

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

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

Etapa 4: Próximas etapas

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

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

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 a página de grupos do Auto Scaling do console do Amazon EC2.
2. Marque a caixa de seleção ao lado do seu grupo do Auto Scaling.
3. Escolha Delete.
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 Excluir.

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

Para excluir o balanceador de carga

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

Para excluir seu grupo de destino

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

Recursos relacionados

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

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

Modelos de inicialização

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

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

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

Com modelos de execução, você também pode usar recursos mais recentes do Amazon EC2. Isso inclui, parâmetros do Systems Manager (ID de AMI), a atual geração de volumes de IOPS provisionadas do EBS (io2), a marcação de volume do EBS, instâncias T2 ilimitadas, Reservas de Capacidade, blocos de capacidade e hosts dedicados, entre outros.

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

Conteúdo

- Permissões para trabalhar com modelos de lançamento

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

Permissões para trabalhar com modelos de lançamento

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

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

Para ver exemplos de políticas do IAM que permitem chamar as operações de `CreateAutoScalingGroup` `UpdateAutoScalingGroup`, e de `RunInstances` API com um modelo de lançamento, consulteSuporte a modelo de execução.

Operações de API compatíveis com os modelos de execução

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

Criar um modelo de execução para um grupo do Auto Scaling

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

Siga os procedimentos abaixo para criar novos modelos de execução.

Conteúdo

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

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

Criar seu modelo de execução (console)

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

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

Para criar um modelo de execução

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

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

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

      - 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.
      - Expanda Optional instance type attributes (Atributos de tipo de instância opcionais) e escolha Add attribute (Adicionar atributo) para limitar ainda mais os tipos de instâncias que podem ser usadas para atender à capacidade desejada. Para obter informações sobre cada atributo, consulte InstanceRequirementsRequesta Referência de API do Amazon EC2.

- Tipos de instância resultantes: 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.

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 informaçõe, consulte Grupos de segurança padrão para sua VPCs no Guia do usuário da Amazon VPC.

e. Execute um destes procedimentos:

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

f. Execute um destes procedimentos:

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

g. Em Resource tags (Etiquetas de recurso), especifique as etiquetas fornecendo combinações de chave e valor. Se você especificar tags de instância em seu modelo de execução e optar por propagar tags de seu grupo do Auto Scaling para suas instâncias,

todas as tags serão mescladas. Se a mesma chave da etiqueta for especificada para uma etiqueta no modelo de execução e uma etiqueta no grupo do Auto Scaling, então, o valor da etiqueta do grupo terá precedência.

6. Definir configurações avançadas (opcional) Por exemplo, você pode escolher 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. Para ter mais informações, consulte Criar um modelo de execução usando configurações avançadas.
7. Quando você estiver pronto para criar seu modelo de execução, escolha Create launch template (Criar modelo de execução).
8. Para criar um grupo do Auto Scaling, escolha Create Auto Scaling group (Criar grupo do Auto Scaling) na página de confirmação.

Alterar as configurações da interface de rede padrão (console)

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

Esta seção mostra como alterar as configurações padrão da interface de rede. Por exemplo, você pode definir se deseja atribuir um 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 fazer isso, você deve usar o AWS CLI ou um SDK para criar o grupo 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.
- 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 e de machine learning, altere a interface de rede para uma interface de rede do Elastic Fabric Adapter. Para obter mais informações, consulte Elastic Fabric Adapter no Manual do usuário do Amazon EC2 para instâncias do Linux.

i. Network card index (Índice da placa de rede): escolha 0para 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.

j. ENA Express: Para tipos de instância compatíveis com ENA Express, escolha Ativar para ativar o ENA Express ou Desativar para desativá-lo. Para obter mais informações, consulte Melhorar o desempenho de rede com o ENA Express em instâncias do Linux no Guia do usuário do Amazon EC2 para instâncias do Linux.

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

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

Modificar a configuração do armazenamento (console)

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

Para modificar a configuração do armazenamento

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

   Se o valor especificado para o tamanho do volume estiver fora dos limites do tipo de volume ou menor que o tamanho do snapshot, uma mensagem de erro será exibida. Para 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 os volumes do Amazon EBS no Guia do usuário do Amazon EBS e o armazenamento de instâncias do Amazon EC2 no Guia do usuário do Amazon EC2 para instâncias 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 sobre o tamanho e a configuração de um volume do EBS no Guia do usuário do Amazon EBS.

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

f. IOPS: se você 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 Criptografia do Amazon EBS e Ativar criptografia por padrão no Guia do usuário do Amazon EBS.

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

   Resultados da criptografia

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

Se o parâmetro **`Encrypted`** estiver definido como...	E se a origem de volume for...	O estado de criptografia padrão será...	Observações
	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 escolheuYes (Sim) para Encrypted (Criptografado), deve selecionar uma chave gerenciada pelo cliente a ser usada para criptografar o volume. Se tiver habilitado a criptografia por padrão nessa região, a chave gerenciada pelo cliente padrão será selecionada para você. Você pode selecionar uma chave diferente ou especificar o ARN de qualquer chave gerenciada pelo cliente que você criou anteriormente usando o AWS Key Management Service.

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

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

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

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

Recursos relacionados

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

Para obter informações sobre modelos de execução, consulte Executar uma instância de um modelo de execução no Guia do usuário do Amazon EC2 para instâncias do Linux.

Limitações

- Embora você possa especificar uma sub-rede em um modelo de execução, isso não é necessário se você usar o modelo de execução somente para criar grupos do Auto Scaling. Você não pode especificar a sub-rede para um grupo do Auto Scaling especificando a sub-rede em um modelo de execução. As sub-redes do grupo do Auto Scaling são retiradas da própria definição de recursos do grupo do Auto Scaling.
- Sobre outras limitações em interfaces de rede definidas pelo usuário, consulte Alterar as configurações da interface de rede padrão (console).

Criar um modelo de execução usando configurações avançadas

Este tópico descreve como criar um modelo de lançamento com configurações avançadas a partir do AWS Management Console.

Para criar um modelo de lançamento usando configurações avançadas

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

Configurações necessárias

Ao criar um modelo de lançamento, você deve incluir as seguintes configurações obrigatórias.

Nome do modelo de lançamento

Insira um nome exclusivo que descreva o modelo de lançamento.

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

Escolha a Amazon Machine Image (AMI) que você deseja usar. Você pode pesquisar ou procurar a AMI que deseja usar. Para obter a melhor eficiência de escalabilidade, escolha uma AMI

personalizada que esteja totalmente configurada para iniciar uma instância com o código do seu aplicativo e que exija poucas modificações na inicialização.

Tipo de instância

Escolha um tipo de instância que seja compatível com sua AMI. Você pode pular a adição de um tipo de instância ao seu modelo de execução se planeja usar vários tipos de instâncias incorporados na própria definição de recursos do grupo Auto Scaling. Um tipo de instância só é necessário se você não planeja criar um grupo misto de instâncias.

Configurações avançadas

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

Amplie a seção Detalhes avançados para ver as configurações avançadas. As seções a seguir descrevem as configurações avançadas mais úteis nas quais se concentrar ao criar um modelo de execução para um grupo de Auto Scaling. Para obter mais informações, consulte Detalhes avançados no Guia do usuário do Amazon EC2 para instâncias Linux.

Perfil de instância do IAM

O perfil da instância contém a função do IAM que você deseja usar. Quando seu grupo de Auto Scaling inicia uma instância do EC2, as permissões definidas na função do IAM associada são concedidas aos aplicativos em execução na instância. Para ter mais informações, consulte Funções do IAM para aplicações que são executadas em instâncias do Amazon EC2.

Termination protection

Quando ativado, esse recurso impede que os usuários encerrem uma instância usando o console do Amazon EC2, os comandos da CLI e as operações de API. A proteção contra rescisão fornece uma proteção extra contra rescisão acidental. Isso não impede que o Amazon EC2 Auto Scaling encerre uma instância. Para controlar quais instâncias o Amazon EC2 Auto Scaling pode encerrar, consulte. Usar proteção de redução na escala na horizontal de instâncias

CloudWatch Monitoramento detalhado

Você pode ativar o monitoramento detalhado de suas instâncias do EC2 para permitir que elas enviem dados métricos para a Amazon CloudWatch em intervalos de 1 minuto. Por padrão, as instâncias do EC2 enviam dados métricos em intervalos CloudWatch de 5 minutos. Aplicam-

se cobranças adicionais. Para ter mais informações, consulte Configurar monitoramento para instâncias do Auto Scaling.

Especificação de crédito

O Amazon EC2 fornece instâncias de desempenho intermitentes, como T2, T3 e T3a, que permitem que os aplicativos ultrapassem o desempenho básico da CPU quando necessário. Por padrão, essas instâncias podem estourar por um tempo limitado antes que o uso da CPU seja limitado. Opcionalmente, você pode ativar o modo ilimitado para que as instâncias possam ultrapassar a linha de base pelo tempo que for necessário. Isso permite que os aplicativos mantenham o alto desempenho da CPU quando necessário. Podem se aplicar cobranças adicionais. Para obter mais informações, consulte Usar um grupo de Auto Scaling para iniciar uma instância de desempenho com capacidade de intermitência como ilimitada no Guia do usuário do Amazon EC2 para instâncias Linux.

Nome do placement group

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

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

- Se um grupo de posicionamento for especificado no modelo de lançamento e no grupo Auto Scaling, o grupo de posicionamento do grupo de Auto Scaling terá precedência. Depois que o grupo é criado, o grupo de posicionamento especificado nas configurações do grupo Auto Scaling não pode ser alterado.
- Em AWS CloudFormation, tenha cuidado ao definir um grupo de posicionamento no modelo de lançamento. O Amazon EC2 Auto Scaling lançará instâncias no grupo de posicionamento especificado. No entanto, não CloudFormation receberá sinais dessas instâncias se você usar um UpdatePolicycom seu grupo de Auto Scaling (embora isso possa mudar no futuro).

Opção de compra

Você pode escolher Solicitar instâncias spot para solicitar instâncias spot pelo preço spot, limitado ao preço sob demanda, e escolher 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 ter mais informações, consulte Solicitar instâncias spot para aplicações flexíveis e com tolerância a falhas. Esta configuração pode ser útil em circunstâncias especiais,

mas, em geral, é melhor não especificá-la e, em seu lugar, é melhor criar um grupo misto de instâncias. Para ter mais informações, consulte Grupos do Auto Scaling com vários tipos de instâncias e opções de compra.

Se você especificar uma solicitação de instância spot em seu modelo de execução, não poderá criar um grupo misto de instâncias. Se você tentar usar um modelo de execução que solicite instâncias spot com um grupo misto de instâncias, você receberá a seguinte mensagem de erro: `Incompatible launch template: You cannot use a launch template that is set to request Spot Instances (InstanceMarketOptions) when you configure an Auto Scaling group with a mixed instances policy. Add a different launch template to the group and try again.`

Capacity Reservation

As reservas de capacidade permitem que você reserve capacidade para suas instâncias do Amazon EC2 em uma zona de disponibilidade específica por qualquer período. Para obter mais informações, consulte Como trabalhar com Reservas de Capacidade no Guia do usuário do Amazon EC2 para instâncias do Linux.

Você pode escolher se deseja executar instâncias em:

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

Para atingir uma reserva de capacidade específica, o tipo de instância em seu modelo de execução deve corresponder ao tipo de instância da reserva. Ao criar seu grupo de Auto Scaling, use a mesma zona de disponibilidade da reserva de capacidade. Dependendo do Região da AWS que você escolher, você pode escolher como alvo um Bloco de Capacidade. Para ter mais informações, consulte Use blocos de capacidade para cargas de trabalho de aprendizado de máquina.

Para atingir um grupo de reservas de capacidade, consulteUse reservas de capacidade sob demanda para reservar capacidade em zonas de disponibilidade específicas. Ao segmentar um grupo de reservas de capacidade, você pode ter a capacidade distribuída em várias zonas de disponibilidade para melhorar a resiliência.

Locação

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

- Compartilhado (compartilhado) — Vários Contas da AWS podem compartilhar o mesmo hardware físico. Essa é a opção de locação padrão ao iniciar uma instância.
- Instâncias dedicadas (dedicadas) — Sua instância é executada em hardware de inquilino único. Nenhum outro AWS cliente compartilha o mesmo servidor físico. Para obter mais informações, consulte Instâncias dedicadas no Guia do usuário do Amazon EC2 para instâncias Linux.
- Hosts dedicados (host dedicado) — A instância é executada em um servidor físico dedicado ao seu uso. O uso de hosts dedicados facilita a transferência de suas próprias licenças (BYOL) que tenham requisitos de hardware dedicados para o EC2 e atendam aos casos de uso de conformidade. Se você escolher essa opção, deverá fornecer um grupo de recursos de host para o grupo de recursos de host de locação. Para obter mais informações, consulte Hosts dedicados no Guia do usuário do Amazon EC2 para instâncias do Linux.

Support for Dedicated Hosts 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.

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

Grupo de recursos do host de locação

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

Configurações de licença

Com essa configuração, você pode especificar uma configuração de licença para suas instâncias sem restringir sua locação a hosts dedicados. A configuração da licença rastreia as licenças de

software implantadas nas instâncias para que você possa monitorar o uso e a conformidade da licença. Para obter mais informações, consulte Criar uma licença autogerenciada no Guia do Usuário do License Manager.

Metadados acessíveis

Você pode escolher se deseja ativar ou desativar o acesso ao endpoint HTTP do serviço de metadados da instância. Por padrão, o endpoint de HTTP está habilitado. Se você optar por desabilitar o endpoint, o acesso aos metadados da instância será desativado. Só é possível especificar a condição para exigir IMDSv2 quando o endpoint HTTP estiver habilitado. Para ter mais informações, consulte Configurar as opções de metadados da instância no Guia do usuário do Amazon EC2 para instâncias Linux.

Versão de metadados

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

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 ter mais informações, consulte Configurar as opções de metadados da instância no Guia do usuário do Amazon EC2 para instâncias Linux.

Dados do usuário

Você pode personalizar e concluir a configuração de suas instâncias no momento da inicialização especificando scripts de shell ou diretivas cloud-init como dados do usuário. Os dados do usuário são executados quando a instância é inicializada inicialmente, permitindo que você instale automaticamente aplicativos, dependências ou personalizações no momento da 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.

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

Solicitar instâncias spot para aplicações flexíveis e com tolerância a falhas

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

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

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

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

- Preço spot. Você paga apenas o preço spot atual pelas instâncias spot que iniciar. Esse preço muda lentamente ao longo do tempo com base em tendências de oferta e demanda no longo prazo. Para mais informações, consulte 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.

Use blocos de capacidade para cargas de trabalho de aprendizado de máquina

Os blocos de capacidade ajudam você a reservar instâncias de GPU muito procuradas em uma data futura para suportar suas cargas de trabalho de aprendizado de máquina (ML) de curta duração.

Para uma visão geral dos blocos de capacidade e de como eles funcionam, consulte Blocos de capacidade para ML no Guia do usuário do Amazon EC2 para instâncias Linux.

Você pode usar blocos de capacidade com os seguintes tipos de instância do EC2 e Regiões da AWS:

Tipos de instância	Regiões
`p5.48xlarge`	Leste dos EUA (Ohio), Leste dos EUA (Norte da Virgínia)
`p4d.24xlarge`	Leste dos EUA (Ohio), Oeste dos EUA (Oregon)

Para começar a usar blocos de capacidade, você cria uma reserva de capacidade em uma zona de disponibilidade específica. Os blocos de capacidade são entregues como reservas de `targeted` capacidade em uma única zona de disponibilidade. Ao criar seu modelo de execução, especifique o ID de reserva e o tipo de instância do Capacity Block. Em seguida, atualize seu grupo de Auto Scaling para usar o modelo de lançamento que você criou e a zona de disponibilidade do Capacity Block. Quando sua reserva de bloco de capacidade começar, use a escalabilidade programada para iniciar o mesmo número de instâncias que sua reserva de bloco de capacidade.

Conteúdo

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

Diretrizes operacionais

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

- Reduza a escala horizontalmente do seu grupo do Auto Scaling até zero mais de 30 minutos antes do horário de término da reserva do bloco de capacidade. O Amazon EC2 encerrará todas as instâncias que ainda estiverem em execução 30 minutos antes do horário final do bloco de capacidade.
- Recomendamos que você use a escalabilidade programada para expandir (adicionar instâncias) e aumentar a escala (remover instâncias) nos horários de reserva apropriados. Para ter mais informações, consulte Escalabilidade programada para o Amazon EC2 Auto Scaling.
- Adicione hooks do ciclo de vida conforme necessário para realizar um desligamento suave do seu aplicativo dentro das instâncias durante a redução da escala horizontalmente. Deixe tempo suficiente para que a ação do ciclo de vida seja concluída antes que o Amazon EC2 comece a encerrar forçosamente suas instâncias 30 minutos antes do horário de término da reserva do bloco de capacidade. Para ter mais informações, consulte Ganchos do ciclo de vida do Amazon EC2 Auto Scaling.
- Certifique-se de que o grupo do Auto Scaling aponte para a versão correta do modelo de execução durante toda a duração da reserva. Recomendamos apontar para uma versão específica do modelo de execução em vez da versão `$Default` ou `$Latest`.

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

Especificar um bloco de capacidade em seu modelo de execução

Para criar um modelo de lançamento que tenha como alvo um bloco de capacidade específico para seu grupo de Auto Scaling, use um dos seguintes métodos:

Console

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

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

AWS CLI

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

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

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

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

Conteúdo de `config.json`.

```
{
    "ImageId": "ami-04d5cc9b88example",
    "InstanceType": "p4d.24xlarge",
    "SecurityGroupIds": [
        "sg-903004f88example"
    ],
    "KeyName": "MyKeyPair",
    "InstanceMarketOptions": {
        "MarketType": "capacity-block"
    },
    "CapacityReservationSpecification": {
        "CapacityReservationTarget": {
            "CapacityReservationId": "cr-02168da1478b509e0"
        }
    }
}
```

A seguir, um exemplo de saída.

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

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

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

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

```
{
    "LaunchTemplateVersions": [
        {
            "LaunchTemplateId": "lt-068f72b724example",
            "LaunchTemplateName": "my-template-for-capacity-block",
            "VersionNumber": 1,
            "CreateTime": "2023-10-27T15:12:44.000Z",
            "CreatedBy": "arn:aws:iam::123456789012:user/Bob",
            "DefaultVersion": true,
            "LaunchTemplateData": {
                "ImageId": "ami-04d5cc9b88example",
                "InstanceType": "p5.48xlarge",
                "SecurityGroupIds": [
                    "sg-903004f88example"
                ],
                "KeyName": "MyKeyPair",
                "InstanceMarketOptions": {
                    "MarketType": "capacity-block"
                },
                "CapacityReservationSpecification": {
                    "CapacityReservationTarget": {
                        "CapacityReservationId": "cr-02168da1478b509e0"
                    }
                }
            }
        }
    ]
}
```

Limitações

- O suporte para blocos de capacidade só está disponível se seu grupo do Auto Scaling tiver uma configuração compatível. Não há suporte para grupos de instâncias mistas e pools aquecidos.
- Você só pode atingir um Bloco de Capacidade por vez.

Recursos relacionados

- Para obter os pré-requisitos e recomendações para o uso de instâncias P5, consulte Comece a usar instâncias P5 no Guia do usuário do Amazon EC2 para instâncias Linux.

- O Amazon EKS oferece suporte ao uso de blocos de capacidade para suportar suas cargas de trabalho de aprendizado de máquina (ML) de curta duração nos clusters do Amazon EKS. Para obter mais informações, consulte Capacity Blocks for ML no Guia do usuário do Amazon EKS.
- Você pode usar blocos de capacidade com tipos de instância e regiões compatíveis. No entanto, as reservas de capacidade sob demanda oferecem flexibilidade para reservar capacidade para outros tipos de instâncias e regiões. Para obter um tutorial que mostra como usar a opção de reserva de capacidade sob demanda, consulteUse reservas de capacidade sob demanda para reservar capacidade em zonas de disponibilidade específicas.

Migre seus grupos de Auto Scaling para modelos de lançamento

A partir de 2023, você não poderá fazer chamadas `CreateLaunchConfiguration` com novos tipos de instância do Amazon EC2 lançados após 31 de dezembro de 2022. Para ter mais informações, consulte Configurações de execução.

Para migrar seus grupos do Auto Scaling das configurações de lançamento para os modelos de lançamento, consulte as etapas a seguir.

> **Important**
>
> Antes de continuar, confirme se você tem as permissões necessárias para trabalhar com modelos de execução. Para ter mais informações, consulte Permissões para trabalhar com modelos de lançamento.

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

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

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

A seguir, um exemplo de saída.

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

```
            {
  "ResourceId": "group-1",
  "ResourceType": "auto-scaling-group",
  "Key": "environment",
  "Value": "production",
  "PropagateAtLaunch": true
            }
        ],
 "TerminationPolicies": [
     "Default"
 ],
 "NewInstancesProtectedFromScaleIn": false,
 "ServiceLinkedRoleARN": "arn",
       "TrafficSources": []
    },

    ... additional groups ...

]
```

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

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

Veja a seguir um exemplo de saída.

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

```
    ... additional groups ...

]
```

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

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

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

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

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

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

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

6. (Opcional) Selecione Criar um grupo do Auto Scaling usando o novo modelo.

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

7. Escolha Copiar.

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

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

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

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

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

1. Abra o console do Amazon EC2 em https://console.aws.amazon.com/ec2/ e escolha 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 do Auto Scaling para usar um modelo de execução (AWS CLI)

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

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

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

Etapa 4: substituir suas instâncias

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

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

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

Mais informações

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

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

Migre AWS CloudFormation pilhas para modelos de lançamento

Você pode migrar seus modelos de 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 do Auto Scaling no modelo de pilha. Em seguida, use seu modelo modificado para atualizar sua pilha.

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

Tópicos

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

Encontre grupos do Auto Scaling que usam uma configuração de execução

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

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

```
aws autoscaling describe-auto-scaling-groups --region REGION \
  --filters Name=tag-key,Values=aws:cloudformation:stack-name \
```

```
  --query 'AutoScalingGroups[?LaunchConfigurationName!
=`null`].AutoScalingGroupName'
```

Veja a seguir um exemplo de saída.

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

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

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

Atualizar uma pilha para usar um modelo de execução

Siga as etapas desta seção para fazer o seguinte:

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

Para modificar o modelo da pilha e atualizar a pilha

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

Exemplo: uma configuração de execução simples

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

Exemplo: O modelo de execução equivalente

```
---
Resources:
  myLaunchTemplate:
    Type: AWS::EC2::LaunchTemplate
    Properties:
      LaunchTemplateName: !Sub ${AWS::StackName}-launch-template
      LaunchTemplateData:
        ImageId: ami-02354e95b3example
        InstanceType: t3.micro
        SecurityGroupIds:
          - Ref! EC2SecurityGroup
        KeyName: MyKeyPair
        BlockDeviceMappings:
          - DeviceName: /dev/xvda
            Ebs:
              VolumeSize: 150
```

```
              DeleteOnTermination: true
        UserData:
          Fn::Base64: !Sub |
            #!/bin/bash -x
            yum install -y aws-cfn-bootstrap
            /opt/aws/bin/cfn-signal -e $? --stack ${AWS::StackName} --resource
myASG --region ${AWS::Region}
```

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

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

3. Se a propriedade **`IamInstanceProfile`** estiver presente em sua configuração de execução, 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 propriedades,**`AssociatePublicIpAddress`** **`InstanceMonitoring`** ou **`PlacementTenancy`** estiverem presentes em sua configuração de execução, você deverá convertê-las em uma estrutura. Para ver exemplos, consulte AWS::EC2::LaunchTemplate.

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

7. Se a propriedade **`SpotPrice`** estiver presente em sua configuração de execução, recomendamos que você a omita do seu modelo de execução. Suas instâncias spot serão executadas pelo preço spot atual. Este preço nunca excederá o preço sob demanda.

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

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

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

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

   Exemplo: sem uma política de instâncias mistas

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

Exemplo: com uma política de instâncias mistas

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

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

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

Compreender atualização de comportamentos de recursos da pilha

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

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

Se você não especificou um `UpdatePolicy` atributo para seu grupo de Auto Scaling, o modelo de lançamento é verificado quanto à exatidão, mas CloudFormation não implanta nenhuma alteração nas instâncias do grupo Auto Scaling. Todas as novas instâncias receberão seu modelo de execução, mas as instâncias existentes continuarão a ser executadas com a configuração de

execução com a qual foram executadas originalmente (apesar da inexistência da configuração de execução). A exceção é quando você altera suas opções de compra, por exemplo, adicionando uma política de instâncias mistas. Neste caso, seu grupo do Auto Scaling substitui gradualmente as instâncias existentes por novas instâncias para corresponder às novas opções de compra.

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

Referência do mapeamento de configuração de execução

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

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

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

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

Exemplos para criar e gerenciar modelos de lançamento com o AWS Command Line Interface (AWS CLI)

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

Conteúdo

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

Exemplo de uso

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

Criar um modelo de execução básico

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

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

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

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

> **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 ter mais informações, consulte Funções do IAM para aplicações que são executadas em instâncias do Amazon EC2.

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

Atribuir um endereço IP público

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

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

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

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

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

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

Especificar um mapeamento de dispositivos de blocos

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

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

Especificar hosts dedicados para trazer licenças de software de fornecedores externos

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

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

Especificar uma interface de rede existente

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

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

Criar várias interfaces de rede

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

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

Se você usar um tipo de instância compatível com várias placas de rede e adaptadores Elastic Fabric (EFAs), poderá adicionar uma interface secundária a uma placa de rede secundária e habilitar o EFA usando o comando a seguir. create-launch-template Para obter mais informações, consulte 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-09d95
```

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

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

Gerenciar modelos de execução

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

Conteúdo

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

Listar e descrever modelos de execução

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

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

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

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

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

A seguir, uma exemplo de resposta.

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

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

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

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

A seguir, uma exemplo de resposta.

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

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

Criar uma versão de modelo de execução

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

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

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

Excluir uma versão de modelo de execução

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

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

Excluir um modelo de execução

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

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

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

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

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

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

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

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

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

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

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

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

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

Para obter mais informações sobre a criação de um parâmetro válido para uma ID de AMI, consulte Criação de parâmetros do Systems Manager.

Crie um modelo de lançamento que especifique um parâmetro para a AMI

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

Console

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

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

4. Em 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:

      Para fazer referência a um parâmetro público:

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

      Para fazer referência a um parâmetro armazenado na mesma conta:

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

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

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

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

AWS CLI

Para criar um modelo de execução que especifique um parâmetro do Systems Manager, você pode usar um dos seguintes exemplos de comandos. Substitua cada *espaço reservado para entrada do usuário* por suas próprias informações.

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

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

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

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

Conteúdo de `config.json`:

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

A seguir, uma exemplo de resposta.

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

Exemplo: criar um modelo de lançamento que especifique um parâmetro armazenado na mesma conta

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

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

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

Conteúdo de `config.json`:

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

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

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

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

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

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

A seguir, um exemplo de saída.

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

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

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

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

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

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

Conteúdo de `config.json`:

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

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

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

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

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

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

A seguir, uma exemplo de resposta.

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

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

Verifique se um modelo de lançamento obtém a ID de AMI correta

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

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

Recursos relacionados

Para obter mais detalhes sobre a especificação de um parâmetro do Systems Manager em seu modelo de lançamento, consulte Use um parâmetro do Systems Manager em vez de um ID de AMI no Guia do usuário do Amazon EC2 para instâncias Linux.

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

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

Limitações

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

- O Amazon EC2 Auto Scaling é compatível apenas com a especificação de IDs de AMI como parâmetros.

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

Configurações de execução

> **Important**
>
> Você não pode chamar `CreateLaunchConfiguration` com novos tipos de instância do Amazon EC2 que sejam lançadas após 31 de dezembro de 2022. Além disso, quaisquer novas contas criadas a partir de 1º de junho de 2023 não terão a opção de criar novas configurações de inicialização por meio do console. No futuro, novas contas não poderão criar novas configurações de lançamento usando o console, a API, a CLI e. CloudFormation Migre para modelos de lançamento para garantir que você não precise criar novas configurações de lançamento agora ou no futuro. Para obter informações sobre como migrar seu grupo do Auto Scaling, para lançar modelos, consulte Migre seus grupos de Auto Scaling para modelos de lançamento.

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.

Conteúdo

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

Criar uma configuração de execução

> **Important**
>
> Você não pode chamar `CreateLaunchConfiguration` com novos tipos de instância do Amazon EC2 que sejam lançadas após 31 de dezembro de 2022. Além disso, quaisquer novas contas criadas a partir de 1º de junho de 2023 não terão a opção de criar novas configurações de inicialização por meio do console. No futuro, novas contas não poderão criar novas configurações de lançamento usando o console, a API, a CLI e. CloudFormation Migre para modelos de lançamento para garantir que você não precise criar novas configurações de lançamento agora ou no futuro. Para obter informações sobre como migrar seu grupo do Auto Scaling, para lançar modelos, consulte Migre seus grupos de Auto Scaling para modelos de lançamento.

Este tópico descreve como criar uma configuração de inicialização.

Depois de criar uma configuração de inicialização, você não pode modificá-la. Em vez disso, você deve criar uma nova configuração de lançamento.

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

Conteúdo

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

Criar uma configuração de execução

Para criar uma configuração de execução (console)

1. Abra o console do Amazon EC2 em https://console.aws.amazon.com/ec2/.
2. Na barra de navegação superior, selecione sua AWS região.
3. No painel de navegação à esquerda, em Auto Scaling, escolha Grupos do Auto Scaling.

4. Escolha Executar configurações próximo ao topo da página. Quando a confirmação for solicitada, escolha Exibir configurações de inicialização para confirmar que deseja visualizar a página Configurações de inicialização.
5. SelecioneCreate launch configuration (Criar uma configuração de execução), e insira um nome para sua configuração de execução.
6. Em Amazon machine image (AMI) (Imagem de máquina da Amazon (AMI)), escolha uma AMI. Para escolher uma AMI específica, você pode encontrar uma AMI adequada, anotar seu ID e inserir o ID como critério de pesquisa.

   Para obter a ID da AMI do Amazon Linux 2:

   a. Abra o console do Amazon EC2 em https://console.aws.amazon.com/ec2/.
   b. No painel de navegação esquerdo, em Instâncias, escolha Instâncias e, em seguida, escolha Iniciar instâncias.
   c. Na guia Quick Start (Início rápido) da página Choose an Amazon Machine Image (Escolha uma Imagem de máquina da Amazon), observe o ID da AMI ao lado de Amazon Linux 2 AMI (HVM).
7. Na etapa Choose Instance Type (Escolher tipo de instância), selecione uma configuração de hardware para suas instâncias.
8. Em Additional configuration (Configuração adicional), preste atenção aos seguintes campos:

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

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

   b. (Opcional) Em IAM instance profile (Perfil de instância do IAM) selecione uma função a ser associada às instâncias. Para ter mais informações, consulte Funções do IAM para aplicações que são executadas em instâncias do Amazon EC2.

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

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

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

9. (Opcional) Em Storage (volumes) (Armazenamento - volumes), se não precisar de armazenamento adicional, ignore esta seção. Caso contrário, para especificar os volumes a serem anexados às instâncias, além dos volumes especificados pela AMI, escolha Add new volume (Adicionar novo volume). Em seguida, escolha as opções desejadas e os valores associados para Devices (Dispositivos), Snapshot, Size (Tamanho), Volume type (Tipo de volume), IOPS, Throughput (Taxa de transferência),Delete on termination (Excluir ao término), e Encrypted (Criptografado).

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

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

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

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

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

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

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

1. Abra o console do Amazon EC2 em https://console.aws.amazon.com/ec2/.
2. Na barra de navegação superior, selecione sua AWS região.
3. No painel de navegação à esquerda, em Auto Scaling, escolha Grupos do Auto Scaling.
4. Escolha Executar configurações próximo ao topo da página. Quando a confirmação for solicitada, escolha Exibir configurações de inicialização para confirmar que deseja visualizar a página Configurações de inicialização.
5. Selecione a configuração de execução e escolha Ações, Copiar configuração de execução. Isso configura uma nova configuração de execução com as mesmas opções da original, mas com "Copy" adicionado ao nome.
6. Na página Copiar configuração de execução, edite as opções de configuração conforme o necessário e escolha Criar configuração de execução.

Para criar uma configuração de execução usando a linha de comando

Você pode usar um dos comandos a seguir:

- create-launch-configuration (AWS CLI)
- Novo como (LaunchConfiguration1)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 o AWS CLI ou um dos SDKs para configurar o IMDS, deverá usar a versão mais recente do AWS CLI ou do SDK para exigir o uso do IMDSv2.

Você pode configurar sua configuração de execução para:

- Exigir o uso 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 ter mais informações, consulte Alterar a configuração de execução de um grupo do Auto Scaling.

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

1. Abra o console do Amazon EC2 em https://console.aws.amazon.com/ec2/.
2. Na barra de navegação superior, selecione sua AWS região.
3. No painel de navegação à esquerda, em Auto Scaling, escolha Grupos do Auto Scaling.
4. Escolha Executar configurações próximo ao topo da página. Quando a confirmação for solicitada, escolha Exibir configurações de inicialização para confirmar que deseja visualizar a página Configurações de inicialização.
5. Escolha Create launch configuration (Criar configuração de execução) e crie a configuração de execução da maneira usual. Inclua o ID da Imagem de máquina da Amazon (AMI), o tipo de instância e, opcionalmente, um par de chaves, um ou mais grupos de segurança e quaisquer volumes do EBS adicionais ou volumes de armazenamento de instâncias para suas instâncias.

6. Para configurar opções de metadados de instância para todas as instâncias associadas a esta configuração de execução, em Additional configuration (Configurações adicionais), em Advanced details (Detalhes avançados), faça o seguinte:

   a. Em Metadata accessible (Metadados acessíveis): escolha se deseja habilitar ou desabilitar o acesso ao endpoint do serviço de metadados da instância. Por padrão, o endpoint de HTTP está habilitado. Se você optar por desabilitar o endpoint, o acesso aos metadados da instância será desativado. 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-configurationcomando a seguir com `--metadata-options` definido como`HttpTokens=required`. Quando você especifica um valor para `HttpTokens`, você também deve definir `HttpEndpoint` como ativado. Como o cabeçalho 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 create-launch-configurationcomando a seguir para desativar o acesso aos metadados da instância. Você pode reativar o acesso posteriormente usando o modify-instance-metadata-optionscomando.

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

Criar uma configuração de execução usando uma instância do EC2

Você também tem a opção de criar uma configuração de execução usando os atributos de uma instância do EC2 em execução.

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

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

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

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

Tópicos

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

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

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

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

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

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

A seguir, uma exemplo de resposta.

```
{
    "LaunchConfigurations": [
        {
            "UserData": null,
            "EbsOptimized": false,
            "LaunchConfigurationARN": "arn",
            "InstanceMonitoring": {
                "Enabled": false
            },
            "ImageId": "ami-05355a6c",
            "CreatedTime": "2014-12-29T16:14:50.382Z",
            "BlockDeviceMappings": [],
            "KeyName": "my-key-pair",
            "SecurityGroups": [
                "sg-8422d1eb"
            ],
```

```
            "LaunchConfigurationName": "my-lc-from-instance",
            "KernelId": "null",
            "RamdiskId": null,
            "InstanceType": "t1.micro",
            "AssociatePublicIpAddress": true
        }
    ]
}
```

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

Por padrão, o Amazon EC2 Auto Scaling usa os atributos da instância do EC2 que você especifica para criar a configuração de execução. No entanto, os dispositivos de blocos são provenientes da AMI usada para iniciar a instância, não a instância. Para adicionar dispositivos de blocos à configuração de execução, substitua o mapeamento de dispositivos de blocos para a configuração de execução.

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

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

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

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

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

```
{
    "LaunchConfigurations": [
        {
            "UserData": null,
            "EbsOptimized": false,
            "LaunchConfigurationARN": "arn",
```

```
            "InstanceMonitoring": {
                "Enabled": false
            },
            "ImageId": "ami-c49c0dac",
            "CreatedTime": "2015-01-07T14:51:26.065Z",
            "BlockDeviceMappings": [
                {
                    "DeviceName": "/dev/sda1",
                    "Ebs": {
                        "SnapshotId": "snap-3decf207"
                    }
                },
                {
                    "DeviceName": "/dev/sdf",
                    "Ebs": {
                        "SnapshotId": "snap-eed6ac86"
                    }
                }
            ],
            "KeyName": "my-key-pair",
            "SecurityGroups": [
                "sg-8637d3e3"
            ],
            "LaunchConfigurationName": "my-lc-from-instance-bdm",
            "KernelId": null,
            "RamdiskId": null,
            "InstanceType": "t1.micro",
            "AssociatePublicIpAddress": true
        }
    ]
}
```

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

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

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

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

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

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

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

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

Alterar a configuração de execução de um grupo do Auto Scaling

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

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

Depois de alterar a configuração de execução, todas as novas instâncias são executadas usando as novas opções de configuração, mas as instâncias existentes não são afetadas. Para ter mais informações, consulte Atualizar instâncias do Auto Scaling.

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

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

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

Para alterar a configuração de inicialização de um grupo do Auto Scaling usando a linha de comando

Você pode usar um dos comandos a seguir:

- update-auto-scaling-group (AWS CLI)
- Atualizar como (AutoScalingGroup)AWS Tools for Windows PowerShell

Grupos do Auto Scaling

> **Note**
>
> Se você é iniciante em grupos de Auto Scaling, siga as etapas do tutorial Criar seu primeiro grupo de Auto Scaling para começar e ver como um grupo de Auto Scaling responde quando uma instância no grupo é encerrada.

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

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

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

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

Ao criar um grupo do Auto Scaling você pode optar por executar 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. Para ter mais informações, consulte Grupos do Auto Scaling com vários tipos de instâncias e opções de compra.

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

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

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

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

Conteúdo

- Criar grupos do Auto Scaling usando modelos de execução
- Criar grupos do Auto Scaling usando configurações de execução
- Atualizar um grupo do Auto Scaling
- Etiquetar grupos e instâncias do Auto Scaling
- Políticas de manutenção de instância
- Ganchos do ciclo de vida do Amazon EC2 Auto Scaling
- Grupos de alta atividade do Amazon EC2 Auto Scaling
- Desanexar ou anexar instâncias
- Remover temporiamente instâncias do grupo do Auto Scaling
- Excluir infraestrutura do Auto Scaling
- Exemplos para criar e gerenciar grupos de Auto Scaling com os SDKs AWS

Criar grupos do Auto Scaling usando modelos de execução

Antes de criar um modelo de execução, você pode criar um grupo do Auto Scaling que use um modelo de execução como modelo de configuração para as instâncias do EC2. O modelo de execução especifica informações, como o ID da AMI, o tipo de instância, o par de chaves, os grupos

de segurança e o mapeamento de dispositivos de blocos para suas instâncias. Para obter mais informações sobre como criar modelos de execução, consulte Criar um modelo de execução para um grupo do Auto Scaling.

Você deve ter permissões suficientes para criar um grupo do 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 por sua própria conta se ela ainda não existir. Para exemplos de políticas do IAM que um administrador pode usar como referência para conceder permissões a você, consulte Exemplos de políticas baseadas em identidade e Suporte a modelo de execução.

Conteúdo

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

Criar um grupo do Auto Scaling usando um modelo de execução

Ao criar um grupo do Auto Scaling, 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 ter mais informações, consulte Criar um modelo de execução para um grupo do Auto Scaling.

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

1. Abra o console do Amazon EC2 em https://console.aws.amazon.com/ec2/ e escolha grupos do Auto Scaling no painel de navegação.
2. Na barra de navegação na parte superior da tela, escolha a mesma Região da AWS que você usou ao criar o modelo de lançamento.
3. Selecione Criar um grupo do Auto Scaling.

4. Na página Choose launch template or configuration (Escolher modelo de execução ou configuração) faça o seguinte:

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

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

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

   d. Verifique se o modelo de execução oferece suporte a todas as opções que você está planejando usar e escolha Next (Próximo).

5. Na página Opções de iniciação de escolha, se você não estiver usando vários tipos de instância, pode pular a seção Requisitos de tipo de instância para usar o tipo de instância do EC2 especificado no modelo de lançamento.

   Para usar vários tipos de instância, consulte Grupos do Auto Scaling com vários tipos de instâncias e opções de compra.

6. Em Network (Rede), para VPC, escolha uma VPC. O grupo do Auto Scaling deve ser criado na mesma VPC do grupo de segurança especificado no modelo de execução.

7. 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 ter mais informações, consulte Considerações sobre a escolha de sub-redes da VPC.

8. 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 ter mais informações, consulte Grupos do Auto Scaling com vários tipos de instâncias e opções de compra.

9. 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).

10. (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 ter mais informações, consulte Usar o Elastic Load Balancing para distribuir tráfego entre as instâncias no grupo do Auto Scaling. 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.

b. (Opcional) Para verificações de integridade e tipos adicionais de verificação de integridade, selecione Ativar verificações de integridade do Elastic Load Balancing.

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

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

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

11. (Opcional) Na página Configure group size and scaling policies (Configurar o tamanho do grupo e as políticas de escalabilidade), configure as seguintes opções e escolha Next (Próximo):

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

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

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

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

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

e. Em Instance scale-in protection (Proteção de redução de instâncias), escolha se deseja habilitar a proteção de redução de instâncias. Para ter mais informações, consulte Usar proteção de redução na escala na horizontal de instâncias.

12. (Opcional) Para receber notificações, em Add notification (Adicionar notificação), configure a notificação e, depois, escolha Next (Próximo). Para ter mais informações, consulte Opções de notificação do Amazon SNS para o Amazon EC2 Auto Scaling.
13. (Opcional) Para adicionar tags, escolha Add tag (Adicionar tag), forneça uma chave e um valor para cada tag e, depois, escolha Next (Próximo). Para ter mais informações, consulte Etiquetar grupos e instâncias do Auto Scaling.
14. 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 como (AutoScalingGroup1)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 área Região da AWS em que planeja criar o grupo 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 a AMI personalizada. Como alternativa, para configurar o modelo de execução do zero ou para obter mais descrição dos parâmetros que você pode configurar para o seu modelo de execução, consulte Criar seu modelo de execução (console).

Para usar uma AMI personalizada como um modelo

1. Abra o console do Amazon EC2 em https://console.aws.amazon.com/ec2/.
2. Na barra de navegação na parte superior da tela, a corrente Região da AWS é exibida. Selecione uma região na qual iniciará o grupo do Auto Scaling.
3. No painel de navegação, escolha Instâncias.

4. Escolha Launch instance (Iniciar instância) e faça o seguinte:

   a. Em Name and tags (Nome e etiquetas), deixe Name (Nome) em branco. O nome não faz parte dos dados usados para criar um modelo de execução.

   b. Em 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 hiperlink consider EC2 Auto Scaling (considerar o EC2 Auto Scaling).

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

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

Para criar um modelo de execução

1. Em Launch template name and description (Nome e descrição do modelo de execução), insira um nome e uma descrição para o modelo de execução.

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

Criar um grupo do Auto Scaling

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

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

Para criar um grupo do Auto Scaling

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

   Em Launch template version (Versão do modelo de execução), indique se o grupo do Auto Scaling usará a versão padrão, a mais recente ou uma versão específica do modelo de execução no aumento da escala na horizontal.
3. Selecione Next (Próximo) para continuar para a próxima etapa.

4. Na página Opções de iniciação de escolha, se você não estiver usando vários tipos de instância, pode pular a seção Requisitos de tipo de instância para usar o tipo de instância do EC2 especificado no modelo de lançamento.

   Para usar vários tipos de instância, consulte Grupos do Auto Scaling com vários tipos de instâncias e opções de compra.

5. Em Network (Rede), para VPC, escolha uma VPC. O grupo do Auto Scaling deve ser criado na mesma VPC do grupo de segurança especificado no modelo de execução.

   > **Tip**
   >
   > Se você não especificou um grupo de segurança no modelo de execução, suas instâncias serão executadas com um grupo de segurança padrão da VPC que você especificar. Por padrão, esse grupo de segurança não permite tráfego de entrada de redes externas.

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

Próximas etapas

Você pode conferir se o grupo do Auto Scaling foi criado corretamente visualizando o histórico de atividades. Na guia Activity (Atividade), em Activity history (Histórico de atividades), a coluna Status exibe se o seu grupo do Auto Scaling lançou instâncias com êxito. Se as instâncias não forem

executadas ou forem executadas, mas terminadas imediatamente, consulte os tópicos a seguir para possíveis causas e resoluções:

- Solucionar problemas do Amazon EC2 Auto Scaling: falhas ao iniciar instâncias do EC2
- Solucionar problemas do Amazon EC2 Auto Scaling: problemas de AMI
- Solucionar problemas com as verificações de integridade do Amazon EC2 Auto Scaling

Agora você pode anexar um balanceador de carga na mesma região do grupo do Auto Scaling, se desejar. Para ter mais informações, consulte Usar o Elastic Load Balancing para distribuir tráfego entre as instâncias no grupo do Auto Scaling.

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

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

As instâncias spot são capacidade ociosa 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. Eles podem ser usados para várias aplicações flexíveis e tolerantes a falhas. Os exemplos incluem servidores web sem estado, endpoints de API, aplicativos de big data e análise, cargas de trabalho em contêineres, pipelines de CI/CD, computação de alto desempenho e alto rendimento (HPC/HTC), cargas de trabalho de renderização e outras cargas de trabalho flexíveis.

Para obter mais informações, consulte Opções de compra de instâncias no Guia do usuário do Amazon EC2 para instâncias Linux.

Tópicos

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

- Usar um modelo de execução diferente para um tipo de instância

Visão geral da configuração

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

Conteúdo

- Visão geral
- Flexibilidade de tipo da instância
- Flexibilidade da zona de disponibilidade
- Preço máximo do spot
- Rebalanceamento proativo de capacidade
- Comportamento do ajuste de escala
- Disponibilidade regional dos tipos de instância
- Recursos relacionados
- Limitações

Visão geral

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

- Seleção de tipo de instância com base em atributos — defina seus requisitos de computação para escolher seus tipos de instância automaticamente com base em seus atributos de instância específicos.
- Seleção manual do tipo de instância — Escolha manualmente os tipos de instância adequados à sua carga de trabalho.

Manual selection

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

1. Escolha um modelo de execução que tenha os parâmetros para executar uma instância do EC2. Os parâmetros nos modelos de execução são opcionais, mas o Amazon EC2 Auto Scaling não pode iniciar uma instância se o ID da imagem de máquina da Amazon (AMI) estiver ausente do modelo de execução.

2. Escolha a opção de substituir o modelo de execução.
3. Escolha manualmente os tipos de instância adequados ao seu workload.
4. Especifique as porcentagens de instâncias sob demanda e de instâncias spot a serem iniciadas.
5. Escolha estratégias de alocação que determinem como o Amazon EC2 Auto Scaling atenderá à capacidade sob demanda e spot com os tipos de instância possíveis.
6. Escolha as zonas de disponibilidade e sub-redes VPC nas quais executar suas instâncias.
7. Especifique o tamanho inicial do grupo (a capacidade desejada) e o tamanho mínimo e máximo do grupo.

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

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

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

Attribute-based selection

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

1. Escolha um modelo de execução que tenha os parâmetros para executar uma instância do EC2. Os parâmetros nos modelos de execução são opcionais, mas o Amazon EC2 Auto

Scaling não pode iniciar uma instância se o ID da imagem de máquina da Amazon (AMI) estiver ausente do modelo de execução.

2. Escolha a opção de substituir o modelo de execução.
3. Especifique atributos de instância que correspondam aos requisitos de computação, como requisitos de vCPUs e memória.
4. Especifique as porcentagens de instâncias sob demanda e de instâncias spot a serem iniciadas.
5. Escolha estratégias de alocação que determinem como o Amazon EC2 Auto Scaling atenderá à capacidade sob demanda e spot com os tipos de instância possíveis.
6. Escolha as zonas de disponibilidade e sub-redes VPC nas quais executar suas instâncias.
7. Especifique o tamanho inicial do grupo (a capacidade desejada) e o tamanho mínimo e máximo do grupo.

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

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

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

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

Flexibilidade de tipo da instância

Para aumentar a disponibilidade, implemente seu aplicativo em vários tipos de instância. É uma prática recomendada usar vários tipos de instância para atender aos requisitos de capacidade. Dessa forma, o Amazon EC2 Auto Scaling pode 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.

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

Flexibilidade da zona de disponibilidade

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

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

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

Preço máximo do spot

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

Quando você especifica o parâmetro `WeightedCapacity` em suas substituições (ou `"DesiredCapacityType": "vcpu"` ou `"DesiredCapacityType": "memory-mib"` no

nível do grupo), o preço máximo representa o preço unitário máximo, não o preço máximo de uma instância inteira.

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

Rebalanceamento proativo de capacidade

Se o seu caso de uso permitir, recomendamos o rebalanceamento de capacidade. O Rebalanceamento de capacidade ajuda a manter a disponibilidade da workload aumentando proativamente sua frota com uma nova instância spot antes que uma instância spot em execução receba o aviso de interrupção de dois minutos.

Quando a capacidade de rebalanceamento está habilitada, o Amazon EC2 Auto Scaling tenta substituir proativamente as instâncias spot 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.

Para ter mais informações, consulte Usar o rebalanceamento de capacidade para lidar com interrupções de spot do Amazon EC2.

Comportamento do ajuste de escala

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

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

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

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

Exemplo: comportamento de escalabilidade

Opções de compra	Tamanho de grupo e número total de instâncias em execução nas opções de compra			
	10	20	30	40
Exemplo 1: base de 10, 50/50% sob demanda/spot				
On-Demand Instances (base amount)	10	10	10	10
On-Demand Instances	0	5	10	15
Spot Instances	0	5	10	15
Exemplo 2: base de 0, 0/100% sob demanda/spot				
On-Demand Instances (base amount)	0	0	0	0
On-Demand Instances	0	0	0	0

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

Opções de compra	Tamanho de grupo e número total de instâncias em execução nas opções de compra			
On-Demand Instances (base amount)	10	12	12	12
On-Demand Instances	0	0	0	0
Spot Instances	0	8	18	28

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

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

Disponibilidade regional dos tipos de instância

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

Recursos relacionados

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

Limitações

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

Estratégias de alocação

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

Para analisar as melhores práticas para um grupo de instâncias mistas, consulteVisão geral da configuração.

Conteúdo

- Instâncias spot
- Instâncias sob demanda
- Como as estratégias de alocação funcionam com pesos

Instâncias spot

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

`price-capacity-optimized` (recomendado)

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

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

`capacity-optimized`

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

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

`capacity-optimized-prioritized`

Você define a ordem dos tipos de instância para as substituições do modelo de execução da prioridade mais alta para a mais baixa (do primeiro ao último na lista). O Amazon EC2 Auto Scaling 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 a preferência por determinados tipos de instância também é importante. Se a estratégia de alocação sob demanda for definida como `prioritized`, a mesma prioridade será aplicada ao atender a capacidade sob demanda.

`lowest-price`

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.

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

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

Instâncias sob demanda

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

`lowest-price`

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

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

`prioritized`

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

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

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

Como as estratégias de alocação funcionam com pesos

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

A única diferença é que, quando você escolhe a `price-capacity-optimized` estratégia `lowest-price` or, suas instâncias vêm dos pools de instâncias com o menor preço por unidade em cada zona de disponibilidade. Para ter mais informações, consulte Configurar um grupo de Auto Scaling para usar pesos de instância.

Por exemplo, imagine 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.

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

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

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

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

- Flexibilidade ideal para instâncias spot — O Amazon EC2 Auto Scaling pode selecionar entre uma ampla variedade de tipos de instância para iniciar instâncias spot. Isso atende à prática recomendada do Spot de ser flexível em relação aos tipos de instância, o que dá ao serviço Spot do Amazon EC2 uma chance melhor de encontrar e alocar a quantidade necessária de capacidade computacional.
- Use facilmente os tipos de instâncias certos: com tantos tipos de instância disponíveis, encontrar os tipos de instância corretos para a workload pode ser demorado. Se você especificar os atributos de instância, os tipos de instância terão automaticamente os atributos necessários para sua workload.
- Uso automático de novos tipos de instância — Seus grupos do Auto Scaling podem usar tipos de instância da nova geração à medida que são lançados. Tipos de instância de geração mais nova são usados automaticamente quando correspondem aos seus requisitos e se alinham com as estratégias de alocação escolhidas para o grupo do Auto Scaling.

Tópicos

- Como funciona a seleção de tipo de instância baseada em atributos
- Proteção de preço
- Pré-requisitos
- Crie um grupo de instâncias mistas com seleção de tipo de instância baseada em atributos (console)

- Crie um grupo de instâncias mistas com seleção de tipo de instância baseada em atributos ()AWS CLI
- Exemplo de configuração
- Pré-visualize os tipos de instância
- Recursos relacionados

Como funciona a seleção de tipo de instância baseada em atributos

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

- Contagem de vCPUs: o número mínimo e máximo de vCPUs por instância.
- Memória — O mínimo e o máximo GiBs de memória por instância.
- Armazenamento local: se o sistema deve usar o EBS ou volumes de armazenamento de instâncias para armazenamento local.
- Desempenho intermitente: se o sistema deve usar a família de instâncias T, incluindo os tipos T4g, T3a, T3 e T2.

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

Quando seu grupo de Auto Scaling precisar iniciar uma instância, ele pesquisará os tipos de instância que correspondam aos atributos especificados e estejam disponíveis nessa zona de disponibilidade. Em seguida, a estratégia de alocação determina quais dos tipos de instância correspondentes devem ser executados. Por padrão, a seleção do tipo de instância baseada em atributos tem um recurso de proteção de preço ativado para impedir que seu grupo de Auto Scaling lance tipos de instância que excedam seus limites orçamentários.

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

Como alternativa, você pode definir o valor da capacidade desejada como o número de vCPUs ou a quantidade de memória. Para fazer isso, use o campo suspenso Tipo de capacidade desejada na `DesiredCapacityType` propriedade AWS Management Console ou na operação da `CreateAutoScalingGroup` `UpdateAutoScalingGroup` API. Em seguida, o Amazon EC2

Auto Scaling inicia o número de instâncias necessárias para atender à capacidade desejada de vCPU ou memória. Por exemplo, se você usar vCPUs como o tipo de capacidade desejado e usar instâncias com 2 vCPUs cada, uma capacidade desejada de 10 vCPUs iniciaria 5 instâncias. Essa é uma alternativa útil aos pesos de instância.

Proteção de preço

Com a proteção de preços, você pode especificar o preço máximo que está disposto a pagar pelas instâncias do EC2 lançadas pelo seu grupo de Auto Scaling. A proteção de preços é um recurso que impede que seu grupo de Auto Scaling use tipos de instância que você consideraria muito caros, mesmo que se encaixem nos atributos que você especificou.

A proteção de preços é ativada por padrão e tem limites de preço separados para instâncias sob demanda e instâncias spot. Quando o Amazon EC2 Auto Scaling precisa lançar novas instâncias, nenhum tipo de instância com preço acima do limite relevante é executado.

Tópicos

- Proteção de preços sob demanda
- Proteção de preço à vista
- Personalize a proteção de preços

Proteção de preços sob demanda

Para instâncias sob demanda, você define o preço máximo sob demanda que está disposto a pagar como uma porcentagem maior do que um preço sob demanda identificado. O preço sob demanda identificado é o preço do tipo de instância C, M ou R da geração atual de menor preço com seus atributos especificados.

Se um valor de proteção de preço sob demanda não for definido explicitamente, será usado um preço padrão máximo sob demanda de 20% maior do que o preço sob demanda identificado.

Proteção de preço à vista

Por padrão, o Amazon EC2 Auto Scaling aplicará automaticamente a melhor proteção de preço de instância spot para selecionar consistentemente entre uma ampla variedade de tipos de instância. Você também pode definir manualmente a proteção de preço. No entanto, deixar que o Amazon EC2 Auto Scaling faça isso por você pode aumentar a probabilidade de que sua capacidade spot seja atendida.

É possível especificar manualmente a proteção de preço usando uma das opções a seguir. Se você definir manualmente a proteção de preço, recomendamos usar a primeira opção.

- Uma porcentagem de um preço sob demanda identificado — O preço sob demanda identificado é o preço do tipo de instância C, M ou R da geração atual de menor preço com seus atributos especificados.
- Uma porcentagem maior do que um preço spot identificado — O preço spot identificado é o preço do tipo de instância C, M ou R da geração atual de menor preço com seus atributos especificados. Não recomendamos o uso dessa opção porque os preços spot podem flutuar e, portanto, seu limite de proteção de preço também pode flutuar.

Personalize a proteção de preços

Você pode personalizar os limites de proteção de preços no console do Amazon EC2 Auto Scaling ou usando os SDKs. AWS CLI

- No console, use as configurações de proteção de preço sob demanda e Proteção de preço spot em Atributos adicionais de instância.
- Na InstanceRequirementsestrutura, para especificar o limite de proteção de preço da instância sob demanda, use a `OnDemandMaxPricePercentageOverLowestPrice` propriedade. Para especificar o limite de proteção de preço da Instância Spot, use o `MaxSpotPriceAsPercentageOfOptimalOnDemandPrice` ou a `SpotMaxPricePercentageOverLowestPrice` propriedade.

Se você definir o tipo de capacidade desejado (`DesiredCapacityType`) como vCPUs ou GiB de memória, a proteção de preço se aplicará com base no preço por vCPU ou por memória, em vez do preço por instância.

Você também pode desativar a proteção de preços. Para indicar que não há limite de proteção de preço, especifique um valor percentual alto, como `999999`.

> **Note**
>
> Se nenhum tipo de instância C, M ou R da geração atual corresponder aos atributos especificados, a proteção de preços ainda será aplicável. Quando nenhuma correspondência é encontrada, o preço identificado é dos tipos de instância da geração atual com menor

preço ou, na falta disso, dos tipos de instância da geração anterior com preços mais baixos, que correspondem aos seus atributos.

Pré-requisitos

- Criar um modelo de execução. Para ter mais informações, consulte Criar um modelo de execução para um grupo do Auto Scaling.
- Verifique se o modelo de execução já não solicita instâncias spot.

Crie um grupo de instâncias mistas com seleção de tipo de instância baseada em atributos (console)

Use o procedimento a seguir para criar um grupo de instâncias mistas usando a seleção de tipo de instância baseada em atributos. Para ajudá-lo a percorrer as etapas com eficiência, algumas seções opcionais são ignoradas.

Para a maioria das cargas de trabalho de uso geral, basta especificar o número de vCPUs e de 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.

Para analisar as melhores práticas para um grupo de instâncias mistas, consulteVisão geral da configuração.

Para criar um grupo de instâncias mistas

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. Na página Escolher opções de inicialização da instância, faça o seguinte:

a. Para Requisitos de tipo de instância, selecione Substituir modelo de execução.

> **Note**
>
> Se você escolher um modelo de execução que já contenha um conjunto de atributos de instância, como vCPUs e memória, os atributos da instância serão exibidos. Esses atributos são adicionados às propriedades do grupo do Auto Scaling, onde você pode atualizá-los do console no Amazon EC2 Auto Scaling, a qualquer momento.

b. Sob Specify instance attributes (Especificar os atributos da instância), comece inserindo seus requisitos de vCPUs e de memória.

- Em vCPUs, insira o número mínimo e máximo desejado de vCPUs. Para não especificar nenhum limite, selecione No minimum (Sem mínimo), No maximum (Sem máximo) ou ambos.
- Em Memory (GiB) (Memória), insira a quantidade mínima e máxima de memória desejada. Para não especificar nenhum limite, selecione No minimum (Sem mínimo), No maximum (Sem máximo) ou ambos.

c. (Opcional) Em Additional instance attributes (Atributos de instância adicionais), você pode, opcionalmente, especificar um ou mais atributos para expressar seus requisitos de computação com mais detalhes. Cada atributo adicional inclui mais restrições à solicitação.

d. Expanda Visualizar tipos de instância correspondentes para ver os tipos de instância que têm seus atributos especificados.

e. Em Opções de compra de instâncias, para Distribuição de instâncias, especifique as porcentagens do grupo para lançamento como instâncias sob demanda e como instâncias spot. Se seu aplicativo for sem estado, tolerante a falhas e puder lidar com a interrupção de uma instância, você poderá especificar uma porcentagem maior de instâncias spot.

f. (Opcional) Quando você especifica uma porcentagem para instâncias spot, selecione Incluir capacidade básica sob demanda e depois especifique a capacidade inicial mínima do grupo do Auto Scaling que deve ser atendido por instâncias sob demanda. Se a capacidade básica for ultrapassada, as configurações Instances distribution (Distribuição de instâncias) serão usadas para determinar quantas instâncias spot e instâncias sob demanda serão executadas.

g. Sob Allocation strategies (Estratégias de alocação), Lowest price (Preço mais baixo) é selecionado automaticamente para a On-Demand allocation strategy (Estratégia de alocação sob demanda), e não pode ser alterado.

h. Para Spot allocation strategy (Estratégia de alocação spot), selecione uma estratégia de alocação. A capacidade de preço otimizada é selecionada por padrão. O preço mais baixo está oculto por padrão e só aparece quando você escolhe Mostrar todas as estratégias. Se você escolher Preço mais baixo, insira o número de pools com preços mais baixos para diversificar para pools com preços mais baixos.

i. Para Rebalanceamento de Capacidade, escolha se deseja ativar ou desativar o Rebalanceamento de Capacidade. Use o Rebalanceamento de capacidade para responder automaticamente quando suas instâncias Spot se aproximarem do encerramento de uma interrupção Spot. Para ter mais informações, consulte Usar o rebalanceamento de capacidade para lidar com interrupções de spot do Amazon EC2.

j. Em Network (Rede), para VPC, escolha uma VPC. O grupo do Auto Scaling deve ser criado na mesma VPC do grupo de segurança especificado no modelo de execução.

k. Para Availability Zones and subnets (Zonas de disponibilidade e sub-redes), selecione uma ou mais sub-redes na VPC especificada. Use sub-redes em várias zonas de disponibilidade para alta disponibilidade. Para ter mais informações, consulte Considerações sobre a escolha de sub-redes da VPC.

l. Escolha Avançar, Avançar.

7. Na etapa Configurar políticas de escalabilidade e tamanho do grupo, faça o seguinte:

a. Para medir a capacidade desejada em unidades que não sejam instâncias, escolha a opção apropriada para Tamanho do grupo, Tipo de capacidade desejada. As opções compatíveis são Unidades, vCPUs e GiBs de memória. Por padrão, o Amazon EC2 Auto Scaling especifica Unidades, o que quer dizer número de instâncias.

b. Para Capacidade desejada, o tamanho inicial do seu grupo do Auto Scaling.

c. Na seção Escalabilidade, em Limites de escalabilidade, se o novo valor para a capacidade desejada for maior que a capacidade mínima desejada e a capacidade máxima desejada,

a capacidade máxima desejada será automaticamente aumentada para o novo valor da capacidade desejada. É possível alterar esses limites conforme necessário. Para ter mais informações, consulte Definir limites de escalabilidade para seu grupo do Auto Scaling.

8. Escolha Skip to review (Ir para revisão).
9. Na página Review (Revisão), escolha Create Auto Scaling group (Criar grupo do Auto Scaling).

Crie um grupo de instâncias mistas com seleção de tipo de instância baseada em atributos ()AWS CLI

Para criar um grupo de instâncias mistas usando a linha de comando

Use um dos seguintes comandos:

- create-auto-scaling-group (AWS CLI)
- Novo como (AutoScalingGroup1)AWS Tools for Windows PowerShell

Exemplo de configuração

Para criar um grupo de Auto Scaling com seleção de tipo de instância baseada em atributos usando o AWS CLI, use o comando a seguir. create-auto-scaling-group

Os seguintes atributos de instância são especificados:

- `VCpuCount` – os tipos de instância devem ter 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-groupcomando a seguir para criar o grupo Auto Scaling. Isso faz referência a um arquivo YAML como o único parâmetro para seu grupo do Auto Scaling.

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

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

```
---
AutoScalingGroupName: my-asg
```

```
DesiredCapacityType: units
MixedInstancesPolicy:
  LaunchTemplate:
    LaunchTemplateSpecification:
      LaunchTemplateName: my-launch-template
      Version: $Default
    Overrides:
    - InstanceRequirements:
        VCpuCount:
          Min: 2
          Max: 4
        MemoryMiB:
          Min: 2048
        CpuManufacturers:
        - intel
  InstancesDistribution:
    OnDemandPercentageAboveBaseCapacity: 50
    SpotAllocationStrategy: price-capacity-optimized
MinSize: 0
MaxSize: 100
DesiredCapacity: 4
DesiredCapacityType: units
VPCZoneIdentifier: subnet-5ea0c127,subnet-6194ea3b,subnet-c934b782
```

Para definir o valor da capacidade desejada como o número de 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.

Pré-visualize os tipos de instância

É possível previsualizar os tipos de instância que correspondem aos requisitos de computação sem iniciá-los e ajustar seus requisitos, se necessário. Ao criar o grupo do Auto Scaling no console do Amazon EC2 Auto Scaling, uma previsualização dos tipos de instância aparece na seção Preview matching instance types (Previsualize os tipos de instância correspondentes) na página Choose instance launch options (Escolha as opções de execução da instância).

Como alternativa, você pode visualizar os tipos de instância fazendo uma chamada de GetInstanceTypesFromInstanceRequirementsAPI do Amazon EC2 usando o AWS CLI ou um SDK. Transmita os parâmetros `InstanceRequirements` na solicitação, no formato exato que você usaria para criar ou atualizar um grupo do Auto Scaling. Para mais informações, consulte Preview instance types with specified attributes (Previsualize tipos de instância com atributos especificados)

no Amazon EC2 User Guide for Linux Instances (Guia do usuário do Amazon EC2 para instâncias do Linux).

Recursos relacionados

Para saber mais sobre a seleção de tipo de instância baseada em atributos, consulte Seleção de tipo de instância baseada em atributos para EC2 Auto Scaling e EC2 Fleet no blog. AWS

Você pode declarar a seleção de tipo de instância baseada em atributos ao criar um grupo do Auto Scaling usando AWS CloudFormation. Para obter mais informações, consulte o trecho de exemplo na seção Trechos de modelo de escalonamento automático do AWS CloudFormation Guia do usuário.

Criar um grupo misto de instâncias escolhendo manualmente os tipos de instância

Este tópico mostra como executar vários tipos de instância em um único grupo do Auto Scaling escolhendo manualmente seus tipos de instância.

Se você preferir usar atributos de instância como critérios para selecionar tipos de instância, consulte Crie um grupo de instâncias mistas usando a seleção de tipo de instância baseada em atributos.

Conteúdos

- Pré-requisitos
- Criar um grupo de instâncias mistas (console)
- Criar um grupo de instâncias mistas (AWS CLI)
- Exemplos de configuração

Pré-requisitos

- Criar um modelo de execução. Para ter mais informações, consulte Criar um modelo de execução para um grupo do Auto Scaling.
- Verifique se o modelo de execução já não solicita instâncias spot.

Criar um grupo de instâncias mistas (console)

Use o procedimento a seguir para criar um grupo de instâncias mistas escolhendo manualmente quais tipos de instância seu grupo pode executar. Para ajudá-lo a percorrer as etapas com eficiência, algumas seções opcionais são ignoradas.

Para analisar as melhores práticas para um grupo de instâncias mistas, consulteVisão geral da configuração.

Para criar um grupo de instâncias mistas

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. Na página Escolha as opções de iniciar uma instância, faça o seguinte:

   a. Para Instance type requirements (Requisitos de tipo de instância), selecione Override launch template (Substituir modelo de execução), e depois escholha Manually add instance types (Adicionar tipos de instância manualmente).

   b. Escolha os tipos de instância. Você pode usar nossas recomendações como ponto de partida. A opção de família e geração flexível é selecionada por padrão.

      - Para alterar a ordem dos tipos de instância, use as setas. Se você escolher uma estratégia de alocação compatível com priorização, a ordem do tipo de instância definirá sua prioridade de execução.
      - Para remover um tipo de instância, escolha X.
      - (Opcional) Para as caixas na coluna Peso, atribua um peso relativo a cada tipo de instância. Para fazer isso, insira o número de unidades que uma instância desse tipo conta para a capacidade desejada do grupo. Isso pode ser útil se os tipos de instância

oferecerem diferentes recursos de vCPU, memória, armazenamento ou largura de banda de rede. Para ter mais informações, consulte Configurar um grupo de Auto Scaling para usar pesos de instância.

Se você optar por usar as recomendações flexíveis de tamanho, todos os tipos de instância que fazem parte desta seção terão automaticamente um valor de peso. Se você não quiser especificar nenhum peso, desmarque as caixas na coluna Peso para todos os tipos de instância.

c. Em Instance purchase options (Opções de compra), para Instances distribution (Distribuição de instâncias), especifique as porcentagens de instâncias do grupo a serem iniciadas como instâncias sob demanda e instâncias spot, respectivamente. Se a aplicação for sem estado, tolerante a falhas e puder lidar com uma interrupção de instância, você poderá especificar uma porcentagem maior de instâncias spot.

d. (Opcional) Quando você especifica uma porcentagem para instâncias spot, selecione Incluir capacidade básica sob demanda e depois especifique a capacidade inicial mínima do grupo do Auto Scaling que deve ser atendido por instâncias sob demanda. Se a capacidade básica for ultrapassada, as configurações Instances distribution (Distribuição de instâncias) serão usadas para determinar quantas instâncias spot e instâncias sob demanda serão executadas.

e. Em Allocation strategies (Estratégias de alocação), para On-Demand allocation strategy (Estratégia de alocação sob demanda), selecione uma estratégia de alocação. Quando você escolhe manualmente seus tipos de instância, a opção Priorizada é selecionada por padrão.

f. Para Spot allocation strategy (Estratégia de alocação spot), selecione uma estratégia de alocação. A capacidade de preço otimizada é selecionada por padrão. O preço mais baixo está oculto por padrão e só aparece quando você escolhe Mostrar todas as estratégias.

- 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 escolheu Capacidade otimizada, você pode, opcionalmente, marcar a caixa Priorizar tipos de instância para permitir que o Amazon EC2 Auto Scaling escolha qual tipo de instância iniciar primeiro com base na ordem em que seus tipos de instância estão listados.

g. Em Rebalanceamento de capacidade, escolha se você deseja habilitar ou desabilitar o rebalanceamento de capacidade. Use o Rebalanceamento de capacidade para responder automaticamente quando suas instâncias Spot se aproximarem do encerramento de

uma interrupção Spot. Para ter mais informações, consulte Usar o rebalanceamento de capacidade para lidar com interrupções de spot do Amazon EC2.

h. Em Network (Rede), para VPC, escolha uma VPC. O grupo do Auto Scaling deve ser criado na mesma VPC do grupo de segurança especificado no modelo de execução.

i. Para Availability Zones and subnets (Zonas de disponibilidade e sub-redes), selecione uma ou mais sub-redes na VPC especificada. Use sub-redes em várias zonas de disponibilidade para alta disponibilidade. Para ter mais informações, consulte Considerações sobre a escolha de sub-redes da VPC.

j. Escolha Avançar, Avançar.

7. Na etapa Configurar políticas de escalabilidade e tamanho do grupo, faça o seguinte:

a. Em tamanho de grupo para Capacidade desejada, insira o número inicial de instâncias a serem executadas.

Por padrão, a capacidade desejada é expressa como o número de instâncias. Se você atribuiu pesos aos seus tipos de instância, deve converter este valor para a mesma unidade de medida usada para atribuir pesos, como o número de vCPUs.

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

8. Escolha Skip to review (Ir para revisão).

9. Na página Review (Revisão), escolha Create Auto Scaling group (Criar grupo do Auto Scaling).

Criar um grupo de instâncias mistas (AWS CLI)

Para criar um grupo de instâncias mistas usando a linha de comando

Use um dos seguintes comandos:

- create-auto-scaling-group (AWS CLI)
- Novo como (AutoScalingGroup1)AWS Tools for Windows PowerShell

Exemplos de configuração

Os exemplos de configuração a seguir mostram como iniciar instâncias spot usando as diferentes estratégias de alocação spot.

> **Note**
>
> Esses exemplos mostram como usar um arquivo de configuração formatado em JSON ou YAML. Se você usar a AWS CLI versão 1, deverá especificar um arquivo de configuração formatado em JSON. Se você usar a AWS CLI versão 2, poderá especificar um arquivo de configuração formatado em YAML ou JSON.

Exemplos

- Exemplo 1: Iniciar instâncias spot usando a estratégia de alocação capacity-optimized
- Exemplo 2: Iniciar instâncias spot usando a estratégia de alocação capacity-optimized-prioritized
- Exemplo 3: Iniciar instâncias spot usando a estratégia de alocação lowest-price diversificada em dois grupos
- Exemplo 4: Iniciar Instâncias spot usando a estratégia de alocação price-capacity-optimized

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

O create-auto-scaling-groupcomando a seguir cria um grupo de Auto Scaling que especifica o seguinte:

- A porcentagem do grupo a ser executado como instâncias sob demanda (`0`) e um número base de instâncias sob demanda com as quais começar (`1`)
- Os tipos de instância a serem iniciadas em ordem de prioridade (`c5.large`, `c5a.large`, `m5.large`, `m5a.large`, `c4.large`, `m4.large`, `c3.large`, `m3.large`)
- As sub-redes nas quais 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
```

O arquivo `config.json` contém o conteúdo a seguir.

```
{
    "AutoScalingGroupName": "my-asg",
    "MixedInstancesPolicy": {
        "LaunchTemplate": {
            "LaunchTemplateSpecification": {
                "LaunchTemplateName": "my-launch-template",
                "Version": "$Default"
            },
            "Overrides": [
                {
                    "InstanceType": "c5.large"
                },
                {
                    "InstanceType": "c5a.large"
                },
                {
                    "InstanceType": "m5.large"
                },
                {
                    "InstanceType": "m5a.large"
                },
                {
                    "InstanceType": "c4.large"
                },
                {
                    "InstanceType": "m4.large"
                },
                {
                    "InstanceType": "c3.large"
                },
                {
                    "InstanceType": "m3.large"
                }
            ]
        },
        "InstancesDistribution": {
            "OnDemandBaseCapacity": 1,
```

```
            "OnDemandPercentageAboveBaseCapacity": 0,
            "SpotAllocationStrategy": "capacity-optimized"
        }
    },
    "MinSize": 1,
    "MaxSize": 5,
    "DesiredCapacity": 3,
    "VPCZoneIdentifier": "subnet-5ea0c127,subnet-6194ea3b,subnet-c934b782"
}
```

YAML

Como alternativa, você pode usar o create-auto-scaling-groupcomando a seguir para criar o grupo Auto Scaling. Isso faz referência a um arquivo YAML como o único parâmetro para o grupo do Auto Scaling.

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

O arquivo `config.yaml` contém o conteúdo a seguir.

```
---
AutoScalingGroupName: my-asg
MixedInstancesPolicy:
  LaunchTemplate:
    LaunchTemplateSpecification:
      LaunchTemplateName: my-launch-template
      Version: $Default
    Overrides:
    - InstanceType: c5.large
    - InstanceType: c5a.large
    - InstanceType: m5.large
    - InstanceType: m5a.large
    - InstanceType: c4.large
    - InstanceType: m4.large
    - InstanceType: c3.large
    - InstanceType: m3.large
  InstancesDistribution:
    OnDemandBaseCapacity: 1
    OnDemandPercentageAboveBaseCapacity: 0
    SpotAllocationStrategy: capacity-optimized
MinSize: 1
MaxSize: 5
```

```
DesiredCapacity: 3
VPCZoneIdentifier: subnet-5ea0c127,subnet-6194ea3b,subnet-c934b782
```

Exemplo 2: Iniciar instâncias spot usando a estratégia de alocação **`capacity-optimized-prioritized`**

O create-auto-scaling-groupcomando a seguir cria um grupo de Auto Scaling que especifica o seguinte:

- A porcentagem do grupo a ser executado como instâncias sob demanda (`0`) e um número base de instâncias sob demanda com as quais começar (1)
- Os tipos de instância a serem iniciadas em ordem de prioridade (`c5.large`, `c5a.large`, `m5.large`, `m5a.large`, `c4.large`, `m4.large`, `c3.large`, `m3.large`)
- As sub-redes nas quais 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
```

O arquivo `config.json` contém o conteúdo a seguir.

```
{
    "AutoScalingGroupName": "my-asg",
    "MixedInstancesPolicy": {
        "LaunchTemplate": {
            "LaunchTemplateSpecification": {
                "LaunchTemplateName": "my-launch-template",
                "Version": "$Latest"
            },
            "Overrides": [
                {
                    "InstanceType": "c5.large"
```

```
                },
                {
                    "InstanceType": "c5a.large"
                },
                {
                    "InstanceType": "m5.large"
                },
                {
                    "InstanceType": "m5a.large"
                },
                {
                    "InstanceType": "c4.large"
                },
                {
                    "InstanceType": "m4.large"
                },
                {
                    "InstanceType": "c3.large"
                },
                {
                    "InstanceType": "m3.large"
                }
            ]
        },
        "InstancesDistribution": {
            "OnDemandBaseCapacity": 1,
            "OnDemandPercentageAboveBaseCapacity": 0,
            "SpotAllocationStrategy": "capacity-optimized-prioritized"
        }
    },
    "MinSize": 1,
    "MaxSize": 5,
    "DesiredCapacity": 3,
    "VPCZoneIdentifier": "subnet-5ea0c127,subnet-6194ea3b,subnet-c934b782"
}
```

YAML

Como alternativa, você pode usar o create-auto-scaling-groupcomando a seguir para criar o grupo Auto Scaling. Isso faz referência a um arquivo YAML como o único parâmetro para o grupo do Auto Scaling.

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

O arquivo `config.yaml` contém o conteúdo a seguir.

```
---
AutoScalingGroupName: my-asg
MixedInstancesPolicy:
  LaunchTemplate:
    LaunchTemplateSpecification:
      LaunchTemplateName: my-launch-template
      Version: $Default
    Overrides:
    - InstanceType: c5.large
    - InstanceType: c5a.large
    - InstanceType: m5.large
    - InstanceType: m5a.large
    - InstanceType: c4.large
    - InstanceType: m4.large
    - InstanceType: c3.large
    - InstanceType: m3.large
  InstancesDistribution:
    OnDemandBaseCapacity: 1
    OnDemandPercentageAboveBaseCapacity: 0
    SpotAllocationStrategy: capacity-optimized-prioritized
MinSize: 1
MaxSize: 5
DesiredCapacity: 3
VPCZoneIdentifier: subnet-5ea0c127,subnet-6194ea3b,subnet-c934b782
```

Exemplo 3: Iniciar instâncias spot usando a estratégia de alocação **`lowest-price`** diversificada em dois grupos

O create-auto-scaling-groupcomando a seguir cria um grupo de Auto Scaling que especifica o seguinte:

- O percentual do grupo a ser iniciado como instâncias sob demanda (`50`) (Isso não especifica um número base de instâncias sob demanda para começar.)
- Os tipos de instância a serem iniciadas em ordem de prioridade (`c5.large`, `c5a.large`, `m5.large`, `m5a.large`, `c4.large`, `m4.large`, `c3.large`, `m3.large`)
- As sub-redes nas quais 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
```

O arquivo `config.json` contém o conteúdo a seguir.

```
{
    "AutoScalingGroupName": "my-asg",
    "MixedInstancesPolicy": {
        "LaunchTemplate": {
            "LaunchTemplateSpecification": {
                "LaunchTemplateName": "my-launch-template",
                "Version": "$Latest"
            },
            "Overrides": [
                {
                    "InstanceType": "c5.large"
                },
                {
                    "InstanceType": "c5a.large"
                },
                {
                    "InstanceType": "m5.large"
                },
                {
                    "InstanceType": "m5a.large"
                },
                {
                    "InstanceType": "c4.large"
                },
                {
                    "InstanceType": "m4.large"
                },
                {
                    "InstanceType": "c3.large"
                },
                {
                    "InstanceType": "m3.large"
```

```
                }
            ]
        },
        "InstancesDistribution": {
            "OnDemandPercentageAboveBaseCapacity": 50,
            "SpotAllocationStrategy": "lowest-price",
            "SpotInstancePools": 2
        }
    },
    "MinSize": 1,
    "MaxSize": 5,
    "DesiredCapacity": 3,
    "VPCZoneIdentifier": "subnet-5ea0c127,subnet-6194ea3b,subnet-c934b782"
}
```

YAML

Como alternativa, você pode usar o create-auto-scaling-groupcomando a seguir para criar o grupo Auto Scaling. Isso faz referência a um arquivo YAML como o único parâmetro para o grupo do Auto Scaling.

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

O arquivo `config.yaml` contém o conteúdo a seguir.

```
---
AutoScalingGroupName: my-asg
MixedInstancesPolicy:
  LaunchTemplate:
    LaunchTemplateSpecification:
      LaunchTemplateName: my-launch-template
      Version: $Default
    Overrides:
    - InstanceType: c5.large
    - InstanceType: c5a.large
    - InstanceType: m5.large
    - InstanceType: m5a.large
    - InstanceType: c4.large
    - InstanceType: m4.large
    - InstanceType: c3.large
    - InstanceType: m3.large
  InstancesDistribution:
    OnDemandPercentageAboveBaseCapacity: 50
```

```
    SpotAllocationStrategy: lowest-price
    SpotInstancePools: 2
MinSize: 1
MaxSize: 5
DesiredCapacity: 3
VPCZoneIdentifier: subnet-5ea0c127,subnet-6194ea3b,subnet-c934b782
```

Exemplo 4: Iniciar Instâncias spot usando a estratégia de alocação **`price-capacity-optimized`**

O create-auto-scaling-groupcomando a seguir cria um grupo de Auto Scaling que especifica o seguinte:

- O percentual do grupo a ser iniciado como instâncias sob demanda (`30`) (Isso não especifica um número base de instâncias sob demanda para começar.)
- Os tipos de instância a serem iniciadas em ordem de prioridade (`c5.large`, `c5a.large`, `m5.large`, `m5a.large`, `c4.large`, `m4.large`, `c3.large`, `m3.large`)
- As sub-redes nas quais 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
```

O arquivo `config.json` contém o conteúdo a seguir.

```
{
    "AutoScalingGroupName": "my-asg",
    "MixedInstancesPolicy": {
        "LaunchTemplate": {
            "LaunchTemplateSpecification": {
                "LaunchTemplateName": "my-launch-template",
                "Version": "$Latest"
            },
```

```
            "Overrides": [
                {
                    "InstanceType": "c5.large"
                },
                {
                    "InstanceType": "c5a.large"
                },
                {
                    "InstanceType": "m5.large"
                },
                {
                    "InstanceType": "m5a.large"
                },
                {
                    "InstanceType": "c4.large"
                },
                {
                    "InstanceType": "m4.large"
                },
                {
                    "InstanceType": "c3.large"
                },
                {
                    "InstanceType": "m3.large"
                }
            ]
        },
        "InstancesDistribution": {
            "OnDemandPercentageAboveBaseCapacity": 30,
            "SpotAllocationStrategy": "price-capacity-optimized"
        }
    },
    "MinSize": 1,
    "MaxSize": 5,
    "DesiredCapacity": 3,
    "VPCZoneIdentifier": "subnet-5ea0c127,subnet-6194ea3b,subnet-c934b782"
}
```

YAML

Como alternativa, você pode usar o create-auto-scaling-groupcomando a seguir para criar o grupo Auto Scaling. Isso faz referência a um arquivo YAML como o único parâmetro para o grupo do Auto Scaling.

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

O arquivo `config.yaml` contém o conteúdo a seguir.

```
---
AutoScalingGroupName: my-asg
MixedInstancesPolicy:
  LaunchTemplate:
    LaunchTemplateSpecification:
      LaunchTemplateName: my-launch-template
      Version: $Default
    Overrides:
    - InstanceType: c5.large
    - InstanceType: c5a.large
    - InstanceType: m5.large
    - InstanceType: m5a.large
    - InstanceType: c4.large
    - InstanceType: m4.large
    - InstanceType: c3.large
    - InstanceType: m3.large
  InstancesDistribution:
    OnDemandPercentageAboveBaseCapacity: 30
    SpotAllocationStrategy: price-capacity-optimized
MinSize: 1
MaxSize: 5
DesiredCapacity: 3
VPCZoneIdentifier: subnet-5ea0c127,subnet-6194ea3b,subnet-c934b782
```

Configurar um grupo de Auto Scaling para usar pesos de instância

Ao usar vários tipos de instância, você pode especificar quantas unidades associar a cada tipo de instância e, em seguida, especificar a capacidade do seu grupo com a mesma unidade de medida. Essa opção de especificação de capacidade é conhecida como pesos.

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
`c5.24xlarge`	96	192

Por padrão, todos os tipos de instância têm o mesmo peso, independentemente do tamanho. Em outras palavras, se o Amazon EC2 Auto Scaling iniciar um tipo de instância grande ou pequeno, cada instância será considerada na capacidade desejada do grupo do Auto Scaling.

Com pesos, no entanto, você atribui um valor numérico que especifica quantas unidades associar a cada tipo de instância. Por exemplo, se as instâncias tiverem tamanhos diferentes, uma instância `c5.2xlarge` poderá ter o peso 2, uma `c5.4xlarge` (que é duas vezes maior) poderá ter o peso 4 e assim por diante. Quando o Amazon EC2 Auto Scaling reduz a escala do grupo, esses pesos se traduzem no número de unidades que cada instância conta para a capacidade desejada.

Os pesos não alteram quais tipos de instância o Amazon EC2 Auto Scaling escolhe executar; em vez disso, as estratégias de alocação fazem isso. Para ter mais informações, consulte Estratégias de alocação.

Important

Para configurar um grupo do Auto Scaling para atender à capacidade desejada usando o número de vCPUs ou a quantidade de memória de cada tipo de instância, recomendamos usar a seleção de tipo de instância baseada em atributos. A configuração do `DesiredCapacityType` parâmetro especifica automaticamente o número de unidades a serem associadas a cada tipo de instância com base no valor definido para esse parâmetro. Para ter mais informações, consulte Crie um grupo de instâncias mistas usando a seleção de tipo de instância baseada em atributos.

Conteúdo

- Considerações
- Comportamentos de peso da instância
- Configurar um grupo do Auto Scaling para usar pesos
- Exemplo de preço spot por unidade hora

Considerações

Esta seção discute as principais considerações para implementar pesos de forma eficaz.

- Escolha alguns tipos de instância que atendam às necessidades de desempenho do seu aplicativo. Decida o peso que cada tipo de instância deve contar para a capacidade desejada do seu grupo de Auto Scaling com base em seus recursos. Esses pesos se aplicam às instâncias atuais e futuras.
- Evite grandes intervalos entre pesos. Por exemplo, não especifique um peso de 1 para um tipo de instância quando o próximo tipo de instância maior tiver um peso de 200. A diferença entre os pesos menores e maiores também não deve ser extrema. Diferenças extremas de peso podem afetar negativamente a otimização de custo-desempenho.
- Especifique a capacidade desejada do grupo em unidades, não em instâncias. Por exemplo, se você usa pesos baseados em vCPU, defina o número desejado de núcleos e também o mínimo e o máximo.
- Defina seus pesos e a capacidade desejada de forma que a capacidade desejada seja pelo menos duas a três vezes maior do que o seu maior peso.

Observe o seguinte ao atualizar grupos existentes:

- Ao adicionar pesos a um grupo existente, inclua pesos para todos os tipos de instância atualmente em uso.
- Quando você adiciona ou altera pesos, o Amazon EC2 Auto Scaling executa ou encerra instâncias para atingir a capacidade desejada com base nos novos valores de peso.
- Se você remover um tipo de instância, as instâncias em execução desse tipo manterão seu último peso, mesmo que não estejam mais definidas.

Comportamentos de peso da instância

Quando você usa pesos de instância, o Amazon EC2 Auto Scaling se comporta da seguinte maneira:

- A capacidade atual será a capacidade desejada ou acima dela. A capacidade atual pode exceder a capacidade desejada se forem lançadas instâncias que excedam as unidades de capacidade desejadas restantes. Por exemplo, vamos supor que você especifique dois tipos de instância, `c5.2xlarge` e `c5.12xlarge`, e atribua pesos de instância de 2 para `c5.2xlarge` e 12 para `c5.12xlarge`. Se houver cinco unidades restantes para atender a capacidade desejada, e o Amazon EC2 Auto Scaling provisionar uma,`c5.12xlarge` a capacidade desejada será excedida em sete unidades.
- Ao lançar instâncias, o Amazon EC2 Auto Scaling prioriza a distribuição da capacidade entre as zonas de disponibilidade e o respeito às estratégias de alocação em vez de exceder a capacidade desejada.
- O Amazon EC2 Auto Scaling pode exceder o limite máximo de capacidade para manter o equilíbrio entre as zonas de disponibilidade, usando suas estratégias de alocação preferidas. O limite rígido imposto pelo Amazon EC2 Auto Scaling é a capacidade desejada mais seu maior peso.

Configurar um grupo do Auto Scaling para usar pesos

Você pode configurar um grupo do Auto Scaling para usar pesos, conforme mostrado nos exemplos AWS CLI a seguir. Para obter instruções sobre como usar o console, consulte Criar um grupo misto de instâncias escolhendo manualmente os tipos de instância.

Para configurar um novo grupo do Auto Scaling para usar pesos (AWS CLI)

Use o comando create-auto-scaling-group. Por exemplo, o comando a seguir cria um novo grupo do Auto Scaling e adiciona pesos especificando o seguinte:

- O percentual do grupo a ser iniciado como instâncias sob demanda (`0`)
- A estratégia de alocação para instâncias spot em cada zona de disponibilidade (`capacity-optimized`)
- Os tipos de instância a serem executados em ordem de prioridade (`m4.16xlarge`, `m5.24xlarge`)
- Os pesos 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
```

O arquivo `config.json` contém o conteúdo a seguir.

```
{
    "AutoScalingGroupName": "my-asg",
    "MixedInstancesPolicy": {
        "LaunchTemplate": {
            "LaunchTemplateSpecification": {
                "LaunchTemplateName": "my-launch-template",
                "Version": "$Latest"
            },
            "Overrides": [
                {
                    "InstanceType": "m4.16xlarge",
                    "WeightedCapacity": "16"
                },
                {
                    "InstanceType": "m5.24xlarge",
                    "WeightedCapacity": "24"
                }
            ]
        },
        "InstancesDistribution": {
            "OnDemandPercentageAboveBaseCapacity": 0,
            "SpotAllocationStrategy": "capacity-optimized"
        }
    },
    "MinSize": 160,
    "MaxSize": 720,
    "DesiredCapacity": 480,
    "VPCZoneIdentifier": "subnet-5ea0c127,subnet-6194ea3b,subnet-c934b782",
    "Tags": []
}
```

Para configurar um grupo do Auto Scaling existente para usar pesos (AWS CLI)

Use o comando update-auto-scaling-group. Por exemplo, o comando a seguir adiciona pesos a tipos de instância em um grupo do Auto Scaling existente especificando o seguinte:

- Os tipos de instância a serem executados em ordem de prioridade (`c5.18xlarge`, `c5.24xlarge`, `c5.2xlarge`, `c5.4xlarge`)
- Os pesos 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
```

O arquivo `config.json` contém o conteúdo a seguir.

```
{
    "AutoScalingGroupName": "my-existing-asg",
    "MixedInstancesPolicy": {
        "LaunchTemplate": {
            "Overrides": [
                {
                    "InstanceType": "c5.18xlarge",
                    "WeightedCapacity": "18"
                },
                {
                    "InstanceType": "c5.24xlarge",
                    "WeightedCapacity": "24"
                },
                {
                    "InstanceType": "c5.2xlarge",
                    "WeightedCapacity": "2"
                },
                {
                    "InstanceType": "c5.4xlarge",
                    "WeightedCapacity": "4"
                }
            ]
        }
    },
    "MinSize": 0,
    "MaxSize": 100,
    "DesiredCapacity": 100
```

```
}
```

Para verificar os pesos usando a linha de comando

Use um dos seguintes comandos:

- describe-auto-scaling-groups (AWS CLI)
- Obter como (AutoScalingGroup)AWS Tools for Windows PowerShell

Exemplo de preço spot por unidade hora

A tabela a seguir compara o preço por hora das instâncias spot em diferentes zonas de disponibilidade no Leste dos EUA (Norte da Virgínia) com o preço das instâncias sob demanda na mesma região. Os preços mostrados são preços de exemplo e não os preços atuais. Estes são seus custos por hora de instância.

Exemplo: preços spot por hora de instância

Tipo de instância	us-east-1a	us-east-1b	us-east-1c	Definição de preço sob demanda
c5.2xlarge	0,180 USD	0,191 USD	0,170 USD	0,34 USD
c5.4xlarge	0,341 USD	0,361 USD	0,318 USD	0,68 USD
c5.12xlarge	0,779 USD	0,777 USD	0,777 USD	2,04 USD
c5.18xlarge	1,207 USD	1,475 USD	1,357 USD	3,06 USD
c5.24xlarge	1,555 USD	1,555 USD	1,555 USD	4,08 USD

Com os pesos de instâncias, você pode avaliar seus custos com base no que você usa por unidade de hora. Você pode determinar o preço por hora dividindo seu preço para um tipo de instância pelo número de unidades que ele representa. Para instâncias sob demanda, o preço por hora ao

implantar um tipo de instância é igual ao que é ao implantar um tamanho diferente do mesmo tipo de instância. Por outro lado, o preço spot por hora varia por grupo spot.

O exemplo a seguir mostra como o cálculo do preço spot por unidade de hora funciona com pesos de instância. Para facilitar o cálculo, digamos que você queira iniciar instâncias spot somente em `us-east-1a`. O preço unitário por hora está capturado na tabela a seguir.

Exemplo: preço spot por unidade hora 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

Usar um modelo de execução diferente para um tipo de instância

Além de usar vários tipos de instância, você também pode usar vários modelos de execução.

Por exemplo, digamos que você configure um grupo do Auto Scaling para aplicações de computação intensiva e queira incluir uma combinação de tipos de instância C5, C5a e C6g. No entanto, as instâncias C6g apresentam um processador AWS Graviton baseado na arquitetura Arm de 64 bits, enquanto as instâncias C5 e C5a são executadas em processadores Intel x86 de 64 bits. A AMI para instâncias C5 e C5a funciona para cada uma dessas instâncias, mas não em instâncias C6g. Para resolver o problema, use um modelo de execução diferente para as instâncias C6g. Você ainda pode usar o mesmo modelo de execução para instâncias C5 e C5a.

Esta seção contém procedimentos para usar o AWS CLI para realizar tarefas relacionadas ao uso de vários modelos de execução. No momento, esse recurso estará disponível somente se você usar a AWS CLI ou um SDK, e não está disponível no console.

Conteúdo

- Configurar um grupo do Auto Scaling para usar vários modelos de execução

- Recursos relacionados

Configurar um grupo do Auto Scaling para usar vários modelos de execução

Você pode configurar um grupo do Auto Scaling para usar vários modelos de execução, conforme mostrado nos exemplos a seguir.

Para configurar um novo grupo do Auto Scaling para usar vários modelos de execução (AWS CLI)

Use o comando create-auto-scaling-group. Por exemplo, o comando a seguir cria um novo grupo do Auto Scaling. Ele especifica os tipos de instância `c5.large`, `c5a.large`, e `c6g.large` e define um novo modelo de execução para o tipo de instância `c6g.large` para garantir que uma AMI apropriada seja usada para iniciar instâncias Arm. O Amazon EC2 Auto Scaling usa a ordem de tipos de instâncias para determinar qual tipo de instância usar primeiro ao atender à capacidade sob demanda.

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

O arquivo `config.json` contém o conteúdo a seguir.

```
{
  "AutoScalingGroupName":"my-asg",
  "MixedInstancesPolicy":{
    "LaunchTemplate":{
      "LaunchTemplateSpecification":{
        "LaunchTemplateName":"my-launch-template-for-x86",
        "Version":"$Latest"
      },
      "Overrides":[
        {
          "InstanceType":"c6g.large",
          "LaunchTemplateSpecification": {
            "LaunchTemplateName": "my-launch-template-for-arm",
            "Version": "$Latest"
          }
        },
        {
          "InstanceType":"c5.large"
        },
        {
          "InstanceType":"c5a.large"
```

```
        }
      ]
    },
    "InstancesDistribution":{
      "OnDemandBaseCapacity": 1,
      "OnDemandPercentageAboveBaseCapacity": 50,
      "SpotAllocationStrategy": "capacity-optimized"
    }
  },
  "MinSize":1,
  "MaxSize":5,
  "DesiredCapacity":3,
  "VPCZoneIdentifier":"subnet-5ea0c127,subnet-6194ea3b,subnet-c934b782",
  "Tags":[ ]
}
```

Para configurar um grupo do Auto Scaling existente para usar vários modelos de execução (AWS CLI)

Use o comando update-auto-scaling-group. Por exemplo, o comando a seguir atribui o modelo de execução chamado *my-launch-template-for-arm* ao tipo de instância *c6g.large* do grupo do Auto Scaling chamado *my-asg*.

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

O arquivo `config.json` contém o conteúdo a seguir.

```
{
  "AutoScalingGroupName":"my-asg",
  "MixedInstancesPolicy":{
    "LaunchTemplate":{
      "Overrides":[
        {
          "InstanceType":"c6g.large",
          "LaunchTemplateSpecification": {
            "LaunchTemplateName": "my-launch-template-for-arm",
            "Version": "$Latest"
          }
        },
        {
          "InstanceType":"c5.large"
        },
```

```
        {
          "InstanceType":"c5a.large"
        }
      ]
    }
  }
}
```

Para verificar os modelos de execução de um grupo do Auto Scaling

Use um dos seguintes comandos:

- describe-auto-scaling-groups (AWS CLI)
- Obter como (AutoScalingGroup)AWS Tools for Windows PowerShell

Recursos relacionados

Você pode encontrar um exemplo de especificação de vários modelos de execução usando a seleção de tipo de instância baseada em atributos em um AWS CloudFormation modelo em re:POST.AWS

Criar grupos do Auto Scaling usando configurações de execução

> **Important**
>
> Você não pode chamar `CreateLaunchConfiguration` com novos tipos de instância do Amazon EC2 que sejam lançadas após 31 de dezembro de 2022. Além disso, quaisquer novas contas criadas a partir de 1º de junho de 2023 não terão a opção de criar novas configurações de inicialização por meio do console. No futuro, novas contas não poderão criar novas configurações de lançamento usando o console, a API, a CLI e. CloudFormation Migre para modelos de lançamento para garantir que você não precise criar novas configurações de lançamento agora ou no futuro. Para obter informações sobre a migração de grupos do Auto Scaling para modelos de execução, consulte Migre seus grupos de Auto Scaling para modelos de lançamento.

Se você criou uma configuração de execução ou uma instância do EC2, poderá criar um grupo do Auto Scaling que use uma configuração de execução como modelo de configuração para suas

instâncias do EC2. A configuração de execução especifica informações como ID da AMI, tipo de instância, par de chaves, grupos de segurança e mapeamento de dispositivos de bloco para suas instâncias. Para obter informações sobre como criar configurações de inicialização, consulte Criar uma configuração de execução.

Você deve ter permissões suficientes para criar um grupo do 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 por sua própria conta se ela ainda não existir. Para ver exemplos de políticas do IAM que um administrador pode usar como referência para lhe conceder permissões, consulte Exemplos de políticas baseadas em identidade.

Conteúdo

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

Criar um grupo do Auto Scaling usando uma configuração de execução

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

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.

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 ter mais informações, consulte Alterar a configuração de execução de um grupo do Auto Scaling.

Pré-requisitos

- É necessário ter criado uma configuração de execução. Para ter mais informações, consulte Criar uma configuração de execução.

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

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

   a. Em Launch Template (Modelo de execução), selecione Switch to launch configuration (Alternar para configuração de execução).

   b. Em Launch configuration (Configuração de execução), escolha uma configuração de execução existente.

   c. Verifique se a configuração de execução oferece suporte a todas as opções que você está planejando usar e escolha Next (Próximo).
6. Na página (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 ter mais informações, consulte Considerações sobre a escolha de sub-redes da VPC.
8. Escolha Próximo.

   Ou é possível aceitar o restante dos padrões e escolher Skip to review (Avançar para análise).
9. (Opcional) Na página Configure advanced options (Configurar opções avançadas), configure as seguintes opções e escolha Next (Próximo):

   a. Para registrar suas instâncias do Amazon EC2 com um balanceador de carga, escolha um load balancer existente ou crie um novo. Para ter mais informações, consulte Usar o Elastic Load Balancing para distribuir tráfego entre as instâncias no grupo do Auto Scaling. 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.

b. (Opcional) Para verificações de integridade e tipos adicionais de verificação de integridade, selecione Ativar verificações de integridade do Elastic Load Balancing.

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

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

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

10. (Opcional) Na página Configure group size and scaling policies (Configurar o tamanho do grupo e as políticas de escalabilidade), configure as seguintes opções e escolha Next (Próximo):

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

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

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

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

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

e. Em Instance scale-in protection (Proteção de redução de instâncias), escolha se deseja habilitar a proteção de redução de instâncias. Para ter mais informações, consulte Usar proteção de redução na escala na horizontal de instâncias.

11. (Opcional) Para receber notificações, em Add notification (Adicionar notificação), configure a notificação e, depois, escolha Next (Próximo). Para ter mais informações, consulte Opções de notificação do Amazon SNS para o Amazon EC2 Auto Scaling.
12. (Opcional) Para adicionar tags, escolha Add tag (Adicionar tag), forneça uma chave e um valor para cada tag e, depois, escolha Next (Próximo). Para ter mais informações, consulte Etiquetar grupos e instâncias do Auto Scaling.
13. Na página Review (Revisão), escolha Create Auto Scaling group (Criar grupo do Auto Scaling).

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

Você pode usar um dos comandos a seguir:

- create-auto-scaling-group (AWS CLI)
- Novo como (AutoScalingGroup1)AWS Tools for Windows PowerShell

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

> **Important**
>
> Fornecemos informações sobre configurações de execução para clientes que ainda não migraram das configurações de execução para os modelos de execução. Para obter informações sobre a migração de grupos do Auto Scaling para modelos de execução, consulte Migre seus grupos de Auto Scaling para modelos de lançamento.

Se esta for a primeira vez que você cria um grupo do Auto Scaling, recomendamos que você use o console para criar um modelo de execução a partir de uma instância do EC2 existente. Em seguida,

use o modelo de execução para criar um novo grupo do Auto Scaling. Para esse procedimento, consulte Criar um grupo do Auto Scaling usando o assistente de execução do Amazon EC2.

O procedimento a seguir mostra como criar um grupo do Auto Scaling especificando uma instância existente a ser usada como base para iniciar outras instâncias. Vários parâmetros são necessárias para criar uma instância do EC2, como o ID do imagem de máquina da Amazon (AMI), o tipo de instância, o par de chaves e o grupo de segurança. Todas essas informações também são usadas pelo Amazon EC2 Auto Scaling para iniciar instâncias em seu nome quando houver necessidade de escalar. Essas informações são armazenadas em um modelo de execução ou uma configuração de execução.

Quando você usa uma instância existente, o Amazon EC2 Auto Scaling cria um grupo do Auto Scaling que inicia instâncias com base em uma configuração de execução criada ao mesmo tempo. A nova configuração de execução tem o mesmo nome do grupo do Auto Scaling e inclui determinados detalhes de configuração da instância identificada.

Os detalhes de configuração a seguir são copiados da instância identificada para a configuração de execução:

- ID 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

A sub-rede VPC e a zona de disponibilidade são copiadas da instância identificada para a própria definição de recursos do grupo do Auto Scaling.

Se a instância identificada estiver em um grupo de posicionamento, o novo grupo do Auto Scaling iniciará instâncias no mesmo grupo de posicionamento da instância identificada. Como as configurações de execução não permitem que um grupo de posicionamento seja especificado, o grupo de posicionamento é copiado para o atributo `PlacementGroup` do novo grupo do Auto Scaling.

Os seguintes detalhes da configuração não são copiados da instância identificada:

- Armazenamento: os dispositivos de bloco (volumes do EBS e volumes de armazenamento de instâncias) não são copiados da instância identificada. Em vez disso, o mapeamento de dispositivos de bloco criado como parte da criação da AMI determina quais dispositivos são usados.
- Número de interfaces de rede: as interfaces de rede não são copiadas da instância identificada. Em vez disso, o Amazon EC2 Auto Scaling usa suas configurações padrão para criar uma interface de rede, que é a interface de rede primária (eth0).
- Opções de metadados da instância: as configurações acessíveis de metadados, versão de metadados e limite de salto de resposta de token não são copiadas da instância identificada. Em vez disso, o Amazon EC2 Auto Scaling usa suas configurações padrão. Para ter mais informações, consulte Configurar as opções de metadados da instância.
- Balanceadores de carga: se a instância identificada estiver registrada em um ou mais balanceadores de carga, as informações sobre o balanceador de carga não serão copiadas para o balanceador de carga nem no atributo do grupo de destino do novo grupo do Auto Scaling.
- Etiquetas: se a instância identificada tiver etiquetas, elas não serão copiadas para o atributo de `Tags` do novo grupo do Auto Scaling.

Pré-requisitos

A instância EC2 deve atender aos seguintes critérios:

- A instância não é um membro de outro grupo do Auto Scaling.
- A instância está no estado `running`.
- A AMI usada para iniciar a instância ainda deve existir.

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).

O procedimento a seguir mostra como usar um comando CLI para criar um grupo do Auto Scaling a partir de uma instância do EC2.

Esse procedimento não adiciona a instância ao grupo do Auto Scaling. Para que a instância seja anexada, 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-groupcomando 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 describe-auto-scaling-groupscomando a seguir para verificar se o grupo Auto Scaling foi criado com êxito.

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

O exemplo de reposta a seguir mostra que a capacidade desejada do grupo é 2, o grupo tem 2 instâncias em execução e a configuração de execução é chamada `my-asg-from-instance`.

```
{
  "AutoScalingGroups":[
    {
      "AutoScalingGroupName":"my-asg-from-instance",
      "AutoScalingGroupARN":"arn",
      "LaunchConfigurationName":"my-asg-from-instance",
      "MinSize":1,
      "MaxSize":2,
      "DesiredCapacity":2,
      "DefaultCooldown":300,
      "AvailabilityZones":[
        "us-west-2a"
      ],
      "LoadBalancerNames":[],
      "TargetGroupARNs":[],
      "HealthCheckType":"EC2",
      "HealthCheckGracePeriod":0,
      "Instances":[
        {
          "InstanceId":"i-06905f55584de02da",
          "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",
      "TrafficSources":[]
    }
  ]
}
```

Para visualizar a configuração de execução

- Use o describe-launch-configurationscomando a seguir para ver os detalhes da configuração de inicialização.

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

  A seguir está um exemplo de saída:

  ```
  {
    "LaunchConfigurations":[
      {
        "LaunchConfigurationName":"my-asg-from-instance",
  ```

```
      "LaunchConfigurationARN":"arn",
      "ImageId":"ami-0528a5175983e7f28",
      "KeyName":"my-key-pair-uswest2",
      "SecurityGroups":[
        "sg-05eaec502fcdadc2e"
      ],
      "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 do Auto Scaling

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

Para atualizar um grupo do Auto Scaling (console)

1. Abra o console do Amazon EC2 em https://console.aws.amazon.com/ec2/ e escolha Grupos do Auto Scaling no painel de navegação.
2. Escolha seu grupo do Auto Scaling para exibir informações sobre o grupo, com guias para Detalhes, Atividade, Escalabilidade automática, Gerenciamento de instâncias, Monitoramento e Atualização de Instâncias.
3. Escolha as guias das áreas de configuração de seu interesse e atualize as configurações conforme necessário. Para cada configuração que você editar, escolha Atualizar para salvar suas alterações na configuração do grupo do Auto Scaling.

   - Guia Detalhes

     Estas são as configurações gerais do seu grupo do Auto Scaling. Você pode editá-las e gerenciá-las da mesma forma que o fez durante a criação do grupo do Auto Scaling.

     A seção Configurações avançadas tem algumas opções que não estão disponíveis ao criar o grupo, como políticas de encerramento, tempo de espera, processos suspensos e vida útil máxima da instância. Você também pode exibir, mas não editar, o grupo de posicionamento e a função vinculada ao serviço do grupo do Auto Scaling.

     Se o grupo estiver associado aos recursos do Elastic Load Balancing, consulte Adicionar e remover zonas de disponibilidade antes de alterar as zonas de disponibilidade. Algumas restrições no balanceador de carga podem impedir que você aplique alterações nas zonas de disponibilidade do seu grupo às zonas de disponibilidade do balanceador de carga.

   - Guia Atividades
     - Notificações de atividades — notificações do Amazon SNS
   - Guia de escalonamento automático
     - Políticas de escalabilidade dinâmica — Políticas de escalabilidade dinâmica
     - Políticas de escalabilidade preditiva — Políticas de escalabilidade preditiva
     - Ações programadas — ações agendadas
   - Guia Gerenciamento de instâncias
     - Ganchos do ciclo de vida — Ganchos do ciclo de vida
     - Piscina aquecida — Piscinas quentes
   - Guia Monitoramento

- Há apenas uma única opção nessa guia, que permite ativar ou desativar a coleta de métricas de CloudWatch grupo.

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

Você pode usar um dos comandos a seguir:

- update-auto-scaling-group (AWS CLI)
- Atualizar como (AutoScalingGroup)AWS Tools for Windows PowerShell

Atualizar instâncias do Auto Scaling

Se você associar um novo modelo de execução ou configuração de execução a um grupo do Auto Scaling, todas as novas instâncias terão a configuração atualizada. As instâncias existentes continuam a ser executadas com a configuração com a qual foram originalmente iniciadas. Para aplicar as alterações às instâncias existentes, você tem as seguintes opções:

- Inicie uma atualização de instância para substituir as instâncias mais antigas. Para ter mais informações, consulte Use uma atualização de instância para atualizar instâncias em um grupo de Auto Scaling.
- Aguarde que as atividades de escalabilidade substituam gradualmente as instâncias mais antigas por instâncias mais novas com base nas suas políticas de encerramento.
- Encerre-as manualmente para que sejam substituídas pelo seu grupo do Auto Scaling.

> **Note**
>
> Você pode alterar os seguintes atributos de instância especificando-os como parte do modelo de execução ou da configuração de execução:
>
> - Imagem de máquina da Amazon (AMI)
> - dispositivos de blocos
> - key pair (par de chaves)
> - instance type (tipo de instância)
> - security groups
> - dados do usuário

- monitoramento
- Perfil de instância do IAM
- Locação de localização
- kernel
- disco ram
- Indica se a instância tem um endereço IP público.

Etiquetar grupos e instâncias do Auto Scaling

Uma tag é um rótulo de atributo personalizado que você atribui ou AWS atribui a um AWS recurso. Cada tag tem duas partes:

- Uma chave de etiqueta (por exemplo, `costcenter`, `environment` ou `project`)
- Um campo opcional conhecido como um valor de etiqueta (por exemplo, `111122223333` ou `production`)

As tags ajudam a:

- Acompanhe seus AWS custos. Você ativa essas tags no AWS Billing and Cost Management painel. AWS usa as tags para categorizar seus custos e entregar um relatório mensal de alocação de custos para você. Para obter mais informações, consulte Uso de tags de alocação de custos no Guia do usuário do AWS Billing .
- Controle o acesso a grupos do Auto Scaling com base em tags. É possível usar condições em suas políticas do IAM para controlar o acesso aos grupos do Auto Scaling com base nas tags desse grupo. Para ter mais informações, consulte Etiquetas para segurança.
- Filtre e pesquise por grupos do Auto Scaling com base nas tags adicionadas. Para ter mais informações, consulte Usar etiquetas para filtrar grupos do Auto Scaling.
- Identifique e organize seus AWS recursos. Muitos Serviços da AWS oferecem suporte à marcação, então você pode atribuir a mesma tag a recursos de serviços diferentes para indicar que os recursos estão relacionados.

Você pode marcar grupos do Auto Scaling novos ou existentes. Você também pode propagar tags de um grupo do Auto Scaling para as instâncias do EC2 que ele executa.

As tags não são propagadas para volumes do Amazon EBS. Para adicionar tags a volumes do Amazon EBS, especifique as tags em um modelo de execução. Para ter mais informações, consulte Criar um modelo de execução para um grupo do Auto Scaling.

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

Conteúdo

- Restrições de nomeação e uso de tags
- Ciclo de vida de marcação de instâncias do EC2
- Marcar seus grupos do Auto Scaling
- Excluir tags
- Etiquetas para segurança
- Controlar o acesso usando etiquetas
- Usar etiquetas para filtrar grupos do Auto Scaling

Restrições de nomeação e uso de tags

As restrições básicas a seguir se aplicam a tags:

- O número máximo de tags por recurso é 50.
- O número máximo de tags que você pode adicionar ou remover usando uma única chamada é 25.
- O comprimento máximo da chave é 128 caracteres Unicode.
- O número máximo de tags que você pode atribuir a um recurso é 50.
- 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 `aws:` prefixo nos nomes ou valores de suas tags, pois ele está reservado para AWS uso. Você não pode editar nem excluir nomes ou valores de tags com esse prefixo, e elas não são contadas em sua quota de tags por recurso.

Ciclo de vida de marcação de instâncias do EC2

Se você tiver optado por propagar tags para suas instâncias do EC2, as tags serão gerenciadas da seguinte forma:

- Quando um grupo do Auto Scaling executa instâncias, ele adiciona tags às instâncias durante a criação do recurso, e não após o recurso ser criado.
- O grupo do Auto Scaling adiciona automaticamente uma etiqueta às instâncias com uma chave do `aws:autoscaling:groupName` e um valor do nome do grupo do Auto Scaling.
- Se você especificar tags de instância em seu modelo de execução e optar por propagar tags de seu grupo para suas instâncias, todas as tags serão mescladas. Se a mesma chave da etiqueta for especificada para uma etiqueta no modelo de execução e uma etiqueta no grupo do Auto Scaling, então, o valor da etiqueta do grupo terá precedência.
- Quando você anexa instâncias existentes, o grupo do Auto Scaling adiciona as tags às instâncias substituindo todas as tags existentes pela mesma chave de tag. Ele também adiciona uma etiqueta com uma chave do `aws:autoscaling:groupName` e um valor do nome do grupo do Auto Scaling.
- Quando você desvincula uma instância de um grupo do Auto Scaling, ele remove apenas a tag `aws:autoscaling:groupName`.

Marcar seus grupos do Auto Scaling

Quando você adiciona uma tag a seu grupo do Auto Scaling, você pode especificar se ela deve ser adicionada às instâncias iniciadas no grupo do Auto Scaling. Se você modificar uma tag, a versão atualizada da tag será adicionada às instâncias executadas no grupo do Auto Scaling depois da alteração. Se você criar ou modificar uma tag em um grupo do Auto Scaling, essas alterações não serão feitas em instâncias que já estão em execução no grupo do Auto Scaling.

Conteúdo

- Adicionar ou modificar tags (console)
- Adicionar ou modificar tags (AWS CLI)

Adicionar ou modificar tags (console)

Para marcar um grupo do Auto Scaling na criação

Ao usar o console do Amazon EC2 para criar um grupo do Auto Scaling, você pode especificar valores e chaves de tags na página Add tags (Configurar tags) do assistente de criação de grupo do Auto Scaling. Para propagar uma tag às instâncias executadas no grupo do Auto Scaling, mantenha a opção Tag New Instances (Marcar novas instâncias) para essa tag selecionada. Caso contrário, desmarque-a.

Para adicionar ou modificar tags de um grupo do Auto Scaling existente

1. Abra o console do Amazon EC2 em https://console.aws.amazon.com/ec2/ e escolha 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 o para adicionar tags AWS CLI ao criar grupos de Auto Scaling e para adicionar ou modificar tags para grupos de Auto Scaling existentes.

Para marcar um grupo do Auto Scaling na criação

Use o create-auto-scaling-groupcomando para criar um novo grupo de Auto Scaling e adicionar uma tag, por exemplo**`environment=production`**, ao grupo Auto Scaling. A tag também é adicionada a todas as instâncias executadas no grupo do Auto Scaling.

```
aws autoscaling create-auto-scaling-group --auto-scaling-group-name my-asg \
  --launch-configuration-name my-launch-config --min-size 1 --max-size 3 \
  --vpc-zone-identifier "subnet-5ea0c127,subnet-6194ea3b,subnet-c934b782" \
  --tags Key=environment,Value=production,PropagateAtLaunch=true
```

Para criar ou modificar tags de um grupo do Auto Scaling existente

Use o create-or-update-tagscomando 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 ver uma lista das tags do grupo especificado.

  ```
  aws autoscaling describe-tags --filters Name=auto-scaling-group,Values=my-asg
  ```

  A seguir, uma exemplo de resposta.

  ```
  {
      "Tags": [
          {
              "ResourceType": "auto-scaling-group",
              "ResourceId": "my-asg",
              "PropagateAtLaunch": true,
              "Value": "production",
              "Key": "environment"
          }
      ]
  }
  ```

- describe-auto-scaling-groups— Você fornece o nome do grupo do Auto Scaling para visualizar os atributos do grupo especificado, incluindo quaisquer tags.

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

  A seguir, uma exemplo de resposta.

```
{
    "AutoScalingGroups": [
        {
            "AutoScalingGroupName": "my-asg",
            "AutoScalingGroupARN": "arn",
            "LaunchTemplate": {
                "LaunchTemplateId": "lt-0b97f1e282EXAMPLE",
                "LaunchTemplateName": "my-launch-template",
                "Version": "$Latest"
            },
            "MinSize": 1,
            "MaxSize": 5,
            "DesiredCapacity": 1,
            ...
            "Tags": [
                {
                    "ResourceType": "auto-scaling-group",
                    "ResourceId": "my-asg",
                    "PropagateAtLaunch": true,
                    "Value": "production",
                    "Key": "environment"
                }
            ],
            ...
        }
    ]
}
```

Excluir tags

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

Conteúdo

- Excluir tags (console)
- Excluir tags (AWS CLI)

Excluir tags (console)

Para excluir uma tag

1. Abra o console do Amazon EC2 em https://console.aws.amazon.com/ec2/ e escolha 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 informações sobre o uso das chaves de condição para controlar o acesso aos grupos do Auto Scaling, consulte Como o Amazon EC2 Auto Scaling funciona com o IAM.

Controlar o acesso usando etiquetas

Use etiquetas para verificar se o solicitante (como um usuário ou perfil do IAM) tem permissões para adicionar, modificar ou excluir etiquetas de grupos do Auto Scaling.

O exemplo de política do IAM a seguir fornece a permissão principal para remover apenas o tag com a chave **`temporary`** dos grupos do Auto Scaling.

```
{
    "Version": "2012-10-17",
    "Statement": [
        {
            "Effect": "Allow",
            "Action": "autoscaling:DeleteTags",
```

```
            "Resource": "*",
            "Condition": {
                "ForAllValues:StringEquals": { "aws:TagKeys": ["temporary"] }
            }
        }
    ]
}
```

Para ver mais exemplos de políticas do IAM que impõem restrições nas tags especificadas para grupos do Auto Scaling, consulte Controlar quais chaves de tag e valores de tag podem ser usados.

> **Note**
>
> Mesmo que você tenha uma política que restrinja os usuários de executar uma operação de marcação (ou desmarcação) em um grupo do Auto Scaling, isso não os impede de alterar manualmente as marcações nas instâncias após elas serem 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 describe-auto-scaling-groupscomando para descrever grupos de Auto Scaling com tags específicas. A filtragem por tags é limitada ao AWS CLI ou a um SDK e não está disponível no console.

Considerações sobre filtragem

- É possível especificar vários filtros e vários valores de filtro em uma única solicitação.
- Não é possível usar curingas com os valores de filtro.
- Os valores do filtro diferenciam maiúsculas de minúsculas.

Exemplo: descreva grupos do Auto Scaling com um par de chave e valor de etiqueta específicos

O comando a seguir mostra como filtrar resultados para mostrar apenas grupos do Auto Scaling com a chave de etiqueta e o par de valores de **`environment=production`**.

```
aws autoscaling describe-auto-scaling-groups \
```

```
  --filters Name=tag-key,Values=environment Name=tag-value,Values=production
```

A seguir, uma exemplo de resposta.

```
{
    "AutoScalingGroups": [
        {
            "AutoScalingGroupName": "my-asg",
            "AutoScalingGroupARN": "arn",
            "LaunchTemplate": {
                "LaunchTemplateId": "lt-0b97f1e282EXAMPLE",
                "LaunchTemplateName": "my-launch-template",
                "Version": "$Latest"
            },
            "MinSize": 1,
            "MaxSize": 5,
            "DesiredCapacity": 1,
            ...
            "Tags": [
                {
                    "ResourceType": "auto-scaling-group",
                    "ResourceId": "my-asg",
                    "PropagateAtLaunch": true,
                    "Value": "production",
                    "Key": "environment"
                }
            ],
            ...
        },

    ... additional groups ...

    ]
}
```

Como alternativa, você pode especificar etiquetas usando um filtro `tag:<key>`. Por exemplo, o comando a seguir mostra como filtrar resultados para mostrar apenas grupos do Auto Scaling com a chave de etiqueta e o par de valores de **`environment=production`**. Este filtro é formatado da seguinte maneira: `Name=tag:<key>,Values=<value>`, com **`<key>`** e **`<value>`** representando uma etiqueta de chave e par de valor.

```
aws autoscaling describe-auto-scaling-groups \
```

```
  --filters Name=tag:environment,Values=production
```

Você também pode filtrar a AWS CLI saída usando a `--query` opção. O exemplo a seguir mostra como limitar a AWS CLI saída do comando anterior somente ao nome do grupo, tamanho mínimo, tamanho máximo e atributos de capacidade desejados.

```
aws autoscaling describe-auto-scaling-groups \
  --filters Name=tag:environment,Values=production \
  --query "AutoScalingGroups[].{AutoScalingGroupName: AutoScalingGroupName, MinSize:
 MinSize, MaxSize: MaxSize, DesiredCapacity: DesiredCapacity}"
```

A seguir, uma exemplo de resposta.

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

    ... additional groups ...

]
```

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

Exemplo: descreva grupos do Auto Scaling com etiquetas que correspondam à chave de etiqueta especificada

O comando a seguir mostra como filtrar resultados para mostrar apenas grupos do Auto Scaling com a etiqueta **`environment`**, independentemente do valor de etiqueta.

```
aws autoscaling describe-auto-scaling-groups \
  --filters Name=tag-key,Values=environment
```

Exemplo: descreva grupos do Auto Scaling com etiquetas que correspondam ao conjunto de chaves de etiquetas especificado

O comando a seguir mostra como filtrar resultados para mostrar apenas grupos do Auto Scaling com etiquetas para **`environment`** e **`project`**, independentemente dos valores das etiquetas.

```
aws autoscaling describe-auto-scaling-groups \
  --filters Name=tag-key,Values=environment Name=tag-key,Values=project
```

Exemplo: descreva grupos do Auto Scaling com etiquetas que correspondam a pelo menos uma das chaves de etiquetas especificadas

O comando a seguir mostra como filtrar resultados para mostrar apenas grupos do Auto Scaling com etiquetas para **`environment`** ou **`project`**, independentemente dos valores das etiquetas.

```
aws autoscaling describe-auto-scaling-groups \
  --filters Name=tag-key,Values=environment,project
```

Exemplo: descreva grupos do Auto Scaling com o valor de etiqueta especificado

O comando a seguir mostra como filtrar resultados para mostrar apenas grupos do Auto Scaling com o valor de etiqueta de **`production`**, independentemente da chave de etiqueta.

```
aws autoscaling describe-auto-scaling-groups \
  --filters Name=tag-value,Values=production
```

Exemplo: descreva grupos do Auto Scaling com o conjunto de valores de etiquetas especificado

O comando a seguir mostra como filtrar resultados para mostrar apenas grupos do Auto Scaling com os valores de **`production`** e **`development`**, independentemente da chave de etiqueta.

```
aws autoscaling describe-auto-scaling-groups \
  --filters Name=tag-value,Values=production Name=tag-value,Values=development
```

Exemplo: descreva grupos do Auto Scaling com etiquetas que correspondam a pelo menos um dos valores das etiquetas especificados

O comando a seguir mostra como filtrar resultados para mostrar apenas grupos do Auto Scaling com o valor de etiqueta de **`production`** ou **`development`**, independentemente da chave de etiqueta.

```
aws autoscaling describe-auto-scaling-groups \
  --filters Name=tag-value,Values=production,development
```

Exemplo: descreva grupos do Auto Scaling com etiquetas que correspondam a várias chaves e valores de etiquetas

Você também pode combinar filtros para criar lógicas AND e OR personalizadas para fazer uma filtragem mais complexa.

O comando a seguir mostra como filtrar resultados para mostrar apenas grupos do Auto Scaling com um conjunto específico de etiquetas. Uma chave de tag é **`environment`** AND o valor da tag é (**`production`** OR **`development`**) AND a outra chave de tag é **`costcenter`** AND o valor da tag é **`cc123`**.

```
aws autoscaling describe-auto-scaling-groups \
  --filters Name=tag:environment,Values=production,development
 Name=tag:costcenter,Values=cc123
```

Políticas de manutenção de instância

Você pode configurar uma política de manutenção de instâncias para seu grupo do Auto Scaling para atender aos requisitos específicos de capacidade durante eventos que fazem com que as instâncias sejam substituídas, como uma atualização da instância ou o processo de verificação de integridade.

Por exemplo, suponha que você tenha um grupo do Auto Scaling com um pequeno número de instâncias. Você quer evitar possíveis interrupções decorrentes do encerramento e substituição de uma instância quando as verificações de integridade indicarem uma instância com defeito. Com uma política de manutenção de instâncias, você pode garantir que o Amazon EC2 Auto Scaling primeiro execute uma nova instância e depois espere que ela esteja totalmente pronta antes de encerrar a instância não íntegra.

Uma política de manutenção de instâncias também ajuda a minimizar possíveis interrupções nos casos em que várias instâncias são substituídas ao mesmo tempo. Você define os parâmetros de porcentagem de integridade mínima e máxima para a política, e seu grupo do Auto Scaling só pode aumentar e diminuir a capacidade dentro desse intervalo mínimo-máximo ao substituir instâncias. Um intervalo maior aumenta o número de instâncias que podem ser substituídas ao mesmo tempo.

Conteúdo

- Visão geral da política de manutenção de instâncias
- Definir uma política de manutenção de instâncias no seu grupo do Auto Scaling

Visão geral da política de manutenção de instâncias

Este tópico dá uma visão geral das opções disponíveis e descreve o que deve ser considerado ao criar uma política de manutenção de instâncias.

Conteúdo

- Visão geral
- Conceitos principais
- Aquecimento da instância
- Período de carência da verificação de integridade
- Dimensionar o grupo do Auto Scaling
- Cenários de exemplo

Visão geral

Quando você cria uma política de manutenção de instâncias para seu grupo do Auto Scaling, a política afeta os eventos do Amazon EC2 Auto Scaling que fazem com que as instâncias sejam substituídas. Isso resulta em comportamentos de substituição mais consistentes dentro do mesmo grupo do Auto Scaling. Também permite otimizar seu grupo quanto à disponibilidade ou ao custo, dependendo de suas necessidades.

No console, as seguintes opções de configuração estão disponíveis:

- Iniciar antes de encerrar – uma nova instância deve ser provisionada primeiro antes que uma instância existente possa ser encerrada. Essa abordagem é uma boa opção para aplicativos que favorecem a disponibilidade em detrimento da redução de custos.
- Encerrar e executar – novas instâncias são provisionadas ao mesmo tempo em que suas instâncias existentes são encerradas. Essa abordagem é uma boa opção para aplicativos que favorecem a redução de custos em relação à disponibilidade. Também é uma boa opção para aplicativos que não devem lançar mais capacidade do que a disponível atualmente, mesmo ao substituir instâncias.
- Política personalizada – essa opção permite que você configure sua política com um intervalo mínimo e máximo personalizado para o nível de capacidade que você deseja disponibilizar ao substituir instâncias. Essa abordagem pode ajudá-lo a alcançar o equilíbrio certo entre custo e disponibilidade.

O padrão para um grupo do Auto Scaling é não ter uma política de manutenção de instâncias, o que faz com que ele responda aos eventos de manutenção de instâncias com os comportamentos padrão. Os comportamentos padrão estão descritos na tabela a seguir.

Comportamentos padrão do evento de manutenção de instâncias

Evento	Descrição	Comportamento padrão
Falhas de verificação de integridade	Acontece automaticamente quando as instâncias falham nas verificações de integrida de. O Amazon EC2 Auto Scaling substitui instâncias que apresentam falhas de verificação de integridade. Para entender as causas das falhas de verificação de integridade, consulte Verificaç ões de integridade para instâncias em um grupo do Auto Scaling.	Encerrar e iniciar.
Atualização de instância	O que acontece quando você inicia uma atualização de instância. Dependendo de sua configuração, uma atualização de instância substitui instância s uma de cada vez, várias por vez ou todas de uma vez. Para ter mais informações, consulte Use uma atualizaç ão de instância para atualizar instâncias em um grupo de Auto Scaling.	Encerrar e iniciar.
Vida útil máxima da instância	Acontece automaticamente quando as instâncias atingem a vida útil máxima que você	Encerrar e iniciar.

Evento	Descrição	Comportamento padrão
	especifica para seu grupo do Auto Scaling. O Amazon EC2 Auto Scaling substitui instâncias que atingem sua vida útil máxima. Para ter mais informações, consulte Substituir instâncias do Auto Scaling com base na vida útil máxima da instância.	

Evento	Descrição	Comportamento padrão
Rebalanceamento	Acontece automaticamente se houver mudanças subjacentes que façam com que o grupo fique desequilibrado. O Amazon EC2 Auto Scaling reequilibra o grupo nas seguintes situações: • Uma zona de disponibilidade que antes tinha capacidade insuficiente se recupera, ou você adiciona ou remove uma zona de disponibilidade do grupo. Quando isso acontece, seu grupo do Auto Scaling tenta se equilibrar uniformemente entre as zonas de disponibilidade. Para ter mais informações, consulte Atividades de rebalanceamento. • Você ativa o rebalanceamento de capacidade em seu grupo do Auto Scaling e ele tenta iniciar novas instâncias spot antes que as existentes sejam interrompidas conforme a mudança de disponibilidade das instâncias spot. Para ter mais informações, consulte Usar o rebalanceamento de capacidade para lidar	Iniciar antes de encerrar. O Amazon EC2 Auto Scaling pode exceder os limites de tamanho do seu grupo em até 10% da capacidade máxima. Porém, se você estiver usando o rebalanceamento de capacidade, ele só poderá exceder esses limites em até 10% da capacidade desejada.

Evento	Descrição	Comportamento padrão
	com interrupções de spot do Amazon EC2. • Você atualiza seu grupo do Auto Scaling e ele substitui gradualmente as instâncias de acordo com as novas opções de compra que você escolheu ao atualizar uma política de instâncias mistas. Para ter mais informações, consulte Atualizar um grupo do Auto Scaling.	

O Amazon EC2 Auto Scaling continuará usando como padrão o encerramento e o lançamento nas seguintes situações. Portanto, quando uma dessas situações ocorre, a capacidade do seu grupo pode ser menor que o limite inferior da sua política de manutenção da instância.

- Quando uma instância é encerrada inesperadamente, por exemplo, devido à ação humana. O Amazon EC2 Auto Scaling substitui imediatamente instâncias que não estão mais em execução. Para ter mais informações, consulte Verificações de integridade do Amazon EC2.
- Quando o Amazon EC2 reinicia, interrompe ou desativa uma instância como parte de um evento programado antes que o Amazon EC2 Auto Scaling possa iniciar a instância substituta. Para obter mais informações sobre eventos programados, consulte Eventos programados para suas instâncias no Guia do usuário do Amazon EC2 para Instâncias Linux.
- Quando o Amazon EC2 Spot Service inicia uma interrupção de Instância Spot e uma Instância Spot é encerrada à força.

Com as instâncias spot, se você habilitou o rebalanceamento de capacidade em seu grupo do Auto Scaling, talvez a instância já tenha uma instância pendente de um pool spot diferente que lançamos antes de iniciarmos a interrupção spot. Para ver mais detalhes sobre como funciona o rebalanceamento de capacidade, consulte Usar o rebalanceamento de capacidade para lidar com interrupções de spot do Amazon EC2.

Porém, como não é garantido que as Instâncias Spot permaneçam disponíveis e possam ser encerradas com um aviso de interrupção da Instância Spot de dois minutos, o limite inferior da sua política de manutenção de instâncias pode ser excedido se as instâncias forem interrompidas antes do lançamento de suas novas instâncias.

Conceitos principais

Antes de começar, familiarize-se com os seguintes conceitos e termos básicos:

Capacidade desejada

A capacidade desejada é a capacidade do grupo do Auto Scaling no momento da criação. É também a capacidade que o grupo tenta manter quando não há condições de escalabilidade associadas ao grupo.

Políticas de manutenção de instâncias

Uma política de manutenção de instâncias controla se uma instância é provisionada primeiro antes do encerramento de uma instância em eventos de manutenção de instâncias. Também determina até que ponto seu grupo do Auto Scaling pode ir abaixo e acima da capacidade desejada para substituir várias instâncias ao mesmo tempo.

Porcentagem máxima de integridade

A porcentagem máxima de integridade é a porcentagem da capacidade desejada que seu grupo do Auto Scaling pode aumentar ao substituir instâncias. Ela representa a porcentagem máxima do grupo que pode estar em serviço e íntegra, ou pendente, para suportar sua workload. No console, você pode definir a porcentagem máxima de integridade ao usar a opção Iniciar antes de encerrar ou a opção Política personalizada. Os valores válidos são 100 a 200%.

Percentual mínimo de integridade

A porcentagem mínima de integridade é a porcentagem da capacidade desejada para se manter em serviço, íntegra e pronta para ser usada a fim de suportar sua workload ao substituir instâncias. Uma instância é considerada íntegra e pronta para uso depois de concluir com êxito sua primeira verificação de integridade e após o término do tempo de aquecimento especificado. No console, você pode definir a porcentagem mínima de integridade ao usar a opção Encerrar e iniciar ou a opção Política personalizada. Os valores válidos são 0 a 100%.

> **Note**
>
> Para substituir instâncias mais rapidamente, você pode especificar uma porcentagem mínima íntegra baixa. Porém, se não houver instâncias íntegras suficientes em execução, a disponibilidade pode ser reduzida. Recomendamos selecionar um valor razoável para manter a disponibilidade em situações em que várias instâncias serão substituídas.

Aquecimento da instância

Se suas instâncias precisarem de tempo para inicializar depois de entrarem no estado`InService`, ative o aquecimento padrão da instância para seu grupo do Auto Scaling. Com o aquecimento padrão da instância, você pode evitar que as instâncias sejam contabilizadas na porcentagem mínima de integridade antes de estarem prontas. Isso garante que o Amazon EC2 Auto Scaling considere quanto tempo é necessário para ter capacidade suficiente para suportar a workload antes de encerrar as instâncias existentes.

Como benefício adicional, você pode melhorar as CloudWatch métricas da Amazon usadas para escalabilidade dinâmica ao ativar o aquecimento padrão da instância. Se seu grupo de Auto Scaling tiver alguma política de escalabilidade, quando o grupo for expandido, ele usará o mesmo período de aquecimento padrão para evitar que as instâncias sejam contabilizadas nas CloudWatch métricas antes de concluírem a inicialização.

Para ter mais informações, consulte Definir o aquecimento padrão da instância para um grupo do Auto Scaling.

Período de carência da verificação de integridade

O Amazon EC2 Auto Scaling determina se a instância está íntegra com base no status das verificações de integridade que o grupo do Auto Scaling usa. Para ter mais informações, consulte Verificações de integridade para instâncias em um grupo do Auto Scaling.

Para garantir que essas verificações de integridade comecem o mais rápido possível, não defina um período de carência da verificação de integridade do grupo muito alto, mas alto o suficiente para que suas verificações de integridade do Elastic Load Balancing consigam determinar se um destino está disponível para lidar com solicitações. Para ter mais informações, consulte Definir um período de carência da verificação de integridade para um grupo do Auto Scaling.

Dimensionar o grupo do Auto Scaling

Uma política de manutenção de instâncias só se aplica a eventos de manutenção de instâncias e não impede que o grupo seja escalado manual ou automaticamente.

Quando há políticas de escalabilidade ou ações programadas anexadas ao seu grupo do Auto Scaling, elas podem ser executadas paralelamente enquanto os eventos de manutenção da instância estão ocorrendo. Nesse caso, eles poderiam aumentar ou diminuir a capacidade desejada do grupo, mas somente dentro dos limites de escalabilidade que você definiu. Para obter mais informações sobre esses limites, consulte Definir limites de escalabilidade para seu grupo do Auto Scaling.

Cenários de exemplo

Em um cenário típico, a política de manutenção da instância e a capacidade desejada podem ser mais ou menos assim:

- Porcentagem mínima de integridade = 90%
- Porcentagem máxima de integridade = 120%
- Capacidade desejada = 100

Durante qualquer evento de manutenção de instância, seu grupo do Auto Scaling pode ter no mínimo 90 instâncias e no máximo 120. Depois do evento, o grupo volta a ter 100 instâncias.

Quando você usa uma política de manutenção de instância com um grupo do Auto Scaling que tem um grupo de aquecimento, as porcentagens de integridade mínima e máxima são aplicadas separadamente ao grupo do Auto Scaling e ao grupo de aquecimento.

Por exemplo, suponha que esta seja sua configuração:

- Porcentagem mínima de integridade = 90%
- Porcentagem máxima de integridade = 120%
- Capacidade desejada = 100
- Tamanho do grupo de aquecimento = 10

Se você iniciar uma atualização de instância para reciclar as instâncias do grupo, o Amazon EC2 Auto Scaling substituirá primeiro as instâncias no grupo do Auto Scaling e depois as instâncias

no grupo de aquecimento. Embora o Amazon EC2 Auto Scaling ainda esteja trabalhando na substituição de instâncias no grupo do Auto Scaling, o grupo pode ter no mínimo 90 instâncias e no máximo 120. Depois de terminar com o grupo, o Amazon EC2 Auto Scaling pode trabalhar na substituição de instâncias no grupo de aquecimento. Enquanto isso acontece, o grupo de aquecimento pode ter no mínimo 9 instâncias e no máximo 12.

Definir uma política de manutenção de instâncias no seu grupo do Auto Scaling

É possível criar uma política de manutenção de instâncias ao criar um grupo do Auto Scaling. Também é possível criá-la para grupos existentes.

Ao definir uma política de manutenção de instância no seu grupo do Auto Scaling, não é mais necessário especificar valores de parâmetros de porcentagem mínima e máxima de integridade para o recurso de atualização da instância a não ser que queira substituir a política de manutenção de instâncias.

No console, o Amazon EC2 Auto Scaling fornece opções para ajudar você a começar.

Conteúdo

- Definir uma política de manutenção de instâncias
- Remover uma política de manutenção de instância

Definir uma política de manutenção de instâncias

Para definir uma política de manutenção de instâncias em um grupo do Auto Scaling, use um dos seguintes métodos:

Console

Para definir uma política de manutenção de instâncias em um novo grupo (console)

1. Siga as instruções em Criar um grupo do Auto Scaling usando um modelo de execução e conclua cada etapa do procedimento, até a etapa 11.
2. Em Configurar tamanho do grupo e políticas de escalabilidade, em Capacidade desejada, insira o número inicial de instâncias a serem executadas.
3. Na seção Escalabilidade, em Limites de escalabilidade, se o novo valor para a capacidade desejada for maior que a capacidade mínima desejada e a capacidade máxima desejada,

a capacidade máxima desejada será automaticamente aumentada para o novo valor da capacidade desejada. Você pode alterar esses limites conforme necessário.

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

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

5. Na seção Política de manutenção de instâncias, escolha uma das opções disponíveis:

   - Iniciar antes de encerrar: uma nova instância deve ser provisionada primeiro antes que uma instância existente possa ser encerrada. Essa é uma boa opção para aplicativos que favorecem a disponibilidade em detrimento da redução de custos.
   - Encerrar e executar: novas instâncias são provisionadas ao mesmo tempo em que as instâncias existentes são encerradas. Esta é uma boa opção para aplicações que favorecem a economia de custos em detrimento da disponibilidade. Também é uma boa opção para aplicativos que não devem lançar mais capacidade do que a disponível atualmente.
   - Política personalizada: essa opção permite que você configure sua política com um intervalo mínimo e máximo personalizado para o nível de capacidade que você deseja disponibilizar ao substituir instâncias. Isso pode ajudá-lo a alcançar o equilíbrio certo entre custo e disponibilidade.

6. Em Definir porcentagem de integridade, insira valores para um ou ambos os campos a seguir. Os campos habilitados variam de acordo com a opção escolhida na etapa anterior.

   - Mínimo: define a porcentagem mínima de integridade necessária para continuar com a substituição de instâncias.
   - Máximo: define a porcentagem máxima de integridade possível ao substituir instâncias.

7. Expanda a seção Exibir capacidade durante as substituições com base na seção de capacidade desejada para confirmar como os valores de Mínimo e Máximo são aplicados ao seu grupo. Os valores exatos usados dependem do valor de capacidade desejado, que mudará se o grupo for ampliado.

8. Continue com as etapas em Criar um grupo do Auto Scaling usando um modelo de execução.

AWS CLI

Para definir uma política de manutenção de instância em um novo grupo (AWS CLI)

Adicione a `--instance-maintenance-policy` opção ao create-auto-scaling-groupcomando. O exemplo a seguir define uma política de manutenção de instâncias em um novo grupo do Auto Scaling chamado *my-asg*.

```
aws autoscaling create-auto-scaling-group \
  --launch-template LaunchTemplateName=my-launch-template,Version='1' \
  --auto-scaling-group-name my-asg \
  --min-size 1 \
  --max-size 10 \
  --desired-capacity 5 \
  --default-instance-warmup 20 \
  --instance-maintenance-policy '{
      "MinHealthyPercentage": 90,
      "MaxHealthyPercentage": 120
    }' \
  --vpc-zone-identifier "subnet-5e6example,subnet-613example,subnet-c93example"
```

Console

Para definir uma política de manutenção de instância em um grupo existente (console)

1. Abra o console do Amazon EC2 em https://console.aws.amazon.com/ec2/ e escolha 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 Política de manutenção de instâncias, Editar.
5. Para definir uma política de manutenção de instância no grupo, escolha uma das opções disponíveis:

   - Iniciar antes de encerrar: uma nova instância deve ser provisionada primeiro antes que uma instância existente possa ser encerrada. Essa é uma boa opção para aplicativos que favorecem a disponibilidade em detrimento da redução de custos.

- Encerrar e executar: novas instâncias são provisionadas ao mesmo tempo em que as instâncias existentes são encerradas. Esta é uma boa opção para aplicações que favorecem a economia de custos em detrimento da disponibilidade. Também é uma boa opção para aplicativos que não devem lançar mais capacidade do que a disponível atualmente.
- Política personalizada: essa opção permite que você configure sua política com um intervalo mínimo e máximo personalizado para o nível de capacidade que você deseja disponibilizar ao substituir instâncias. Isso pode ajudá-lo a alcançar o equilíbrio certo entre custo e disponibilidade.

6. Em Definir porcentagem de integridade, insira valores para um ou ambos os campos a seguir. Os campos habilitados variam de acordo com a opção escolhida na etapa anterior.
   - Mínimo: define a porcentagem mínima de integridade necessária para continuar com a substituição de instâncias.
   - Máximo: define a porcentagem máxima de integridade possível ao substituir instâncias.
7. Expanda a seção Exibir capacidade durante as substituições com base na seção de capacidade desejada para confirmar como os valores de Mínimo e Máximo são aplicados ao seu grupo. Os valores exatos usados dependem do valor de capacidade desejado, que mudará se o grupo for ampliado.
8. Escolha Atualizar.

AWS CLI

Para definir uma política de manutenção de instância em um grupo existente (AWS CLI)

Adicione a `--instance-maintenance-policy` opção ao update-auto-scaling-groupcomando. O exemplo a seguir define uma política de manutenção de instância no grupo do Auto Scaling especificado.

```
aws autoscaling update-auto-scaling-group --auto-scaling-group-name my-asg \
  --instance-maintenance-policy '{
      "MinHealthyPercentage": 90,
      "MaxHealthyPercentage": 120
    }'
```

Remover uma política de manutenção de instância

Se você quiser parar de usar uma política de manutenção de instâncias com seu grupo do Auto Scaling, você pode removê-la.

Console

Para remover uma política de manutenção de instância (console)

1. Abra o console do Amazon EC2 em https://console.aws.amazon.com/ec2/ e escolha 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 Política de manutenção de instâncias, Editar.
5. Escolha Nenhuma política de manutenção de instâncias.
6. Escolha Atualizar.

AWS CLI

Para remover uma política de manutenção de instâncias (AWS CLI)

Adicione a `--instance-maintenance-policy` opção ao update-auto-scaling-groupcomando. O exemplo a seguir remove a política de manutenção de instâncias do grupo do Auto Scaling especificado.

```
aws autoscaling update-auto-scaling-group --auto-scaling-group-name my-asg \
  --instance-maintenance-policy '{
      "MinHealthyPercentage": -1,
      "MaxHealthyPercentage": -1
    }'
```

Ganchos do ciclo de vida do Amazon EC2 Auto Scaling

O Amazon EC2 Auto Scaling oferece a capacidade de adicionar ganchos do ciclo de vida aos seus grupos do Auto Scaling. Esses 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 escalabilidade, um gancho de ciclo de vida pausa a instância antes que ela seja encerrada e envia uma notificação usando a Amazon. EventBridge Enquanto a instância estiver em estado de espera, você pode invocar uma AWS Lambda função ou conectar-se à instância para baixar registros ou outros dados antes que a instância seja totalmente encerrada.

Um uso popular de ganchos do ciclo de vida é controlar quando as instâncias são registradas com o Elastic Load Balancing. Ao adicionar um gancho do ciclo de vida de execução ao seu grupo do Auto Scaling, você pode garantir que seus scripts de bootstrap foram completados com êxito e que as aplicações nas instâncias estejam prontas para aceitar tráfego antes de serem registradas no balanceador de carga no final do gancho do ciclo de vida.

Conteúdo

- Disponibilidade de ganchos do ciclo de vida
- Considerações e limitações dos ganchos do ciclo de vida
- Recursos relacionados
- Como os ganchos do ciclo de vida funcionam
- Preparar para adicionar um gancho do ciclo de vida a um grupo do Auto Scaling
- Recuperar o estado de destino do ciclo de vida por meio de metadados de instância
- Adicionar ganchos do ciclo de vida
- Concluir uma ação do ciclo de vida

- Tutorial: configurar dados do usuário para recuperar o estado de destino do ciclo de vida por meio de metadados de instância
- Tutorial: Configurar um gancho do ciclo de vida que invoca uma função do Lambda

Disponibilidade de ganchos do ciclo de vida

A tabela a seguir lista os ganchos do ciclo de vida disponíveis para vários cenários.

Evento	Início ou término da instância[1]	Maximum Instance Lifetime (Tempo de vida máximo da instância): instâncias de substituição	Instance Refresh (Atualização ds instância): instâncias de substituição	Capacity Rebalancing (Rebalanc eamento de capacidade): instâncias de substituição	Warm Pools (Grupos de alta atividade): instância s entrando e saindo do grupo de alta atividade
Início de instâncias	✓	✓	✓	✓	✓
Término de instâncias	✓	✓	✓	✓	✓

[1] Aplica-se a todos as execuções e encerramentos, sejam eles iniciados automática ou manualmente, por exemplo, quando você chama as operações `SetDesiredCapacity` ou `TerminateInstanceInAutoScalingGroup`. Não se aplica quando você anexa ou desvincula instâncias, move instâncias dentro e fora do modo de espera ou exclui o grupo com a opção force delete (forçar exclusão).

Considerações e limitações dos ganchos do ciclo de vida

Ao trabalhar com hooks do ciclo de vida, tenha em mente as seguintes notas e limitações:

- O Amazon EC2 Auto Scaling fornece seu próprio ciclo de vida para ajudar no gerenciamento de grupos do Auto Scaling. Esse ciclo de vida é diferente do de outras instâncias do EC2. Para ter mais informações, consulte Ciclo de vida das instâncias do Amazon EC2 Auto Scaling. As

instâncias em um grupo de alta atividade também têm seu próprio ciclo de vida, conforme descrito em Transições de estado do ciclo de vida para instâncias em um grupo de alta atividade.

- 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 ter mais informações, consulte Usar o rebalanceamento de capacidade para lidar com interrupções de spot do Amazon EC2.
- 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 hook do ciclo de vida pode ser abandonar ou continuar. Se uma instância estiver sendo executada, continue indica que suas ações foram bem-sucedidas e que o Amazon EC2 Auto Scaling pode colocar a instância em serviço. Caso contrário, abandonar indica que suas ações personalizadas não tiveram êxito e que podemos encerrar e substituir a instância. Se uma instância estiver sendo encerrada, abandone e continue permita que a instância seja encerrada. No entanto, abandon (abandonar) interrompe quaisquer ações restantes, como outros ganchos do ciclo de vida, e continue (continuar) permite que quaisquer outros ganchos 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.
- Criar e atualizar ganchos de ciclo de vida usando o AWS CLI, AWS CloudFormation, ou um SDK fornece opções não disponíveis ao criar um gancho de ciclo de vida a partir do. 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 Amazon. EventBridge Esses eventos podem ser filtrados e redirecionados para AWS serviços como Lambda, Amazon SNS e Amazon SQS, conforme necessário.

- Você pode adicionar vários ganchos de ciclo de vida a um grupo do Auto Scaling enquanto o cria, chamando a CreateAutoScalingGroupAPI usando o AWS CLI,, ou um SDK. AWS CloudFormation 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 por vez em chamadas separadas para a API. PutLifecycleHook
- Se você adicionar um hook do ciclo de vida para a execução da instância, o período de carência da verificação de integridade será iniciado assim que a instância atingir o estado`InService`. Para obter mais informações, consulte Definir um período de carência da verificação de integridade para um grupo do Auto Scaling.

Considerações sobre dimensionamento

- As políticas de escalabilidade dinâmica aumentam e diminuem em resposta a dados CloudWatch métricos, como CPU e E/S de rede, que são agregados em várias instâncias. Quando há uma expansão, o Amazon EC2 Auto Scaling não conta imediatamente uma nova instância para as métricas agregadas de instância do grupo do Auto Scaling. Ele espera até que a instância atinja o estado `InService` e o aquecimento da instância seja concluído. Para obter mais informações, consulte Considerações sobre o desempenho de escalabilidade o tópico de aquecimento da instância padrão.
- Na redução da escala na horizontal, talvez as métricas agregadas da instância não reflitam instantaneamente a remoção de uma instância de encerramento. A instância de encerramento para de contabilizar as métricas agregadas de instância do grupo pouco após o início do fluxo de trabalho de encerramento do Amazon EC2 Auto Scaling.
- Na maioria dos casos, quando os hooks do ciclo de vida são invocados, as atividades de escalabilidade devido a políticas de escalabilidade simples são pausadas até que as ações do ciclo de vida sejam concluídas e o período de esfriamento expire. A definição de um intervalo longo para o período de desaquecimento significa que a retomada da escalabilidade levará mais tempo. Para obter mais informações, consulte hooks do ciclo de vida podem causar mais atrasos o tópico de resfriamento. Em geral, não recomendamos o uso de políticas de escalabilidade simples se você puder usar políticas de escalabilidade por etapas ou rastreamento de metas.

Recursos relacionados

Para ver um vídeo de introdução, consulte AWS re:Invent 2018: Gerenciamento de capacidade facilitado com o Amazon EC2 Auto Scaling ativado. YouTube

Fornecemos alguns trechos de modelos JSON e YAML que você pode usar para entender como declarar ganchos de ciclo de vida em seus modelos de pilha. AWS CloudFormation Para obter mais informações, consulte a AWS::AutoScaling::LifecycleHookreferência no Guia AWS CloudFormation do usuário.

Você também pode visitar nosso GitHubrepositório para baixar exemplos de modelos e scripts de dados do usuário para ganchos de ciclo de vida.

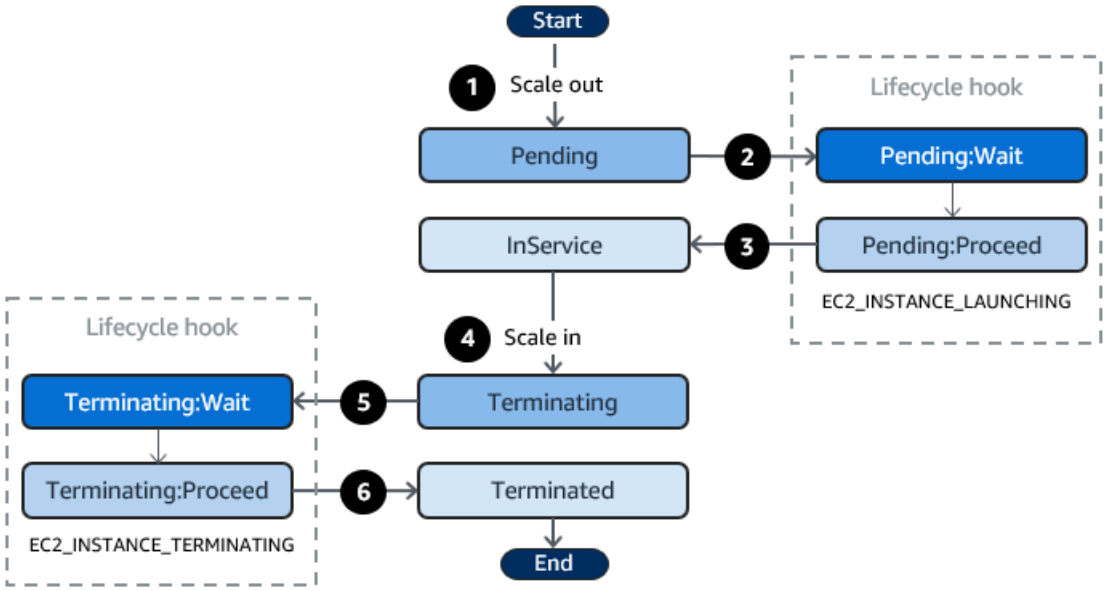
Para obter exemplos do uso de hooks do ciclo de vida, consulte as seguintes postagens no blog.

- Construir um sistema de backup para instâncias escaláveis usando o comando de execução Lambda e Amazon EC2
- Execute o código antes de encerrar uma instância do EC2 Auto Scaling.

Como os ganchos do ciclo de vida funcionam

Uma instância do Amazon EC2 passa por diferentes estados do momento em que é iniciada até seu término. Você pode criar ações personalizadas para que seu grupo do Auto Scaling atue quando uma instância transitar para um estado de espera devido a um hook do ciclo de vida.

A ilustração a seguir mostra as transições entre os estados da instância do Auto Scaling quando você usa ganchos de ciclo de vida para expandir e aumentar a escala.



Conforme mostrado no diagrama anterior:

1. O grupo do Auto Scaling responde a um evento de aumento de escala na horizontal e começa a iniciar uma instância.
2. O gancho do ciclo de vida coloca a instância em um estado de espera (`Pending:Wait`) e, em seguida, executa uma ação personalizada.

   A instância permanece em um estado de espera até que você conclua a ação do ciclo de vida ou até o período de tempo limite terminar. Por padrão, a instância permanece em estado de espera por uma hora e, em seguida, o grupo do Auto Scaling continua o processo de início (`Pending:Proceed`). Se precisar de mais tempo, você poderá reiniciar o período de tempo limite registrando uma pulsação. Se você concluir a ação do ciclo de vida quando a ação personalizada estiver concluída e o período de tempo limite ainda não tiver expirado, o período terminará e o grupo do Auto Scaling continuará o processo de execução.
3. A instância entra no estado `InService` e o período de carência da verificação de integridade é iniciado. Contudo, antes da instância atingir o estado `InService`, se o grupo do Auto Scaling estiver associado a um balanceador de carga Elastic Load Balancing, a instância será registrada no balanceador de carga e o balanceador de carga começará a verificar sua integridade. Após o término do período de carência da verificação de integridade, o Amazon EC2 Auto Scaling começa a verificar o estado de integridade da instância.
4. O grupo do Auto Scaling responde a um evento de redução de escala na horizontal e começa a terminar uma instância. Se o grupo do Auto Scaling estiver sendo usado com o Elastic Load Balancing, primeiro é cancelado o registro da instância em término no balanceador de carga. Se a descarga da conexão estiver habilitada para o balanceador de carga, a instância deixará de aceitar novas conexões e aguardará até que as conexões existentes sejam descarregadas antes de concluir o processo de cancelamento do registro.
5. O gancho do ciclo de vida coloca a instância em um estado de espera (`Terminating:Wait`) e, em seguida, executa uma ação personalizada.

   A instância permanece em um estado de espera até que você conclua a ação do ciclo de vida, ou até o período de tempo limite terminar (uma hora, por padrão). Depois de concluir o gancho do ciclo de vida ou do período de tempo limite expirar, a instância passa para o próximo estado (`Terminating:Proceed`).
6. A instância está terminada.

> **Important**
>
> As instâncias em um 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.

Preparar para adicionar um gancho do ciclo de vida a um grupo do Auto Scaling

Antes de adicionar um gancho do ciclo de vida ao grupo do Auto Scaling, certifique-se de que o script de dados do usuário ou o destino de notificação esteja configurado corretamente.

- Não é necessário configurar um destino de notificação para usar um script de dados do usuário a fim de executar ações personalizadas em suas instâncias enquanto elas estão sendo iniciadas. No entanto, você já deverá ter criado o modelo de execução ou a configuração de execução que especifica o script de dados do usuário e associado ao seu grupo do Auto Scaling. Para obter mais informações sobre scripts de dados do usuário, consulte Executar comandos na instância do Linux na inicialização no Guia do usuário do Amazon EC2 para instâncias do Linux.
- Para sinalizar ao Amazon EC2 Auto Scaling quando a ação do ciclo de vida for concluída, você deve adicionar CompleteLifecycleActiona chamada de API ao script e criar manualmente uma função do IAM com uma política que permita que as instâncias do Auto Scaling chamem essa API. Seu modelo de execução ou configuração de execução deve especificar essa função usando um perfil de instância do IAM que é anexado às suas instâncias do Amazon EC2 na inicialização. Para obter mais informações, consulte Concluir uma ação do ciclo de vida e Funções do IAM para aplicações que são executadas em instâncias do Amazon EC2.
- Para usar um serviço como o Lambda para realizar uma ação personalizada, você já deve ter criado uma EventBridge regra e especificado uma função do Lambda como destino. Para ter mais informações, consulte Configurar um destino de notificação para notificações de ciclo de vida.
- Para permitir que o Lambda sinalize o Amazon EC2 Auto Scaling quando a ação do ciclo de vida for concluída, você deve CompleteLifecycleActionadicionar a chamada de API ao código da função. Você também deve ter anexado uma política do IAM à função de execução da função que concede permissão ao Lambda para concluir ações de ciclo de vida. Para ter mais informações, consulte Tutorial: Configurar um gancho do ciclo de vida que invoca uma função do Lambda.
- Para usar um serviço como um Amazon SNS ou Amazon SQS para executar uma ação personalizada, você já deverá ter criado o tópico do SNS ou a fila do SQS e ter pronto o nome do

recurso da Amazon (ARN). Você também já deverá ter criado 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 ter mais informações, consulte Configurar um destino de notificação para notificações de ciclo de vida.

> **Note**
>
> Por padrão, quando você adiciona um gancho de ciclo de vida no console, o Amazon EC2 Auto Scaling envia notificações de eventos de ciclo de vida para a Amazon. EventBridge Usar EventBridge ou usar um script de dados do usuário é uma prática recomendada. Para criar um gancho de ciclo de vida que envie notificações diretamente para o Amazon SNS ou o Amazon SQS, use o, AWS CloudFormation, ou um SDK para adicionar AWS CLI o gancho de 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 EventBridge a Amazon para 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 prática recomendada, recomendamos que você use EventBridge. As notificações enviadas para o Amazon SNS e o Amazon SQS contêm as mesmas informações que as notificações para as quais o Amazon EC2 Auto Scaling envia. EventBridge Antes EventBridge, a prática padrão era enviar uma notificação ao SNS ou ao SQS e integrar outro serviço ao SNS ou SQS para realizar ações programáticas. Hoje, EventBridge oferece mais opções para quais serviços você pode segmentar e facilita o tratamento de eventos usando a arquitetura sem servidor.

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.

Conteúdo

- Encaminhe notificações para o Lambda usando EventBridge
- Receba notificações usando o Amazon SNS
- Receba notificações usando o Amazon SQS
- Exemplo de mensagem de notificação para o Amazon SNS e o Amazon SQS

Important

A EventBridge regra, a função Lambda, o tópico do Amazon SNS e a fila do Amazon SQS que você usa com ganchos de ciclo de vida devem estar sempre na mesma região em que você criou seu grupo de Auto Scaling.

Encaminhe notificações para o Lambda usando EventBridge

Você pode configurar uma EventBridge regra para invocar uma função Lambda quando uma instância entra em um estado de espera. O Amazon EC2 Auto Scaling emite uma notificação de evento do ciclo de vida sobre EventBridge a instância que está sendo iniciada ou encerrada e um token que você pode usar para controlar a ação do ciclo de vida. Para obter exemplos desses eventos, consulte Referência de eventos do Amazon EC2 Auto Scaling.

Note

Quando você usa o AWS Management Console para criar uma regra de evento, o console adiciona automaticamente as permissões do IAM necessárias para conceder EventBridge permissão para chamar sua função Lambda. Caso esteja criando uma regra de evento usando a AWS CLI, você precisa conceder essa permissão explicitamente.
Para obter informações sobre como criar regras de eventos no EventBridge console, consulte Criação de EventBridge regras da Amazon que reagem a eventos no Guia EventBridge do usuário da Amazon.
- ou -
Para um tutorial introdutório direcionado a usuários do console, consulte Tutorial: Configurar um gancho do ciclo de vida que invoca uma função do Lambda. Este tutorial mostra como

criar uma função Lambda simples que escuta os eventos de lançamento e os grava em um CloudWatch registro de registros.

Para criar uma EventBridge regra que invoque uma função Lambda

1. Crie uma função do Lambda usando o console do Lambda e observe seu nome de recurso da Amazon (ARN). Por exemplo, `arn:aws:lambda:region:123456789012:function:my-function`. Você precisa do ARN para criar um EventBridge destino. Para obter mais informações, consulte Conceitos básicos do Lambda no Guia do desenvolvedor do AWS Lambda .
2. Para criar uma regra que faça a correspondência com eventos para inicialização de instâncias, use o seguinte comando put-rule.

   ```
   aws events put-rule --name my-rule --event-pattern file://pattern.json --state
    ENABLED
   ```

   O exemplo a seguir mostra o `pattern.json` de uma ação de ciclo de vida de execício ução de instância. Substitua o texto em *itálico* 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 responderá com o ARN da regra. Anote esse ARN. Você vai precisar dele na etapa 4.

   Para criar uma regra que faça a correspondência com outros eventos, modifique o padrão de evento. Para ter mais informações, consulte Use EventBridge para lidar com eventos do Auto Scaling.
3. Para especificar a função do Lambda a ser usada como destino para a regra, use o seguinte comando put-targets.

```
aws events put-targets --rule my-rule --targets
 Id=1,Arn=arn:aws:lambda:region:123456789012:function:my-function
```

No comando anterior, *my-rule* é o nome que você especificou para a regra na etapa 2, 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 no EventBridge service principal (`events.amazonaws.com`) e define o escopo das permissões para a regra especificada.

```
aws lambda add-permission --function-name my-function --statement-id my-unique-id \
  --action 'lambda:InvokeFunction' --principal events.amazonaws.com --source-arn
 arn:aws:events:region:123456789012:rule/my-rule
```

No comando anterior:

- *my-function* é o nome da função 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 da função Lambda.
- `source-arn`é o ARN da regra. EventBridge

Se o comando for executado com êxito, você receberá um resultado semelhante a este.

```
{
  "Statement": "{\"Sid\":\"my-unique-id\",
    \"Effect\":\"Allow\",
    \"Principal\":{\"Service\":\"events.amazonaws.com\"},
    \"Action\":\"lambda:InvokeFunction\",
    \"Resource\":\"arn:aws:lambda:us-west-2:123456789012:function:my-function\",
    \"Condition\":
      {\"ArnLike\":
        {\"AWS:SourceArn\":
         \"arn:aws:events:us-west-2:123456789012:rule/my-rule\"}}}"
}
```

O valor de `Statement` é uma versão da string JSON da instrução adicionada à política da função do Lambda.

5. Depois que você tiver seguido estas instruções, prossiga para Adicionar ganchos do ciclo de vida 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 IAM em https://console.aws.amazon.com/iam/.
   b. No painel de navegação à esquerda, escolha Roles (Funções).
   c. Selecione 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) 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 IAM em https://console.aws.amazon.com/iam/.

   b. No painel de navegação à esquerda, escolha Roles (Funções).

   c. Selecione 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) 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 Conta da AWS ID.
- `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 entregues pelo Amazon EC2 Auto Scaling EventBridge para, consulte. Referência de eventos do Amazon EC2 Auto Scaling

Recuperar o estado de destino do ciclo de vida por meio de metadados de instância

Cada instância do Auto Scaling que você inicia passa por vários estados do ciclo de vida. Para invocar ações personalizadas de dentro de uma instância que atuem em transições específicas de estado do ciclo de vida, você deve recuperar o estado do ciclo de vida de destino por meio de metadados da instância.

Por exemplo, você pode precisar de um mecanismo para detectar o encerramento da instância de dentro da instância para executar algum código na instância antes que ela seja encerrada. Você pode fazer isso escrevendo um código que pesquise o estado do ciclo de vida de uma instância diretamente da instância. Em seguida, será possível adicionar um hook do ciclo de vida ao grupo do Auto Scaling para manter a instância em execução até que seu código envie o comando complete-lifecycle-action para continuar.

O ciclo de vida de instância do Auto Scaling tem dois estados estáveis primários (`InService` e `Terminated`) e dois estados estáveis paralelos (`Detached` e `Standby`). Se você usar o 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 de destino na 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 Instance Metadata Service é um componente na instância que o código local usa para acessar os metadados da instância. O código local pode incluir scripts de dados de usuário ou aplicações em execução na instância.

O código local pode acessar 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"` \
&& 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, 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.

> **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.

Depois de criar um grupo de Auto Scaling a partir do AWS Management Console, você pode adicionar um ou mais ganchos de ciclo de vida a ele, até um total de 50 ganchos de ciclo de vida. Você também pode usar o AWS CLI, AWS CloudFormation, ou um SDK para adicionar ganchos de ciclo de vida a um grupo de Auto Scaling ao criá-lo.

Por padrão, quando você adiciona um gancho de ciclo de vida no console, o Amazon EC2 Auto Scaling envia notificações de eventos de ciclo de vida para a Amazon. EventBridge Usar EventBridge ou usar um script de dados do usuário é uma prática recomendada. Para criar um gancho de ciclo de vida que envie notificações diretamente para o Amazon SNS ou o Amazon SQS, você pode usar put-lifecycle-hooko comando, conforme mostrado nos exemplos deste tópico.

Conteúdo

- Adicionar ganchos do ciclo de vida (console)
- Adicionar ganchos do ciclo de vida (AWS CLI)

Adicionar ganchos do ciclo de vida (console)

Siga estas etapas para adicionar hooks do ciclo de vida ao seu grupo do Auto Scaling. Para adicionar hooks do ciclo de vida para expansão (inicialização de instâncias) e expansão (terminação de instâncias ou retorno para um pool quente), você deve criar dois hooks separados.

Antes de começar, confirme se você configurou uma ação personalizada, conforme necessário, conforme descrito em Preparar para adicionar um gancho do ciclo de vida a um grupo do Auto Scaling.

Para adicionar um hook do ciclo de vida 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. 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 hook do ciclo de vida para expansão (execução de instâncias), faça o seguinte:
    a. Em Lifecycle Hook Name (Nome do gancho do ciclo de vida), especifique um nome para o gancho do ciclo de vida.

b. Em Lifecycle transition (Transição do ciclo de vida), escolha Instance launch (Início da instância).

c. Para Tempo limite de pulsação, especifique o tempo, em segundos, para que as instâncias permaneçam em estado de espera durante a expansão antes que o hook expire. O intervalo é de `30` a `7200` segundos. Definir um longo período de tempo limite fornece mais tempo para a conclusão da sua ação personalizada. Em seguida, se você terminar antes que o período de tempo limite termine, envie o complete-lifecycle-actioncomando para permitir que a instância prossiga para o próximo estado.

d. Em Resultado padrão, especifique a ação a ser adotada mediante o término do tempo limite do hook do ciclo de vida ou quando houver uma falha inesperada. Você pode escolher CONTINUAR ou ABANDONAR.

   - Se você escolher CONTINUAR, o grupo do Auto Scaling poderá prosseguir com qualquer outro hook do ciclo de vida e depois colocar a instância em serviço.
   - Se você escolher ABANDON, o grupo do Auto Scaling interromperá todas as ações restantes e encerrará a instância imediatamente.

e. (Opcional) Em Metadados de notificação, especifique outras informações que você deseja incluir quando o Amazon EC2 Auto Scaling enviar uma mensagem ao destino da notificação.

5. Escolha Criar.

Para adicionar um hook do ciclo de vida para redução

1. Escolha Criar hook do ciclo de vida para continuar de onde você parou depois de criar um hook do ciclo de vida para expansão.
2. Para definir um hook do ciclo de vida para redução (instâncias que terminam ou retornam a um grupo de aquecimento), faça o seguinte:

   a. Em Lifecycle Hook Name (Nome do gancho do ciclo de vida), especifique um nome para o gancho do ciclo de vida.

   b. Em Lifecycle transition (Transição do ciclo de vida), escolha Instance terminate (Término da instância).

   c. Para Tempo limite de pulsação, especifique o tempo, em segundos, para que as instâncias permaneçam em estado de espera durante a expansão antes que o hook expire. Recomendamos um curto período de tempo limite de `30` até `120` segundos, dependendo de

quanto tempo você precisa para realizar qualquer tarefa final, como extrair registros do EC2. CloudWatch

d. Em Default result (Resultado padrão), especifique a ação que o grupo do Auto Scaling executa quando o tempo limite se esgota ou quando há uma falha inesperada. ABANDON (ABANDONAR) e CONTINUE (CONTINUAR) permitem que a instância termine.

   - Se você escolher CONTINUE (CONTINUAR), o grupo do Auto Scaling poderá prosseguir com todas as ações restantes, como outros ganchos do ciclo de vida, antes do término.
   - Se você escolher ABANDON, o grupo do Auto Scaling encerrará a instância imediatamente.

e. (Opcional) Em Metadados de notificação, especifique outras informações que você deseja incluir quando o Amazon EC2 Auto Scaling enviar uma mensagem ao destino da notificação.

3. Escolha Criar.

Adicionar ganchos do ciclo de vida (AWS CLI)

Crie e atualize ganchos do 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, o put-lifecycle-hookcomando cria um gancho de ciclo de vida com um tempo limite de pulsação de `3600` segundos (uma hora).

Para alterar o tempo limite de pulsação de um 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 o comando CLI record-lifecycle-action-heartbeat. Isso estende o tempo limite pelo valor especificado quando você criou o gancho de ciclo de vida. Se você terminar antes do término do período de tempo limite, poderá enviar o comando da complete-lifecycle-actionCLI para permitir que a instância prossiga para o próximo estado. Para ter mais informações e exemplos, consulte Concluir uma ação do ciclo de vida.

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.

Concluir a ação do ciclo de vida com um resultado de `CONTINUE` é útil se você terminar antes que o tempo limite expire. Se você não concluir a ação do ciclo de vida, o hook do ciclo de vida vai para o status que você especificou para Resultado padrão após o término do tempo limite.

Conteúdo

- Concluir uma ação do ciclo de vida (manual)
- Concluir uma ação do ciclo de vida (automática)

Concluir uma ação do ciclo de vida (manual)

O procedimento a seguir é para a interface de linha de comando e não tem suporte 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 record-lifecycle-action-heartbeatcomando 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, conforme é mostrado no comando a seguir.

   ```
   aws autoscaling record-lifecycle-action-heartbeat --lifecycle-hook-name my-launch-hook \
     --auto-scaling-group-name my-asg --lifecycle-action-token bcd2f1b8-9a78-44d3-8a7a-4dd07d7cf635
   ```

   Como alternativa, é possível especificar o ID da instância, recebido com a notificação, conforme mostrado no comando a seguir.

   ```
   aws autoscaling record-lifecycle-action-heartbeat --lifecycle-hook-name my-launch-hook \
     --auto-scaling-group-name my-asg --instance-id i-1a2b3c4d
   ```

2. Se você concluir a ação personalizada antes que o período de tempo limite termine, use o complete-lifecycle-actioncomando para que o grupo do Auto Scaling possa continuar iniciando ou encerrando a instância. Você pode especificar o token da ação 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 concluit manualmente as ações do ciclo de vida. Você pode adicionar o complete-lifecycle-actioncomando ao script. O script pode recuperar o ID da instância dos metadados da instância e sinalizar ao Amazon EC2 Auto Scaling quando os scripts de bootstrap tiverem sido concluídos com êxito.

Se você já não estiver fazendo isso, atualize seu script para recuperar o ID da instância nos metadados da instância. Para mais informações, consulte 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 ter mais informações, consulte Tutorial: Configurar um gancho do ciclo de vida que invoca uma função do Lambda.

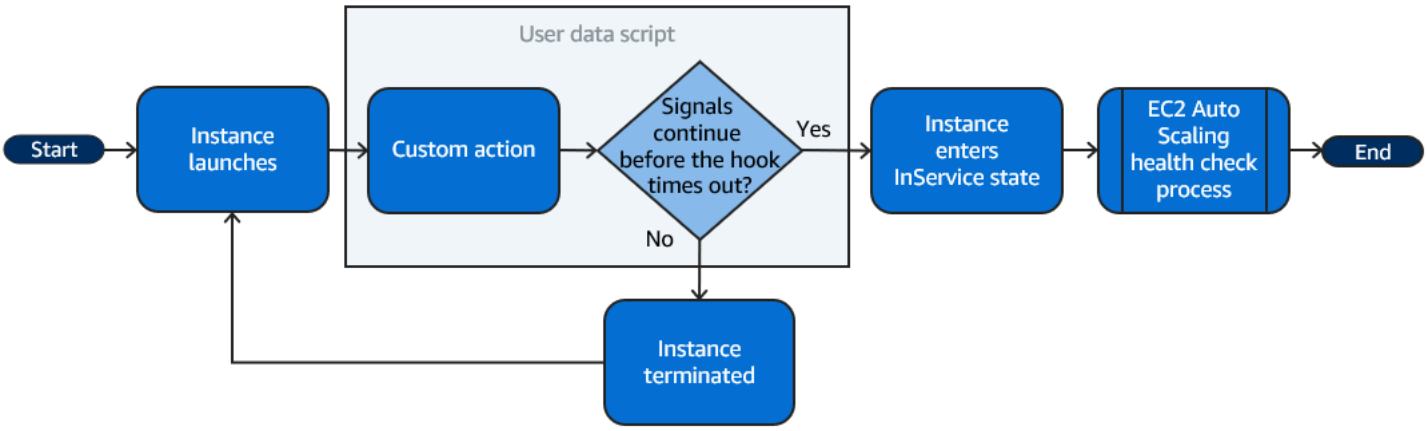
Tutorial: configurar dados do usuário para recuperar o estado de destino do ciclo de vida por meio de metadados de instância

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 a Amazon. EventBridge No entanto, usando um script de dados do usuário para mover o código que configura instâncias e conclui a ação do ciclo de vida para as próprias instâncias, você pode evitar a necessidade de criar infraestrutura adicional.

O tutorial a seguir mostra como começar a usar um script de dados do usuário e metadados de instância. Você cria uma configuração básica de grupo do Auto Scaling com um script de dados do usuário que lê o estado de destino do ciclo de vida das instâncias em seu grupo e executa uma ação

de retorno de chamada em uma fase específica do ciclo de vida de uma instância para continuar o processo de execução.

A ilustração a seguir resume o fluxo de um evento de expansão quando você usa um script de dados do usuário para realizar uma ação personalizada. Depois que uma instância é iniciada, o ciclo de vida da instância é pausado até que o gancho do ciclo de vida seja concluído, seja por meio do tempo limite limite ou pelo Amazon EC2 Auto Scaling recebendo um sinal para continuar.



Conteúdo

- Etapa 1: criar uma função do IAM com permissões para concluir ações de ciclo de vida
- Etapa 2: criar um modelo de execução e incluir a função do IAM e um script de dados de usuário
- Etapa 3: criar um grupo do Auto Scaling
- Etapa 4: Adicionar um gancho do ciclo de vida
- Etapa 5: testar e verificar a funcionalidade
- Etapa 6: limpar
- Recursos relacionados

Etapa 1: criar uma função do IAM com permissões para concluir ações de ciclo de vida

Ao usar o AWS CLI ou um AWS SDK para enviar um retorno de chamada para concluir as ações do ciclo de vida, você deve usar uma função do IAM com permissões para concluir as ações do ciclo de vida.

Para criar a política

1. Abra a página Policies (Políticas) do console do IAM e escolha Create policy (Criar política).

2. Selecione a guia JSON.
3. Na caixa Policy Document (Documento de política), copie e cole o seguinte documento de política. 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 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

1. No painel de navegação à esquerda, escolha Roles (Funções).
2. Selecione 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. Em Adicionar permissões, escolha a política que você criou (TestAutoScalingEvent-policy). Em seguida, clique em 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 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 o perfil da instância do IAM, escolha o nome do perfil da instância do IAM da sua função do IAM (TestAutoScalingEvent-role). Um perfil de instância é um contêiner para 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 pelo `group_name` nome do grupo de Auto Scaling que você deseja criar e `region` pelo que Região da AWS você deseja que seu grupo de Auto Scaling use.

```
#!/bin/bash

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

> **Note**
>
> Para outros exemplos que você pode usar como referência para desenvolver seu script de dados de usuário, consulte o GitHub repositório do Amazon EC2 Auto Scaling.

Etapa 3: criar um grupo do Auto Scaling

Depois de criar seu modelo de execução, crie um grupo do Auto Scaling.

Para criar um grupo do Auto Scaling

1. Na página Choose launch template or configuration (Escolher o modelo ou a configuração de execução), em Auto Scaling group name (Nome do grupo do Auto Scaling), insira um nome para o grupo do Auto Scaling (**`TestAutoScalingEvent-group`**).
2. Escolha Next (Próximo) e acesse a página Choose instance launch options (Escolher as opções de execução de instância).
3. Em Network (Rede), selecione uma VPC.
4. Em Availability Zones and subnets (Zonas de disponibilidade e sub-redes), escolha uma ou mais sub-redes de uma ou mais zonas de disponibilidade.
5. Na seção Instance type requirements (Requisitos de tipo de instância), use a configuração padrão para simplificar essa etapa. (Não substitua o modelo de execução.) Neste tutorial, você fará o execução de apenas uma das Instâncias sob demanda usando o tipo de instância especificado no modelo de execução.
6. Selecione Skip to review (Pular para a revisão) na parte inferior da tela.

7. Na página Review (Revisar), reveja as configurações do grupo do Auto Scaling e escolha Create Auto Scaling group (Criar grupo do Auto Scaling).

Etapa 4: Adicionar um gancho do ciclo de vida

Adicione um gancho do ciclo de vida para manter a instância em um estado de espera até que a ação do ciclo de vida esteja concluída.

Para adicionar um gancho de ciclo de vida

1. Abra a página de grupos do Auto Scaling do console do Amazon EC2.
2. Marque a caixa de seleção ao lado do seu grupo do Auto Scaling. Um painel dividido é aberto na parte inferior da página.
3. No painel inferior, na guia Instance management (Gerenciamento de instâncias), em Lifecycle hooks (Ganchos do ciclo de vida), escolha Create lifecycle hook (Criar gancho do ciclo de vida).
4. Para definir um hook do ciclo de vida para expansão (execução 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 Criar.

Etapa 5: testar e verificar a funcionalidade

Para testar a funcionalidade, atualize o grupo do Auto Scaling aumentando em 1 a capacidade desejada do grupo do Auto Scaling. O script de dados de usuário é executado e começa a verificar o estado de destino do ciclo de vida da instância logo após a execução da instância. O script altera o nome do host e envia uma ação de retorno de chamada quando o estado de destino do ciclo de vida for `InService`. Isso geralmente leva apenas alguns segundos para terminar.

Para aumentar o tamanho de grupo do Auto Scalin

1. Abra a página de grupos do Auto Scaling do console do Amazon EC2.
2. Marque a caixa de seleção ao lado do seu grupo do Auto Scaling. Veja os detalhes em um painel inferior enquanto ainda vê as linhas superiores do painel superior.
3. No painel inferior, na guia Details (Detalhes), escolha Group details (Detalhes do grupo, Edit (Editar).
4. Em Desired capacity (Capacidade desejada), aumente o valor atual em 1.
5. 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 á página Auto Scaling groups (Grupos do Auto Scaling) e selecione seu grupo.
2. Na guia Activity (Atividade), em Activity history (Histórico de atividades), a coluna Status exibe se o seu grupo do Auto Scaling iniciou uma instância com êxito.
3. Se o script de dados de usuário falhar, você observará uma ação de escalabilidade com um status de `Canceled` e uma mensagem de status de `Instance failed to complete user's Lifecycle Action: Lifecycle Action with token e85eb647-4fe0-4909-b341-a6c42EXAMPLE was abandoned: Lifecycle Action Completed with ABANDON Result` após o término do período de tempo limite.

Etapa 6: limpar

Se tiver terminado de trabalhar com os recursos que criou exclusivamente para este tutorial, siga as etapas abaixo para excluí-los.

Para excluir o gancho do ciclo de vida

1. Abra a página de grupos do Auto Scaling do console do Amazon EC2.
2. Marque a caixa de seleção ao lado do seu grupo do Auto Scaling.
3. Na guia Instance management (Gerenciamento de instâncias), em Lifecycle hooks (Ganchos do ciclo de vida), escolha o gancho do ciclo de vida (`TestAutoScalingEvent-hook`).

4. Escolha Ações, Excluir.
5. Para confirmar, escolha Delete (Excluir) novamente.

Para excluir o modelo de execução

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

Se tiver terminado de trabalhar com o grupo de exemplo do Auto Scaling, exclua-o. Você também pode excluir a função do IAM e a política de permissões que criou.

Para excluir o grupo do Auto Scaling

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

   Um ícone de carregamento na coluna Name (Nome) indica que o grupo do Auto Scaling está sendo excluído. É necessário aguardar alguns minutos para encerrar a instância e excluir o grupo.

Para excluir 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.
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 Ações, Excluir.
4. Quando for solicitada confirmação, digite o nome da política e escolha Excluir.

Recursos relacionados

Os tópicos relacionados a seguir podem ser úteis à medida que você desenvolve um código que invoca ações em instâncias com base nos dados disponíveis nos metadados da instância.

- Recuperar o estado de destino do ciclo de vida por meio de metadados de instância. Esta seção descreve o estado do ciclo de vida de outros casos de uso, como o encerramento da instância.
- Adicionar ganchos do ciclo de vida (console). Este procedimento mostra como adicionar hooks do ciclo de vida tanto para expansão (execução de instâncias) quanto para redução (instâncias encerrando ou retornando a um grupo de aquecimento).
- Categorias de metadados de instância no Guia do usuário do Amazon EC2 para instâncias do Linux. Este tópico lista todas as categorias de metadados de instância que você pode usar para invocar ações em instâncias do EC2.

Para ver um tutorial que mostra como usar a Amazon EventBridge para criar regras que invocam funções Lambda com base em eventos que acontecem com as instâncias em seu grupo de Auto Scaling, consulte. Tutorial: Configurar um gancho do ciclo de vida que invoca uma função do Lambda

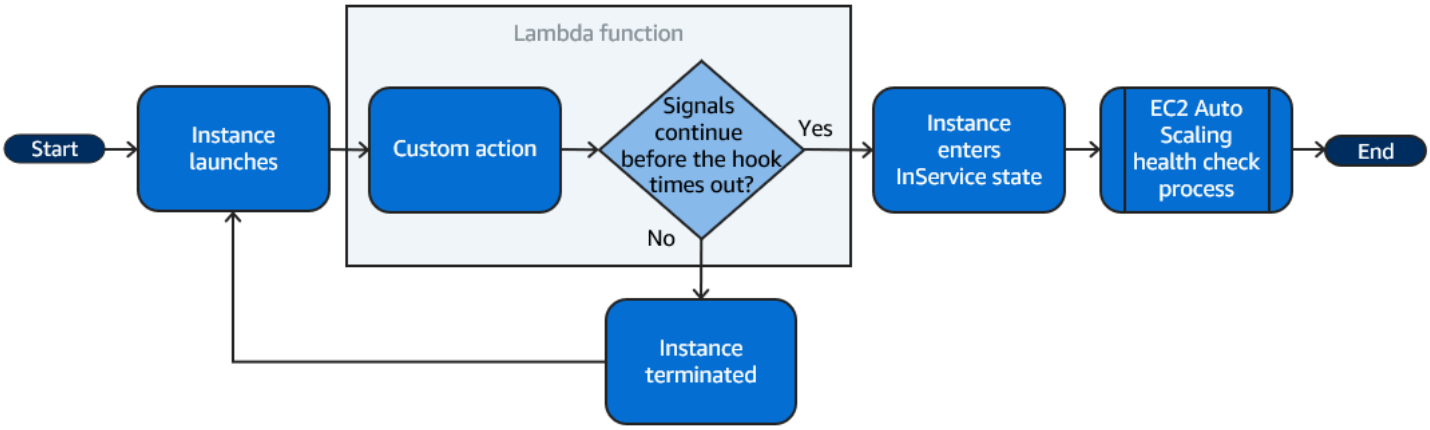
Tutorial: Configurar um gancho do ciclo de vida que invoca uma função do Lambda

Neste exercício, você cria uma EventBridge regra da Amazon que inclui um padrão de filtro que, quando combinado, invoca uma AWS Lambda função como destino da regra. Nós fornecemos o padrão de filtro e código de função de exemplo a ser usada.

Se tudo estiver configurado corretamente, no final deste tutorial, a função do Lambda executará uma ação personalizada quando as instâncias forem iniciadas. A ação personalizada simplesmente registra o evento no stream de CloudWatch registros de registros associado à função Lambda.

A função do Lambda também executa um retorno de chamada para permitir que o ciclo de vida da instância prossiga se essa ação for bem-sucedida, mas permite que a instância abandone o início e termine se a ação falhar.

A ilustração a seguir resume o fluxo de um evento de expansão quando você usa uma função Lambda para realizar uma ação personalizada. Depois que uma instância é iniciada, o ciclo de vida da instância é pausado até que o gancho do ciclo de vida seja concluído, seja por meio do tempo limite limite ou pelo Amazon EC2 Auto Scaling recebendo um sinal para continuar.



Conteúdos

- Pré-requisitos
- Etapa 1: criar uma função do IAM com permissões para concluir ações de ciclo de vida
- Etapa 2: Criar uma função do Lambda
- Etapa 3: criar uma EventBridge regra
- Etapa 4: Adicionar um gancho do ciclo de vida
- Etapa 5: Testar e verificar o evento
- Etapa 6: limpar
- Recursos relacionados

Pré-requisitos

Antes de iniciar este tutorial, crie um grupo do Auto Scaling, se você ainda não tiver um. Para criar um grupo do Auto Scaling, abra a página Grupos do Auto Scaling do console do Amazon EC2 e escolha Criar grupo de Auto Scaling.

Etapa 1: criar 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. Selecione a guia JSON.
3. Na caixa Policy Document (Documento da política), cole o documento de política a seguir na caixa, substituindo o texto em *itálico* 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 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

1. No painel de navegação à esquerda, escolha Roles (Funções).
2. Selecione 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. Em Adicionar permissões, escolha a política que você criou (LogAutoScalingEvent-policy) e a política nomeada AWSLambdaBasicExecutionRole. Em seguida, clique em Próximo.

> **Note**
>
> A AWSLambdaBasicExecutionRolepolítica tem as permissões que a função precisa para gravar registros em CloudWatch Logs.

6. Na página Name, review, and create (Nomear, revisar e criar), em Role name (Nome da função), insira **`LogAutoScalingEvent-role`** e escolha Create role (Criar função).

Etapa 2: Criar uma função do Lambda

Crie uma função do Lambda para servir como destino para eventos. A função Lambda de amostra, escrita em Node.js, é invocada EventBridge quando um evento correspondente é emitido pelo Amazon EC2 Auto Scaling.

Criar uma função do Lambda

1. Abra a página Functions (Funções) no console do Lambda.
2. Escolha Create function (Criar função) e Author from scratch (Criar desde o início).
3. Em Basic information (Informações básicas), em Function name (Nome da função), insira **`LogAutoScalingEvent`**.
4. Em Runtime, selecione Node.js 18.x.
5. Role para baixo e escolha Alterar função de execução padrão e, em seguida, para Função de execução, escolha Usar uma função existente.
6. Em Função existente, escolha LogAutoScalingEvent-role.
7. Deixe os outros valores padrão.
8. Escolha a opção Criar função. Você é retornado ao código e configuração da função.
9. Com sua `LogAutoScalingEvent` função ainda aberta no console, em Código-fonte, no editor, cole o código de exemplo a seguir no arquivo denominado index.mjs.

```
import { AutoScalingClient, CompleteLifecycleActionCommand } from "@aws-sdk/client-
auto-scaling";
export const handler = async(event) => {
```

```
  console.log('LogAutoScalingEvent');
  console.log('Received event:', JSON.stringify(event, null, 2));
  var autoscaling = new AutoScalingClient({ region: event.region });
  var eventDetail = event.detail;
  var params = {
    AutoScalingGroupName: eventDetail['AutoScalingGroupName'], /* required */
    LifecycleActionResult: 'CONTINUE', /* required */
    LifecycleHookName: eventDetail['LifecycleHookName'], /* required */
    InstanceId: eventDetail['EC2InstanceId'],
    LifecycleActionToken: eventDetail['LifecycleActionToken']
  };
  var response;
  const command = new CompleteLifecycleActionCommand(params);
  try {
    var data = await autoscaling.send(command);
    console.log(data); // successful response
    response = {
      statusCode: 200,
      body: JSON.stringify('SUCCESS'),
    };
  } catch (err) {
    console.log(err, err.stack); // an error occurred
    response = {
      statusCode: 500,
      body: JSON.stringify('ERROR'),
    };
  }
  return response;
};
```

Esse código simplesmente registra o evento para que, no final deste tutorial, você possa ver um evento aparecer no stream de CloudWatch registros de registros associado a essa função Lambda.

10. Escolha Implantar.

Etapa 3: criar uma EventBridge regra

Crie uma EventBridge regra para executar sua função Lambda. Para obter mais informações sobre o uso EventBridge, consulte[Use EventBridge para lidar com eventos do Auto Scaling](#).

Como criar uma regra usando o console

1. Abra o console de EventBridge.
2. No painel de navegação, escolha Regras.
3. Selecione Criar regra.
4. Em Define rule detail (Definir detalhe da regra), faça o seguinte:
   a. Em Nome, digite **`LogAutoScalingEvent-rule`**.
   b. Em Event Bus (Barramento de eventos), escolha default (padrão). Quando um AWS service (Serviço da AWS) em sua conta gera um evento, ele sempre vai para o ônibus de eventos padrão da sua conta.
   c. Em Rule type (Tipo de regra), selecione Rule with an event pattern (Regra com um padrão de evento).
   d. Selecione Next (Próximo).
5. Em Build event pattern (Criar padrão de evento), faça o seguinte:
   a. Em Origem do evento, escolha AWS eventos ou eventos de EventBridge parceiros.
   b. Role para baixo até o Padrão de eventos e faça o seguinte:
   c. i. Em Event source, escolha Serviços da AWS.
      ii. Em AWS service (Serviço da AWS), escolha Auto Scaling.
      iii. Em Event type (Tipo de evento), selecione Instance Launch and Terminate (Inicialização e encerramento de instância).
      iv. Por padrão, a regra faz a correspondência com qualquer evento de aumento ou redução horizontal da escala. Para criar uma regra que notifique você quando houver um evento de aumento horizontal da escala e uma instância for colocada em estado de espera devido a um gancho do ciclo de vida, escolhaSpecific instance event(s) (Eventos específicos de instância) e selecione EC2 Instance-launch Lifecycle Action (Ação 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 específico do Auto Scaling, escolha Nome(s) de grupo específico(s) e selecione o grupo.
      vi. Escolha Próximo.
6. Em Select target(s) (Selecionar destino(s)), faça o seguinte:
   a. Em Target types (Tipos de destino), escolha AWS service (Serviço da AWS).

b. Em Select a target (Selecionar um destino), escolha Lambda function (Função do Lambda).

c. Em Função, escolha LogAutoScalingEvent.

d. Escolha Next (Próximo) duas vezes.

7. Na página Revisar e criar, escolha Criar regra.

Etapa 4: Adicionar um gancho do ciclo de vida

Nesta seção, você adicionará um gancho do ciclo de vida para que o Lambda execute sua função em instâncias no início.

Para adicionar um gancho de ciclo de vida

1. Abra a página de grupos do Auto Scaling do console do Amazon EC2.
2. Marque a caixa de seleção ao lado do seu grupo do Auto Scaling. Um painel dividido é aberto na parte inferior da página.
3. No painel inferior, na guia Instance management (Gerenciamento de instâncias), em Lifecycle hooks (Ganchos do ciclo de vida), escolha Create lifecycle hook (Criar gancho do ciclo de vida).
4. Para definir um hook do ciclo de vida para expansão (execução 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 passamos EventBridge contêm todas as informações necessárias para invocar a função Lambda.

5. Escolha Criar.

Etapa 5: Testar e verificar o evento

Para testar o evento, atualize o grupo do Auto Scaling aumentando a capacidade desejada do grupo do Auto Scaling em 1. Sua função do Lambda é invocada dentro de alguns segundos depois do aumento da capacidade desejada.

Para aumentar o tamanho de grupo do Auto Scalin

1. Abra a página de grupos do Auto Scaling do console do Amazon EC2.
2. Marque a caixa de seleção ao lado do seu grupo do Auto Scaling para visualizar detalhes em um painel inferior e ainda ver as linhas superiores do painel superior.
3. No painel inferior, na guia Details (Detalhes), escolha Group details (Detalhes do grupo, Edit (Editar).
4. Em Desired capacity (Capacidade desejada), aumente o valor atual em 1.
5. 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 Grupos de registros do CloudWatch console.
2. Selecione o nome do grupo de logs para sua função do Lambda (`/aws/lambda/LogAutoScalingEvent`).
3. Selecione o nome do fluxo de logs para visualizar os dados fornecidos pela função para a ação do ciclo de vida.

Em seguida, é possível verificar se a instância foi iniciada com êxito a partir da descrição das atividades de escalabilidade.

Para visualizar as atividades de escalabilidade

1. Retorne á página Auto Scaling groups (Grupos do Auto Scaling) e selecione seu grupo.
2. Na guia Activity (Atividade), em Activity history (Histórico de atividades), a coluna Status exibe se o seu grupo do Auto Scaling iniciou uma instância com êxito.

- Se a ação foi bem-sucedida, a atividade de escalabilidade terá o status “Successful” (Sucesso).
- Se falhar, depois de esperar alguns minutos, você observará uma atividade de escalabilidade com o status “Cancelled” (Cancelado) e uma mensagem de status "Instance failed to complete user's Lifecycle Action: Lifecycle Action with token e85eb647-4fe0-4909-b341-a6c42EXAMPLE was abandoned: Lifecycle Action Completed with ABANDON Result" (Instância falhou ao concluir a ação do ciclo de vida do usuário: ação do ciclo de vida com token e85eb647-4fe0-4909-b341-a6c42EXAMPLE foi abandonada: ação do ciclo de vida concluída com o resultado ABANDONAR).

Para reduzir o tamanho do grupo do Auto Scaling

Se não for necessária a instância adicional iniciada para este teste, você pode abrir a guia Details (Detalhes) e reduzir Desired capacity (Capacidade desejada) em 1.

Etapa 6: limpar

Se você tiver terminado de trabalhar com os recursos que você criou apenas para este tutorial, use as seguintes etapas para excluí-los.

Para excluir o gancho do ciclo de vida

1. Abra a página de grupos do Auto Scaling do console do Amazon EC2.
2. Marque a caixa de seleção ao lado do seu grupo do Auto Scaling.
3. Na guia Instance management (Gerenciamento de instâncias), em Lifecycle hooks (Ganchos do ciclo de vida), escolha o gancho do ciclo de vida (`LogAutoScalingEvent-hook`).
4. Escolha Ações, Excluir.
5. Para confirmar, escolha Delete (Excluir) novamente.

Para excluir a EventBridge regra da Amazon

1. Abra a página de regras no EventBridge console da Amazon.
2. Em Event bus (Barramento de eventos), escolha o barramento de eventos associado à regra (`Default`).
3. Marque a caixa de seleção ao lado da sua regra (`LogAutoScalingEvent-rule`).

4. Escolha Delete.
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 Ações, Excluir.
4. Quando for solicitada confirmação, digite **`delete`** para confirmar a exclusão do modelo de execução especificado e, em seguida, escolha Excluir.

Para excluir o grupo de logs

1. Abra a página Grupos de registros do CloudWatch console.
2. Selecione o grupo de logs da função (`/aws/lambda/LogAutoScalingEvent`).
3. Selecione Actions (Ações), Delete log group(s) (Excluir grupo(s) de log).
4. Na caixa de diálogo Delete log group(s) (Excluir grupo(s) de logs), escolha Delete (Excluir).

Para excluir a função de execução

1. Abra a página Roles (Funções) no console do IAM.
2. Selecione o papel da função (`LogAutoScalingEvent-role`).
3. Escolha Delete.
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 Ações, Excluir.

4. Quando for solicitada confirmação, digite o nome da política e escolha Excluir.

Recursos relacionados

Os tópicos relacionados a seguir podem ser úteis à medida que você cria EventBridge regras com base em eventos que acontecem nas instâncias do seu grupo de Auto Scaling.

- Use EventBridge para lidar com eventos do Auto Scaling. Esta seção mostra exemplos de eventos para outros casos de uso, incluindo eventos para redução.
- Adicionar ganchos do ciclo de vida (console). Este procedimento mostra como adicionar hooks do ciclo de vida tanto para expansão (execução de instâncias) quanto para redução (instâncias encerrando ou retornando a um grupo de aquecimento).

Para ver um tutorial que mostra como usar o serviço de metadados de instância (IMDS) para invocar uma ação de dentro da própria instância, consulte. Tutorial: configurar dados do usuário para recuperar o estado de destino do ciclo de vida por meio de metadados de instância

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.

Tópicos

- Conceitos principais

- Pré-requisitos
- Atualização das instâncias em um grupo de aquecimento
- Recursos relacionados
- Limitações
- Usar ganchos do ciclo de vida com um grupo de alta atividade
- Criar um grupo de alta atividade para seu grupo do Auto Scaling
- Visualizar o status e o motivo de falhas da verificação de integridade
- Exemplos para criar e gerenciar piscinas aquecidas com o AWS CLI

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

Você usa hooks do ciclo de vida para colocar instâncias em um estado de espera para poder executar ações personalizadas nas instâncias. Ações personalizadas são executadas à medida que as instâncias são iniciadas ou antes de serem terminadas.

Em uma configuração de pool quente, os hooks do ciclo de vida atrasam a interrupção ou hibernação das instâncias e a colocação em serviço durante um evento de expansão até que concluam a inicialização. Se você adicionar um 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 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

Antes de criar um grupo de aquecimento para seu grupo do Auto Scaling, decida como você usará hooks do ciclo de vida para inicializar novas instâncias com um estado inicial apropriado.

Para realizar ações personalizadas em instâncias enquanto elas estão em estado de espera devido a um hook do ciclo de vida, você tem duas opções:

- Para cenários simples em que você deseja executar comandos em suas instâncias no início, você pode incluir um script de dados do usuário ao criar um modelo de execução ou configuração de execução para o grupo do Auto Scaling. Os scripts de dados do usuário são apenas scripts de shell normais ou diretivas 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 AWS Lambda fazer algo enquanto as instâncias entram ou saem do pool aquecido, você pode criar um gancho de ciclo de vida para seu grupo de Auto Scaling e configurar o serviço de destino para realizar ações

personalizadas com base nas notificações do ciclo de vida. Para ter mais informações, consulte Destinos de notificação compatíveis.

Preparar instâncias para hibernação

Para preparar instâncias do Auto Scaling para usar o estado de grupo `Hibernated`, crie um novo modelo de execução ou configuração de execução configurada corretamente para oferecer suporte à hibernação de instância, conforme descrito no tópico Pré-requisitos de hibernação no Guia do usuário do Amazon EC2 para instâncias do Linux. Em seguida, associe o novo modelo de execução ou a configuração de execução ao grupo do Auto Scaling e inicie uma atualização de instância para substituir as instâncias associadas a um modelo de execução ou a uma configuração de execução anterior. Para ter mais informações, consulte Use uma atualização de instância para atualizar instâncias em um grupo de Auto Scaling.

Atualização das instâncias em um grupo de aquecimento

Para atualizar as instâncias em um grupo de aquecimento, você cria um novo modelo de execução ou configuração de execução e o associa ao grupo do Auto Scaling. Todas as novas instâncias serão iniciadas usando a nova AMI e outras atualizações especificadas no modelo de execução ou na configuração de execução, mas as instâncias existentes não serão afetadas.

Para forçar as instâncias do grupo de aquecimento de substituição a executar esse uso, o modelo de execução ou a configuração de execução, é possível iniciar uma atualização de instância para fazer uma atualização contínua de seu grupo. Uma atualização de instância substitui primeiro as instâncias `InService`. Em seguida, ela substitui as instâncias no grupo de alta atividade. Para ter mais informações, consulte Use uma atualização de instância para atualizar instâncias em um grupo de Auto Scaling.

Recursos relacionados

Você pode visitar nosso GitHubrepositório para ver exemplos de ganchos de ciclo de vida para piscinas aquecidas.

Limitações

- Não há suporte para grupos de instâncias mistas. Não é possível adicionar um grupo de aquecimento aos grupos do Auto Scaling que substituem o tipo de instância especificado em um modelo de execução ou que estão configurados para executar 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 uma instância dentro do pool quente encontrar um problema durante o processo de inicialização, impedindo que ela atinja o `InService` estado, a instância será considerada uma falha na inicialização e encerrada. Isso se aplica independentemente da causa subjacente, como um erro de capacidade insuficiente ou qualquer outro fator.
- 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 todas as lojas em Regiões da AWS que o Amazon EC2 Auto Scaling e a hibernação estão disponíveis, exceto o seguinte:
  - Ásia-Pacífico (Hyderabad)
  - Ásia-Pacífico (Melbourne)
  - Oeste do Canadá (Calgary)
  - Região da China (Pequim)
  - Região da China (Ningxia)
  - Europa (Espanha)

- Israel (Tel Aviv)

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. Para obter mais informações sobre a conclusão de uma ação do ciclo de vida, consulte Concluir uma ação do ciclo de vida.

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.

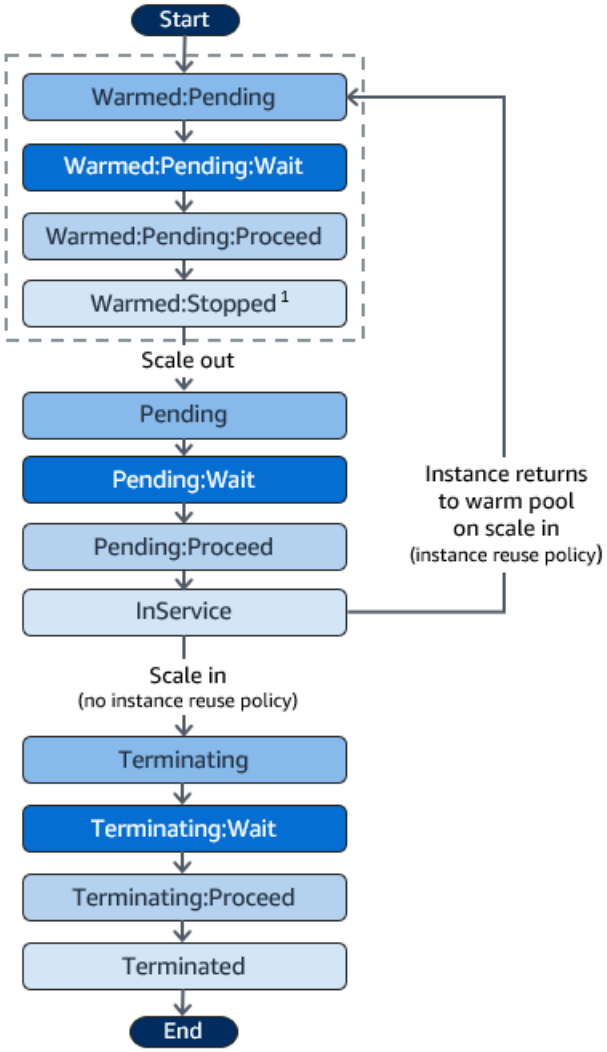
Para instâncias que saem do grupo de alta atividade, talvez você precise de um gancho do ciclo de vida por um dos seguintes motivos:

- 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:

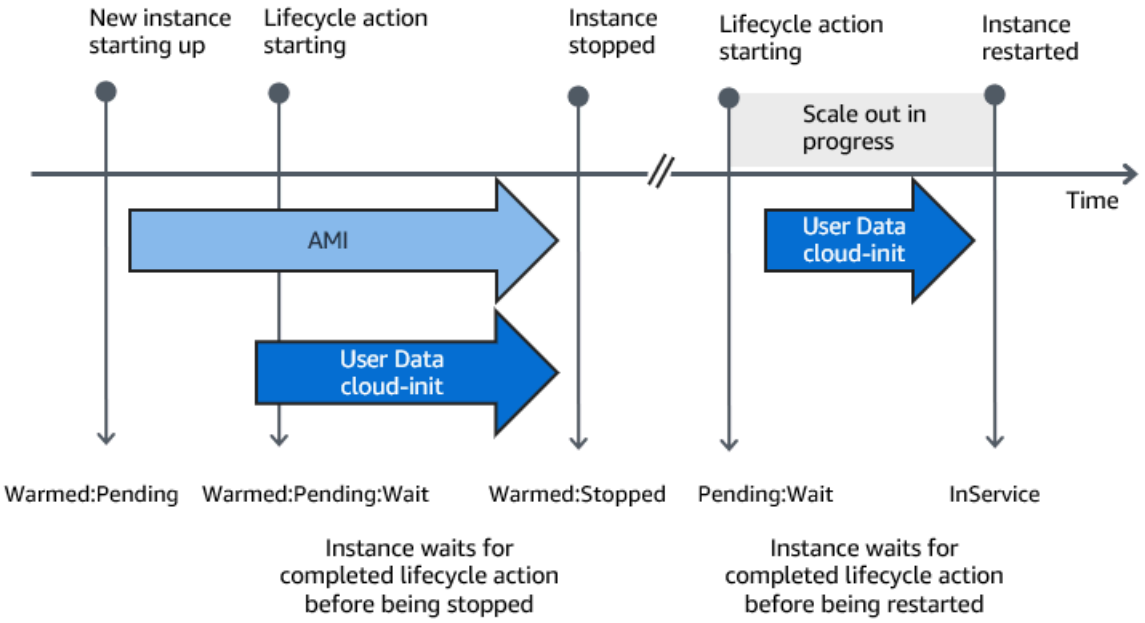


[1] Esse estado varia de acordo com a configuração do estado do grupo de alta atividade. Se o estado do grupo estiver definido como `Running`, então esse estado será `Warmed:Running`, em vez disso. Se o estado do grupo estiver definido como `Hibernated`, então esse estado será `Warmed:Hibernated`, em 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 primeira pausa para realizar uma ação personalizada quando atinge o estado `Warmed:Pending:Wait` e, novamente, quando a instância for reiniciada e atingir o estado `Pending:Wait`.
- Quando um gancho do ciclo de vida é configurado para a ação `EC2_INSTANCE_TERMINATING` 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 ter mais informações, consulte Exemplos de eventos e padrões de grupo de aquecimento.

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:

- EventBridge regras
- 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 regras: para executar código quando o Amazon EC2 Auto Scaling coloca uma instância em estado de espera, você pode criar EventBridge uma regra e especificar uma função Lambda como 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 ter mais informações, consulte Crie EventBridge regras para eventos de piscina aquecida.

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 ter mais informações, consulte Receba notificações usando o Amazon SNS.

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 ter mais informações, consulte Receba notificações usando o Amazon SQS.

Criar um grupo de alta atividade para seu grupo do Auto Scaling

Este tópico descreve como criar um grupo de aquecimento para seu grupo do Auto Scaling.

> **Important**
>
> Antes de continuar, preencha os pré-requisitos para criar um grupo de aquecimento e confirme se você criou um hook do ciclo de vida para o grupo do Auto Scaling.

Criar um grupo de alta atividade

Use o procedimento a seguir para criar um grupo de aquecimento 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.

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 Reutilização de instâncias, marque a caixa de seleção Reutilizar em escala em para permitir que as instâncias do grupo Auto Scaling retornem ao pool aquecido em escala.

   d. Para o tamanho da piscina quente, escolha uma das opções disponíveis:

      - Especificação padrão: O tamanho da piscina aquecida é determinado pela diferença entre a capacidade máxima e a desejada do grupo Auto Scaling. Essa opção simplifica o gerenciamento de piscinas aquecidas. Depois de criar a piscina aquecida, seu tamanho pode ser facilmente atualizado apenas ajustando a capacidade máxima do grupo.
      - Especificação personalizada: o tamanho da piscina aquecida é determinado pela diferença entre um valor personalizado e a capacidade desejada do grupo Auto Scaling. Essa opção oferece flexibilidade para gerenciar o tamanho da sua piscina aquecida independentemente da capacidade máxima do grupo.

6. Veja a seção Tamanho estimado da piscina aquecida com base nas configurações atuais para confirmar como a especificação padrão ou personalizada se aplica ao tamanho da piscina aquecida. Lembre-se de que o tamanho da piscina aquecida depende da capacidade desejada do grupo Auto Scaling, que mudará se o grupo for ampliado.
7. Escolha Criar.

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.
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 Excluir.

Visualizar o status e o motivo de falhas da verificação de integridade

As verificações de integridade permitem que o Amazon EC2 Auto Scaling determine quando uma instância não está íntegra e deve ser terminada. Para instâncias de 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 ter mais informações, consulte Veja o motivo das falhas na verificação de integridade.

Verificações de integridade personalizadas também são aceitas. Isso poderá ser útil se você tiver seu próprio sistema de verificação de integridade capaz de detectar a integridade de uma instância e enviar essas informações para o Amazon EC2 Auto Scaling. Para ter mais informações, consulte Verificações de integridade personalizadas.

No console do Amazon EC2 Auto Scaling, é possível visualizar o status (íntegra ou não íntegra) das instâncias do grupo de alta atividade. Você também pode ver o status de saúde deles usando o SDKs AWS CLI ou um deles.

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 comando a seguir describe-warm-pool.

```
aws autoscaling describe-warm-pool --auto-scaling-group-name my-asg
```

Saída de exemplo.

```
{
    "WarmPoolConfiguration": {
        "MinSize": 0,
        "PoolState": "Stopped"
    },
    "Instances": [
        {
            "InstanceId": "i-0b5e5e7521cfaa46c",
            "InstanceType": "t2.micro",
            "AvailabilityZone": "us-west-2a",
            "LifecycleState": "Warmed:Stopped",
            "HealthStatus": "Healthy",
            "LaunchTemplate": {
                "LaunchTemplateId": "lt-08c4cd42f320d5dcd",
                "LaunchTemplateName": "my-template-for-auto-scaling",
                "Version": "1"
            }
        },
        {
            "InstanceId": "i-0e21af9dcfb7aa6bf",
            "InstanceType": "t2.micro",
            "AvailabilityZone": "us-west-2a",
            "LifecycleState": "Warmed:Stopped",
            "HealthStatus": "Healthy",
            "LaunchTemplate": {
                "LaunchTemplateId": "lt-08c4cd42f320d5dcd",
                "LaunchTemplateName": "my-template-for-auto-scaling",
                "Version": "1"
            }
        }
    ]
}
```

Para visualizar o motivo das falhas de verificação de integridade (AWS CLI)

Use o seguinte comando describe-scaling-activities:

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

Esta é uma resposta de exemplo, em que `Description` indica que seu grupo do Auto Scaling encerrou uma instância e `Cause` indica o motivo da falha na verificação de integridade.

As ações de escalabilidade são ordenadas por horário de início. As atividades ainda em andamento são descritas primeiro.

```
{
  "Activities": [
    {
      "ActivityId": "4c65e23d-a35a-4e7d-b6e4-2eaa8753dc12",
      "AutoScalingGroupName": "my-asg",
      "Description": "Terminating EC2 instance: i-04925c838b6438f14",
      "Cause": "At 2021-04-01T21:48:35Z an instance was taken out of service in
 response to EBS volume health check failure.",
      "StartTime": "2021-04-01T21:48:35.859Z",
      "EndTime": "2021-04-01T21:49:18Z",
      "StatusCode": "Successful",
      "Progress": 100,
      "Details": "{\"Subnet ID\":\"subnet-5ea0c127\",\"Availability Zone\":\"us-west-2a
\"...}",
      "AutoScalingGroupARN": "arn:aws:autoscaling:us-
west-2:123456789012:autoScalingGroup:283179a2-
f3ce-423d-93f6-66bb518232f7:autoScalingGroupName/my-asg"
    },
...
  ]
}
```

Exemplos para criar e gerenciar piscinas aquecidas com o AWS CLI

Você pode criar e gerenciar pools quentes usando o AWS Management Console, 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.

Conteúdo

- Exemplo 1: manter instâncias no estado Stopped

- Exemplo 2: manter instâncias no estado Running
- Exemplo 3: manter instâncias no estado Hibernated
- Exemplo 4: retornar instâncias para o grupo de alta atividade ao reduzir a escala na horizontal
- Exemplo 5: especificar o número mínimo de instâncias no grupo de alta atividade
- Exemplo 6: Defina o tamanho da piscina aquecida usando uma especificação personalizada
- Exemplo 7: definir um tamanho de grupo de alta atividade absoluto
- Exemplo 8: exclusão um grupo de alta atividade

Exemplo 1: manter instâncias no estado **`Stopped`**

O put-warm-poolexemplo a seguir cria um pool aquecido que mantém as instâncias em um `Stopped` estado.

```
aws autoscaling put-warm-pool --auto-scaling-group-name my-asg /
  --pool-state Stopped
```

Exemplo 2: manter instâncias no estado **`Running`**

O put-warm-poolexemplo a seguir cria um pool aquecido que mantém as instâncias em um `Running` estado em vez de em um `Stopped` estado.

```
aws autoscaling put-warm-pool --auto-scaling-group-name my-asg /
  --pool-state Running
```

Exemplo 3: manter instâncias no estado **`Hibernated`**

O put-warm-poolexemplo a seguir cria um pool aquecido que mantém as instâncias em um `Hibernated` estado em vez de em um `Stopped` estado. Isso permite interromper instâncias sem excluir o conteúdo da memória (RAM).

```
aws autoscaling put-warm-pool --auto-scaling-group-name my-asg /
  --pool-state Hibernated
```

Exemplo 4: retornar instâncias para o grupo de alta atividade ao reduzir a escala na horizontal

O put-warm-poolexemplo a seguir cria um pool aquecido que mantém as instâncias em um `Stopped` estado e inclui a `--instance-reuse-policy` opção. O valor da política de reutilização de instâncias `'{"ReuseOnScaleIn": true}'` informa ao Amazon EC2 Auto Scaling para devolver as instâncias ao 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-poolexemplo a seguir cria um pool aquecido que mantém no mínimo 4 instâncias, para que haja pelo menos 4 instâncias disponíveis para lidar com picos de tráfego.

```
aws autoscaling put-warm-pool --auto-scaling-group-name my-asg /
  --pool-state Stopped --min-size 4
```

Exemplo 6: Defina o tamanho da piscina aquecida usando uma especificação personalizada

Por padrão, o Amazon EC2 Auto Scaling gerencia o tamanho da sua piscina aquecida como a diferença entre a capacidade máxima e a desejada do grupo Auto Scaling. No entanto, você pode gerenciar o tamanho da piscina aquecida independentemente da capacidade máxima do grupo usando a `--max-group-prepared-capacity` opção.

O put-warm-poolexemplo a seguir cria um pool aquecido e define o número máximo de instâncias que podem existir simultaneamente no pool aquecido e no grupo Auto Scaling. Se o grupo tiver uma capacidade desejada de 800, o pool aquecido inicialmente terá um tamanho de 100 à medida que for inicializado após a execução desse comando.

```
aws autoscaling put-warm-pool --auto-scaling-group-name my-asg /
  --pool-state Stopped --max-group-prepared-capacity 900
```

Para manter um número mínimo de instâncias no 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-poolexemplo a seguir cria um pool quente que mantém um tamanho constante de pool quente de 10 instâncias.

```
aws autoscaling put-warm-pool --auto-scaling-group-name my-asg /
  --pool-state Stopped --min-size 10 --max-group-prepared-capacity 10
```

Exemplo 8: exclusão um grupo de alta atividade

Use o delete-warm-poolcomando a seguir para excluir uma piscina aquecida.

```
aws autoscaling delete-warm-pool --auto-scaling-group-name my-asg
```

Se houver instâncias no pool aquecido ou se atividades de escalonamento estiverem em andamento, use o delete-warm-poolcomando com a `--force-delete` opção. Essa opção também terminará as instâncias do Amazon EC2 e quaisquer ações de ciclo de vida pendentes.

```
aws autoscaling delete-warm-pool --auto-scaling-group-name my-asg --force-delete
```

Desanexar ou anexar instâncias

Você pode separar instâncias do seu grupo de Auto Scaling. Depois que uma instância é desanexada, ela se torna independente e pode ser gerenciada sozinha ou anexada a um grupo diferente do Auto Scaling, separado do grupo original ao qual ela pertencia. Isso pode ser útil, por exemplo, quando você deseja realizar testes usando instâncias existentes que já estão executando seu aplicativo.

Este tópico fornece instruções sobre como desanexar e anexar instâncias. Ao anexar instâncias, você também pode usar uma instância existente em vez de uma desanexada.

Em vez de desanexar e reanexar uma instância ao mesmo grupo, recomendamos usar o procedimento de espera para remover temporariamente a instância do grupo. Para ter mais informações, consulte Remover temporiamente instâncias do grupo do Auto Scaling.

Conteúdo

- Considerações sobre a separação de instâncias
- Considerações para anexar instâncias
- Mova uma instância para um grupo diferente usando desanexar e anexar

Considerações sobre a separação de instâncias

Ao separar instâncias, lembre-se dos seguintes pontos:

- Você pode desanexar uma instância somente quando ela estiver no `InService` estado.
- Depois de desanexar uma instância, ela continua em execução e incorrendo em cobranças. Para evitar cobranças desnecessárias, reconecte ou encerre as instâncias desconectadas quando elas não forem mais necessárias.
- Você pode optar por diminuir a capacidade desejada pelo número de instâncias que você está desanexando. Se você optar por não diminuir a capacidade, o Amazon EC2 Auto Scaling lançará novas instâncias para substituir as desconectadas e manter a capacidade desejada.
- Se o número de instâncias que você está separando fizer com que o grupo de Auto Scaling fique abaixo de sua capacidade mínima, você deverá diminuir a capacidade mínima.
- Se você separar várias instâncias da mesma zona de disponibilidade sem diminuir a capacidade desejada, o grupo se reequilibrará, a menos que você suspenda o processo. `AZRebalance` Para ter mais informações, consulte Suspender e retomar os processos do Amazon EC2 Auto Scaling.
- Se você desvincular uma instância de um grupo do Auto Scaling que tenha um grupo de destino de balanceador de carga ou um Classic Load Balancer anexado, a instância será cancelada no balanceador de carga. Se a drenagem da conexão (atraso no cancelamento do registro) estiver habilitada para seu balanceador de carga, o Amazon EC2 Auto Scaling aguardará a conclusão das solicitações em andamento.

> **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.

Considerações para anexar instâncias

Observe o seguinte ao anexar instâncias:

- O Amazon EC2 Auto Scaling trata as instâncias anexadas da mesma forma que as instâncias lançadas pelo próprio grupo. Isso significa que as instâncias anexadas podem ser encerradas durante eventos de expansão se forem selecionadas.
- Quando você anexa instâncias, a capacidade desejada do grupo aumente de acordo com o número de instâncias que estão sendo anexadas. Se a capacidade desejada após a adição das novas instâncias exceder o tamanho máximo do grupo, a solicitação para anexar mais instâncias falhará.
- Se você adicionar instâncias ao seu grupo causando uma distribuição desigual entre as zonas de disponibilidade, o Amazon EC2 Auto Scaling reequilibra o grupo para restabelecer uma distribuição uniforme, a menos que você suspenda o processo. `AZRebalance` Para ter mais informações, consulte Suspender e retomar os processos do Amazon EC2 Auto Scaling.
- Se você anexar uma instância a um grupo do Auto Scaling que tenha um grupo de destino de balanceador de carga ou um Classic Load Balancer anexado, a instância será registrada no balanceador de carga.

Para que uma instância seja anexada, ela deve atender aos seguintes critérios:

- A instância está no estado `running` com o Amazon EC2.
- A AMI usada para ativar a instância ainda deve existir.
- A instância não é um membro de outro grupo do Auto Scaling.
- A instância é iniciada em uma das zonas de disponibilidade definidas no grupo 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.

Mova uma instância para um grupo diferente usando desanexar e anexar

Use um dos procedimentos a seguir para separar uma instância do seu grupo de Auto Scaling e anexá-la a outro grupo de Auto Scaling.

Para criar um novo grupo de Auto Scaling a partir de uma instância separada, consulte Criar um grupo do Auto Scaling usando parâmetros de uma instância existente (não recomendado, cria uma configuração de execução).

Console

Para separar uma instância de 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 Instance management (Gerenciamento de instâncias), em Instances (Instâncias), selecione uma instância e escolha Actions (Ações) e Detach (Desvincular).
4. Na caixa de diálogo Desanexar instância, mantenha a caixa de seleção Substituir instância marcada para iniciar uma instância substituta. Desmarque a caixa de seleção para diminuir a capacidade desejada.
5. Quando a confirmação for solicitada, digite **`detach`** para confirmar a exclusão da instância especificada do grupo do Auto Scaling e, em seguida, escolha Desvincular instância.

Agora você pode anexar a instância a um grupo diferente do Auto Scaling.

Para anexar uma instância a um grupo do Auto Scaling

1. Abra o console do Amazon EC2 em https://console.aws.amazon.com/ec2/.
2. (Opcional) No painel de navegação, em Auto Scaling, escolha Grupos do Auto Scaling. Selecione o grupo do Auto Scaling e verifique se o tamanho máximo do grupo do Auto Scaling é grande o suficiente para que você possa adicionar outra instância. Caso contrário, na guia Detalhes aumente a capacidade máxima.
3. No painel de navegação, em Instances (Instâncias), escolha Instances (Instâncias) e selecione uma instância.
4. Escolha Actions (Ações), Instance settings (Configurações da instância), Attach to Auto Scaling Group (Anexar ao grupo do Auto Scaling).
5. Na página Attach to Auto Scaling group (Anexar ao grupo do Auto Scaling), em Auto Scaling Group (Grupo do Auto Scaling), selecione o grupo do Auto Scaling e, em seguida, escolha Attach (Anexar).
6. Se a instância não atender aos critérios, você receberá uma mensagem de erro com os detalhes. Por exemplo, a instância pode não estar na mesma zona de disponibilidade que

o grupo do Auto Scaling. Escolha Fechar e tente novamente com um grupo de Auto Scaling que atenda aos critérios.

AWS CLI

Para desanexar e anexar uma instância, use os comandos de exemplo a seguir. Substitua cada *espaço reservado para entrada do usuário* por suas próprias informações.

Para separar uma instância de um grupo do Auto Scaling

1. Para descrever as instâncias atuais, use o describe-auto-scaling-instancescomando a seguir.

```
aws autoscaling describe-auto-scaling-instances \
  --query 'AutoScalingInstances[?AutoScalingGroupName==`my-asg`]'
```

O exemplo a seguir mostra a saída produzida quando você executa esse comando.

Anote o ID da instância que você pretende remover do grupo. Você precisará desse ID na próxima etapa.

```
{
    "AutoScalingInstances": [
        {
            "ProtectedFromScaleIn": false,
            "AvailabilityZone": "us-west-2a",
            "LaunchTemplate": {
                "LaunchTemplateName": "my-launch-template",
                "Version": "1",
                "LaunchTemplateId": "lt-050555ad16a3f9c7f"
            },
            "InstanceId": "i-05b4f7d5be44822a6",
            "InstanceType": "t3.micro",
            "AutoScalingGroupName": "my-asg",
            "HealthStatus": "HEALTHY",
            "LifecycleState": "InService"
        },
        {
            "ProtectedFromScaleIn": false,
            "AvailabilityZone": "us-west-2a",
            "LaunchTemplate": {
                "LaunchTemplateName": "my-launch-template",
```

```
                "Version": "1",
                "LaunchTemplateId": "lt-050555ad16a3f9c7f"
            },
            "InstanceId": "i-0c20ac468fa3049e8",
            "InstanceType": "t3.micro",
            "AutoScalingGroupName": "my-asg",
            "HealthStatus": "HEALTHY",
            "LifecycleState": "InService"
        },
        {
            "ProtectedFromScaleIn": false,
            "AvailabilityZone": "us-west-2a",
            "LaunchTemplate": {
                "LaunchTemplateName": "my-launch-template",
                "Version": "1",
                "LaunchTemplateId": "lt-050555ad16a3f9c7f"
            },
            "InstanceId": "i-0787762faf1c28619",
            "InstanceType": "t3.micro",
            "AutoScalingGroupName": "my-asg",
            "HealthStatus": "HEALTHY",
            "LifecycleState": "InService"
        },
        {
            "ProtectedFromScaleIn": false,
            "AvailabilityZone": "us-west-2a",
            "LaunchTemplate": {
                "LaunchTemplateName": "my-launch-template",
                "Version": "1",
                "LaunchTemplateId": "lt-050555ad16a3f9c7f"
            },
            "InstanceId": "i-0f280a4c58d319a8a",
            "InstanceType": "t3.micro",
            "AutoScalingGroupName": "my-asg",
            "HealthStatus": "HEALTHY",
            "LifecycleState": "InService"
        }
    ]
}
```

2. Para desanexar uma instância sem diminuir a capacidade desejada, use o comando detach-instances a seguir.

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

Para separar uma instância e diminuir a capacidade desejada, inclua a `--should-decrement-desired-capacity` opção.

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

Agora você pode anexar a instância a um grupo diferente do Auto Scaling.

Para anexar uma instância a um grupo do Auto Scaling

1. Para anexar a instância a um grupo diferente do Auto Scaling, use o comando attach-instances a seguir.

```
aws autoscaling attach-instances --instance-ids i-05b4f7d5be44822a6 --auto-
scaling-group-name my-asg-for-testing
```

2. Para verificar o tamanho do grupo Auto Scaling depois de anexar uma instância, use o comando a seguir. describe-auto-scaling-groups

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

O exemplo de resposta a seguir mostra que o grupo tem duas instâncias em execução, uma das quais é a instância que você anexou.

```
{
    "AutoScalingGroups": [
        {
            "AutoScalingGroupName": "my-asg-for-testing",
            "AutoScalingGroupARN": "arn",
            "LaunchTemplate": {
                "LaunchTemplateName": "my-launch-template",
                "Version": "2",
                "LaunchTemplateId": "lt-050555ad16a3f9c7f"
            },
            "MinSize": 1,
```

```
            "MaxSize": 5,
            "DesiredCapacity": 2,
            ...
            "Instances": [
                {
                    "ProtectedFromScaleIn": false,
                    "AvailabilityZone": "us-west-2a",
                    "LaunchTemplate": {
                        "LaunchTemplateName": "my-launch-template",
                        "Version": "1",
                        "LaunchTemplateId": "lt-050555ad16a3f9c7f"
                    },
                    "InstanceId": "i-05b4f7d5be44822a6",
                    "InstanceType": "t3.micro",
                    "HealthStatus": "Healthy",
                    "LifecycleState": "InService"
                },
                {
                    "ProtectedFromScaleIn": false,
                    "AvailabilityZone": "us-west-2a",
                    "LaunchTemplate": {
                        "LaunchTemplateName": "my-launch-template",
                        "Version": "2",
                        "LaunchTemplateId": "lt-050555ad16a3f9c7f"
                    },
                    "InstanceId": "i-00dcdfffdf5175890",
                    "InstanceType": "t3.micro",
                    "HealthStatus": "Healthy",
                    "LifecycleState": "InService"
                }
            ],
            ...
        }
    ]
}
```

Remover temporiamente instâncias do grupo do Auto Scaling

Você pode colocar uma instância que está no estado `InService` no estado `Standby`, atualize ou solucione problemas da instância e, em seguida, devolva a instância ao serviço. As instâncias que

estão em espera ainda fazem parte do grupo do Auto Scaling, mas não lidam ativamente com o tráfego do balanceador de carga.

Esse recurso ajuda a interromper e iniciar as instâncias ou reiniciá-las sem se preocupar com o término das instâncias do Amazon EC2 Auto Scaling como parte de suas verificações de saúde ou durante eventos de redução de escala na horizontal.

Por exemplo, você pode alterar a imagem de máquina da Amazon (AMI) para um grupo do Auto Scaling a qualquer momento alterando o modelo de execução ou a configuração de execução. Todas as instâncias subsequentes iniciadas pelo grupo do Auto Scaling usam essa AMI. No entanto, o grupo do Auto Scaling não atualiza as instâncias que estão em serviço atualmente. Você pode terminar essas instâncias e permitir que o Amazon EC2 Auto Scaling as substitua ou usar o recurso de atualização de instância para terminar e substituir as instâncias. Você também pode colocar as instâncias em espera, atualizar o software e, em seguida, colocar as instâncias de volta em serviço.

A desvinculação de instâncias de um grupo do Auto Scaling é semelhante a colocar instâncias em espera. Desanexar instâncias pode ser útil se você quiser anexá-las a um grupo diferente ou gerenciar as instâncias, como instâncias autônomas do EC2, e possivelmente encerrá-las. Para ter mais informações, consulte Desanexar ou anexar instâncias.

Conteúdo

- Como o estado de espera funciona
- Considerações
- Status de integridade de uma instância em um estado de espera
- Remova temporariamente uma instância configurando-a como espera

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 balanceador de carga ou um Classic Load Balancer anexado ao seu grupo do Auto Scaling, a instância será registrada no balanceador de carga.

Para obter mais informações sobre o ciclo de vida de instâncias em um grupo do Auto Scaling, consulte Ciclo de vida das instâncias do Amazon EC2 Auto Scaling.

Considerações

Veja a seguir algumas considerações ao mover instâncias para dentro e para fora do estado de espera:

- Ao colocar uma instância em espera, você pode diminuir a capacidade desejada por meio dessa operação ou mantê-la no mesmo valor.
  - Se você optar por não reduzir a capacidade desejada do grupo do Auto Scaling, o Amazon EC2 Auto Scaling iniciará uma instância para substituir a que está em espera. A intenção é ajudar você a manter a capacidade para a aplicação enquanto uma ou mais instâncias estão em espera.
  - Se você optar por diminuir a capacidade desejada do grupo do Auto Scaling, isso impedirá a execução de uma instância para substituir a que está em espera.
- Depois de colocar a instância novamente em serviço, a capacidade desejada é incrementada para refletir quantas instâncias estão no grupo do Auto Scaling.
- Para aumentar (e diminuir), a nova capacidade desejada deve estar entre o tamanho mínimo e máximo do grupo. Caso contrário, haverá falha na operação.
- Se, a qualquer momento, após colocar uma instância em espera ou retornar a instância ao serviço ao sair do estado de espera, descobrir que seu grupo do Auto Scaling não está equilibrado entre as zonas de disponibilidade, o Amazon EC2 Auto Scaling compensa reequilibrando as zonas de disponibilidade, a menos que você suspenda o processo. `AZRebalance` Para ter mais informações, consulte Suspender e retomar os processos do Amazon EC2 Auto Scaling.
- Você é cobrado por instâncias que estão em estado de espera.

Status de integridade de uma instância em um estado de espera

O Amazon EC2 Auto Scaling não executa verificações de integridade em instâncias que estão em um estado de espera. Enquanto a instância está em estado de espera, seu status de integridade reflete o status que ela tinha antes de ser colocada em espera. O Amazon EC2 Auto Scaling não executa uma verificação de integridade na instância até você colocá-la em serviço.

Por exemplo, se você colocar uma instância íntegra em espera e, em seguida, terminá-la, o Amazon EC2 Auto Scaling continuará a relatar a instância como íntegra. Se você tentar colocar uma instância terminada que estava em espera em funcionamento novamente, o Amazon EC2 Auto Scaling executará uma verificação de integridade na instância, determinará que ela está sendo terminada e que não está íntegra e iniciará uma instância de substituição. Para ter mais informações, consulte Verificações de integridade para instâncias em um grupo do Auto Scaling.

Remova temporariamente uma instância configurando-a como espera

Use um dos procedimentos a seguir para tirar temporariamente uma instância de serviço colocando-a no estado de espera.

Console

Para remover uma instância temporariamente

1. Abra o console do Amazon EC2 em https://console.aws.amazon.com/ec2/ e escolha 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 Colocar em espera, mantenha a caixa de seleção Substituir instância para iniciar uma instância substituta. Desmarque a caixa de seleção para diminuir a capacidade desejada.
6. Quando a confirmação for solicitada, digite **`standby`** para confirmar a colocação da instância especificada no estado `Standby` e, em seguida, escolha Colocar em espera.
7. Você pode atualizar ou solucionar problemas de uma instância, conforme necessário. Quando tiver concluído, continue com a próxima etapa para retornar a instância para serviço.

8. Selecione a instância, escolha Ações, Definir como InService. Na caixa de InService diálogo Definir como, escolha Definir como InService.

AWS CLI

Para remover temporariamente uma instância do seu grupo de Auto Scaling, use os seguintes exemplos de comandos. Substitua cada *espaço reservado para entrada do usuário* por suas próprias informações.

Para remover uma instância temporariamente

1. Use o describe-auto-scaling-instancescomando a seguir para identificar a instância a ser atualizada.

```
aws autoscaling describe-auto-scaling-instances \
  --query 'AutoScalingInstances[?AutoScalingGroupName==`my-asg`]'
```

O exemplo a seguir mostra a saída produzida quando você executa esse comando.

Anote o ID da instância que você pretende remover do grupo. Você precisará desse ID na próxima etapa.

```
{
    "AutoScalingInstances": [
        {
            "ProtectedFromScaleIn": false,
            "AvailabilityZone": "us-west-2a",
            "LaunchTemplate": {
                "LaunchTemplateName": "my-launch-template",
                "Version": "1",
                "LaunchTemplateId": "lt-050555ad16a3f9c7f"
            },
            "InstanceId": "i-05b4f7d5be44822a6",
            "InstanceId": "t3.micro",
            "AutoScalingGroupName": "my-asg",
            "HealthStatus": "HEALTHY",
            "LifecycleState": "InService"
        },
        ...
    ]
```

```
}
```

2. Mude a instância para o estado `Standby` usando o seguinte comando enter-standby. A opção `--should-decrement-desired-capacity` reduz a capacidade desejada para que o grupo do Auto Scaling não execute uma instância de substituição.

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

A seguir, uma exemplo de resposta.

```
{
    "Activities": [
        {
            "ActivityId": "3b1839fe-24b0-40d9-80ae-bcd883c2be32",
            "AutoScalingGroupName": "my-asg",
            "Description": "Moving EC2 instance to Standby:
 i-05b4f7d5be44822a6",
            "Cause": "At 2023-12-15T21:31:26Z instance i-05b4f7d5be44822a6 was
 moved to standby
              in response to a user request, shrinking the capacity from 4 to
 3.",
            "StartTime": "2023-12-15T21:31:26.150Z",
            "StatusCode": "InProgress",
            "Progress": 50,
            "Details": "{\"Subnet ID\":\"subnet-c934b782\",\"Availability Zone
\":\"us-west-2a\"}"
        }
    ]
}
```

3. (Opcional) Verifique se a instância está em funcionamento `Standby` usando o describe-auto-scaling-instancescomando a seguir.

```
aws autoscaling describe-auto-scaling-instances --instance-
ids i-05b4f7d5be44822a6
```

A seguir, uma exemplo de resposta. Observe que o status da instância agora é `Standby`.

```
{
    "AutoScalingInstances": [
```

```
        {
            "ProtectedFromScaleIn": false,
            "AvailabilityZone": "us-west-2a",
            "LaunchTemplate": {
                "LaunchTemplateName": "my-launch-template",
                "Version": "1",
                "LaunchTemplateId": "lt-050555ad16a3f9c7f"
            },
            "InstanceId": "i-05b4f7d5be44822a6",
            "InstanceType": "t3.micro",
            "AutoScalingGroupName": "my-asg",
            "HealthStatus": "HEALTHY",
            "LifecycleState": "Standby"
        },
        ...
    ]
}
```

4. Você pode atualizar ou solucionar problemas de uma instância, conforme necessário. Quando tiver concluído, continue com a próxima etapa para retornar a instância para serviço.
5. Coloque a instância de volta em serviço usando o seguinte comando exit-standby.

```
aws autoscaling exit-standby --instance-ids i-05b4f7d5be44822a6 --auto-scaling-
group-name my-asg
```

A seguir, uma exemplo de resposta.

```
{
    "Activities": [
        {
            "ActivityId": "db12b166-cdcc-4c54-8aac-08c5935f8389",
            "AutoScalingGroupName": "my-asg",
            "Description": "Moving EC2 instance out of Standby:
 i-05b4f7d5be44822a6",
            "Cause": "At 2023-12-15T21:46:14Z instance i-05b4f7d5be44822a6 was
 moved out of standby in
               response to a user request, increasing the capacity from 3 to
 4.",
            "StartTime": "2023-12-15T21:46:14.678Z",
            "StatusCode": "PreInService",
            "Progress": 30,
```

```
            "Details": "{\"Subnet ID\":\"subnet-c934b782\",\"Availability Zone
\":\"us-west-2a\"}"
        }
    ]
}
```

6. (Opcional) Verifique se a instância está de volta em serviço usando o seguinte comando `describe-auto-scaling-instances`.

```
aws autoscaling describe-auto-scaling-instances --instance-
ids i-05b4f7d5be44822a6
```

A seguir, uma exemplo de resposta. Observe que o status da instância é `InService`.

```
{
    "AutoScalingInstances": [
        {
            "ProtectedFromScaleIn": false,
            "AvailabilityZone": "us-west-2a",
            "LaunchTemplate": {
                "LaunchTemplateName": "my-launch-template",
                "Version": "1",
                "LaunchTemplateId": "lt-050555ad16a3f9c7f"
            },
            "InstanceId": "i-05b4f7d5be44822a6",
            "InstanceType": "t3.micro",
            "AutoScalingGroupName": "my-asg",
            "HealthStatus": "HEALTHY",
            "LifecycleState": "InService"
        },
       ...
    ]
}
```

Excluir infraestrutura do Auto Scaling

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

Tarefas

- Excluir seu grupo do Auto Scaling

- (Opcional) Excluir a configuração de execução
- (Opcional) Excluir o modelo de execução
- (Opcional) Excluir o balanceador de carga e grupos de destino
- (Opcional) Excluir CloudWatch alarmes

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 do Auto Scaling e escolha Ações, Excluir.
3. Quando a confirmação for solicitada, digite **`delete`** para confirmar a exclusão do grupo do Auto Scaling especificado e, em seguida, escolha Excluir.

   Um ícone de carregamento na coluna Name (Nome) indica que o grupo do Auto Scaling está sendo excluído. As colunas 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 delete-auto-scaling-groupcomando a seguir para excluir o grupo Auto Scaling. Essa operação não funciona se o grupo tiver alguma instância do EC2; é somente para grupos com zero instâncias.

```
aws autoscaling delete-auto-scaling-group --auto-scaling-group-name my-asg
```

Se o grupo tiver instâncias ou atividades de escalabilidade em andamento, use o delete-auto-scaling-groupcomando com a `--force-delete` opção. 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 à esquerda, em Auto Scaling, escolha Grupos do Auto Scaling.
3. Escolha Executar configurações próximo ao topo da página. Quando solicitada a confirmação, escolha Exibir configurações de execução para confirmar que você deseja exibir a página de Configurações de execução.
4. Selecione sua configuração de inicialização e escolha Ações, Excluir configuração de inicialização.
5. Quando a confirmação for solicitada, escolha Excluir.

Para excluir a configuração de ativação (AWS CLI)

Use o seguinte comando delete-launch-configuration:

```
aws autoscaling delete-launch-configuration --launch-configuration-name my-launch-config
```

(Opcional) Excluir o modelo de execução

Você pode excluir o modelo de execução ou apenas uma versão do seu modelo de execução. Ao excluir um modelo de execução, todas as suas versões são excluídas.

É possível ignorar esta etapa para manter o modelo de execução para uso futuro.

Para excluir seu modelo de execução (console)

1. Abra o console do Amazon EC2 em https://console.aws.amazon.com/ec2/.
2. No painel de navegação, escolha Instances e, em seguida, Launch Templates.

3. Selecione o modelo de execução e, depois, execute uma das seguintes ações:

   - Escolha Actions (Ações), Delete template (Excluir modelo). Quando a confirmação for solicitada, digite **`Delete`** para confirmar a exclusão do modelo de execução especificado e, em seguida, escolha Excluir.
   - Escolha Actions (Ações), Delete template version (Excluir versão do modelo). Selecione a versão a ser excluída e escolha Delete (Excluir).

Para excluir o modelo de execução (AWS CLI)

Use o delete-launch-templatecomando 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 delete-launch-template-versionscomando para excluir uma versão específica de um modelo de lançamento.

```
aws ec2 delete-launch-template-versions --launch-template-id lt-068f72b72934aff71 --
versions 1
```

(Opcional) Excluir o balanceador de carga e grupos de destino

Ignore esta etapa se seu grupo do Auto Scaling não estiver associado a um balanceador de carga Elastic Load Balancing ou se desejar manter o balanceador de carga para uso futuro.

Para excluir o balanceador de carga (console)

1. Abra o console do Amazon EC2 em https://console.aws.amazon.com/ec2/.
2. No painel de navegação, em Load Balancing (Balanceamento de carga), escolha Load balancers (Balanceadores de carga).
3. Selecione o balanceador de carga e Actions (Ações), Delete (Excluir).
4. Quando a confirmação for solicitada, escolha Sim, excluir.

Para excluir o grupo de destino (console)

1. No painel de navegação, em Load Balancing (Balanceamento de carga), escolha Grupos de destino.

2. Selecione o grupo de destino e escolha Actions (Ações), Delete (Excluir).
3. Quando a confirmação for solicitada, escolha Sim, excluir.

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

Para balanceadores de carga de aplicativos e balanceadores de carga de rede, use os comandos a seguir delete-load-balancer. delete-target-group

```
aws elbv2 delete-load-balancer --load-balancer-arn my-load-balancer-arn
aws elbv2 delete-target-group --target-group-arn my-target-group-arn
```

Para balanceadores de carga clássicos, use o delete-load-balancercomando a seguir.

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

(Opcional) Excluir CloudWatch alarmes

Para excluir os CloudWatch alarmes associados ao seu grupo de Auto Scaling, conclua as etapas a seguir. Por exemplo, você pode ter alarmes associados à escalabilidade por etapas ou às políticas de escalabilidade simples.

> **Note**
>
> A exclusão de um grupo do Auto Scaling exclui automaticamente os alarmes que CloudWatch o Amazon EC2 Auto Scaling gerencia para uma política de escalabilidade de rastreamento alvo.

Você pode pular essa etapa se o grupo do Auto Scaling não estiver associado a CloudWatch nenhum alarme ou se quiser manter os alarmes para uso futuro.

Para excluir os CloudWatch alarmes (console)

1. Abra o CloudWatch console em https://console.aws.amazon.com/cloudwatch/.
2. No painel de navegação, escolha Alarms (Alarmes).
3. Selecione os alarmes e escolha Action (Ação), Delete (Excluir).
4. Quando a confirmação for solicitada, escolha Excluir.

Para excluir os CloudWatch alarmes ()AWS CLI

Use o comando delete-alarms. É possível excluir um ou mais alarmes por vez. Por exemplo, use o comando a seguir para excluir os alarmes `Step-Scaling-AlarmHigh-AddCapacity` e `Step-Scaling-AlarmLow-RemoveCapacity`.

```
aws cloudwatch delete-alarms --alarm-name Step-Scaling-AlarmHigh-AddCapacity Step-Scaling-AlarmLow-RemoveCapacity
```

Exemplos para criar e gerenciar grupos de Auto Scaling com os SDKs AWS

Você pode criar um grupo de Auto Scaling usando o AWS Management Console, o AWS CLI, um AWS SDK e. AWS CloudFormation

Os exemplos de código a seguir mostram como criar, atualizar, descrever e excluir um grupo do Auto Scaling na sua linguagem de programação compatível favorita usando os AWS SDKs.

Conteúdo

- Crie um grupo de Auto Scaling usando um SDK AWS
- Atualizar um grupo do Auto Scaling usando um SDK AWS
- Descrever um grupo de Auto Scaling usando um SDK AWS
- Excluir um grupo do Auto Scaling usando um SDK AWS

Crie um grupo de Auto Scaling usando um SDK AWS

Os exemplos de código a seguir mostram como usar`CreateAutoScalingGroup`.

.NET

AWS SDK for .NET

> **Note**
>
> Tem mais sobre GitHub. Encontre o exemplo completo e saiba como configurar e executar no AWS Code Examples Repository.

```
    /// <summary>
    /// Create a new Amazon EC2 Auto Scaling group.
    /// </summary>
    /// <param name="groupName">The name to use for the new Auto Scaling
    /// group.</param>
    /// <param name="launchTemplateName">The name of the Amazon EC2 Auto Scaling
    /// launch template to use to create instances in the group.</param>
    /// <returns>A Boolean value indicating the success of the action.</returns>
    public async Task<bool> CreateAutoScalingGroupAsync(
        string groupName,
        string launchTemplateName,
        string availabilityZone)
    {
        var templateSpecification = new LaunchTemplateSpecification
        {
            LaunchTemplateName = launchTemplateName,
        };

        var zoneList = new List<string>
            {
                availabilityZone,
            };

        var request = new CreateAutoScalingGroupRequest
        {
            AutoScalingGroupName = groupName,
            AvailabilityZones = zoneList,
            LaunchTemplate = templateSpecification,
            MaxSize = 6,
            MinSize = 1
        };

        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 CreateAutoScalingGroupa Referência AWS SDK for .NET da API.

C++

SDK for C++

> **Note**
>
> Tem mais sobre GitHub. Encontre o exemplo completo e saiba como configurar e executar no AWS Code Examples Repository.

```
        Aws::Client::ClientConfiguration clientConfig;
        // Optional: Set to the AWS Region (overrides config file).
        // clientConfig.region = "us-east-1";

    Aws::AutoScaling::AutoScalingClient autoScalingClient(clientConfig);

        Aws::AutoScaling::Model::CreateAutoScalingGroupRequest request;
        request.SetAutoScalingGroupName(groupName);
        Aws::Vector<Aws::String> availabilityGroupZones;
        availabilityGroupZones.push_back(
                availabilityZones[availabilityZoneChoice - 1].GetZoneName());
        request.SetAvailabilityZones(availabilityGroupZones);
        request.SetMaxSize(1);
        request.SetMinSize(1);

        Aws::AutoScaling::Model::LaunchTemplateSpecification
launchTemplateSpecification;
        launchTemplateSpecification.SetLaunchTemplateName(templateName);
        request.SetLaunchTemplate(launchTemplateSpecification);

        Aws::AutoScaling::Model::CreateAutoScalingGroupOutcome outcome =
                autoScalingClient.CreateAutoScalingGroup(request);

        if (outcome.IsSuccess()) {
            std::cout << "Created Auto Scaling group '" << groupName << "'..."
                      << std::endl;
        }
        else if (outcome.GetError().GetErrorType() ==
                 Aws::AutoScaling::AutoScalingErrors::ALREADY_EXISTS_FAULT) {
            std::cout << "Auto Scaling group '" << groupName << "' already
exists."
                      << std::endl;
```

```
        }
        else {
            std::cerr << "Error with AutoScaling::CreateAutoScalingGroup. "
                      << outcome.GetError().GetMessage()
                      << std::endl;

        }
```

- Para obter detalhes da API, consulte CreateAutoScalingGroupa Referência AWS SDK for C ++ da API.

CLI

AWS CLI

Exemplo 1: como criar um grupo do Auto Scaling

O exemplo de `create-auto-scaling-group` a seguir cria um grupo do Auto Scaling em sub-redes de várias zonas de disponibilidade de uma região. As instâncias são executadas com a versão padrão do modelo de execução especificado. Observe que os padrões são usados na maioria das outras configurações, como nas políticas de encerramento e na configuração de verificação de integridade.

```
aws autoscaling create-auto-scaling-group \
    --auto-scaling-group-name my-asg \
    --launch-template LaunchTemplateId=lt-1234567890abcde12 \
    --min-size 1 \
    --max-size 5 \
    --vpc-zone-identifier "subnet-5ea0c127,subnet-6194ea3b,subnet-c934b782"
```

Este comando não produz saída.

Para obter mais informações, consulte Grupos do Auto Scaling no Guia do usuário do Amazon EC2 Auto Scaling.

Exemplo 2: como anexar o Application Load Balancer, o Network Load Balancer ou o Gateway Load Balancer

Este exemplo especifica o ARN de um grupo de destino para um balanceador de carga compatível com o tráfego esperado. O tipo de verificação de integridade especifica o ELB.

Desta forma, quando o Elastic Load Balancing reportar uma instância como não íntegra, o grupo do Auto Scaling a substitui. O comando também define um período de carência de 600 segundos para a verificação de integridade. O período de carência ajuda a evitar o encerramento prematuro de instâncias recém-iniciadas.

```
aws autoscaling create-auto-scaling-group \
    --auto-scaling-group-name my-asg \
    --launch-template LaunchTemplateId=lt-1234567890abcde12 \
    --target-group-arns arn:aws:elasticloadbalancing:us-
west-2:123456789012:targetgroup/my-targets/943f017f100becff \
    --health-check-type ELB \
    --health-check-grace-period 600 \
    --min-size 1 \
    --max-size 5 \
    --vpc-zone-identifier "subnet-5ea0c127,subnet-6194ea3b,subnet-c934b782"
```

Este comando não produz saída.

Para obter mais informações, consulte Elastic Load Balancing e o Auto Scaling do Amazon EC2 no Guia do usuário do Auto Scaling do Amazon EC2.

Exemplo 3: como especificar um grupo de posicionamento e usar a versão mais recente do modelo de execução

Este exemplo executa instâncias em um grupo de posicionamento dentro de uma única zona de disponibilidade. Isso pode ser útil para grupos de baixa latência com workloads de HPC. Esse exemplo também especifica o tamanho mínimo e máximo e a capacidade desejada do grupo.

```
aws autoscaling create-auto-scaling-group \
    --auto-scaling-group-name my-asg \
    --launch-template LaunchTemplateId=lt-1234567890abcde12,Version='$Latest' \
    --min-size 1 \
    --max-size 5 \
    --desired-capacity 3 \
    --placement-group my-placement-group \
    --vpc-zone-identifier "subnet-6194ea3b"
```

Este comando não produz saída.

Para obter mais informações, consulte Grupos de posicionamento no Guia do usuário para instâncias do Linux do Amazon EC2.

Exemplo 4: como especificar um grupo do Auto Scaling de instância única e usar uma versão específica para iniciar o modelo

Este exemplo cria um grupo do Auto Scaling com capacidade mínima e máxima definida como 1 para impor que uma apenas instância seja executada. O comando também especifica a v1 de um modelo de execução no qual o ID de um ENI existente é especificado. Ao usar um modelo de execução que especifica um ENI existente para eth0, é necessário especificar uma zona de disponibilidade para o grupo do Auto Scaling que corresponda à interface de rede, mas sem especificar um ID de sub-rede na solicitação.

```
aws autoscaling create-auto-scaling-group \
    --auto-scaling-group-name my-asg-single-instance \
    --launch-template LaunchTemplateName=my-template-for-auto-scaling,Version='1'
 \
    --min-size 1 \
    --max-size 1 \
    --availability-zones us-west-2a
```

Este comando não produz saída.

Para obter mais informações, consulte Grupos do Auto Scaling no Guia do usuário do Amazon EC2 Auto Scaling.

Exemplo 5: como especificar uma política de encerramento diferente

Este exemplo cria um grupo do Auto Scaling usando uma configuração de execução e define a política de encerramento para encerrar as instâncias mais antigas primeiro. O comando também aplica uma tag ao grupo e suas instâncias, com uma chave `Role` e valor de `WebServer`.

```
aws autoscaling create-auto-scaling-group \
    --auto-scaling-group-name my-asg \
    --launch-configuration-name my-lc \
    --min-size 1 \
    --max-size 5 \
    --termination-policies "OldestInstance" \
    --tags "ResourceId=my-asg,ResourceType=auto-scaling-
group,Key=Role,Value=WebServer,PropagateAtLaunch=true" \
    --vpc-zone-identifier "subnet-5ea0c127,subnet-6194ea3b,subnet-c934b782"
```

Este comando não produz saída.

Para obter mais informações, consulte Trabalhar com políticas de término do Amazon EC2 Auto Scaling no Guia do usuário do Amazon EC2 Auto Scaling.

Exemplo 6: como especificar um gancho do ciclo de vida de lançamento

Este exemplo a seguir cria um grupo do Auto Scaling com um gancho do ciclo de vida que oferece suporte a uma ação personalizada na inicialização da instância.

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

Conteúdo do arquivo `config.json`:

```
{
    "AutoScalingGroupName": "my-asg",
    "LaunchTemplate": {
        "LaunchTemplateId": "lt-1234567890abcde12"
    },
    "LifecycleHookSpecificationList": [{
        "LifecycleHookName": "my-launch-hook",
        "LifecycleTransition": "autoscaling:EC2_INSTANCE_LAUNCHING",
        "NotificationTargetARN": "arn:aws:sqs:us-west-2:123456789012:my-sqs-
queue",
        "RoleARN": "arn:aws:iam::123456789012:role/my-notification-role",
        "NotificationMetadata": "SQS message metadata",
        "HeartbeatTimeout": 4800,
        "DefaultResult": "ABANDON"
    }],
    "MinSize": 1,
    "MaxSize": 5,
    "VPCZoneIdentifier": "subnet-5ea0c127,subnet-6194ea3b,subnet-c934b782",
    "Tags": [{
        "ResourceType": "auto-scaling-group",
        "ResourceId": "my-asg",
        "PropagateAtLaunch": true,
        "Value": "test",
        "Key": "environment"
    }]
}
```

Este comando não produz saída.

Para obter mais informações, consulte Ganchos do ciclo de vida do Amazon EC2 Auto Scaling no Manual do usuário do Amazon EC2 Auto Scaling.

Exemplo 7: como especificar um gancho do ciclo de vida de encerramento

Este exemplo a seguir cria um grupo do Auto Scaling com um gancho do ciclo de vida que oferece suporte a uma ação personalizada no encerramento da instância.

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

Conteúdo de `config.json`:

```
{
    "AutoScalingGroupName": "my-asg",
    "LaunchTemplate": {
        "LaunchTemplateId": "lt-1234567890abcde12"
    },
    "LifecycleHookSpecificationList": [{
        "LifecycleHookName": "my-termination-hook",
        "LifecycleTransition": "autoscaling:EC2_INSTANCE_TERMINATING",
        "HeartbeatTimeout": 120,
        "DefaultResult": "CONTINUE"
    }],
    "MinSize": 1,
    "MaxSize": 5,
    "TargetGroupARNs": [
        "arn:aws:elasticloadbalancing:us-west-2:123456789012:targetgroup/my-
targets/73e2d6bc24d8a067"
    ],
    "VPCZoneIdentifier": "subnet-5ea0c127,subnet-6194ea3b,subnet-c934b782"
}
```

Este comando não produz saída.

Para obter mais informações, consulte Ganchos do ciclo de vida do Amazon EC2 Auto Scaling no Manual do usuário do Amazon EC2 Auto Scaling.

Exemplo 8: como especificar uma política de encerramento personalizada

Este exemplo cria um grupo do Auto Scaling que especifica uma política de encerramento da função do Lambda personalizada que diz ao Amazon EC2 Auto Scaling quais instâncias podem ser encerradas com segurança em escala.

```
aws autoscaling create-auto-scaling-group \
    --auto-scaling-group-name my-asg-single-instance \
    --launch-template LaunchTemplateName=my-template-for-auto-scaling \
    --min-size 1 \
    --max-size 5 \
    --termination-policies "arn:aws:lambda:us-
west-2:123456789012:function:HelloFunction:prod" \
    --vpc-zone-identifier "subnet-5ea0c127,subnet-6194ea3b,subnet-c934b782"
```

Este comando não produz saída.

Para obter mais informações, consulte Criar uma política de término personalizada com o Lambda no Guia do usuário do Amazon EC2 Auto Scaling.

- Para obter detalhes da API, consulte CreateAutoScalingGroupna Referência de AWS CLI Comandos.

Java

SDK para Java 2.x

> **Note**
>
> Tem mais sobre GitHub. Encontre o exemplo completo e saiba como configurar e executar no AWS Code Examples Repository.

```
import software.amazon.awssdk.core.waiters.WaiterResponse;
import software.amazon.awssdk.regions.Region;
import software.amazon.awssdk.services.autoscaling.AutoScalingClient;
import software.amazon.awssdk.services.autoscaling.model.AutoScalingException;
import
 software.amazon.awssdk.services.autoscaling.model.CreateAutoScalingGroupRequest;
import
 software.amazon.awssdk.services.autoscaling.model.DescribeAutoScalingGroupsRequest;
import
 software.amazon.awssdk.services.autoscaling.model.DescribeAutoScalingGroupsResponse;
import
 software.amazon.awssdk.services.autoscaling.model.LaunchTemplateSpecification;
import software.amazon.awssdk.services.autoscaling.waiters.AutoScalingWaiter;
```

```
/**
 * Before running this SDK for Java (v2) code example, set up your development
 * environment, including your credentials.
 *
 * For more information, see the following documentation:
 *
 * https://docs.aws.amazon.com/sdk-for-java/latest/developer-guide/get-
started.html
 */
public class CreateAutoScalingGroup {
    public static void main(String[] args) {
        final String usage = """

                Usage:
                    <groupName> <launchTemplateName> <serviceLinkedRoleARN>
 <vpcZoneId>

                Where:
                    groupName - The name of the Auto Scaling group.
                    launchTemplateName - The name of the launch template.\s
                    vpcZoneId - A subnet Id for a virtual private cloud (VPC)
 where instances in the Auto Scaling group can be created.
                """;

        if (args.length != 3) {
            System.out.println(usage);
            System.exit(1);
        }

        String groupName = args[0];
        String launchTemplateName = args[1];
        String vpcZoneId = args[2];
        AutoScalingClient autoScalingClient = AutoScalingClient.builder()
                .region(Region.US_EAST_1)
                .build();

        createAutoScalingGroup(autoScalingClient, groupName, launchTemplateName,
 vpcZoneId);
        autoScalingClient.close();
    }

    public static void createAutoScalingGroup(AutoScalingClient
 autoScalingClient,
            String groupName,
```

```
            String launchTemplateName,
            String vpcZoneId) {

        try {
            AutoScalingWaiter waiter = autoScalingClient.waiter();
            LaunchTemplateSpecification templateSpecification =
 LaunchTemplateSpecification.builder()
                    .launchTemplateName(launchTemplateName)
                    .build();

            CreateAutoScalingGroupRequest request =
 CreateAutoScalingGroupRequest.builder()
                    .autoScalingGroupName(groupName)
                    .availabilityZones("us-east-1a")
                    .launchTemplate(templateSpecification)
                    .maxSize(1)
                    .minSize(1)
                    .vpcZoneIdentifier(vpcZoneId)
                    .build();

            autoScalingClient.createAutoScalingGroup(request);
            DescribeAutoScalingGroupsRequest groupsRequest =
 DescribeAutoScalingGroupsRequest.builder()
                    .autoScalingGroupNames(groupName)
                    .build();

            WaiterResponse<DescribeAutoScalingGroupsResponse> waiterResponse =
 waiter
                    .waitUntilGroupExists(groupsRequest);
            waiterResponse.matched().response().ifPresent(System.out::println);
            System.out.println("Auto Scaling Group created");

        } catch (AutoScalingException e) {
            System.err.println(e.awsErrorDetails().errorMessage());
            System.exit(1);
        }
    }
}
```

- Para obter detalhes da API, consulte CreateAutoScalingGroupa Referência AWS SDK for Java 2.x da API.

Kotlin

SDK for Kotlin

> **Note**
>
> Tem mais sobre GitHub. Encontre o exemplo completo e saiba como configurar e executar no AWS Code Examples Repository.

```
suspend fun createAutoScalingGroup(groupName: String, launchTemplateNameVal:
 String, serviceLinkedRoleARNVal: String, vpcZoneIdVal: String) {
    val templateSpecification = LaunchTemplateSpecification {
        launchTemplateName = launchTemplateNameVal
    }

    val request = CreateAutoScalingGroupRequest {
        autoScalingGroupName = groupName
        availabilityZones = listOf("us-east-1a")
        launchTemplate = templateSpecification
        maxSize = 1
        minSize = 1
        vpcZoneIdentifier = vpcZoneIdVal
        serviceLinkedRoleArn = serviceLinkedRoleARNVal
    }

    // This object is required for the waiter call.
    val groupsRequestWaiter = DescribeAutoScalingGroupsRequest {
        autoScalingGroupNames = listOf(groupName)
    }

    AutoScalingClient { region = "us-east-1" }.use { autoScalingClient ->
        autoScalingClient.createAutoScalingGroup(request)
        autoScalingClient.waitUntilGroupExists(groupsRequestWaiter)
        println("$groupName was created!")
    }
}
```

- Para obter detalhes da API, consulte a CreateAutoScalingGroupreferência da API AWS SDK for Kotlin.

PHP

SDK para PHP

> **Note**
>
> Tem mais sobre GitHub. Encontre o exemplo completo e saiba como configurar e executar no AWS Code Examples Repository.

```
public function createAutoScalingGroup(
    $autoScalingGroupName,
    $availabilityZones,
    $minSize,
    $maxSize,
    $launchTemplateId
) {
    return $this->autoScalingClient->createAutoScalingGroup([
        'AutoScalingGroupName' => $autoScalingGroupName,
        'AvailabilityZones' => $availabilityZones,
        'MinSize' => $minSize,
        'MaxSize' => $maxSize,
        'LaunchTemplate' => [
            'LaunchTemplateId' => $launchTemplateId,
        ],
    ]);
}
```

- Para obter detalhes da API, consulte CreateAutoScalingGroupa Referência AWS SDK for PHP da API.

PowerShell

Ferramentas para PowerShell

Exemplo 1: Este exemplo cria um grupo de Auto Scaling com o nome e os atributos especificados. A capacidade padrão desejada é o tamanho mínimo. Portanto, esse grupo de Auto Scaling inicia duas instâncias, uma em cada uma das duas zonas de disponibilidade especificadas.

```
New-ASAutoScalingGroup -AutoScalingGroupName my-asg -LaunchConfigurationName my-
lc -MinSize 2 -MaxSize 6 -AvailabilityZone @("us-west-2a", "us-west-2b")
```

- Para obter detalhes da API, consulte CreateAutoScalingGroupem Referência de AWS Tools for PowerShell cmdlet.

Python

SDK para Python (Boto3)

> **Note**
>
> Tem mais sobre GitHub. Encontre o exemplo completo e saiba como configurar e executar no AWS Code Examples Repository.

```
class AutoScalingWrapper:
    """Encapsulates Amazon EC2 Auto Scaling actions."""

    def __init__(self, autoscaling_client):
        """
        :param autoscaling_client: A Boto3 Amazon EC2 Auto Scaling client.
        """
        self.autoscaling_client = autoscaling_client


    def create_group(
        self, group_name, 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 a CreateAutoScalingGroupReferência da API AWS SDK for Python (Boto3).

Rust

SDK for Rust

> **Note**
>
> Tem mais sobre GitHub. Encontre o exemplo completo e saiba como configurar e executar no AWS Code Examples Repository.

```
async fn create_group(client: &Client, name: &str, id: &str) -> Result<(), Error>
 {
    client
```

```
        .create_auto_scaling_group()
        .auto_scaling_group_name(name)
        .instance_id(id)
        .min_size(1)
        .max_size(5)
        .send()
        .await?;

    println!("Created AutoScaling group");

    Ok(())
}
```

- Para obter detalhes da API, consulte a CreateAutoScalingGroupreferência da API AWS SDK for Rust.

Para ver exemplos que você pode usar ao criar grupos de instâncias mistas, consulte os recursos a seguir.

- AWS SDK for .NET
- AWS SDK for Go
- AWS SDK para JavaScript
- AWS SDK para PHP V3
- AWS SDK para Python
- AWS SDK para Ruby V3

Atualizar um grupo do Auto Scaling usando um SDK AWS

Os exemplos de código a seguir mostram como usar`UpdateAutoScalingGroup`.

.NET

AWS SDK for .NET

> **Note**
>
> Tem mais sobre GitHub. Encontre o exemplo completo e saiba como configurar e executar no AWS Code Examples Repository.

```
    /// <summary>
    /// Update the capacity of an Auto Scaling group.
    /// </summary>
    /// <param name="groupName">The name of the Auto Scaling group.</param>
    /// <param name="launchTemplateName">The name of the EC2 launch template.</
param>
    /// <param name="maxSize">The maximum number of instances that can be
    /// created for the Auto Scaling group.</param>
    /// <returns>A Boolean value indicating the success of the action.</returns>
    public async Task<bool> UpdateAutoScalingGroupAsync(
        string groupName,
        string launchTemplateName,
        int maxSize)
    {
        var templateSpecification = new LaunchTemplateSpecification
        {
            LaunchTemplateName = launchTemplateName,
        };

        var groupRequest = new UpdateAutoScalingGroupRequest
        {
            MaxSize = maxSize,
            AutoScalingGroupName = groupName,
            LaunchTemplate = templateSpecification,
        };

        var response = await
 _amazonAutoScaling.UpdateAutoScalingGroupAsync(groupRequest);
        if (response.HttpStatusCode == System.Net.HttpStatusCode.OK)
        {
            Console.WriteLine($"You successfully updated the Auto Scaling group
 {groupName}.");
```

```
            return true;
        }
        else
        {
            return false;
        }
    }
```

- Para obter detalhes da API, consulte UpdateAutoScalingGroupa Referência AWS SDK for .NET da API.

C++

SDK for C++

> **Note**
>
> Tem mais sobre GitHub. Encontre o exemplo completo e saiba como configurar e executar no AWS Code Examples Repository.

```
        Aws::Client::ClientConfiguration clientConfig;
        // Optional: Set to the AWS Region (overrides config file).
        // clientConfig.region = "us-east-1";

    Aws::AutoScaling::AutoScalingClient autoScalingClient(clientConfig);

        Aws::AutoScaling::Model::UpdateAutoScalingGroupRequest request;
        request.SetAutoScalingGroupName(groupName);
        request.SetMaxSize(3);

        Aws::AutoScaling::Model::UpdateAutoScalingGroupOutcome outcome =
                autoScalingClient.UpdateAutoScalingGroup(request);

        if (!outcome.IsSuccess()) {
            std::cerr << "Error with AutoScaling::UpdateAutoScalingGroup. "
                      << outcome.GetError().GetMessage()
                      << std::endl;
```

```
        }
```

- Para obter detalhes da API, consulte UpdateAutoScalingGroupa Referência AWS SDK for C ++ da API.

CLI

AWS CLI

Exemplo 1: como atualizar os limites de tamanho de um grupo do Auto Scaling

Este exemplo atualiza o grupo do Auto Scaling especificado com um tamanho mínimo de 2 e máximo de 10.

```
aws autoscaling update-auto-scaling-group \
    --auto-scaling-group-name my-asg \
    --min-size 2 \
    --max-size 10
```

Este comando não produz saída.

Para obter mais informações, consulte Definir limites de escalabilidade para seu grupo do Auto Scaling no Guia do usuário do Amazon EC2 Auto Scaling.

Exemplo 2: como adicionar verificações de integridade do Elastic Load Balancing e especificar quais zonas de disponibilidade e sub-redes usar

Este exemplo atualiza o grupo do Auto Scaling especificado para adicionar verificações de integridade do Elastic Load Balancing. Esse comando também atualiza o valor de `--vpc-zone-identifier` com uma lista de IDs de sub-rede em várias zonas de disponibilidade.

```
aws autoscaling update-auto-scaling-group \
    --auto-scaling-group-name my-asg \
    --health-check-type ELB \
    --health-check-grace-period 600 \
    --vpc-zone-identifier "subnet-5ea0c127,subnet-6194ea3b,subnet-c934b782"
```

Este comando não produz saída.

Para obter mais informações, consulte Elastic Load Balancing e o Auto Scaling do Amazon EC2 no Guia do usuário do Auto Scaling do Amazon EC2.

Exemplo 3: como atualizar o grupo de posicionamento e a política de encerramento

Este exemplo atualiza o grupo de posicionamento e a política de encerramento que devem ser usados.

```
aws autoscaling update-auto-scaling-group \
    --auto-scaling-group-name my-asg \
    --placement-group my-placement-group \
    --termination-policies "OldestInstance"
```

Este comando não produz saída.

Para obter mais informações, consulte Grupos do Auto Scaling no Guia do usuário do Amazon EC2 Auto Scaling.

Exemplo 4: como usar a versão mais recente do modelo de execução

Este exemplo atualiza o grupo do Auto Scaling especificado para que use a versão mais recente do modelo de execução especificado.

```
aws autoscaling update-auto-scaling-group \
    --auto-scaling-group-name my-asg \
    --launch-template LaunchTemplateId=lt-1234567890abcde12,Version='$Latest'
```

Este comando não produz saída.

Para obter mais informações, consulte Modelos de execução no Manual do usuário do Amazon EC2 Auto Scaling.

Exemplo 5: como usar uma versão específica do modelo de execução

Este exemplo atualiza o grupo do Auto Scaling especificado para que use uma versão específica do modelo de execução em vez da versão mais recente ou padrão.

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

Este comando não produz saída.

Para obter mais informações, consulte Modelos de execução no Manual do usuário do Amazon EC2 Auto Scaling.

Exemplo 6: como definir uma política de instâncias mistas e habilitar o rebalanceamento de capacidade

Este exemplo atualiza o grupo do Auto Scaling especificado para que use uma política de instâncias mistas e permita o rebalanceamento de capacidade. Essa estrutura permite especificar grupos com capacidades spot e sob demanda e usar modelos de execução diferentes para arquiteturas diferentes.

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

Conteúdo de `config.json`:

```
{
    "AutoScalingGroupName": "my-asg",
    "CapacityRebalance": true,
    "MixedInstancesPolicy": {
        "LaunchTemplate": {
            "LaunchTemplateSpecification": {
                "LaunchTemplateName": "my-launch-template-for-x86",
                "Version": "$Latest"
            },
            "Overrides": [
                {
                    "InstanceType": "c6g.large",
                    "LaunchTemplateSpecification": {
                        "LaunchTemplateName": "my-launch-template-for-arm",
                        "Version": "$Latest"
                    }
                },
                {
                    "InstanceType": "c5.large"
                },
                {
                    "InstanceType": "c5a.large"
                }
            ]
```

```
        },
        "InstancesDistribution": {
            "OnDemandPercentageAboveBaseCapacity": 50,
            "SpotAllocationStrategy": "capacity-optimized"
        }
    }
}
```

Este comando não produz saída.

Para obter mais informações, consulte Grupos de Auto Scaling com vários tipos de instância e opções de compra no Manual do usuário do Amazon EC2 Auto Scaling.

- Para obter detalhes da API, consulte UpdateAutoScalingGroupna Referência de AWS CLI Comandos.

Java

SDK para Java 2.x

> **Note**
>
> Tem mais sobre GitHub. Encontre o exemplo completo e saiba como configurar e executar no AWS Code Examples Repository.

```
   public static void updateAutoScalingGroup(AutoScalingClient
autoScalingClient, String groupName,
           String launchTemplateName) {
       try {
           AutoScalingWaiter waiter = autoScalingClient.waiter();
           LaunchTemplateSpecification templateSpecification =
LaunchTemplateSpecification.builder()
                   .launchTemplateName(launchTemplateName)
                   .build();

           UpdateAutoScalingGroupRequest groupRequest =
UpdateAutoScalingGroupRequest.builder()
                   .maxSize(3)
                   .autoScalingGroupName(groupName)
                   .launchTemplate(templateSpecification)
                   .build();
```

```
            autoScalingClient.updateAutoScalingGroup(groupRequest);
            DescribeAutoScalingGroupsRequest groupsRequest =
 DescribeAutoScalingGroupsRequest.builder()
                    .autoScalingGroupNames(groupName)
                    .build();

            WaiterResponse<DescribeAutoScalingGroupsResponse> waiterResponse =
 waiter
                    .waitUntilGroupInService(groupsRequest);
            waiterResponse.matched().response().ifPresent(System.out::println);
            System.out.println("You successfully updated the auto scaling group
 " + groupName);

        } catch (AutoScalingException e) {
            System.err.println(e.awsErrorDetails().errorMessage());
            System.exit(1);
        }
    }
```

- Para obter detalhes da API, consulte UpdateAutoScalingGroupa Referência AWS SDK for Java 2.x da API.

Kotlin

SDK for Kotlin

> **Note**
>
> Tem mais sobre GitHub. Encontre o exemplo completo e saiba como configurar e executar no AWS Code Examples Repository.

```
suspend fun updateAutoScalingGroup(groupName: String, launchTemplateNameVal:
 String, serviceLinkedRoleARNVal: String) {
    val templateSpecification = LaunchTemplateSpecification {
        launchTemplateName = launchTemplateNameVal
    }

    val groupRequest = UpdateAutoScalingGroupRequest {
```

```
        maxSize = 3
        serviceLinkedRoleArn = serviceLinkedRoleARNVal
        autoScalingGroupName = groupName
        launchTemplate = templateSpecification
    }

    val groupsRequestWaiter = DescribeAutoScalingGroupsRequest {
        autoScalingGroupNames = listOf(groupName)
    }

    AutoScalingClient { region = "us-east-1" }.use { autoScalingClient ->
        autoScalingClient.updateAutoScalingGroup(groupRequest)
        autoScalingClient.waitUntilGroupExists(groupsRequestWaiter)
        println("You successfully updated the Auto Scaling group  $groupName")
    }
}
```

- Para obter detalhes da API, consulte a UpdateAutoScalingGroupreferência da API AWS SDK for Kotlin.

PHP

SDK para PHP

> **Note**
>
> Tem mais sobre GitHub. Encontre o exemplo completo e saiba como configurar e executar no AWS Code Examples Repository.

```
    public function updateAutoScalingGroup($autoScalingGroupName, $args)
    {
        if (array_key_exists('MaxSize', $args)) {
            $maxSize = ['MaxSize' => $args['MaxSize']];
        } else {
            $maxSize = [];
        }
        if (array_key_exists('MinSize', $args)) {
            $minSize = ['MinSize' => $args['MinSize']];
        } else {
```

```
            $minSize = [];
        }
        $parameters = ['AutoScalingGroupName' => $autoScalingGroupName];
        $parameters = array_merge($parameters, $minSize, $maxSize);
        return $this->autoScalingClient->updateAutoScalingGroup($parameters);
    }
```

- Para obter detalhes da API, consulte UpdateAutoScalingGroupa Referência AWS SDK for PHP da API.

PowerShell

Ferramentas para PowerShell

Exemplo 1: Este exemplo atualiza o tamanho mínimo e máximo do grupo de Auto Scaling especificado.

```
Update-ASAutoScalingGroup -AutoScalingGroupName my-asg -MaxSize 5 -MinSize 1
```

Exemplo 2: Este exemplo atualiza o período de espera padrão do grupo de Auto Scaling especificado.

```
Update-ASAutoScalingGroup -AutoScalingGroupName my-asg -DefaultCooldown 10
```

Exemplo 3: Este exemplo atualiza as zonas de disponibilidade do grupo de Auto Scaling especificado.

```
Update-ASAutoScalingGroup -AutoScalingGroupName my-asg -AvailabilityZone @("us-
west-2a", "us-west-2b")
```

Exemplo 4: Este exemplo atualiza o grupo de Auto Scaling especificado para usar as verificações de saúde do Elastic Load Balancing.

```
Update-ASAutoScalingGroup -AutoScalingGroupName my-asg -HealthCheckType ELB -
HealthCheckGracePeriod 60
```

- Para obter detalhes da API, consulte UpdateAutoScalingGroupem Referência de AWS Tools for PowerShell cmdlet.

Python

SDK para Python (Boto3)

> **Note**
>
> Tem mais sobre GitHub. Encontre o exemplo completo e saiba como configurar e executar no AWS Code Examples Repository.

```
class AutoScalingWrapper:
    """Encapsulates Amazon EC2 Auto Scaling actions."""

    def __init__(self, autoscaling_client):
        """
        :param autoscaling_client: A Boto3 Amazon EC2 Auto Scaling client.
        """
        self.autoscaling_client = autoscaling_client


    def update_group(self, group_name, **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 a UpdateAutoScalingGroupReferência da API AWS SDK for Python (Boto3).

Rust

SDK for Rust

> **Note**
>
> Tem mais sobre GitHub. Encontre o exemplo completo e saiba como configurar e executar no AWS Code Examples Repository.

```
async fn update_group(client: &Client, name: &str, size: i32) -> Result<(),
 Error> {
    client
        .update_auto_scaling_group()
        .auto_scaling_group_name(name)
        .max_size(size)
        .send()
        .await?;

    println!("Updated AutoScaling group");

    Ok(())
}
```

- Para obter detalhes da API, consulte a UpdateAutoScalingGroupreferência da API AWS SDK for Rust.

Descrever um grupo de Auto Scaling usando um SDK AWS

Os exemplos de código a seguir mostram como usar`DescribeAutoScalingGroups`.

.NET

AWS SDK for .NET

> **Note**
>
> Tem mais sobre GitHub. Encontre o exemplo completo e saiba como configurar e executar no AWS Code Examples Repository.

```
    /// <summary>
    /// Get data about the instances in an Amazon EC2 Auto Scaling group.
    /// </summary>
    /// <param name="groupName">The name of the Amazon EC2 Auto Scaling group.</
param>
    /// <returns>A list of Amazon EC2 Auto Scaling details.</returns>
    public async Task<List<AutoScalingInstanceDetails>>
 DescribeAutoScalingInstancesAsync(
        string groupName)
    {
        var groups = await DescribeAutoScalingGroupsAsync(groupName);
        var instanceIds = new List<string>();
        groups!.ForEach(group =>
        {
            if (group.AutoScalingGroupName == groupName)
            {
                group.Instances.ForEach(instance =>
                {
                    instanceIds.Add(instance.InstanceId);
                });
            }
        });

        var scalingGroupsRequest = new DescribeAutoScalingInstancesRequest
        {
            MaxRecords = 10,
            InstanceIds = instanceIds,
        };

        var response = await
 _amazonAutoScaling.DescribeAutoScalingInstancesAsync(scalingGroupsRequest);
```

```
        var instanceDetails = response.AutoScalingInstances;

        return instanceDetails;
    }
```

- Para obter detalhes da API, consulte DescribeAutoScalingGroupsa Referência AWS SDK for .NET da API.

C++

SDK for C++

> **Note**
>
> Tem mais sobre GitHub. Encontre o exemplo completo e saiba como configurar e executar no AWS Code Examples Repository.

```
    Aws::Client::ClientConfiguration clientConfig;
    // Optional: Set to the AWS Region (overrides config file).
    // clientConfig.region = "us-east-1";

Aws::AutoScaling::AutoScalingClient autoScalingClient(clientConfig);

Aws::AutoScaling::Model::DescribeAutoScalingGroupsRequest request;
Aws::Vector<Aws::String> groupNames;
groupNames.push_back(groupName);
request.SetAutoScalingGroupNames(groupNames);

Aws::AutoScaling::Model::DescribeAutoScalingGroupsOutcome outcome =
        client.DescribeAutoScalingGroups(request);

if (outcome.IsSuccess()) {
    autoScalingGroup = outcome.GetResult().GetAutoScalingGroups();
}
else {
    std::cerr << "Error with AutoScaling::DescribeAutoScalingGroups. "
              << outcome.GetError().GetMessage()
              << std::endl;
```

```
    }
```

- Para obter detalhes da API, consulte DescribeAutoScalingGroupsa Referência AWS SDK for C++ da API.

CLI

AWS CLI

Exemplo 1: como descrever o grupo do Auto Scaling especificado

Este exemplo descreve o grupo do Auto Scaling especificado.

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

Saída:

```
{
    "AutoScalingGroups": [
        {
            "AutoScalingGroupName": "my-asg",
            "AutoScalingGroupARN": "arn:aws:autoscaling:us-
west-2:123456789012:autoScalingGroup:930d940e-891e-4781-
a11a-7b0acd480f03:autoScalingGroupName/my-asg",
            "LaunchTemplate": {
                "LaunchTemplateName": "my-launch-template",
                "Version": "1",
                "LaunchTemplateId": "lt-1234567890abcde12"
            },
            "MinSize": 0,
            "MaxSize": 1,
            "DesiredCapacity": 1,
            "DefaultCooldown": 300,
            "AvailabilityZones": [
                "us-west-2a",
                "us-west-2b",
                "us-west-2c"
            ],
            "LoadBalancerNames": [],
            "TargetGroupARNs": [],
```

```
            "HealthCheckType": "EC2",
            "HealthCheckGracePeriod": 0,
            "Instances": [
                {
                    "InstanceId": "i-06905f55584de02da",
                    "InstanceType": "t2.micro",
                    "AvailabilityZone": "us-west-2a",
                    "HealthStatus": "Healthy",
                    "LifecycleState": "InService",
                    "ProtectedFromScaleIn": false,
                    "LaunchTemplate": {
                        "LaunchTemplateName": "my-launch-template",
                        "Version": "1",
                        "LaunchTemplateId": "lt-1234567890abcde12"
                    }
                }
            ],
            "CreatedTime": "2023-10-28T02:39:22.152Z",
            "SuspendedProcesses": [],
            "VPCZoneIdentifier": "subnet-5ea0c127,subnet-6194ea3b,subnet-
c934b782",
            "EnabledMetrics": [],
            "Tags": [],
            "TerminationPolicies": [
                "Default"
            ],
            "NewInstancesProtectedFromScaleIn": false,
            "ServiceLinkedRoleARN":"arn",
            "TrafficSources": []
        }
    ]
}
```

Exemplo 2: como descrever os primeiros 100 grupos do Auto Scaling especificados

Este exemplo descreve os grupos do Auto Scaling especificados. Ele permite especificar até cem nomes de grupos.

```
aws autoscaling describe-auto-scaling-groups \
    --max-items 100 \
    --auto-scaling-group-name "group1" "group2" "group3" "group4"
```

Consulte um exemplo de saída no exemplo 1.

Exemplo 3: como descrever um grupo do Auto Scaling na região especificada

Este exemplo descreve até 75 grupos do Auto Scaling na região especificada.

```
aws autoscaling describe-auto-scaling-groups \
    --max-items 75 \
    --region us-east-1
```

Consulte um exemplo de saída no exemplo 1.

Exemplo 4: como descrever o número especificado do grupo do Auto Scaling

Use a opção `--max-items` para retornar um número específico de grupos do Auto Scaling.

```
aws autoscaling describe-auto-scaling-groups \
    --max-items 1
```

Consulte um exemplo de saída no exemplo 1.

Se a saída incluir um campo `NextToken`, há mais grupos. Para obter os grupos adicionais, use o valor desse campo com a opção `--starting-token` em uma chamada subsequente da seguinte maneira.

```
aws autoscaling describe-auto-scaling-groups \
    --starting-token Z3M3LMPEXAMPLE
```

Consulte um exemplo de saída no exemplo 1.

Exemplo 5: Para descrever grupos de Auto Scaling que usam configurações de inicialização

Este exemplo usa a `--query` opção para descrever grupos de Auto Scaling que usam configurações de inicialização.

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

Saída:

```
[
    {
        "AutoScalingGroupName": "my-asg",
```

```
        "AutoScalingGroupARN": "arn:aws:autoscaling:us-
west-2:123456789012:autoScalingGroup:930d940e-891e-4781-
a11a-7b0acd480f03:autoScalingGroupName/my-asg",
        "LaunchConfigurationName": "my-lc",
        "MinSize": 0,
        "MaxSize": 1,
        "DesiredCapacity": 1,
        "DefaultCooldown": 300,
        "AvailabilityZones": [
            "us-west-2a",
            "us-west-2b",
            "us-west-2c"
        ],
        "LoadBalancerNames": [],
        "TargetGroupARNs": [],
        "HealthCheckType": "EC2",
        "HealthCheckGracePeriod": 0,
        "Instances": [
            {
                "InstanceId": "i-088c57934a6449037",
                "InstanceType": "t2.micro",
                "AvailabilityZone": "us-west-2c",
                "HealthStatus": "Healthy",
                "LifecycleState": "InService",
                "LaunchConfigurationName": "my-lc",
                "ProtectedFromScaleIn": false
            }
        ],
        "CreatedTime": "2023-10-28T02:39:22.152Z",
        "SuspendedProcesses": [],
        "VPCZoneIdentifier": "subnet-5ea0c127,subnet-6194ea3b,subnet-c934b782",
        "EnabledMetrics": [],
        "Tags": [],
        "TerminationPolicies": [
            "Default"
        ],
        "NewInstancesProtectedFromScaleIn": false,
        "ServiceLinkedRoleARN":"arn",
        "TrafficSources": []
    }
]
```

Para obter mais informações, consulte a saída da AWS CLI do filtro no Guia do usuário da interface de linha de AWS comando.

- Para obter detalhes da API, consulte DescribeAutoScalingGroupsna Referência de AWS CLI Comandos.

Java

SDK para Java 2.x

> **Note**
>
> Tem mais sobre GitHub. Encontre o exemplo completo e saiba como configurar e executar no AWS Code Examples Repository.

```
import software.amazon.awssdk.regions.Region;
import software.amazon.awssdk.services.autoscaling.AutoScalingClient;
import software.amazon.awssdk.services.autoscaling.model.AutoScalingException;
import software.amazon.awssdk.services.autoscaling.model.AutoScalingGroup;
import
 software.amazon.awssdk.services.autoscaling.model.DescribeAutoScalingGroupsResponse;
import
 software.amazon.awssdk.services.autoscaling.model.DescribeAutoScalingGroupsRequest;
import software.amazon.awssdk.services.autoscaling.model.Instance;
import java.util.List;

/**
 * Before running this SDK for Java (v2) code example, set up your development
 * environment, including your credentials.
 *
 * For more information, see the following documentation:
 *
 * https://docs.aws.amazon.com/sdk-for-java/latest/developer-guide/get-
started.html
 */
public class DescribeAutoScalingInstances {
    public static void main(String[] args) {
        final String usage = """

                Usage:
                    <groupName>
```

```
                Where:
                    groupName - The name of the Auto Scaling group.
                """;

        if (args.length != 1) {
            System.out.println(usage);
            System.exit(1);
        }

        String groupName = args[0];
        AutoScalingClient autoScalingClient = AutoScalingClient.builder()
                .region(Region.US_EAST_1)
                .build();

        String instanceId = getAutoScaling(autoScalingClient, groupName);
        System.out.println(instanceId);
        autoScalingClient.close();
    }

    public static String getAutoScaling(AutoScalingClient autoScalingClient,
String groupName) {
        try {
            String instanceId = "";
            DescribeAutoScalingGroupsRequest scalingGroupsRequest =
DescribeAutoScalingGroupsRequest.builder()
                    .autoScalingGroupNames(groupName)
                    .build();

            DescribeAutoScalingGroupsResponse response = autoScalingClient
                    .describeAutoScalingGroups(scalingGroupsRequest);
            List<AutoScalingGroup> groups = response.autoScalingGroups();
            for (AutoScalingGroup group : groups) {
                System.out.println("The group name is " +
group.autoScalingGroupName());
                System.out.println("The group ARN is " +
group.autoScalingGroupARN());

                List<Instance> instances = group.instances();
                for (Instance instance : instances) {
                    instanceId = instance.instanceId();
                }
            }
            return instanceId;
```

```
        } catch (AutoScalingException e) {
            System.err.println(e.awsErrorDetails().errorMessage());
            System.exit(1);
        }
        return "";
    }
}
```

- Para obter detalhes da API, consulte DescribeAutoScalingGroupsa Referência AWS SDK for Java 2.x da API.

Kotlin

SDK for Kotlin

> **Note**
>
> Tem mais sobre GitHub. Encontre o exemplo completo e saiba como configurar e executar no AWS Code Examples Repository.

```
suspend fun getAutoScalingGroups(groupName: String) {
    val scalingGroupsRequest = DescribeAutoScalingGroupsRequest {
        autoScalingGroupNames = listOf(groupName)
    }

    AutoScalingClient { region = "us-east-1" }.use { autoScalingClient ->
        val response =
 autoScalingClient.describeAutoScalingGroups(scalingGroupsRequest)
        response.autoScalingGroups?.forEach { group ->
            println("The group name is ${group.autoScalingGroupName}")
            println("The group ARN is ${group.autoScalingGroupArn}")
            group.instances?.forEach { instance ->
                println("The instance id is ${instance.instanceId}")
                println("The lifecycle state is " + instance.lifecycleState)
            }
        }
    }
}
```

- Para obter detalhes da API, consulte a DescribeAutoScalingGroupsreferência da API AWS SDK for Kotlin.

PHP

SDK para PHP

> **Note**
>
> Tem mais sobre GitHub. Encontre o exemplo completo e saiba como configurar e executar no AWS Code Examples Repository.

```
    public function describeAutoScalingGroups($autoScalingGroupNames)
    {
        return $this->autoScalingClient->describeAutoScalingGroups([
            'AutoScalingGroupNames' => $autoScalingGroupNames
        ]);
    }
```

- Para obter detalhes da API, consulte DescribeAutoScalingGroupsa Referência AWS SDK for PHP da API.

PowerShell

Ferramentas para PowerShell

Exemplo 1: Este exemplo lista os nomes dos seus grupos do Auto Scaling.

```
Get-ASAutoScalingGroup | format-table -property AutoScalingGroupName
```

Saída:

```
AutoScalingGroupName
--------------------
my-asg-1
my-asg-2
my-asg-3
```

```
my-asg-4
my-asg-5
my-asg-6
```

Exemplo 2: Este exemplo descreve o grupo de Auto Scaling especificado.

```
Get-ASAutoScalingGroup -AutoScalingGroupName my-asg-1
```

Saída:

```
AutoScalingGroupARN     : arn:aws:autoscaling:us-
west-2:123456789012:autoScalingGroup:930d940e-891e-4781-a11a-7b0acd480
                          f03:autoScalingGroupName/my-asg-1
AutoScalingGroupName    : my-asg-1
AvailabilityZones       : {us-west-2b, us-west-2a}
CreatedTime             : 3/1/2015 9:05:31 AM
DefaultCooldown         : 300
DesiredCapacity         : 2
EnabledMetrics          : {}
HealthCheckGracePeriod  : 300
HealthCheckType         : EC2
Instances               : {my-lc}
LaunchConfigurationName : my-lc
LoadBalancerNames       : {}
MaxSize                 : 0
MinSize                 : 0
PlacementGroup          :
Status                  :
SuspendedProcesses      : {}
Tags                    : {}
TerminationPolicies     : {Default}
VPCZoneIdentifier       : subnet-e4f33493,subnet-5264e837
```

Exemplo 3: Este exemplo descreve os dois grupos de Auto Scaling especificados.

```
Get-ASAutoScalingGroup -AutoScalingGroupName @("my-asg-1", "my-asg-2")
```

Exemplo 4: Este exemplo descreve as instâncias do Auto Scaling para o grupo de Auto Scaling especificado.

```
(Get-ASAutoScalingGroup -AutoScalingGroupName my-asg-1).Instances
```

Exemplo 5: Este exemplo descreve todos os seus grupos de Auto Scaling.

```
Get-ASAutoScalingGroup
```

Exemplo 6: Este exemplo descreve todos os seus grupos de Auto Scaling, em lotes de 10.

```
$nextToken = $null
do {
  Get-ASAutoScalingGroup -NextToken $nextToken -MaxRecord 10
  $nextToken = $AWSHistory.LastServiceResponse.NextToken
} while ($nextToken -ne $null)
```

Exemplo 7: Este LaunchTemplate exemplo descreve o grupo de Auto Scaling especificado. Este exemplo pressupõe que as “Opções de compra de instância” estejam definidas como “Aderir ao modelo de lançamento”. Caso essa opção esteja definida como “Combinar opções de compra e tipos de instância”, LaunchTemplate pode ser acessada usando "MixedInstancesPolicy. LaunchTemplate” propriedade.

```
(Get-ASAutoScalingGroup -AutoScalingGroupName my-ag-1).LaunchTemplate
```

Saída:

```
LaunchTemplateId     LaunchTemplateName   Version
----------------     ------------------   -------
lt-06095fd619cb40371 test-launch-template $Default
```

- Para obter detalhes da API, consulte DescribeAutoScalingGroupsem Referência de AWS Tools for PowerShell cmdlet.

Python

SDK para Python (Boto3)

> **Note**
>
> Tem mais sobre GitHub. Encontre o exemplo completo e saiba como configurar e executar no AWS Code Examples Repository.

```
class AutoScalingWrapper:
    """Encapsulates Amazon EC2 Auto Scaling actions."""

    def __init__(self, autoscaling_client):
        """
        :param autoscaling_client: A Boto3 Amazon EC2 Auto Scaling client.
        """
        self.autoscaling_client = autoscaling_client


    def describe_group(self, group_name):
        """
        Gets information about an Auto Scaling group.

        :param group_name: The name of the group to look up.
        :return: Information about the group, if found.
        """
        try:
            response = self.autoscaling_client.describe_auto_scaling_groups(
                AutoScalingGroupNames=[group_name]
            )
        except ClientError as err:
            logger.error(
                "Couldn't describe group %s. Here's why: %s: %s",
                group_name,
                err.response["Error"]["Code"],
                err.response["Error"]["Message"],
            )
            raise
        else:
            groups = response.get("AutoScalingGroups", [])
            return groups[0] if len(groups) > 0 else None
```

- Para obter detalhes da API, consulte a DescribeAutoScalingGroupsReferência da API AWS SDK for Python (Boto3).

Rust

SDK for Rust

> **Note**
>
> Tem mais sobre GitHub. Encontre o exemplo completo e saiba como configurar e executar no AWS Code Examples Repository.

```
async fn list_groups(client: &Client) -> Result<(), Error> {
    let resp = client.describe_auto_scaling_groups().send().await?;

    println!("Groups:");

    let groups = resp.auto_scaling_groups();

    for group in groups {
        println!(
            "Name:  {}",
            group.auto_scaling_group_name().unwrap_or("Unknown")
        );
        println!(
            "Arn:   {}",
            group.auto_scaling_group_arn().unwrap_or("unknown"),
        );
        println!("Zones: {:?}", group.availability_zones(),);
        println!();
    }

    println!("Found {} group(s)", groups.len());

    Ok(())
}
```

- Para obter detalhes da API, consulte a DescribeAutoScalingGroupsreferência da API AWS SDK for Rust.

Excluir um grupo do Auto Scaling usando um SDK AWS

Os exemplos de código a seguir mostram como usar`DeleteAutoScalingGroup`.

.NET

AWS SDK for .NET

> **Note**
>
> Tem mais sobre GitHub. Encontre o exemplo completo e saiba como configurar e executar no AWS Code Examples Repository.

Atualizar o tamanho mínimo de um grupo do Auto Scaling para zero, encerrar todas as instâncias no grupo e excluir o grupo.

```
    /// <summary>
    /// Try to terminate an instance by its Id.
    /// </summary>
    /// <param name="instanceId">The Id of the instance to terminate.</param>
    /// <returns>Async task.</returns>
    public async Task TryTerminateInstanceById(string instanceId)
    {
        var stopping = false;
        Console.WriteLine($"Stopping {instanceId}...");
        while (!stopping)
        {
            try
            {
                await
_amazonAutoScaling.TerminateInstanceInAutoScalingGroupAsync(
                    new TerminateInstanceInAutoScalingGroupRequest()
                    {
                        InstanceId = instanceId,
                        ShouldDecrementDesiredCapacity = false
                    });
                stopping = true;
            }
            catch (ScalingActivityInProgressException)
            {
```

```
                Console.WriteLine($"Scaling activity in progress for
{instanceId}. Waiting...");
                Thread.Sleep(10000);
            }
        }
    }

    /// <summary>
    /// Tries to delete the EC2 Auto Scaling group. If the group is in use or in
progress,
    /// waits and retries until the group is successfully deleted.
    /// </summary>
    /// <param name="groupName">The name of the group to try to delete.</param>
    /// <returns>Async task.</returns>
    public async Task TryDeleteGroupByName(string groupName)
    {
        var stopped = false;
        while (!stopped)
        {
            try
            {
                await _amazonAutoScaling.DeleteAutoScalingGroupAsync(
                    new DeleteAutoScalingGroupRequest()
                    {
                        AutoScalingGroupName = groupName
                    });
                stopped = true;
            }
            catch (Exception e)
                when ((e is ScalingActivityInProgressException)
                      || (e is Amazon.AutoScaling.Model.ResourceInUseException))
            {
                Console.WriteLine($"Some instances are still running.
Waiting...");
                Thread.Sleep(10000);
            }
        }
    }

    /// <summary>
    /// Terminate instances and delete the Auto Scaling group by name.
    /// </summary>
    /// <param name="groupName">The name of the group to delete.</param>
    /// <returns>Async task.</returns>
```

```
    public async Task TerminateAndDeleteAutoScalingGroupWithName(string
 groupName)
    {
        var describeGroupsResponse = await
 _amazonAutoScaling.DescribeAutoScalingGroupsAsync(
            new DescribeAutoScalingGroupsRequest()
            {
                AutoScalingGroupNames = new List<string>() { groupName }
            });
        if (describeGroupsResponse.AutoScalingGroups.Any())
        {
            // Update the size to 0.
            await _amazonAutoScaling.UpdateAutoScalingGroupAsync(
                new UpdateAutoScalingGroupRequest()
                {
                    AutoScalingGroupName = groupName,
                    MinSize = 0
                });
            var group = describeGroupsResponse.AutoScalingGroups[0];
            foreach (var instance in group.Instances)
            {
                await TryTerminateInstanceById(instance.InstanceId);
            }

            await TryDeleteGroupByName(groupName);
        }
        else
        {
            Console.WriteLine($"No groups found with name {groupName}.");
        }
    }
```

```
    /// <summary>
    /// Delete an Auto Scaling group.
    /// </summary>
    /// <param name="groupName">The name of the Amazon EC2 Auto Scaling group.</
param>
    /// <returns>A Boolean value indicating the success of the action.</returns>
    public async Task<bool> DeleteAutoScalingGroupAsync(
        string groupName)
    {
```

```
        var deleteAutoScalingGroupRequest = new DeleteAutoScalingGroupRequest
        {
            AutoScalingGroupName = groupName,
            ForceDelete = true,
        };

        var response = await
_amazonAutoScaling.DeleteAutoScalingGroupAsync(deleteAutoScalingGroupRequest);
        if (response.HttpStatusCode == System.Net.HttpStatusCode.OK)
        {
            Console.WriteLine($"You successfully deleted {groupName}");
            return true;
        }

        Console.WriteLine($"Couldn't delete {groupName}.");
        return false;
    }
```

- Para obter detalhes da API, consulte DeleteAutoScalingGroupa Referência AWS SDK for .NET da API.

C++

SDK for C++

> **Note**
>
> Tem mais sobre GitHub. Encontre o exemplo completo e saiba como configurar e executar no AWS Code Examples Repository.

```
        Aws::Client::ClientConfiguration clientConfig;
        // Optional: Set to the AWS Region (overrides config file).
        // clientConfig.region = "us-east-1";

    Aws::AutoScaling::AutoScalingClient autoScalingClient(clientConfig);

            Aws::AutoScaling::Model::DeleteAutoScalingGroupRequest request;
            request.SetAutoScalingGroupName(groupName);
```

```
            Aws::AutoScaling::Model::DeleteAutoScalingGroupOutcome outcome =
                    autoScalingClient.DeleteAutoScalingGroup(request);

            if (outcome.IsSuccess()) {
                std::cout << "Auto Scaling group '" << groupName << "' was
 deleted."
                          << std::endl;
            }
            else {
                std::cerr << "Error with AutoScaling::DeleteAutoScalingGroup. "
                          << outcome.GetError().GetMessage()
                          << std::endl;
                result = false;
            }
        }
```

- Para obter detalhes da API, consulte DeleteAutoScalingGroupa Referência AWS SDK for C ++ da API.

CLI

AWS CLI

Exemplo 1: como excluir o grupo do Auto Scaling especificado

Este exemplo exclui o grupo do Auto Scaling especificado.

```
aws autoscaling delete-auto-scaling-group \
    --auto-scaling-group-name my-asg
```

Este comando não produz saída.

Para obter mais informações, consulte Excluir infraestrutura do Auto Scaling no Guia do usuário do Amazon EC2 Auto Scaling.

Exemplo 2: como forçar a exclusão do grupo do Auto Scaling especificado

Use a opção `--force-delete` para excluir o grupo do Auto Scaling sem precisar esperar que as instâncias do grupo sejam encerradas.

```
aws autoscaling delete-auto-scaling-group \
    --auto-scaling-group-name my-asg \
    --force-delete
```

Este comando não produz saída.

Para obter mais informações, consulte Excluir infraestrutura do Auto Scaling no Guia do usuário do Amazon EC2 Auto Scaling.

- Para obter detalhes da API, consulte DeleteAutoScalingGroupna Referência de AWS CLI Comandos.

Java

SDK para Java 2.x

> **Note**
>
> Tem mais sobre GitHub. Encontre o exemplo completo e saiba como configurar e executar no AWS Code Examples Repository.

```
import software.amazon.awssdk.regions.Region;
import software.amazon.awssdk.services.autoscaling.AutoScalingClient;
import software.amazon.awssdk.services.autoscaling.model.AutoScalingException;
import
 software.amazon.awssdk.services.autoscaling.model.DeleteAutoScalingGroupRequest;

/**
 * Before running this SDK for Java (v2) code example, set up your development
 * environment, including your credentials.
 *
 * For more information, see the following documentation:
 *
 * https://docs.aws.amazon.com/sdk-for-java/latest/developer-guide/get-
started.html
 */
public class DeleteAutoScalingGroup {
    public static void main(String[] args) {
        final String usage = """
```

```
                Usage:
                    <groupName>

                Where:
                    groupName - The name of the Auto Scaling group.
                """;

        if (args.length != 1) {
            System.out.println(usage);
            System.exit(1);
        }

        String groupName = args[0];
        AutoScalingClient autoScalingClient = AutoScalingClient.builder()
                .region(Region.US_EAST_1)
                .build();

        deleteAutoScalingGroup(autoScalingClient, groupName);
        autoScalingClient.close();
    }

    public static void deleteAutoScalingGroup(AutoScalingClient
 autoScalingClient, String groupName) {
        try {
            DeleteAutoScalingGroupRequest deleteAutoScalingGroupRequest =
 DeleteAutoScalingGroupRequest.builder()
                    .autoScalingGroupName(groupName)
                    .forceDelete(true)
                    .build();


 autoScalingClient.deleteAutoScalingGroup(deleteAutoScalingGroupRequest);
            System.out.println("You successfully deleted " + groupName);

        } catch (AutoScalingException e) {
            System.err.println(e.awsErrorDetails().errorMessage());
            System.exit(1);
        }
    }
}
```

- Para obter detalhes da API, consulte DeleteAutoScalingGroupa Referência AWS SDK for Java 2.x da API.

Kotlin

SDK for Kotlin

> **Note**
>
> Tem mais sobre GitHub. Encontre o exemplo completo e saiba como configurar e executar no AWS Code Examples Repository.

```
suspend fun deleteSpecificAutoScalingGroup(groupName: String) {
    val deleteAutoScalingGroupRequest = DeleteAutoScalingGroupRequest {
        autoScalingGroupName = groupName
        forceDelete = true
    }

    AutoScalingClient { region = "us-east-1" }.use { autoScalingClient ->
        autoScalingClient.deleteAutoScalingGroup(deleteAutoScalingGroupRequest)
        println("You successfully deleted $groupName")
    }
}
```

- Para obter detalhes da API, consulte a DeleteAutoScalingGroupreferência da API AWS SDK for Kotlin.

PHP

SDK para PHP

> **Note**
>
> Tem mais sobre GitHub. Encontre o exemplo completo e saiba como configurar e executar no AWS Code Examples Repository.

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

- Para obter detalhes da API, consulte DeleteAutoScalingGroupa Referência AWS SDK for PHP da API.

PowerShell

Ferramentas para PowerShell

Exemplo 1: Este exemplo exclui o grupo de Auto Scaling especificado se ele não tiver instâncias em execução. Você será solicitado a confirmar antes que a operação continue.

```
Remove-ASAutoScalingGroup -AutoScalingGroupName my-asg
```

Saída:

```
Confirm
Are you sure you want to perform this action?
Performing operation "Remove-ASAutoScalingGroup (DeleteAutoScalingGroup)" on
 Target "my-asg".
[Y] Yes  [A] Yes to All  [N] No  [L] No to All  [S] Suspend  [?] Help (default is
 "Y"):
```

Exemplo 2: Se você especificar o parâmetro Force, não será solicitada a confirmação antes que a operação continue.

```
Remove-ASAutoScalingGroup -AutoScalingGroupName my-asg -Force
```

Exemplo 3: Este exemplo exclui o grupo de Auto Scaling especificado e encerra todas as instâncias em execução que ele contém.

```
Remove-ASAutoScalingGroup -AutoScalingGroupName my-asg -ForceDelete $true -Force
```

- Para obter detalhes da API, consulte DeleteAutoScalingGroupem Referência de AWS Tools for PowerShell cmdlet.

Python

SDK para Python (Boto3)

> **Note**
>
> Tem mais sobre GitHub. Encontre o exemplo completo e saiba como configurar e executar no AWS Code Examples Repository.

Atualizar o tamanho mínimo de um grupo do Auto Scaling para zero, encerrar todas as instâncias no grupo e excluir o grupo.

```
class AutoScaler:
    """
    Encapsulates Amazon EC2 Auto Scaling and EC2 management actions.
    """

    def __init__(
        self,
        resource_prefix,
        inst_type,
        ami_param,
        autoscaling_client,
        ec2_client,
        ssm_client,
        iam_client,
    ):
        """
        :param resource_prefix: The prefix for naming AWS resources that are
 created by this class.
        :param inst_type: The type of EC2 instance to create, such as t3.micro.
        :param ami_param: The Systems Manager parameter used to look up the AMI
 that is
                          created.
        :param autoscaling_client: A Boto3 EC2 Auto Scaling client.
        :param ec2_client: A Boto3 EC2 client.
        :param ssm_client: A Boto3 Systems Manager client.
```

```
        :param iam_client: A Boto3 IAM client.
        """
        self.inst_type = inst_type
        self.ami_param = ami_param
        self.autoscaling_client = autoscaling_client
        self.ec2_client = ec2_client
        self.ssm_client = ssm_client
        self.iam_client = iam_client
        self.launch_template_name = f"{resource_prefix}-template"
        self.group_name = f"{resource_prefix}-group"
        self.instance_policy_name = f"{resource_prefix}-pol"
        self.instance_role_name = f"{resource_prefix}-role"
        self.instance_profile_name = f"{resource_prefix}-prof"
        self.bad_creds_policy_name = f"{resource_prefix}-bc-pol"
        self.bad_creds_role_name = f"{resource_prefix}-bc-role"
        self.bad_creds_profile_name = f"{resource_prefix}-bc-prof"
        self.key_pair_name = f"{resource_prefix}-key-pair"


    def _try_terminate_instance(self, inst_id):
        stopping = False
        log.info(f"Stopping {inst_id}.")
        while not stopping:
            try:
                self.autoscaling_client.terminate_instance_in_auto_scaling_group(
                    InstanceId=inst_id, ShouldDecrementDesiredCapacity=True
                )
                stopping = True
            except ClientError as err:
                if err.response["Error"]["Code"] == "ScalingActivityInProgress":
                    log.info("Scaling activity in progress for %s. Waiting...",
inst_id)
                    time.sleep(10)
                else:
                    raise AutoScalerError(f"Couldn't stop instance {inst_id}:
{err}.")

    def _try_delete_group(self):
        """
        Tries to delete the EC2 Auto Scaling group. If the group is in use or in
progress,
        the function waits and retries until the group is successfully deleted.
        """
        stopped = False
```

```
        while not stopped:
            try:
                self.autoscaling_client.delete_auto_scaling_group(
                    AutoScalingGroupName=self.group_name
                )
                stopped = True
                log.info("Deleted EC2 Auto Scaling group %s.", self.group_name)
            except ClientError as err:
                if (
                    err.response["Error"]["Code"] == "ResourceInUse"
                    or err.response["Error"]["Code"] ==
 "ScalingActivityInProgress"
                ):
                    log.info(
                        "Some instances are still running. Waiting for them to
 stop..."
                    )
                    time.sleep(10)
                else:
                    raise AutoScalerError(
                        f"Couldn't delete group {self.group_name}: {err}."
                    )

    def delete_group(self):
        """
        Terminates all instances in the group, deletes the EC2 Auto Scaling
 group.
        """
        try:
            response = self.autoscaling_client.describe_auto_scaling_groups(
                AutoScalingGroupNames=[self.group_name]
            )
            groups = response.get("AutoScalingGroups", [])
            if len(groups) > 0:
                self.autoscaling_client.update_auto_scaling_group(
                    AutoScalingGroupName=self.group_name, MinSize=0
                )
                instance_ids = [inst["InstanceId"] for inst in groups[0]
["Instances"]]
                for inst_id in instance_ids:
                    self._try_terminate_instance(inst_id)
                self._try_delete_group()
            else:
```

```
                log.info("No groups found named %s, nothing to do.",
 self.group_name)
        except ClientError as err:
            raise AutoScalerError(f"Couldn't delete group {self.group_name}:
 {err}.")
```

- Para obter detalhes da API, consulte a DeleteAutoScalingGroupReferência da API AWS SDK for Python (Boto3).

Rust

SDK for Rust

> **Note**
>
> Tem mais sobre GitHub. Encontre o exemplo completo e saiba como configurar e executar no AWS Code Examples Repository.

```
async fn delete_group(client: &Client, name: &str, force: bool) -> Result<(),
 Error> {
    client
        .delete_auto_scaling_group()
        .auto_scaling_group_name(name)
        .set_force_delete(if force { Some(true) } else { None })
        .send()
        .await?;

    println!("Deleted Auto Scaling group");

    Ok(())
}
```

- Para obter detalhes da API, consulte a DeleteAutoScalingGroupreferência da API AWS SDK for Rust.

Recicle as instâncias em seu grupo do Auto Scaling

O Amazon EC2 Auto Scaling oferece recursos que permitem substituir as instâncias do Amazon EC2 em seu grupo de Auto Scaling após fazer atualizações, como adicionar um novo modelo de lançamento por uma nova Amazon Machine Image (AMI) ou adicionar novos tipos de instância. 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.
- Use pontos de verificação para atualizar 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 algum CloudWatch alarme da Amazon que você especificar entrar no `ALARM` estado.
- Limitar a vida útil das instâncias para fornecer versões de software consistentes e configurações de instância em todo o grupo do Auto Scaling.

Conteúdo

- Use uma atualização de instância para atualizar instâncias em um grupo de Auto Scaling
- Substituir instâncias do Auto Scaling com base na vida útil máxima da instância

Use uma atualização de instância para atualizar instâncias em um grupo de Auto Scaling

Você pode usar uma atualização de instância para atualizar as instâncias em seu grupo de Auto Scaling. Esse recurso pode ser útil quando uma alteração na configuração exige que você substitua instâncias, especialmente se seu grupo de Auto Scaling contiver um grande número de instâncias.

Algumas situações em que uma atualização de instância pode ajudar incluem:

- Implantação de uma nova Amazon Machine Image (AMI) ou script de dados do usuário em todo o seu grupo de Auto Scaling. Você pode criar um novo modelo de execução com as alterações e, em seguida, usar uma atualização de instância para implantar as atualizações imediatamente.
- Migrar suas instâncias para novos tipos de instância para aproveitar as melhorias e otimizações mais recentes.
- Mudando seus grupos de Auto Scaling do uso de uma configuração de lançamento para o uso de um modelo de lançamento. Você pode copiar suas configurações de execução para modelos de execução e, em seguida, usar uma atualização de instância para atualizar suas instâncias para os novos modelos. Para obter mais informações sobre a migração para modelos de lançamento, consulte Migre seus grupos de Auto Scaling para modelos de lançamento.

Conteúdo

- Como funciona uma atualização de instância
- Entender os valores padrão de uma atualização de instância
- Iniciar uma atualização de instância
- Monitore uma atualização de instância
- Cancelar uma atualização de instância
- Desfazer alterações com uma reversão
- Usar uma atualização de instância com opção de ignorar correspondência
- Adicionar pontos de verificação a uma atualização de instância

Como funciona uma atualização de instância

Este tópico descreve como a atualização de uma instância funciona e apresenta os principais conceitos que você precisa entender para usá-la com eficiência.

Conteúdo

- Como funcionam
- Conceitos principais
- Período de carência da verificação de integridade
- Compatibilidade de tipo de instância
- Limitações

Como funcionam

Para atualizar instâncias em um grupo de Auto Scaling, você pode definir uma nova configuração que contenha a versão mais recente do seu aplicativo e quaisquer outras atualizações que você queira fazer. Em seguida, inicie uma atualização da instância para substituir as existentes por novas com base nessa configuração.

Para realizar uma atualização da instância:

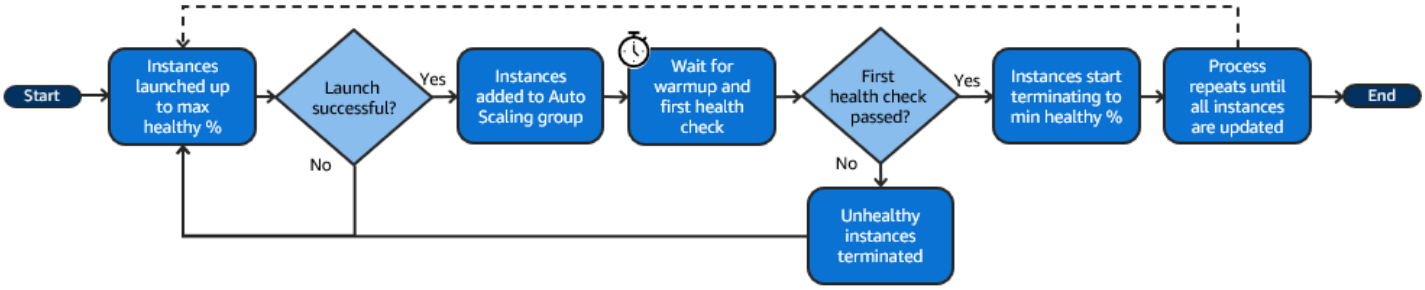
1. Crie um novo modelo de lançamento ou atualize o modelo existente com as alterações de configuração desejadas, como uma nova Amazon Machine Image (AMI). Para ter mais informações, consulte Criar um modelo de execução para um grupo do Auto Scaling.
2. Inicie a atualização da instância usando o console AWS CLI do Amazon EC2 Auto Scaling, ou SDK:
   - Especifique o novo modelo de lançamento ou a versão do modelo de lançamento que você criou. Isso será usado para iniciar novas instâncias.
   - Defina a porcentagem saudável mínima e máxima preferida. Isso controla quantas instâncias são substituídas simultaneamente e se novas instâncias são iniciadas antes de encerrar as antigas.
   - Defina todas as configurações opcionais, como:
     - Pontos de verificação — pause a atualização da instância após uma certa porcentagem de substituições para verificar o progresso.
     - Ignorar a correspondência — compare as instâncias antigas com a nova configuração e substitua somente aquelas que não correspondem. Quando você inicia uma atualização de instância no console, a opção ignorar a correspondência está ativada por padrão.
     - Vários tipos de instância — aplique uma política de instâncias mistas nova ou atualizada como parte da configuração desejada.

Quando a atualização da instância for iniciada, o Amazon EC2 Auto Scaling irá:

- Substitua as instâncias em lotes com base nas porcentagens mínimas e máximas de integridade.
- Inicie as novas instâncias antes de encerrar as antigas, se a porcentagem mínima íntegra estiver definida como 100 por cento. Isso garante que a capacidade desejada seja mantida em todos os momentos.
- Verifique o estado de saúde das instâncias e aguarde um tempo para que elas se aqueçam antes que mais instâncias sejam substituídas.

- Encerre e substitua instâncias consideradas insalubres.
- Atualize automaticamente as configurações do grupo Auto Scaling com as novas alterações de configuração após a atualização da instância ser bem-sucedida.
- Se o seu grupo tiver um pool aquecido, o Amazon EC2 Auto Scaling substituirá primeiro as instâncias. `InService` Em seguida, ela substitui as instâncias no grupo de alta atividade.

O fluxograma a seguir ilustra o comportamento de lançamento antes do encerramento quando você define a porcentagem íntegra mínima como 100 por cento.



> **Note**
>
> As porcentagens mínimas e máximas de integridade de uma atualização de instância só precisam ser especificadas se você não tiver definido uma política de manutenção da instância ou se precisar substituir a política existente. Para ter mais informações, consulte Políticas de manutenção de instância.
> Da mesma forma, você só precisa especificar o período de aquecimento da instância para uma atualização da instância se não tiver ativado o aquecimento padrão ou se precisar substituir o padrão. Para ter mais informações, consulte Definir o aquecimento padrão da instância para um grupo do Auto Scaling.

Conceitos principais

Antes de começar, familiarize-se com os seguintes conceitos básicos de atualização de instância:

Percentual mínimo de integridade

A porcentagem mínima de integridade é a porcentagem da capacidade desejada para se manter em serviço, íntegra e pronta para uso durante a atualização de uma instância para que a atualização possa continuar. Por exemplo, se a porcentagem mínima de integridade for 90% e a porcentagem máxima de integridade for 100%, 10% da capacidade será substituída por vez. Se

as novas instâncias não passarem nas verificações de integridade, o Amazon EC2 Auto Scaling as encerrará e substituirá. Se a atualização da instância não puder iniciar nenhuma instância íntegra, ela eventualmente falhará, deixando os outros 90% do grupo intactos. Se as novas instâncias permanecerem saudáveis e concluírem seu 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 por vez, defina a porcentagem mínima e máxima de integridade como 100%. Isso altera o comportamento de uma atualização de instância para ser iniciada antes do encerramento, o que evita que a capacidade do grupo fique abaixo de 100% da capacidade desejada. Para substituir todas as instâncias de uma vez, defina uma porcentagem mínima de integridade de 0%.

Porcentagem máxima de integridade

A porcentagem máxima íntegra é a porcentagem da capacidade desejada que seu grupo do Auto Scaling pode aumentar ao substituir instâncias. A diferença entre o mínimo e o máximo não pode ser maior que 100. Um intervalo maior aumenta o número de instâncias que podem ser substituídas ao mesmo tempo.

Aquecimento da instância

O aquecimento da instância é o período de tempo desde a mudança do estado de uma nova instância até o momento em que a `InService` inicialização é considerada concluída. Durante uma atualização de instância, se as instâncias passam na verificação de integridade, o Amazon EC2 Auto Scaling não avança imediatamente para substituir a próxima instância após determinar que uma instância recém-iniciada está íntegra. Ele aguarda o período de aquecimento antes de passar a substituir a próxima instância. Isso pode ser útil quando o aplicativo ainda precisar de um tempo de inicialização antes de responder às solicitações.

O aquecimento da instância funciona da mesma forma que o aquecimento de instâncias padrão. Portanto, as mesmas considerações de escalabilidade são aplicadas. Para ter mais informações, consulte Definir o aquecimento padrão da instância para um grupo do Auto Scaling.

Configuração desejada

A configuração desejada é a nova configuração que você deseja que o Amazon EC2 Auto Scaling implante no grupo do Auto Scaling. Por exemplo, você pode especificar um novo modelo de execução e novos tipos de instância para suas instâncias. Durante uma atualização de instância, o Amazon EC2 Auto Scaling atualiza o grupo do Auto Scaling para a configuração desejada. Se um evento aumento da escala na horizontal ocorrer durante uma atualização de

instância, o Amazon EC2 Auto Scaling iniciará novas instâncias com a configuração desejada em vez das configurações atuais do grupo. Depois que a atualização de instância tem êxito, o Amazon EC2 Auto Scaling atualiza as configurações do grupo do Auto Scaling para refletir a nova configuração desejada que você especificou como parte da atualização de instância.

Ignorar correspondência

Ignorar a correspondência diz ao Amazon EC2 Auto Scaling para ignorar as instâncias que já tenham as atualizações mais recentes. Assim, você não substituirá mais instâncias do que o necessário. Isso é útil quando você deseja garantir que o grupo do Auto Scaling usará uma versão específica de seu modelo de execução e substituirá apenas as instâncias que usam outra versão.

Pontos de verificação

Um ponto de verificação é um ponto no tempo em que a atualização de instância é interrompida por um período especificado. Uma atualização de instância pode conter vários pontos de verificação. O Amazon EC2 Auto Scaling emite eventos para cada ponto de verificação. Portanto, você pode adicionar uma EventBridge regra para enviar os eventos para um destino, como o Amazon SNS, para ser notificado quando um ponto de verificação for alcançado. Depois que um ponto de verificação é atingido, você tem a oportunidade de verificar sua implantação. Se algum problema for identificado, você poderá cancelar a atualização de instância ou revertê-la. A capacidade de implantar atualizações em fases é um benefício fundamental dos pontos de verificação. Se você não usar pontos de verificação, as substituições continuas serão executadas ininterruptamente.

Para saber mais sobre todas as configurações padrão que você pode definir ao iniciar uma atualização de instância, consulte Entender os valores padrão de uma atualização de instância.

Período de carência da verificação de integridade

O Amazon EC2 Auto Scaling determina se a instância está íntegra com base no status das verificações de integridade que o grupo do Auto Scaling usa. Para ter mais informações, consulte Verificações de integridade para instâncias em um grupo do Auto Scaling.

Para garantir que essas verificações de integridade comecem o mais rápido possível, não defina um período de carência da verificação de integridade do grupo muito alto, mas alto o suficiente para que suas verificações de integridade do Elastic Load Balancing consigam determinar se um destino está disponível para lidar com solicitações. Para ter mais informações, consulte Definir um período de carência da verificação de integridade para um grupo do Auto Scaling.

Compatibilidade de tipo de instância

Antes de alterar o tipo de instância, convém verificar se ela funciona com seu modelo de execução. Isso confirma a compatibilidade com a AMI especificada. Por exemplo, digamos que você iniciou suas instâncias originais com base em uma AMI paravirtual (PV), mas deseja alterar para um tipo de instância da geração atual que tenha suporte apenas em uma AMI de máquina virtual (HVM). Nesse caso, é necessário usar uma AMI HVM no modelo de execução.

Para confirmar a compatibilidade do tipo de instância sem iniciar instâncias, use o comando run-instances com a opção `--dry-run`, conforme mostrado no exemplo a seguir.

```
aws ec2 run-instances --launch-template LaunchTemplateName=my-template,Version='1' --
dry-run
```

Para obter informações sobre como a compatibilidade é determinada, consulte Compatibilidade para alterar o tipo de instância no Guia do usuário do Amazon EC2 para instâncias do Linux.

Limitações

- Duração total: o tempo máximo que uma atualização de instância pode permanecer ativamente substituindo instâncias é 14 dias.
- Diferença no comportamento específico de grupos ponderados: se um grupo de instâncias mistas estiver configurado com um peso de instância maior ou igual à capacidade desejada do grupo, o Amazon EC2 Auto Scaling poderá substituir todas as instâncias `InService` de uma só vez. Para evitar essa situação, siga a recomendação do tópico Configurar um grupo de Auto Scaling para usar pesos de instância. Especifique uma capacidade desejada que seja maior do que seu maior peso ao usar pesos com seu grupo do Auto Scaling.
- Tempo limite de uma hora: quando uma atualização de instância é incapaz de continuar fazendo substituições porque a aplicação está aguardando para substituir instâncias em espera ou protegidas contra a redução da escala 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.
- Implantação de código por meio de dados do usuário: Ignorar a correspondência não verifica as alterações de código implantadas a partir de um script de dados do usuário. Se você usa dados do usuário para extrair um novo código e instalar essas atualizações em novas instâncias, recomendamos que você desative a correspondência para garantir que todas as instâncias

recebam seu código mais recente, mesmo sem uma atualização da versão do modelo de lançamento.

Entender os valores padrão de uma atualização de instância

Antes de iniciar uma atualização de instância, é possível personalizar diversas preferências que afetam a atualização de instância. Alguns padrões de preferência são diferentes dependendo se você usa o console ou a linha de comando (AWS CLI ou AWS SDK).

A tabela a seguir lista os valores padrão das configurações de atualização de instância.

Configuração	AWS CLI ou AWS SDK	Console do Amazon EC2 Auto Scaling
CloudWatch alarme	Desativado (nulo)	Desabilitado
Reversão automática	Desabilitado (`false`)	Desabilitado
Pontos de verificação	Desabilitado (`false`)	Desabilitado
Atraso no ponto de verificação	1 hora (`3600` segundos)	1 hora
Aquecimento da instância	O aquecimento de instâncias padrão, se estiver definido, ou o período de carência da verificação de integridade, se não estiver.	O aquecimento de instâncias padrão, se estiver definido, ou o período de carência da verificação de integridade, se não estiver.
Porcentagem máxima de integridade	Varia com base em sua política de manutenção de instâncias. Se não houver política de manutenção de instâncias, o padrão é 100% (nulo).	Varia com base em sua política de manutenção de instâncias. Se não houver política de manutenção de instâncias, o padrão é 100% (nulo).
Percentual mínimo de integridade	Varia com base em sua política de manutenção de instâncias. Se não houver	Varia com base em sua política de manutenção de instâncias. Se não houver

Configuração	AWS CLI ou AWS SDK	Console do Amazon EC2 Auto Scaling
	política de manutenção de instâncias, o padrão é 90%.	política de manutenção de instâncias, o padrão é 90%.
Instâncias protegidas contra redução da escala na horizontal	Aguardar	Ignorar
Ignorar correspondência	Desabilitado (`false`)	Habilitado
Instâncias em espera	Aguardar	Ignorar

Segue uma descrição de cada configuração:

CloudWatch alarme (**`AlarmSpecification`**)

A especificação do CloudWatch alarme. CloudWatch os alarmes podem ser usados para identificar quaisquer problemas e falhar na operação se um alarme entrar no `ALARM` estado. Para ter mais informações, consulte Iniciar uma atualização de instância com reversão automática.

Reversão automática (**`AutoRollback`**)

Controla se o Amazon EC2 Auto Scaling reverte o grupo do Auto Scaling para sua configuração anterior se a atualização da instância falhar. Para ter mais informações, consulte Desfazer alterações com uma reversão.

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 ter mais informações, consulte Adicionar pontos de verificação a uma atualização de instância.

Atraso no ponto de verificação (**`CheckpointDelay`**)

A quantidade de tempo, em segundos, para aguardar após um ponto de verificação antes de continuar. Para ter mais informações, consulte Adicionar pontos de verificação a uma atualização de instância.

Aquecimento da instância (**`InstanceWarmup`**)

Um período, em segundos, durante o qual o Amazon EC2 Auto Scaling espera até que uma nova instância seja considerada como inicialização concluída antes de substituir a próxima instância. Se você já definiu corretamente um aquecimento de instâncias padrão para o grupo do Auto Scaling, não é necessário alterar o aquecimento da instância (a menos que deseje substituir o padrão). Para ter mais informações, consulte Definir o aquecimento padrão da instância para um grupo do Auto Scaling.

Porcentagem máxima de integridade (**`MaxHealthyPercentage`**)

A porcentagem da capacidade desejada do grupo do Auto Scaling que seu grupo pode aumentar ao substituir instâncias.

Percentual mínimo de integridade (**`MinHealthyPercentage`**)

A porcentagem da capacidade desejada do grupo do Auto Scaling que deve estar em serviço, íntegra e pronta para uso antes que a operação possa continuar.

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

Controla o que o Amazon EC2 Auto Scaling faz se forem encontradas instâncias protegidas contra redução de escala. Para obter mais informações sobre essas instâncias, consulte Usar proteção de redução na escala na horizontal de instâncias.

O Amazon EC2 Auto Scaling fornece estas opções:

- Replace (`Refresh`) — Substitui as instâncias que estão protegidas da escalabilidade.
- Ignore (`Ignore`) — Ignora as instâncias que estão protegidas da escalabilidade e continua substituindo as instâncias que não estão protegidas.
- Espere (`Wait`) — Espera uma hora para que você remova a proteção Scale-In. 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 ter mais informações, consulte Usar uma atualização de instância com opção de ignorar correspondência.

Instâncias em espera (**`StandbyInstances`**)

Controla o que o Amazon EC2 Auto Scaling faz se as instâncias forem encontradas no estado`Standby`. Para obter mais informações sobre essas instâncias, consulte Remover temporiamente instâncias do grupo do Auto Scaling.

O Amazon EC2 Auto Scaling fornece estas opções:

- Terminate (`Terminate`) — Encerra as instâncias que estão em. `Standby`
- Ignore (`Ignore`) — Ignora as instâncias que estão dentro `Standby` e continua substituindo as instâncias que estão no `InService` estado.
- Wait (**`Wait`**) — Espera uma hora para que você retorne as instâncias ao serviço. Se você não fizer isso, a atualização de instância falhará.

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 ter mais informações, consulte Desfazer alterações com uma reversão.

Os procedimentos a seguir ajudam você a iniciar uma atualização de instância usando o AWS Management Console ou AWS CLI.

Iniciar uma atualização de instância (console)

Se esta for a primeira vez que inicia uma atualização de instância, fazer isso usando o console ajudará você a entender os recursos e as opções disponíveis.

Iniciar uma atualização de instância no console (procedimento básico)

Use o procedimento a seguir se você não tiver definido anteriormente uma política de instâncias mistas para seu grupo do Auto Scaling. Se você já definiu uma política de instâncias mistas, consulte Iniciar uma atualização de instância no console (grupo de instâncias mistas) para iniciar uma atualização de instância.

Para iniciar uma atualização de instância

1. Abra o console do Amazon EC2 em https://console.aws.amazon.com/ec2/ e escolha 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 Atualização de instância, em Atualização ativa de instância, escolha Iniciar atualização de instância.
4. Para configurações de disponibilidade, faça o seguinte:

   a. Para o método de substituição de instância:

      - Se você não definiu uma política de manutenção de instâncias no grupo do Auto Scaling, a configuração padrão para o método de substituição de instância é Encerrar e iniciar. Esse é o comportamento padrão legado de uma atualização de instância.
      - Se você definir uma política de manutenção de instância no grupo do Auto Scaling, ela fornecerá valores padrão para o método de substituição de instância. Para substituir a política de manutenção da instância, escolha Substituir. A substituição é aplicada somente à atualização de instância atual. Na próxima vez que você iniciar uma atualização de instância, esses valores serão redefinidos para os padrões da política de manutenção de instâncias.

      O procedimento a seguir explica como atualizar o método de substituição de instância.

      i. Escolha um dos seguintes métodos de substituição de instância:

         - Iniciar antes de encerrar: uma nova instância deve ser provisionada primeiro antes que uma instância existente possa ser encerrada. Essa é uma boa opção para aplicativos que favorecem a disponibilidade em detrimento da redução de custos.
         - Encerrar e executar: novas instâncias são provisionadas ao mesmo tempo em que as instâncias existentes são encerradas. Esta é uma boa opção para aplicações que favorecem a economia de custos em detrimento da disponibilidade. Também é uma boa opção para aplicativos que não devem lançar mais capacidade do que a disponível atualmente.

- Comportamento personalizado: esta opção permite configurar um intervalo mínimo e máximo personalizado para a quantidade de capacidade que você deseja disponibilizar ao substituir instâncias. Isso pode ajudá-lo a alcançar o equilíbrio certo entre custo e disponibilidade.

ii. Em Definir porcentagem de integridade, insira valores para um ou ambos os campos a seguir. Os campos de ativação variam de acordo com a opção escolhida para o método de substituição de instância.

- Mínimo: define a porcentagem mínima de integridade necessária para continuar com a atualização de instâncias.
- Máximo: Define a porcentagem máxima íntegra possível durante a atualização da instância.

iii. Expanda a seção Exibir capacidade temporária estimada durante substituições com base no tamanho atual do grupo para confirmar como os valores de Mínimo e Máximo se aplicam ao seu grupo. Os valores exatos usados dependem do valor de capacidade desejado, que mudará se o grupo for ampliado.

iv. Expanda a seção Definir comportamento alternativo para tamanhos de reposição inválidos e, em seguida, escolha se deseja violar a porcentagem máxima de integridade para priorizar a disponibilidade ou violar a porcentagem mínima de integridade.

Manter a opção padrão de Violar porcentagem mínima de integridade não é recomendado para grupos muito pequenos. Se houver apenas uma instância no grupo do Auto Scaling, iniciar uma atualização de instância poderá resultar em uma interrupção.

Essa etapa configura o comportamento de fallback se você estiver usando um grupo do Auto Scaling que ainda não tem uma política de manutenção de instâncias. Essa opção não está disponível e não aparece quando seu grupo tem uma política de manutenção de instâncias. Essa opção também está disponível somente para o método de substituição Encerrar e iniciar. Outros métodos de substituição violarão a porcentagem máxima de integridade para priorizar a disponibilidade.

b. Em Aquecimento da instância, insira o número de segundos desde a mudança do estado de uma nova instância até o `InService` término da inicialização. O Amazon EC2 Auto Scaling aguarda esse tempo antes de substituir a próxima instância.

Durante o aquecimento, instâncias recém-iniciadas também não são contabilizadas nas métricas agregadas do grupo do Auto Scaling (como `CPUUtilization`, `NetworkIn`, `NetworkOut` etc.). Se você adicionou políticas de escalabilidade ao grupo do Auto Scaling, as ações de escalabilidade serão executadas em paralelo. Se você definir um intervalo longo para o período de aquecimento da atualização da instância, levará mais tempo para que as instâncias recém-lançadas apareçam nas métricas. Portanto, um período de aquecimento adequado impede que o Amazon EC2 Auto Scaling escale com base em dados métricos obsoletos.

Se você já definiu corretamente um aquecimento de instâncias padrão para o grupo do Auto Scaling, não é necessário alterar o aquecimento da instância. Porém, se quiser substituir o padrão, você pode definir um valor para essa opção. Para obter mais informações sobre como configurar o aquecimento de instâncias, consulte Definir o aquecimento padrão da instância para um grupo do Auto Scaling.

5. Para configurações de atualização, faça o seguinte:

   a. (Opcional) Em Pontos de verificação, escolha Habilitar pontos de verificação para substituir instâncias usando uma abordagem incremental ou faseada para uma atualização de instância. Isso fornece tempo adicional para verificação entre conjuntos de substituições. Se você optar por não ativar pontos de verificação, as instâncias serão substituídas em uma operação quase contínua.

      Se você habilitar pontos de verificação, consulte Habilitar pontos de verificação (console) para obter etapas adicionais.

   b. Habilitar ou desativar Ignorar correspondência :

      - Para ignorar a substituição de instâncias que já correspondem ao modelo de execução, mantenha a caixa de seleção Habilitar opção de ignorar correspondência marcada.
      - Se você desativar ignorar correspondência desmarcando essa caixa de seleção, todas as instâncias poderão ser substituídas.

      Ao ativar a correspondência ignorada, você pode definir um novo modelo de execução ou uma nova versão do modelo de execução em vez de usar o existente. Faça isso na seção Configuração desejada da página Iniciar atualização de instância.

> **Note**
>
> Para usar o recurso de ignorar correspondência para atualizar um grupo do Auto Scaling que atualmente use uma configuração de execução, é necessário selecionar um modelo de execução em Configuração desejada. Ignorar correspondência com uma configuração de inicialização não é compatível.

c. Em Instâncias em espera, escolha Ignorar, Terminar ou Aguardar. Isso determina o que acontecerá se as instâncias forem encontradas no estado `Standby`. Para ter mais informações, consulte Remover temporiamente instâncias do grupo do Auto Scaling.

   Se você escolher Aguardar, deverá realizar outras ações para retornar essas instâncias ao serviço. Senão, a atualização de instância substituirá todas as instâncias `InService` e aguardará uma hora. Então, se alguma instância `Standby` permanecer, a atualização de instância falhará. Para evitar essa situação, escolha Ignorar ou Terminar as instâncias.

d. Para Instâncias protegidas de redução da escala na horizontal, escolha Ignorar, Substituir ou Aguardar. Isso determina o que acontecerá se instâncias protegidas contra redução da escala na horizontal forem encontradas. Para ter mais informações, consulte Usar proteção de redução na escala na horizontal de instâncias.

   Se você escolher Aguardar, deverá realizar outras ações para remover a proteção contra redução da escala na horizontal dessas instâncias. Senão, a atualização de instância substituirá todas as instâncias não protegidas e aguardará uma hora. Então, se alguma instância protegida contra redução da escala na horizontal permanecer, a atualização de instância falhará. Para evitar essa situação, escolha Ignorar ou Substituir as instâncias.

6. (Opcional) Para CloudWatch alarme, escolha Ativar CloudWatch alarmes e, em seguida, escolha um ou mais alarmes. CloudWatch os alarmes podem ser usados para identificar quaisquer problemas e falhar na operação se um alarme entrar no `ALARM` estado. Para ter mais informações, consulte Iniciar uma atualização de instância com reversão automática.

7. (Opcional) Expanda a seção Configuração desejada para especificar as atualizações que você deseja fazer no grupo do Auto Scaling.

   Nesta etapa, você pode optar por usar a sintaxe JSON ou YAML para editar valores de parâmetros em vez de fazer seleções na interface do console. Para isso, escolha Usar editor de código em vez de Usar a interface do console. O procedimento a seguir explica como fazer seleções usando a interface do console.

a. Para Atualizar o modelo de execução:

- Se você não criou um novo modelo de execução ou uma nova versão de modelo de execução para seu grupo do Auto Scaling, não marque essa caixa de seleção.
- Se você criou um novo modelo de execução ou uma nova versão do modelo de execução, marque esta caixa de seleção. Quando você seleciona essa opção, o Amazon EC2 Auto Scaling exibe o modelo de execução atual e a versão atual do modelo de execução. Também lista todas as outras versões disponíveis. Escolha o modelo de lançamento e, em seguida, escolha a versão.

  Após escolher uma versão, você poderá visualizar as informações da versão. Esta é a versão do modelo de execução que será usada ao substituir instâncias como parte de uma atualização de instância. Se a atualização da instância tiver êxito, essa versão do modelo de execução também será usada sempre que novas instâncias forem iniciadas, como quando o grupo for dimensionado.

b. Em Choose a set of instance types and purchase options to override the instance type in the launch template (Escolha um conjunto de tipos de instância e opções de compra para substituir o tipo de instância no modelo de execução):

- Não marque essa caixa de seleção se quiser usar o tipo de instância e a opção de compra que você especificou no modelo de execução.
- Marque esta caixa de seleção se quiser substituir o tipo de instância no modelo de execução ou executar instâncias spot. É possível adicionar manualmente cada tipo de instância ou escolher um tipo de instância primária e uma opção de recomendação que recupere outros tipos de instância correspondentes para você. Se você pretende iniciar instâncias spot, recomendamos adicionar alguns tipos diferentes de instância. Dessa forma, o Amazon EC2 Auto Scaling pode executar outro tipo de instância se houver capacidade de instância insuficiente nas zonas de disponibilidade escolhidas. Para ter mais informações, consulte Grupos do Auto Scaling com vários tipos de instâncias e opções de compra.

  > ⚠ Warning
  >
  > Não use instâncias spot com aplicações que não conseguem lidar com uma interrupção de instância spot. As interrupções poderão ocorrer se o serviço do Amazon EC2 Spot precisar recuperar a capacidade.

Se você marcar essa caixa de seleção, verifique se o modelo de execução já não solicita instâncias spot. Não é possível usar um modelo de execução que solicite instâncias spot para criar um grupo do Auto Scaling que use vários tipos de instância e execute instâncias spot e sob demanda.

> **Note**
>
> Para configurar essas opções em um grupo do Auto Scaling que atualmente use uma configuração de execução, é necessário selecionar um modelo de execução em Update launch template (Atualizar modelo de execução). Não há suporte à substituição do tipo de instância na configuração de execução.

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

   Essa configuração só pode ser habilitada quando o grupo do Auto Scaling atende aos pré-requisitos para usar reversões.

   Para ter mais informações, consulte Desfazer alterações com uma reversão.
9. Revise todas as seleções para confirmar que tudo esteja configurado corretamente.

   Nesse ponto, é bom verificar se as diferenças entre as alterações atuais e propostas não afetarão sua aplicação de maneiras inesperadas ou indesejadas. Para confirmar se o tipo de instância é compatível com o modelo de execução, consulte Compatibilidade de tipo de instância.
10. Quando estiver satisfeito com as seleções de atualização da instância, escolha Iniciar atualização da instância.

Iniciar uma atualização de instância no console (grupo de instâncias mistas)

Use o procedimento a seguir se você criou um grupo do Auto Scaling com política de instâncias mistas. Se você não definiu ainda uma política de instâncias mistas para seu grupo, consulte Iniciar uma atualização de instância no console (procedimento básico) para iniciar uma atualização de instância.

Para iniciar uma atualização de instância

1. Abra o console do Amazon EC2 em https://console.aws.amazon.com/ec2/ e escolha 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 Atualização de instância, em Atualização ativa de instância, escolha Iniciar atualização de instância.
4. Para configurações de disponibilidade, faça o seguinte:

   a. Para o método de substituição de instância:

      - Se você não definiu uma política de manutenção de instâncias no grupo do Auto Scaling, a configuração padrão para o método de substituição de instância é Encerrar e iniciar. Esse é o comportamento padrão legado de uma atualização de instância.
      - Se você definir uma política de manutenção de instância no grupo do Auto Scaling, ela fornecerá valores padrão para o método de substituição de instância. Para substituir a política de manutenção da instância, escolha Substituir. A substituição é aplicada somente à atualização de instância atual. Na próxima vez que você iniciar uma atualização de instância, esses valores serão redefinidos para os padrões da política de manutenção de instâncias.

      O procedimento a seguir explica como atualizar o método de substituição de instância.

      i. Escolha um dos seguintes métodos de substituição de instância:

         - Iniciar antes de encerrar: uma nova instância deve ser provisionada primeiro antes que uma instância existente possa ser encerrada. Essa é uma boa opção para aplicativos que favorecem a disponibilidade em detrimento da redução de custos.
         - Encerrar e executar: novas instâncias são provisionadas ao mesmo tempo em que as instâncias existentes são encerradas. Esta é uma boa opção para aplicações que favorecem a economia de custos em detrimento da disponibilidade. Também é uma boa opção para aplicativos que não devem lançar mais capacidade do que a disponível atualmente.

- Comportamento personalizado: esta opção permite configurar um intervalo mínimo e máximo personalizado para a quantidade de capacidade que você deseja disponibilizar ao substituir instâncias. Isso pode ajudá-lo a alcançar o equilíbrio certo entre custo e disponibilidade.

ii. Em Definir porcentagem de integridade, insira valores para um ou ambos os campos a seguir. Os campos de ativação variam de acordo com a opção escolhida para o método de substituição de instância.

- Mínimo: define a porcentagem mínima de integridade necessária para continuar com a atualização de instâncias.
- Máximo: Define a porcentagem máxima íntegra possível durante a atualização da instância.

iii. Expanda a seção Exibir capacidade temporária estimada durante substituições com base no tamanho atual do grupo para confirmar como os valores de Mínimo e Máximo se aplicam ao seu grupo. Os valores exatos usados dependem do valor de capacidade desejado, que mudará se o grupo for ampliado.

iv. Expanda a seção Definir comportamento alternativo para tamanhos de reposição inválidos e, em seguida, escolha se deseja violar a porcentagem máxima de integridade para priorizar a disponibilidade ou violar a porcentagem mínima de integridade.

Manter a opção padrão de Violar porcentagem mínima de integridade não é recomendado para grupos muito pequenos. Se houver apenas uma instância no grupo do Auto Scaling, iniciar uma atualização de instância poderá resultar em uma interrupção.

Essa etapa configura o comportamento de fallback se você estiver usando um grupo do Auto Scaling que ainda não tem uma política de manutenção de instâncias. Essa opção não está disponível e não aparece quando seu grupo tem uma política de manutenção de instâncias. Essa opção também está disponível somente para o método de substituição Encerrar e iniciar. Outros métodos de substituição violarão a porcentagem máxima de integridade para priorizar a disponibilidade.

b. Em Aquecimento da instância, insira o número de segundos desde a mudança do estado de uma nova instância até o `InService` término da inicialização. O Amazon EC2 Auto Scaling aguarda esse tempo antes de substituir a próxima instância.

Durante o aquecimento, instâncias recém-iniciadas também não são contabilizadas nas métricas agregadas do grupo do Auto Scaling (como `CPUUtilization`, `NetworkIn`, `NetworkOut` etc.). Se você adicionou políticas de escalabilidade ao grupo do Auto Scaling, as ações de escalabilidade serão executadas em paralelo. Se você definir um intervalo longo para o período de aquecimento da atualização da instância, levará mais tempo para que as instâncias recém-lançadas apareçam nas métricas. Portanto, um período de aquecimento adequado impede que o Amazon EC2 Auto Scaling escale com base em dados métricos obsoletos.

Se você já definiu corretamente um aquecimento de instâncias padrão para o grupo do Auto Scaling, não é necessário alterar o aquecimento da instância. Porém, se quiser substituir o padrão, você pode definir um valor para essa opção. Para obter mais informações sobre como configurar o aquecimento de instâncias, consulte Definir o aquecimento padrão da instância para um grupo do Auto Scaling.

5. Para configurações de atualização, faça o seguinte:

   a. (Opcional) Em Pontos de verificação, escolha Habilitar pontos de verificação para substituir instâncias usando uma abordagem incremental ou faseada para uma atualização de instância. Isso fornece tempo adicional para verificação entre conjuntos de substituições. Se você optar por não ativar pontos de verificação, as instâncias serão substituídas em uma operação quase contínua.

      Se você habilitar pontos de verificação, consulte Habilitar pontos de verificação (console) para obter etapas adicionais.

   b. Habilitar ou desativar Ignorar correspondência:

      - Para ignorar a substituição de instâncias que já correspondem ao modelo de execução e quaisquer substituições de tipo de instância, mantenha a caixa de seleção Habilitar opção de ignorar correspondência marcada.
      - Se você optar por desativar ignorar correspondência desmarcando essa caixa de seleção, todas as instâncias poderão ser substituídas.

      Ao ativar a correspondência ignorada, você pode definir um novo modelo de execução ou uma nova versão do modelo de execução em vez de usar o existente. Faça isso na seção Configuração desejada da página Iniciar atualização de instância. Você também pode

atualizar suas substituições de tipo de instância em Desired configuration (Configuração desejada).

c. Em Instâncias em espera, escolha Ignorar, Terminar ou Aguardar. Isso determina o que acontecerá se as instâncias forem encontradas no estado `Standby`. Para ter mais informações, consulte Remover temporiamente instâncias do grupo do Auto Scaling.

   Se você escolher Aguardar, deverá realizar outras ações para retornar essas instâncias ao serviço. Do contrário, a atualização de instância substituirá todas as instâncias `InService` e aguardará uma hora. Então, se alguma instância `Standby` permanecer, a atualização de instância falhará. Para evitar essa situação, escolha Ignorar ou Terminar as instâncias.

d. Para Instâncias protegidas de redução da escala na horizontal, escolha Ignorar, Substituir ou Aguardar. Isso determina o que acontecerá se instâncias protegidas contra redução da escala na horizontal forem encontradas. Para ter mais informações, consulte Usar proteção de redução na escala na horizontal de instâncias.

   Se você escolher Aguardar, deverá realizar outras ações para remover a proteção contra redução da escala na horizontal dessas instâncias. Senão, a atualização de instância substituirá todas as instâncias não protegidas e aguardará uma hora. Então, se alguma instância protegida contra redução da escala na horizontal permanecer, a atualização de instância falhará. Para evitar essa situação, escolha Ignorar ou Substituir as instâncias.

6. (Opcional) Para CloudWatch alarme, escolha Ativar CloudWatch alarmes e, em seguida, escolha um ou mais alarmes. CloudWatch os alarmes podem ser usados para identificar quaisquer problemas e falhar na operação se um alarme entrar no `ALARM` estado. Para ter mais informações, consulte Iniciar uma atualização de instância com reversão automática.

7. Na seção Desired configuration (Configuração desejada), faça o seguinte:

   Nesta etapa, você pode optar por usar a sintaxe JSON ou YAML para editar valores de parâmetros em vez de fazer seleções na interface do console. Para isso, escolha Usar editor de código em vez de Usar a interface do console. O procedimento a seguir explica como fazer seleções usando a interface do console.

   a. Para Atualizar o modelo de execução:

      - Se você não criou um novo modelo de execução ou uma nova versão de modelo de execução para seu grupo do Auto Scaling, não marque essa caixa de seleção.
      - Se você criou um novo modelo de execução ou uma nova versão do modelo de execução, marque esta caixa de seleção. Quando você seleciona essa opção, o Amazon

EC2 Auto Scaling exibe o modelo de execução atual e a versão atual do modelo de execução. Também lista todas as outras versões disponíveis. Escolha o modelo de lançamento e, em seguida, escolha a versão.

Após escolher uma versão, você poderá visualizar as informações da versão. Esta é a versão do modelo de execução que será usada ao substituir instâncias como parte de uma atualização de instância. Se a atualização da instância tiver êxito, essa versão do modelo de execução também será usada sempre que novas instâncias forem iniciadas, como quando o grupo for dimensionado.

b. Em Use these settings to override the instance type and purchase option defined in the launch template (Use estas configurações para substituir o tipo de instância e a opção de compra definidas no modelo de execução):

Por padrão, esta caixa de seleção está marcada. O Amazon EC2 Auto Scaling preenche cada parâmetro com o valor que está atualmente definido na política de instâncias mistas para o grupo do Auto Scaling. Atualize somente os valores dos parâmetros que você deseja alterar. Para obter orientações sobre essas configurações, consulte Grupos do Auto Scaling com vários tipos de instâncias e opções de compra.

> **Warning**
>
> Recomendamos não desmarcar essa caixa de seleção. Apenas a desmarque se desejar parar de usar uma política de instâncias mistas. Após o término com êxito da atualização de instância, o Amazon EC2 Auto Scaling atualiza seu grupo para corresponder à Desired configuration (Configuração desejada). Se não incluir mais uma política de instâncias mistas, o Amazon EC2 Auto Scaling terminará gradualmente todas as instâncias spot que estejam em execução no momento e as substituirá por instâncias sob demanda. Ou, se seu modelo de execução solicitar instâncias spot, o Amazon EC2 Auto Scaling terminará gradualmente todas as instâncias sob demanda que estejam em execução no momento e as substituirá por instâncias spot.

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

Essa configuração só pode ser habilitada quando o grupo do Auto Scaling atende aos pré-requisitos para usar reversões.

Para ter mais informações, consulte Desfazer alterações com uma reversão.

9. Revise todas as seleções para confirmar que tudo esteja configurado corretamente.

   Nesse ponto, é bom verificar se as diferenças entre as alterações atuais e propostas não afetarão sua aplicação de maneiras inesperadas ou indesejadas. Para confirmar se o tipo de instância é compatível com o modelo de execução, consulte Compatibilidade de tipo de instância.

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

Iniciar uma atualização de instância (AWS CLI)

Para iniciar uma atualização de instância

Use o start-instance-refreshcomando a seguir para iniciar uma atualização de instância a AWS CLI partir do. Você pode especificar as preferências que deseja alterar em um arquivo de configuração JSON. Ao referenciar o arquivo de configuração, forneça o caminho e o nome do arquivo, conforme mostrado no exemplo a seguir.

```
aws autoscaling start-instance-refresh --cli-input-json file://config.json
```

Conteúdo de `config.json`:

```
{
    "AutoScalingGroupName": "my-asg",
    "Preferences": {
      "InstanceWarmup": 60,
      "MinHealthyPercentage": 50,
      "AutoRollback": true,
      "ScaleInProtectedInstances": Ignore,
      "StandbyInstances": Terminate
    }
}
```

Se as preferências não forem fornecidas, serão usados os valores padrão. Para ter mais informações, consulte Entender os valores padrão de uma atualização de instância.

Resultado do exemplo:

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

Monitore uma atualização de instância

Você pode monitorar uma atualização de instância em andamento ou consultar o status de atualizações de instâncias anteriores nas últimas seis semanas usando o ou. AWS Management Console AWS CLI

Monitore e verifique o status de uma atualização de instância

Para monitorar e verificar o status de uma atualização de instância, use um dos seguintes métodos:

Console

> **Tip**
>
> Nesse procedimento, as colunas nomeadas já devem ser exibidas. Para exibir colunas ocultas ou alterar o número de linhas mostradas, escolha o ícone de engrenagem no canto superior direito da seção para abrir o modal de preferências. Atualize as configurações, conforme necessário, e escolha Confirmar.

Para monitorar e 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.
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 em `Pending` status enquanto está sendo inicializada. Depois, o status deve mudar rapidamente para `InProgress`. Quando todas as instâncias estão atualizadas, o status muda para `Successful`.
4. Você pode monitorar ainda mais o sucesso ou o fracasso das atividades em andamento visualizando as atividades de escalonamento do grupo. Na guia Atividade, em Histórico de

atividades, quando a atualização da instância for iniciada, você observará entradas quando instâncias forem encerradas e outro conjunto de entradas quando instâncias forem iniciadas. Se você tiver várias atividades de escalonamento, poderá ver mais delas escolhendo o ícone > na parte superior do histórico de atividades. Para obter informações sobre a solução de problemas que podem causar falhas nas atividades, consulteSolucionar problemas do Amazon EC2 Auto Scaling.

5. (Opcional) Na guia Gerenciamento de instâncias, em Instâncias, você pode analisar o progresso de instâncias específicas conforme necessário.

AWS CLI

Para monitorar e verificar o status de uma instância refresh ()AWS CLI

Use o seguinte comando describe-instance-refreshes:

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

A seguir, um exemplo de saída.

As atualizações de instâncias são ordenadas pelo horário de início. As atualizações de instâncias ainda em andamento são descritas primeiro.


```
{
  "InstanceRefreshes":[
    {
      "InstanceRefreshId":"08b91cf7-8fa6-48af-b6a6-d227f40f1b9b",
      "AutoScalingGroupName":"my-asg",
      "Status":"InProgress",
      "StatusReason":"Waiting for instances to warm up before continuing. For
 example: i-0645704820a8e83ff is warming up.",
      "StartTime":"2023-11-24T16:46:52+00:00",
      "PercentageComplete":50,
      "InstancesToUpdate":0,
      "Preferences":{
        "MaxHealthyPercentage":120,
        "MinHealthyPercentage":90,
        "InstanceWarmup":60,
        "SkipMatching":false,
        "AutoRollback":true,
        "ScaleInProtectedInstances":"Ignore",
```

```
        "StandbyInstances":"Ignore"
      }
    },
    {
      "InstanceRefreshId":"0e151305-1e57-4a32-a256-1fd14157c5ec",
      "AutoScalingGroupName":"my-asg",
      "Status":"Successful",
      "StartTime":"2023-11-22T13:53:37+00:00",
      "EndTime":"2023-11-22T13:59:45+00:00",
      "PercentageComplete":100,
      "InstancesToUpdate":0,
      "Preferences":{
        "MaxHealthyPercentage":120,
        "MinHealthyPercentage":90,
        "InstanceWarmup":60,
        "SkipMatching":false,
        "AutoRollback":true,
        "ScaleInProtectedInstances":"Ignore",
        "StandbyInstances":"Ignore"
      }
    }
  ]
}
```


Você pode monitorar ainda mais o sucesso ou o fracasso das atividades em andamento visualizando as atividades de escalonamento do grupo. As atividades de escalabilidade também ajudam você a se aprofundar para obter mais detalhes para ajudar a solucionar problemas com a atualização de uma instância. Para ter mais informações, consulte Solucionar problemas do Amazon EC2 Auto Scaling.

Status de atualização de instância

Quando uma atualização de instância é iniciada, ela entra no status Pending. Ele passa de Pendente para InProgressaté atingir Sucesso, Falha, Cancelado ou RollbackFailed. RollbackSuccessful

A atualização de instância pode ter os seguintes status:

Status	Descrição
Pendente	A solicitação foi criada, mas a atualização de instância não foi iniciada.

Status	Descrição
InProgress	Uma atualização de instância está em andamento.
Com êxito	Uma atualização de instância foi concluída com êxito.
Com falha	Falha ao concluir uma atualização de instância. É possível solucionar problemas usando o motivo do status e as ações de escalabilidade.
Cancelando	Uma atualização de instância em andamento está sendo cancelada.
Cancelado	A atualização de instância foi cancelada.
RollbackInProgress	Uma atualização de instância está sendo revertida.
RollbackFailed	Falha ao concluir a reversão. É possível solucionar problemas usando o motivo do status e as ações de escalabilidade.
RollbackSuccessful	A reversão foi concluída com êxito.

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 ter mais informações, consulte Desfazer alterações com uma reversão.

Tópicos

- Cancelar uma atualização de instância (console)
- Cancelar uma atualização de instância (AWS CLI)

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 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-refreshcomando do AWS CLI e forneça o nome do grupo Auto Scaling.

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

Resultado do exemplo:

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

Desfazer 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 CloudWatch se algum alarme que você especificar entrar no estado. `ALARM`

Conteúdo

- Considerações
- Iniciar uma reversão manualmente
- Iniciar uma atualização de instância com reversão automática

Considerações

As seguintes considerações se aplicam ao usar uma reversão:

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

Iniciar uma reversão manualmente

Console

Para iniciar manualmente a reversão de uma atualização de instância (console)

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

AWS CLI

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

Use o rollback-instance-refreshcomando do AWS CLI e forneça o nome do grupo Auto Scaling.

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

Resultado do exemplo:

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

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

Iniciar uma atualização de instância com reversão automática

Usando o recurso de reversão automática, você pode reverter automaticamente a atualização da instância quando ela falhar, como quando há erros ou quando um CloudWatch alarme específico da Amazon entra no estado. `ALARM`

Se você ativar a reversão automática e houver erros ao substituir as instâncias, a atualização da instância tentará concluir todas as substituições por uma hora antes de falhar e reverter. Esses erros geralmente são causados por coisas como falhas de inicialização do EC2, verificações de integridade mal configuradas ou por não ignorar ou permitir o encerramento de instâncias que estão no estado `Standby` ou protegidas contra redução.

A especificação de CloudWatch alarmes é opcional. Para especificar um alarme, primeiro você precisa criá-lo. Você pode especificar alarmes métricos e alarmes compostos. Para obter informações sobre como criar o alarme, consulte o Guia CloudWatch do usuário da Amazon. Usando as métricas do Elastic Load Balancing como exemplo, se você usar um Application Load Balancer, poderá usar as métricas `HTTPCode_ELB_5XX_Count` e`HTTPCode_ELB_4XX_Count`.

Considerações

- Se você especificar um CloudWatch alarme, mas não ativar a reversão automática, e o estado do alarme continuar`ALARM`, a atualização da instância falhará sem reverter.
- Você pode escolher no máximo 10 alarmes ao iniciar uma atualização da instância.
- Ao escolher um CloudWatch alarme, o alarme deve estar em um estado compatível. Se o estado do alarme for `INSUFFICIENT_DATA` ou`ALARM`, você receberá um erro ao tentar iniciar a atualização da instância.
- Ao criar um alarme para uso do Amazon EC2 Auto Scaling, o alarme deve incluir como tratar pontos de dados ausentes. Se uma métrica tiver frequentemente pontos de dados ausentes por projeto, o estado do alarme será `INSUFFICIENT_DATA` durante esses períodos. Quando isso acontece, o Amazon EC2 Auto Scaling não pode substituir instâncias até que novos pontos de dados sejam encontrados. Para forçar o alarme a manter o estado `ALARM` ou `OK` anterior, você pode optar por ignorar os dados ausentes. Para obter mais informações, consulte Configurando como os alarmes tratam os dados perdidos no Guia CloudWatch do usuário da Amazon.

Console

Para iniciar uma atualização de instância com reversão automática (console)

1. Abra o console do Amazon EC2 em https://console.aws.amazon.com/ec2/ e escolha 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 ativa de instância, escolha Iniciar atualização de instância.
4. Siga o Iniciar uma atualização de instância (console) procedimento e defina as configurações de atualização da instância conforme necessário.
5. (Opcional) Em Atualizar configurações, para CloudWatch alarme, escolha Ativar CloudWatch alarmes e, em seguida, escolha um ou mais alarmes para identificar quaisquer problemas e falhar na operação se um alarme entrar nesse estado. `ALARM`
6. Em Configurações de reversão, escolha Habilitar a reversão automática para reverter automaticamente uma atualização de instância com falha para a configuração que você salvou pela última vez no grupo do Auto Scaling antes de iniciar a atualização de instância.
7. Revise suas seleções e escolha Iniciar atualização da instância.

AWS CLI

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

Use o start-instance-refreshcomando e especifique `true` a `AutoRollback` opção no`Preferences`.

O exemplo a seguir mostra como iniciar uma atualização de instância que será revertida automaticamente se ocorrer uma falha. Substitua os valores dos *`italicized`* parâmetros pelos seus próprios.

```
aws autoscaling start-instance-refresh --cli-input-json file://config.json
```

Conteúdo de `config.json`.

```
{
    "AutoScalingGroupName": "my-asg",
    "DesiredConfiguration": {
      "LaunchTemplate": {
          "LaunchTemplateName": "my-launch-template",
          "Version": "1"
       }
    },
    "Preferences": {
      "AutoRollback": true
    }
}
```

Como alternativa, para reverter automaticamente quando a atualização da instância falhar ou quando um CloudWatch alarme especificado estiver no `ALARM` estado, especifique a `AlarmSpecification` opção no `Preferences` e forneça o nome do alarme, como no exemplo a seguir. Substitua os valores dos *`italicized`* parâmetros pelos seus próprios.

```
{
    "AutoScalingGroupName": "my-asg",
    "DesiredConfiguration": {
      "LaunchTemplate": {
          "LaunchTemplateName": "my-launch-template",
          "Version": "1"
       }
    },
    "Preferences": {
```

```
      "AutoRollback": true,
      "AlarmSpecification": { "Alarms": [ "my-alarm" ] }
    }
}
```

Se houver êxito, o comando gerará uma saída semelhante à seguinte.

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

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

Usar uma atualização de instância com opção de ignorar correspondência

Ignorar a correspondência diz ao Amazon EC2 Auto Scaling para ignorar as instâncias que já tenham as atualizações mais recentes. Assim, você não substituirá mais instâncias do que o necessário. Isso é útil quando você deseja garantir que seu grupo do Auto Scaling usará uma versão específica de seu modelo de execução e substituirá apenas as instâncias que usam outra versão.

As seguintes considerações se aplicam à opção ignorar a correspondência:

- Se você iniciar uma atualização de instância com a opção de ignorar correspondência e uma configuração desejada, o Amazon EC2 Auto Scaling verificará se alguma instância corresponde à configuração desejada. Em seguida, ele substituirá apenas as instâncias que não correspondam à configuração desejada. Depois que a atualização de instância tem êxito, o Amazon EC2 Auto Scaling atualiza o grupo para refletir a configuração desejada.
- Se você iniciar uma atualização de instância com opção de ignorar correspondência, mas não especificar a configuração desejada, o Amazon EC2 Auto Scaling verificará se alguma instância corresponde à configuração que você salvou pela última vez no grupo do Auto Scaling. Em seguida, ele substituirá apenas as instâncias que não correspondam à última configuração salva.
- Você pode usar a opção de ignorar correspondência com um novo modelo de execução, uma nova versão do modelo de execução ou um conjunto de tipos de instância. Se você habilitar ignorar a correspondência, mas nenhum deles for alterado, a atualização da instância será bem-

sucedida imediatamente sem substituir nenhuma instância. Se você tiver feito outras alterações na configuração desejada (como alterar a estratégia de alocação spot), o Amazon EC2 Auto Scaling aguardará a atualização de instância ser concluída com êxito. Em seguida, ele atualizará as configurações do grupo do Auto Scaling para refletir a nova configuração desejada.

- Você não pode usar ignorar a correspondência com uma nova configuração de inicialização.
- Quando você inicia uma atualização de instância e fornece a configuração desejada, o Amazon EC2 Auto Scaling garante que todas as instâncias usem a configuração desejada. Portanto, quando você especifica uma `$Default` ou `$Latest` como a versão desejada para seu modelo de execução e, em seguida, cria uma nova versão do modelo de execução enquanto a atualização da instância está em andamento, todas as instâncias que já foram substituídas serão substituídas novamente.
- O Skip Matching não sabe se um script de dados do usuário no modelo de lançamento extrairá o código atualizado e o instalará em novas instâncias. Como resultado, ignorar a correspondência pode ignorar a substituição de instâncias com código desatualizado instalado. Nesse caso, você deve desativar o skip matching para garantir que todas as instâncias recebam seu código mais recente, mesmo sem uma atualização da versão do modelo de lançamento.

Esta seção inclui AWS CLI instruções para iniciar uma atualização de instância com a opção skip matching ativada. Para obter instruções sobre como usar o console, consulte Iniciar uma atualização de instância (console).

Ignorar correspondência (procedimento básico)

Siga as etapas desta seção para usar o AWS CLI para fazer o seguinte:

- Crie o modelo de execução que deseja aplicar às instâncias.
- Inicie uma atualização de instância para aplicar seu modelo de execução ao grupo do Auto Scaling. Se você não habilitar a opção de ignorar correspondência, todas as instâncias serão substituídas. Isso ocorre mesmo que o modelo de execução usado para provisionar a instância seja o mesmo que você especificou para a configuração desejada.

Para usar a opção de ignorar correspondência com um novo modelo de execução

1. Use o create-launch-templatecomando para criar um novo modelo de lançamento para seu grupo de Auto Scaling. Inclua a opção `--launch-template-data` e a entrada JSON que definem os detalhes das instâncias criadas para seu grupo do Auto Scaling.

Por exemplo, use o comando a seguir para criar um modelo de execução básico com o ID de AMI *ami-0123456789abcdef0* e o tipo de instância *t2.micro*.

```
aws ec2 create-launch-template --launch-template-name my-template-for-auto-scaling
 --version-description version1 \
  --launch-template-data
 '{"ImageId":"ami-0123456789abcdef0","InstanceType":"t2.micro"}'
```

Se houver êxito, o comando gerará uma saída semelhante à seguinte.

```
{
   "LaunchTemplate": {
     "LaunchTemplateId": "lt-068f72b729example",
     "LaunchTemplateName": "my-template-for-auto-scaling",
     "CreatedBy": "arn:aws:iam::123456789012:user/Bob",
     "CreateTime": "2023-01-30T18:16:06.000Z",
     "DefaultVersionNumber": 1,
     "LatestVersionNumber": 1
  }
}
```

Para ter mais informações, consulte Exemplos para criar e gerenciar modelos de lançamento com o AWS Command Line Interface (AWS CLI).

2. Use o start-instance-refreshcomando para iniciar o fluxo de trabalho de substituição da instância e aplicar seu novo modelo de lançamento com o ID *lt-068f72b729example*. Por ser novo, o modelo de execução tem apenas uma versão. Isso significa que a versão 1 do modelo de execução é o destino dessa atualização de instâncias. Se ocorrer um evento de aumento da escala na horizontal durante a atualização de instâncias e o Amazon EC2 Auto Scaling provisionar novas instâncias usando a versão 1 desse modelo de execução, elas não serão substituídas. Quando a operação for concluída com êxito, o novo modelo de execução será aplicado com êxito ao grupo do Auto Scaling.

```
aws autoscaling start-instance-refresh --cli-input-json file://config.json
```

Conteúdo de `config.json`.

```
{
    "AutoScalingGroupName": "my-asg",
```

```
    "DesiredConfiguration": {
      "LaunchTemplate": {
          "LaunchTemplateId": "lt-068f72b729example",
          "Version": "$Default"
       }
    },
    "Preferences": {
      "SkipMatching": true
    }
}
```

Se houver êxito, o comando gerará uma saída semelhante à seguinte.

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

Ignorar correspondências (grupos de instâncias mistas)

Se você tiver um grupo de Auto Scaling com uma política de instâncias mistas, siga as etapas nesta seção para usar o AWS CLI para iniciar uma atualização de instância com skip matching. Você tem as seguintes opções:

- Forneça um novo modelo de execução para aplicar a todos os tipos de instância especificados na política.
- Forneça um conjunto atualizado de tipos de instância alterando ou não o modelo de execução na política. Por exemplo, digamos que você faça uma migração de tipos de instância indesejados. Você usaria o modelo de execução como está, sem alterar a AMI, os grupos de segurança ou outras especificidades das instâncias a serem substituídas.

Siga as etapas em uma das seções a seguir, de acordo com a opção que atenda às suas necessidades.

Para usar a opção de ignorar correspondência com um novo modelo de execução

1. Use o create-launch-templatecomando para criar um novo modelo de lançamento para seu grupo de Auto Scaling. Inclua a opção `--launch-template-data` e a entrada JSON que definem os detalhes das instâncias criadas para seu grupo do Auto Scaling.

Por exemplo, use o comando a seguir para criar um modelo de execução com o ID de AMI *ami-0123456789abcdef0*.

```
aws ec2 create-launch-template --launch-template-name my-new-template --version-
description version1 \
  --launch-template-data '{"ImageId":"ami-0123456789abcdef0"}'
```

Se houver êxito, o comando gerará uma saída semelhante à seguinte.

```
{
   "LaunchTemplate": {
     "LaunchTemplateId": "lt-04d5cc9b88example",
     "LaunchTemplateName": "my-new-template",
     "CreatedBy": "arn:aws:iam::123456789012:user/Bob",
     "CreateTime": "2023-01-31T15:56:02.000Z",
     "DefaultVersionNumber": 1,
     "LatestVersionNumber": 1
  }
}
```

Para ter mais informações, consulte Exemplos para criar e gerenciar modelos de lançamento com o AWS Command Line Interface (AWS CLI).

2. Para ver a política de instâncias mistas existente para seu grupo de Auto Scaling, execute o describe-auto-scaling-groupscomando. Você precisará dessas informações na próxima etapa, ao iniciar a atualização de instância.

   O comando de exemplo a seguir retorna a política de instâncias mistas configurada para o grupo do Auto Scaling chamado *my-asg*.

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

   Se houver êxito, o comando gerará uma saída semelhante à seguinte.

   ```
   {
     "AutoScalingGroups":[
       {
         "AutoScalingGroupName":"my-asg",
         "AutoScalingGroupARN":"arn",
         "MixedInstancesPolicy":{
   ```

```
        "LaunchTemplate":{
          "LaunchTemplateSpecification":{
            "LaunchTemplateId":"lt-073693ed27example",
            "LaunchTemplateName":"my-old-template",
            "Version":"$Default"
          },
          "Overrides":[
            {
              "InstanceType":"c5.large"
            },
            {
              "InstanceType":"c5a.large"
            },
            {
              "InstanceType":"m5.large"
            },
            {
              "InstanceType":"m5a.large"
            }
          ]
        },
        "InstancesDistribution":{
          "OnDemandAllocationStrategy":"prioritized",
          "OnDemandBaseCapacity":1,
          "OnDemandPercentageAboveBaseCapacity":50,
          "SpotAllocationStrategy":"price-capacity-optimized"
        }
      },
      "MinSize":1,
      "MaxSize":5,
      "DesiredCapacity":4,
      ...
    }
  ]
}
```

3. Use o start-instance-refreshcomando para iniciar o fluxo de trabalho de substituição da instância e aplicar seu novo modelo de lançamento com o ID *lt-04d5cc9b88example*. Por ser novo, o modelo de execução tem apenas uma versão. Isso significa que a versão 1 do modelo de execução é o destino dessa atualização de instâncias. Se ocorrer um evento de aumento da escala na horizontal durante a atualização de instâncias e o Amazon EC2 Auto Scaling provisionar novas instâncias usando a versão 1 desse modelo de execução, elas não serão

substituídas. Quando a operação for concluída com êxito, a política de instâncias mistas atualizada será aplicada com êxito ao grupo do Auto Scaling.

```
aws autoscaling start-instance-refresh --cli-input-json file://config.json
```

Conteúdo de `config.json`.

```
{
  "AutoScalingGroupName":"my-asg",
  "DesiredConfiguration":{
    "MixedInstancesPolicy":{
      "LaunchTemplate":{
        "LaunchTemplateSpecification":{
          "LaunchTemplateId":"lt-04d5cc9b88example",
          "Version":"$Default"
        },
        "Overrides":[
          {
            "InstanceType":"c5.large"
          },
          {
            "InstanceType":"c5a.large"
          },
          {
            "InstanceType":"m5.large"
          },
          {
            "InstanceType":"m5a.large"
          }
        ]
      },
      "InstancesDistribution":{
        "OnDemandAllocationStrategy":"prioritized",
        "OnDemandBaseCapacity":1,
        "OnDemandPercentageAboveBaseCapacity":50,
        "SpotAllocationStrategy":"price-capacity-optimized"
        }
      }
    }
  },
  "Preferences":{
    "SkipMatching":true
```

```
  }
}
```

Se houver êxito, o comando gerará uma saída semelhante à seguinte.

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

No próximo procedimento, você fornecerá um conjunto atualizado de tipos de instância sem alterar o modelo de execução.

Para usar a opção de ignorar correspondência com um conjunto atualizado de tipos de instância

1. Para ver a política de instâncias mistas existente para seu grupo de Auto Scaling, execute o describe-auto-scaling-groupscomando. Você precisará dessas informações na próxima etapa, ao iniciar a atualização de instância.

   O comando de exemplo a seguir retorna a política de instâncias mistas configurada para o grupo do Auto Scaling chamado *my-asg*.

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

   Se houver êxito, o comando gerará uma saída semelhante à seguinte.

   ```
   {
     "AutoScalingGroups":[
       {
         "AutoScalingGroupName":"my-asg",
         "AutoScalingGroupARN":"arn",
         "MixedInstancesPolicy":{
           "LaunchTemplate":{
             "LaunchTemplateSpecification":{
               "LaunchTemplateId":"lt-073693ed27example",
               "LaunchTemplateName":"my-template-for-auto-scaling",
               "Version":"$Default"
             },
             "Overrides":[
               {
                 "InstanceType":"c5.large"
   ```

```
            },
            {
              "InstanceType":"c5a.large"
            },
            {
              "InstanceType":"m5.large"
            },
            {
              "InstanceType":"m5a.large"
            }
          ]
        },
        "InstancesDistribution":{
          "OnDemandAllocationStrategy":"prioritized",
          "OnDemandBaseCapacity":1,
          "OnDemandPercentageAboveBaseCapacity":50,
          "SpotAllocationStrategy":"price-capacity-optimized"
        }
      },
      "MinSize":1,
      "MaxSize":5,
      "DesiredCapacity":4,
      ...
    }
  ]
}
```

2. Use o start-instance-refreshcomando para iniciar o fluxo de trabalho de substituição de instâncias e aplicar suas atualizações. Para substituir instâncias que usam tipos de instância específicos, a configuração desejada deve especificar a política de instâncias mistas somente com os tipos de instância que você deseja. Você pode escolher se deseja adicionar novos tipos de instância no lugar deles.

   O comando de exemplo a seguir inicia uma atualização de instância sem o tipo de instância indesejado *`m5a.large`*. Quando um tipo de instância de seu grupo não corresponde a um dos três tipos de instância restantes, as instâncias são substituídas. (Uma atualização de instância não escolhe os tipos de instância dos quais provisionar as novas instâncias; são as estratégias de alocação que fazem isso.) Quando a operação for concluída com êxito, a política de instâncias mistas atualizada será aplicada com êxito ao grupo do Auto Scaling.

```
aws autoscaling start-instance-refresh --cli-input-json file://config.json
```

Conteúdo de `config.json`

```
{
  "AutoScalingGroupName":"my-asg",
  "DesiredConfiguration":{
    "MixedInstancesPolicy":{
      "LaunchTemplate":{
        "LaunchTemplateSpecification":{
          "LaunchTemplateId":"lt-073693ed27example",
          "Version":"$Default"
        },
        "Overrides":[
          {
            "InstanceType":"c5.large"
          },
          {
            "InstanceType":"c5a.large"
          },
          {
            "InstanceType":"m5.large"
          }
        ]
      },
      "InstancesDistribution":{
        "OnDemandAllocationStrategy":"prioritized",
        "OnDemandBaseCapacity":1,
        "OnDemandPercentageAboveBaseCapacity":50,
        "SpotAllocationStrategy":"price-capacity-optimized"
        }
      }
    }
  },
  "Preferences":{
    "SkipMatching":true
  }
}
```

Adicionar pontos de verificação a uma atualização de instância

Ao usar uma atualização de instância, você pode escolher substituir instâncias em fases para poder executar verificações em suas instâncias durante o uso. Para fazer uma substituição em fases,

adicione pontos de verificação, que são pontos no tempo em que a atualização da instância pausa. O uso de pontos de verificação dá a você maior controle sobre como escolhe atualizar seu grupo do Auto Scaling. Isso ajuda a confirmar que sua aplicação funcionará de forma confiável e previsível.

Conteúdo

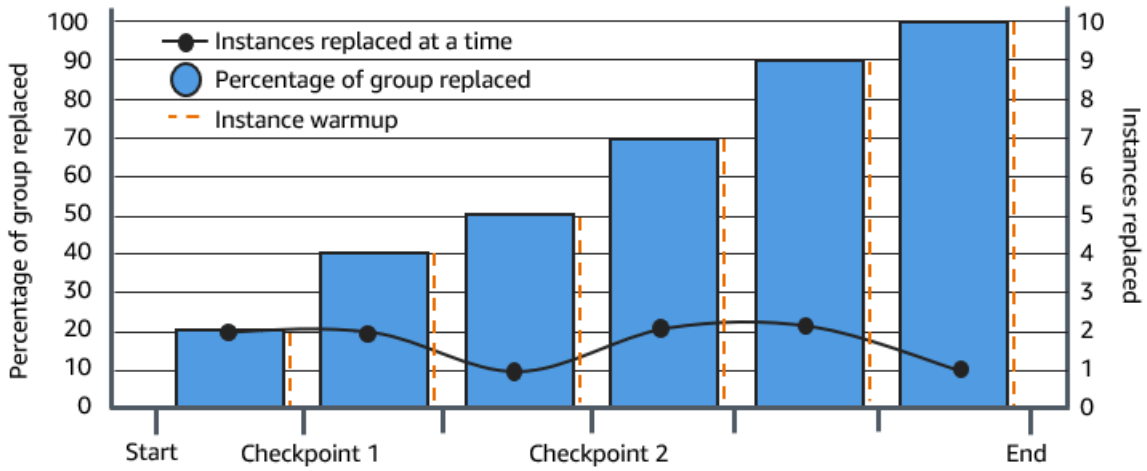
- Como funcionam
- Considerações
- Habilitar pontos de verificação (console)
- Habilitar pontos de verificação (AWS CLI)

Como funcionam

Ao iniciar uma atualização de instância, você especifica pontos de verificação como porcentagens do número total de instâncias no grupo Auto Scaling. Esses pontos de verificação indicam a porcentagem mínima de instâncias no grupo Auto Scaling que devem ser novas instâncias antes que o ponto de verificação seja considerado atingido. Por exemplo, se seus pontos de verificação forem`[20, 50, 100]`, o primeiro ponto de verificação é alcançado quando 20% das instâncias são novas, o segundo quando 50% são novas e o ponto de verificação final quando todas as instâncias são novas.

O Amazon EC2 Auto Scaling acelera as substituições de instâncias para honrar as porcentagens de pontos de verificação especificadas e, ao mesmo tempo, manter a porcentagem mínima saudável do grupo. Para atingir uma porcentagem de pontos de verificação, o Amazon EC2 Auto Scaling às vezes substitui menos, mas nunca mais do que a porcentagem mínima de integridade permitida.

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



No exemplo acima, há um período de aquecimento da instância para cada nova instância iniciada. Você também pode ter um hook do ciclo de vida que coloca uma instância em um estado de espera e, em seguida, executa uma ação personalizada ao iniciar ou encerrar.

O Amazon EC2 Auto Scaling emite eventos para cada ponto de verificação, exceto para o ponto de verificação 100% completo. Você pode adicionar uma EventBridge regra para enviar os eventos para um destino, como o Amazon SNS. Assim, você é notificado quando pode executar as verificações necessárias. Para ter mais informações, consulte Crie EventBridge regras, por exemplo, eventos de atualização.

Considerações

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

- Como os pontos de verificação são baseados em percentuais, o número de instâncias a serem substituídas muda de acordo com o tamanho do grupo. Quando uma atividade de aumento de escala na horizontal ocorre e o tamanho do grupo aumenta, uma operação em andamento pode chegar a um ponto de verificação novamente. Se isso acontecer, o Amazon EC2 Auto Scaling enviará outra notificação e repetirá o tempo de espera entre pontos de verificação antes de continuar.
- É possível pular um ponto de verificação sob certas circunstâncias. Por exemplo, suponha que seu grupo do Auto Scaling tenha duas instâncias e seus percentuais de ponto de verificação sejam `[10,40,100]`. Após a primeira instância ser substituída, o Amazon EC2 Auto Scaling calcula que 50% do grupo foi substituído. Como 50% é maior do que os dois primeiros pontos de verificação, ele ignora o primeiro ponto de verificação (`10`) e envia uma notificação para o segundo ponto de verificação (`40`).

- O cancelamento da operação impede que quaisquer outras substituições sejam feitas. Se a operação for cancelada ou ela falhar antes de atingir o último ponto de verificação, quaisquer instâncias que já tiverem sido substituídas não serão revertidas para a configuração anterior.
- No caso de uma atualização parcial, quando você executa novamente a operação, o Amazon EC2 Auto Scaling não é reiniciado desde o último ponto de verificação, nem para quando apenas as instâncias mais antigas são substituídas. No entanto, ele mira as instâncias mais antigas para substituição primeiro antes de lidar com as instâncias novas.
- A porcentagem real concluída pode ser maior do que a porcentagem desse ponto de verificação quando a porcentagem do ponto de verificação é muito baixa em relação ao número de instâncias no grupo. Por exemplo, suponha que a porcentagem do ponto de verificação seja de 20% e o grupo tenha quatro instâncias. Se o Amazon EC2 Auto Scaling substituir uma das quatro instâncias, a porcentagem real substituída (25%) será maior do que a porcentagem do ponto de verificação (20%).
- Depois que um ponto de verificação é alcançado, a porcentagem geral de conclusão exibida não é atualizada até que as instâncias terminem de aquecer. Por exemplo, as porcentagens do seu ponto de verificação estão `[20,50]` com um atraso de 15 minutos e uma porcentagem mínima saudável de 80%. Seu grupo de Auto Scaling tem 10 instâncias e 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, a porcentagem concluída mostrará 50%.

Habilitar pontos de verificação (console)

Você pode habilitar pontos de verificação antes de iniciar uma atualização de instância para substituir instâncias usando uma abordagem incremental ou em fases. Isso fornece tempo adicional para verificação.

Para iniciar uma atualização de instância que usa pontos de verificação

1. Abra o console do Amazon EC2 em https://console.aws.amazon.com/ec2/ e escolha 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. Na página Start instance refresh (Iniciar atualização de instância), insira os valores aplicáveis para Minimum healthy percentage (Percentual mínimo de integridade) e Instance warmup (Aquecimento da instância).
5. Marque a caixa de seleção Enable checkpoints (Habilitar pontos de verificação).

   Isso exibe uma caixa onde você pode definir o limite percentual para o primeiro ponto de verificação.
6. Em Proceed until ____ % of the group is refreshed (Prosseguir até ____% do grupo ser atualizado), insira um número (1–100). Isso define o percentual para o primeiro ponto de verificação.
7. Para adicionar outro ponto de verificação, escolha Add checkpoint (Adicionar ponto de verificação) e, em seguida, defina o percentual para o próximo ponto de verificação.
8. Para especificar quanto tempo o Amazon EC2 Auto Scaling espera após um ponto de verificação ser atingido, atualize os campos em Wait for **`1 hour`** between checkpoints (Aguardar X Y entre pontos de verificação). A unidade de tempo pode ser horas, minutos ou segundos.
9. Se você tiver concluído as seleções de atualização da instância, escolha Iniciar atualização da instância.

Habilitar pontos de verificação (AWS CLI)

Para iniciar uma atualização de instância com pontos de verificação habilitados usando o AWS CLI, você precisa de um arquivo de configuração que defina os seguintes parâmetros:

- `CheckpointPercentages`: especifica valores de limites para o percentual de instâncias que serão substituídas. Esses valores de limites fornecem os pontos de verificação. Quando o percentual de instâncias substituídas e aquecidas atinge um dos limites especificados, a operação

aguarda por um período especificado. Você especifica o número de segundos para esperar em `CheckpointDelay`. Quando o período de tempo especificado tiver passado, a atualização da instância continuará até atingir o próximo ponto de verificação (se aplicável).

- `CheckpointDelay`: especifica a quantidade de tempo, em segundos, para aguardar após um ponto de verificação ser atingido antes de continuar. Escolha um período que forneça tempo suficiente para executar suas verificações.

O último valor exibido na matriz `CheckpointPercentages` descreve o percentual do grupo do Auto Scaling que precisa ser substituído com êxito. A operação fará a transição para `Successful` após essa porcentagem ser substituída com êxito e cada instância ter concluído a inicialização.

Para criar vários pontos de verificação

Para criar vários pontos de verificação, use o start-instance-refreshcomando de exemplo a seguir. Este exemplo configura uma atualização de instância que atualiza inicialmente 1% do grupo do Auto Scaling. Depois de esperar 10 minutos, ele atualiza os próximos 19% e aguarda mais 10 minutos. Finalmente, ele atualiza o resto do grupo antes de concluir a operação.

```
aws autoscaling start-instance-refresh --cli-input-json file://config.json
```

Conteúdo de `config.json`:

```
{
    "AutoScalingGroupName": "my-asg",
    "Preferences": {
      "InstanceWarmup": 60,
      "MinHealthyPercentage": 80,
      "CheckpointPercentages": [1,20,100],
      "CheckpointDelay": 600
    }
}
```

Para criar um único ponto de verificação

Para criar um único ponto de verificação, use o start-instance-refreshcomando de exemplo a seguir. Este exemplo configura uma atualização de instância que atualiza inicialmente 20% do grupo do Auto Scaling. Depois de aguardar 10 minutos, ele atualiza então o resto do grupo antes de concluir a operação.

```
aws autoscaling start-instance-refresh --cli-input-json file://config.json
```

Conteúdo de `config.json`:

```
{
    "AutoScalingGroupName": "my-asg",
    "Preferences": {
      "InstanceWarmup": 60,
      "MinHealthyPercentage": 80,
      "CheckpointPercentages": [20,100],
      "CheckpointDelay": 600
    }
}
```

Para atualizar parcialmente o grupo do Auto Scaling

Para substituir somente uma parte do seu grupo de Auto Scaling e depois parar completamente, use o comando de exemplo start-instance-refresha seguir. Este exemplo configura uma atualização de instância que atualiza inicialmente 1% do grupo do Auto Scaling. Depois de aguardar 10 minutos, ele atualiza então os próximos 19% antes de concluir a operação.

```
aws autoscaling start-instance-refresh --cli-input-json file://config.json
```

Conteúdo de `config.json`:

```
{
    "AutoScalingGroupName": "my-asg",
    "Preferences": {
      "InstanceWarmup": 60,
      "MinHealthyPercentage": 80,
      "CheckpointPercentages": [1,20],
      "CheckpointDelay": 600
    }
}
```

Substituir instâncias do Auto Scaling com base na vida útil máxima da instância

O tempo de vida máximo da instância especifica o tempo máximo (em segundos) que uma instância pode estar em serviço antes de ser terminada e substituída. Um caso de uso comum pode ser um requisito para substituir as instâncias em uma programação devido a políticas de segurança internas ou a controles de conformidade externos.

É necessário especificar um valor de pelo menos 86.400 segundos (1 dia). Para limpar um valor definido anteriormente, especifique um novo valor de 0. Essa configuração se aplica a todas as instâncias atuais e futuras do grupo do Auto Scaling.

Conteúdo

- Considerações
- Definir o tempo de vida máximo da instância
- Limitações

Considerações

Veja a seguir algumas considerações ao usar esse recurso:

- Sempre que uma instância mais antiga é substituída e uma nova instância é iniciada, a nova instância usa o modelo de execução ou a configuração de execução atualmente associada ao grupo do Auto Scaling. Se seu modelo de lançamento ou configuração de lançamento especificar o ID da Amazon Machine Image (AMI) de uma versão diferente do seu aplicativo, essa versão do seu aplicativo será implantada automaticamente.
- Definir a vida útil máxima da instância muito baixa pode fazer com que as instâncias sejam substituídas mais rápido do que o desejado. O Amazon EC2 Auto Scaling geralmente substitui as instâncias uma de cada vez, com uma pausa entre as substituições. No entanto, se a vida útil máxima especificada da instância não fornecer tempo suficiente para substituir cada instância individualmente, o Amazon EC2 Auto Scaling deverá substituir mais de uma instância por vez. Várias instâncias podem ser substituídas de uma só vez, em até 10% da capacidade atual do grupo do Auto Scaling. Para evitar a substituição de muitas instâncias ao mesmo tempo, defina uma vida útil máxima de instância mais longa ou use a proteção de escalabilidade de instância para impedir temporariamente que instâncias individuais sejam encerradas. Para ter mais informações, consulte Usar proteção de redução na escala na horizontal de instâncias.

- Por padrão, o Amazon EC2 Auto Scaling cria uma nova atividade de escalabilidade para encerrar a instância e, em seguida, finaliza-a. Enquanto a instância estiver sendo encerrada, outra atividade de escalonamento iniciará uma nova instância. Você pode alterar esse comportamento para iniciar antes de encerrar usando uma política de manutenção de instância. Para ter mais informações, consulte Políticas de manutenção de instância.

Definir o tempo de vida máximo da instância

Quando você cria um grupo do Auto Scaling no console, não é possível configurar o tempo de vida máximo da instância. No entanto, depois que o grupo for criado, você poderá editá-lo para definir o tempo de vida máximo da instância.

Para definir o tempo de vida máximo da instância para um grupo (console)

1. Abra o console do Amazon EC2 em https://console.aws.amazon.com/ec2/ e escolha 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 o AWS CLI para definir a vida útil máxima da instância para grupos de Auto Scaling novos ou existentes.

Para novos grupos de Auto Scaling, use o create-auto-scaling-groupcomando.

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

Veja a seguir um arquivo `config.json` de exemplo que mostra um tempo de vida máximo da instância de `2592000` segundos (30 dias).

```
{
    "AutoScalingGroupName": "my-asg",
    "LaunchTemplate": {
        "LaunchTemplateName": "my-launch-template",
        "Version": "$Default"
    },
    "MinSize": 1,
    "MaxSize": 5,
    "MaxInstanceLifetime": 2592000,
    "VPCZoneIdentifier": "subnet-5ea0c127,subnet-6194ea3b,subnet-c934b782",
    "Tags": []
}
```

Para grupos de Auto Scaling existentes, use o update-auto-scaling-groupcomando.

```
aws autoscaling update-auto-scaling-group --auto-scaling-group-name my-existing-asg --
max-instance-lifetime 2592000
```

Para verificar o tempo de vida máximo da instância para um grupo do Auto Scaling

Use o comando describe-auto-scaling-groups.

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

Limitações

- Não há garantia de que tempo de vida máximo será exato para cada instância: não há garantia de que as instâncias serão substituídas apenas no final de sua duração máxima. Em algumas situações, talvez o Amazon EC2 Auto Scaling precise iniciar a substituição de instâncias logo após você atualizar o parâmetro de tempo de vida máximo da instância. A razão para esse comportamento é evitar a substituição de todas as instâncias ao mesmo tempo.
- Proteção de escalabilidade de instância honrada: o Amazon EC2 Auto Scaling fornece proteção de escalabilidade de instâncias para ajudar você a controlar quais instâncias podem ser encerradas. Quando essa proteção é ativada em uma instância, o Amazon EC2 Auto Scaling não encerrará a instância, mesmo que ela tenha atingido sua vida útil máxima.

- Instâncias encerradas antes da execução: quando há apenas uma instância no grupo do Auto Scaling, o recurso de vida útil máxima da instância pode resultar em uma interrupção porque o Amazon EC2 Auto Scaling encerra uma instância e, em seguida, inicia uma nova instância por padrão. Para alterar esse comportamento de iniciar antes de encerrar, consulte Políticas de manutenção de instância.

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 somente pelos recursos que usa. Saber suas restrições de custo ajuda você a decidir quando escalar suas aplicações e por quanto.
- Quais métricas são importantes para sua aplicação? A Amazon CloudWatch oferece suporte a várias métricas diferentes que você pode usar com seu grupo de Auto Scaling.

Conteúdo

- Escolha seu método de escalabilidade
- Definir limites de escalabilidade para seu grupo do Auto Scaling
- Definir o aquecimento padrão da instância para um grupo do Auto Scaling
- Escalabilidade manual para o Amazon EC2 Auto Scaling
- Escalabilidade programada para o Amazon EC2 Auto Scaling
- Escalabilidade dinâmica para o Amazon EC2 Auto Scaling
- Escala preditiva para o Amazon EC2 Auto Scaling
- Controlar quais instâncias do Auto Scaling serão terminadas durante uma redução de escala na horizontal
- Suspender e retomar os processos do Amazon EC2 Auto Scaling

Escolha seu método de escalabilidade

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

Manter um número fixo de instâncias

O padrão para um grupo do Auto Scaling é não ter nenhuma política de escalabilidade anexada ou ação agendada, o que faz com que ele mantenha um tamanho fixo. Depois de você ter criado seu grupo do Auto Scaling, o grupo começa executando instâncias suficientes para atender à sua capacidade desejada. Se não houver condições de escalabilidade associadas ao grupo, ele continuará mantendo a capacidade desejada, mesmo que uma instância não esteja mais íntegra. O Amazon EC2 Auto Scaling monitora a integridade de cada instância do grupo do Auto Scaling. Quando encontra uma instância que não está mais íntegra, ela a substitui por uma nova. Você pode ler uma descrição mais aprofundada deste processo em Verificações de integridade para instâncias em um grupo do Auto Scaling.

Dimensionar manualmente

A escalabilidade manual é a maneira mais básica para escalar seu grupo do Auto Scaling. Você pode atualizar a capacidade desejada do grupo Auto Scaling ou encerrar instâncias no grupo Auto Scaling. Para ter mais informações, consulte Escalabilidade manual para o Amazon EC2 Auto Scaling.

Escala baseada em uma programação

O escalonamento por cronograma significa que as ações de escalonamento são executadas automaticamente em função da data e da 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 ter mais informações, consulte Escalabilidade programada para o Amazon EC2 Auto Scaling.

Dimensione dinamicamente com base na demanda

Uma maneira mais avançada de escalar seus recursos, usando a escalabilidade dinâmica, permite que você defina uma política de escalabilidade que redimensione dinamicamente o grupo do Auto Scaling para atender às alterações na demanda. Por exemplo, vamos supor que você tenha uma aplicação Web que atualmente é executada em duas instâncias e você queira que a utilização da CPU do grupo do Auto Scaling permaneça em cerca de 50% quando a carga na aplicação mudar. Esse método é útil para escalar à medida que ocorrem mudanças no tráfego, quando você não sabe quando o tráfego mudará. É possível configurar políticas de escalabilidade para responder por você. Há vários tipos de políticas (ou uma combinação delas) que você pode usar para escalar em resposta às mudanças de tráfego. Para ter mais informações, consulte Escalabilidade dinâmica para o Amazon EC2 Auto Scaling.

Dimensione de forma proativa

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 ter mais informações, consulte Escala preditiva para o Amazon EC2 Auto Scaling.

Definir limites de escalabilidade para seu grupo do Auto Scaling

Os limites de escalabilidade representam os tamanhos mínimo e máximo do grupo que você deseja para seu grupo do Auto Scaling. Você define limites separadamente para o tamanho mínimo e máximo.

É possível redimensionar a capacidade desejada do grupo para um número que esteja dentro do intervalo dos limites de tamanho mínimo e máximo. A capacidade desejada deve ser maior ou igual ao tamanho mínimo do grupo e menor ou igual ao tamanho máximo do grupo.

- Desired capacity (Capacidade desejada): representa a capacidade inicial do grupo do Auto Scaling no momento da criação. Um grupo do Auto Scaling tenta manter a capacidade desejada. Ele começa executando o número de instâncias especificado para a capacidade desejada e manterá esse número de instâncias desde que não haja políticas de escalabilidade ou ações programadas anexadas ao grupo do Auto Scaling.
- Minimum capacity (Capacidade mínima): representa o tamanho mínimo do grupo. Quando as políticas de escalabilidade estão definidas, um grupo não pode diminuir sua capacidade desejada abaixo do limite de capacidade mínima.
- Maximum capacity (Capacidade máxima): representa o tamanho máximo do grupo. Quando as políticas de escalabilidade estão definidas, um grupo não pode aumentar sua capacidade desejada acima do limite da capacidade máxima.

Os limites de tamanho mínimo e máximo também são aplicáveis aos seguintes cenários:

- Quando você define manualmente a escala do grupo do Auto Scaling mediante a atualização de sua capacidade desejada.
- Quando há a execução de ações agendadas que atualizam a capacidade desejada. Se uma ação agendada for executada sem especificar novos limites de tamanho mínimo e máximo para o grupo, os atuais limites de tamanho mínimo e máximo do grupo serão aplicados.

Um grupo do Auto Scaling sempre tenta manter sua capacidade desejada. Em casos nos quais uma instância seja encerrada inesperadamente (p. ex., devido a uma interrupção da instância spot, uma falha na verificação de integridade ou ação humana), o grupo iniciará automaticamente uma nova instância para manter a capacidade desejada.

Para gerenciar essas configurações no console

1. Abra o console do Amazon EC2 em https://console.aws.amazon.com/ec2/.
2. No painel de navegação, em Auto Scaling, escolha Auto Scaling Groups (Grupos de Auto Scaling).
3. Na página Auto Scaling groups (Grupos do Auto Scaling), marque a caixa de seleção ao lado do seu grupo do Auto Scaling.

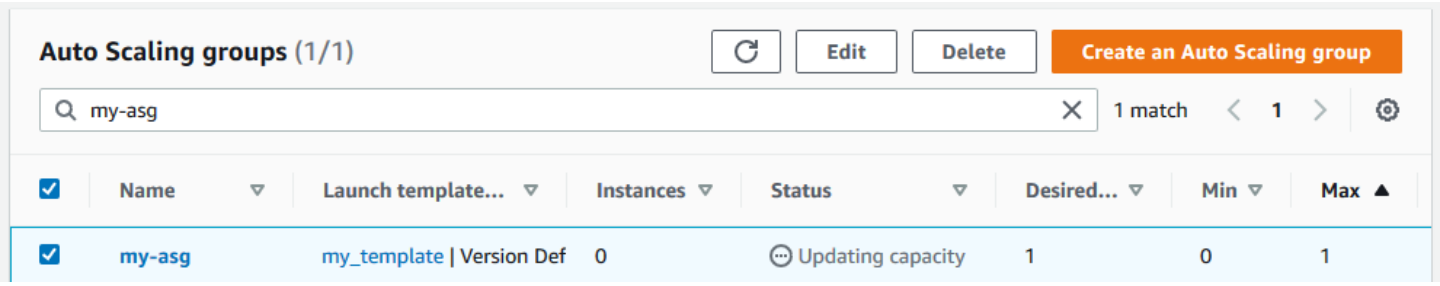
   Um painel dividido é aberto na parte inferior da página.
4. No painel inferior, na guia Detalhes, visualize ou altere as configurações atuais para as capacidades mínima, máxima e desejada do grupo. Para ter mais informações, consulte Alterar a capacidade desejada de um grupo do Auto Scaling existente.

Acima do painel Detalhes, você encontrará informações como o número atual de instâncias no grupo do Auto Scaling, as capacidades mínima, máxima e desejada, além de uma coluna de status. Se o grupo do Auto Scaling usar pesos de instância, você também poderá encontrar o número de unidades de capacidade contribuídas para a capacidade desejada.

Para adicionar ou remover colunas da lista, escolha o ícone de configurações na parte superior da página. Em seguida, em Auto Scaling groups attributes (Atributos dos grupos do Auto Scaling), ative ou desative cada coluna e escolha Confirm (Confirmar).

Para verificar o tamanho de seu grupo do Auto Scaling após fazer alterações.

A coluna Instances (Instâncias) exibe o número de instâncias em execução no momento. Enquanto uma instância está sendo iniciada ou terminada, a coluna Status exibe um status Updating capacity (Atualizando capacidade), conforme mostrado na imagem a seguir.



Aguarde alguns minutos e atualize a visualização para ver o status mais recente. Após a conclusão de uma atividade de escalabilidade, a coluna Instances (Instâncias) exibirá um valor atualizado.

Você pode visualizar o número de instâncias e o status das instâncas que estão em execução no momento aa guia Instance management (Gerenciamento de instâncias), em Instances (Instâncias).

Definir o aquecimento padrão da instância para um grupo do Auto Scaling

CloudWatch coleta e agrega dados de uso, como CPU e E/S de rede, em suas instâncias do Auto Scaling. Use essas métricas para criar políticas de escalabilidade que ajustam o número de instâncias no grupo do Auto Scaling à medida que o valor da métrica selecionada aumenta e diminui.

Você pode especificar quanto tempo depois de uma instância atingir o `InService` estado que espera antes de contribuir com dados de uso para as métricas agregadas. Esse tempo especificado é chamado de aquecimento padrão da instância. Isso evita que a escalabilidade dinâmica seja afetada pelas métricas de instâncias individuais que ainda não estão lidando com o tráfego de aplicativos e que podem estar experimentando um uso temporariamente alto de recursos computacionais.

Para otimizar o desempenho de suas políticas de rastreamento de metas e escalabilidade de etapas, é altamente recomendável que você ative e configure o aquecimento padrão da instância. Ele não está habilitado ou configurado por padrão.

Ao ativar o aquecimento padrão da instância, lembre-se de que, se seu grupo de Auto Scaling estiver configurado para usar uma política de manutenção de instâncias ou se você usar uma atualização de instância para substituir instâncias, você pode evitar que as instâncias sejam contabilizadas na porcentagem íntegra mínima antes de concluírem a inicialização.

Conteúdo

- Considerações sobre o desempenho de escalabilidade
- Escolha o tempo padrão de aquecimento da instância
- Habilitar o aquecimento de instância padrão para um grupo
- Verificar o aquecimento de instância padrão para um grupo
- Encontre políticas de escalabilidade com um tempo de aquecimento de instância definido anteriormente
- Limpe o aquecimento da instância definido anteriormente para uma política de escalabilidade

Considerações sobre o desempenho de escalabilidade

É útil para a maioria dos aplicativos ter um tempo de aquecimento de instância padrão que se aplique a todos os recursos, em vez de tempos de aquecimento diferentes para recursos diferentes. Por exemplo, se você não definir um aquecimento padrão da instância, 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, ela usará 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 terá um 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 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 diminuir, o escalonamento dinâmico se tornará mais conservador para proteger a disponibilidade do seu aplicativo. Isso bloqueia as atividades de expansão para escalabilidade dinâmica até que as novas instâncias terminem de se aquecer.

Durante a escalabilidade horizontal, o Amazon EC2 Auto Scaling considera as instâncias que estão se aquecendo como parte da capacidade do grupo ao decidir quantas instâncias adicionar ao grupo. Portanto, várias violações de alarme que exigem que uma quantidade similar de capacidade seja adicionada resultam em uma única atividade de escalabilidade. A intenção é expandir continuamente, sem fazer isso excessivamente.

Se o aquecimento padrão da instância não estiver ativado, a quantidade de tempo que uma instância espera antes de enviar métricas CloudWatch e contá-las para a capacidade atual variará de instância para instância. Portanto, 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 as políticas de escalabilidade preditiva, as instâncias começam a contribuir imediatamente para as métricas agregadas. Se essas instâncias tiverem maior uso de recursos no startup, a adição de instâncias poderá fazer com que as métricas agregadas tenham um pico. Dependendo do tempo necessário para o uso se estabilizar, isso pode afetar qualquer política de escalabilidade dinâmica que use essas métricas. Se o limite alto de alarme de uma política de escalabilidade dinâmica for violado, o grupo aumentará de tamanho novamente. Enquanto as novas instâncias estiverem se aquecendo, as atividades para reduzir a escala horizontalmente serão bloqueadas.

Escolha o tempo padrão de aquecimento da instância

A chave para definir o aquecimento padrão da instância é determinar quanto tempo suas instâncias precisam terminar a inicialização e para que o consumo de recursos se estabilize após atingirem o estado`InService`. Ao escolher o tempo de aquecimento da instância, tente manter 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 do Auto Scaling vinculado a um balanceador de carga do Elastic Load Balancing. Quando as instâncias concluem seu lançamento, elas são vinculadas ao balanceador de carga antes de entrarem no estado `InService`. Depois que as instâncias entram no estado `InService`, o consumo de recursos ainda pode passar por picos temporários e precisar de tempo para se estabilizar. Por exemplo, o consumo de recursos para um servidor de aplicações que precisa baixar ativos grandes e armazená-los em cache leva mais tempo para se estabilizar do que um servidor Web leve e sem ativos grandes para baixar. O aquecimento de instâncias fornece o tempo de atraso necessário para que o consumo de recursos se estabilize.

> **Important**
>
> Se você não tiver certeza de quanto tempo precisa para o aquecimento, pode começar com 300 segundos. Em seguida, diminua ou aumente gradualmente até obter o melhor desempenho de escalabilidade para seu aplicativo. Talvez seja necessário fazer isso algumas vezes para acertar. Como alternativa, se você tiver alguma política de escalabilidade que tenha seu próprio tempo de aquecimento (`EstimatedInstanceWarmup`), você pode usar esse valor para começar. Para ter mais informações, consulte Encontre políticas de escalabilidade com um tempo de aquecimento de instância definido anteriormente.

Também é necessário considerar o uso de hooks do ciclo de vida para casos de uso em que você tem tarefas de configuração ou scripts para executar no startup. Os hooks do ciclo de vida também podem atrasar a colocação de instâncias em serviço até que elas tenham concluído a inicialização. Eles são especialmente úteis se você tiver scripts de bootstrapping que demoram um pouco para serem concluídos. Se você adicionar um hook do ciclo de vida, será possível reduzir o valor do aquecimento de instância padrão. Para obter mais informações sobre ganchos do ciclo de vida, consulte Ganchos do ciclo de vida do Amazon EC2 Auto Scaling.

Habilitar o aquecimento de instância padrão para um grupo

É possível habilitar o aquecimento padrão da instância ao criar um grupo do Auto Scaling. Também é possível habilitar para grupos existentes.

Ao ativar o recurso padrão de aquecimento de instâncias, você não precisa mais especificar valores para os parâmetros de aquecimento dos seguintes recursos:

- Atualização de instância
- Escalabilidade de rastreamento de destino
- Escalabilidade em etapas

Console

Para habilitar o aquecimento de instância padrão para um novo grupo (console)

Ao criar o grupo do Auto Scaling, na página Configure advanced options (Configurar opções avançadas), em Additional settings (Configurações adicionais), selecione a opção Enable default instance warmup (Habilitar aquecimento de instância padrão). Escolha o tempo de aquecimento necessário para sua aplicação.

AWS CLI

Para habilitar o aquecimento de instância padrão para um novo grupo (AWS CLI)

Para habilitar o aquecimento de instância padrão para um grupo do Auto Scaling, adicione a opção `--default-instance-warmup` e especifique um valor, em segundos, de 0 a 3600. Depois de habilitado, um valor de `-1` desativará essa configuração.

O create-auto-scaling-groupcomando a seguir cria um grupo de Auto Scaling com o nome my-asg e ativa o aquecimento padrão da instância com um valor de 120 segundos.

```
aws autoscaling create-auto-scaling-group --auto-scaling-group-name my-asg --
default-instance-warmup 120 ...
```

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

Console

Para habilitar o aquecimento de instância padrão para um grupo existente (console)

1. Abra o console do Amazon EC2 em https://console.aws.amazon.com/ec2/ e escolha 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 Aquecimento da instância padrão, escolha o tempo de aquecimento necessário para seu aplicativo.
6. Escolha Atualizar.

AWS CLI

Para habilitar o aquecimento de instância padrão para um grupo existente (AWS CLI)

O exemplo a seguir usa o update-auto-scaling-groupcomando para ativar o aquecimento padrão da instância com um valor de 120 segundos para um grupo existente do 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 gerar um erro, verifique se você atualizou o AWS CLI localmente para a versão mais recente.

Verificar o aquecimento de instância padrão para um grupo

Para verificar o aquecimento padrão da instância para um grupo do Auto Scaling (AWS CLI)

Use o seguinte comando describe-auto-scaling-groups: Substitua *my-asg* pelo nome do seu grupo do Auto Scaling.

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

A seguir, uma exemplo de resposta.

```
{
    "AutoScalingGroups": [
        {
            "AutoScalingGroupName": "my-asg",
            "AutoScalingGroupARN": "arn",
            ...
            "DefaultInstanceWarmup": 120
        }
    ]
}
```

Encontre políticas de escalabilidade com um tempo de aquecimento de instância definido anteriormente

Para identificar se você tem políticas que têm seu próprio tempo de aquecimentoEstimatedInstanceWarmup, execute o seguinte comando describe-policies usando o. AWS CLI Substitua *my-asg* pelo nome do seu grupo do Auto Scaling.

```
aws autoscaling describe-policies --auto-scaling-group-name my-asg
  --query 'ScalingPolicies[?EstimatedInstanceWarmup!=`null`]'
```

A seguir, um exemplo de saída.

```
[
  {
    "AutoScalingGroupName":"my-asg",
    "PolicyName":"cpu50-target-tracking-scaling-policy",
    "PolicyARN":"arn",
    "PolicyType":"TargetTrackingScaling",
    "StepAdjustments":[],
    "EstimatedInstanceWarmup":120,
    "Alarms":[{
```

```
        "AlarmARN": "arn:aws:cloudwatch:us-west-2:123456789012:alarm:TargetTracking-my-
asg-AlarmHigh-fc0e4183-23ac-497e-9992-691c9980c38e",
        "AlarmName": "TargetTracking-my-asg-AlarmHigh-
fc0e4183-23ac-497e-9992-691c9980c38e"
      },
      {
            "AlarmARN": "arn:aws:cloudwatch:us-
west-2:123456789012:alarm:TargetTracking-my-asg-AlarmLow-61a39305-ed0c-47af-
bd9e-471a352ee1a2",
            "AlarmName": "TargetTracking-my-asg-AlarmLow-61a39305-ed0c-47af-
bd9e-471a352ee1a2"
    }],
    "TargetTrackingConfiguration":{
      "PredefinedMetricSpecification":{
        "PredefinedMetricType":"ASGAverageCPUUtilization"
      },
      "TargetValue":50.0,
      "DisableScaleIn":false
    },
    "Enabled":true
  },

    ... additional policies ...

]
```

Limpe o aquecimento da instância definido anteriormente para uma política de escalabilidade

Depois de ativar o aquecimento padrão da instância, atualize todas as políticas de escalabilidade que ainda tenham seu próprio tempo de aquecimento para limpar o valor definido anteriormente. Caso contrário, ele substituirá o aquecimento padrão da instância.

Você pode atualizar as políticas de escalabilidade usando o console ou os AWS CLI AWS SDKs. Esta seção aborda as etapas do console. Se você usa 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 Escabilidade automática, em Políticas de escalabilidade dinâmica, escolha a política na qual você está interessado e, em seguida, escolha Ações, Editar.
4. Em Aquecimento da instância, limpe o valor de aquecimento da instância para usar o valor padrão de aquecimento da instância.
5. Escolha Atualizar.

Escalabilidade manual para o Amazon EC2 Auto Scaling

Você pode ajustar manualmente o número de instâncias do EC2 em seu grupo de Auto Scaling a qualquer momento. Esse processo de alterar manualmente a contagem de instâncias é chamado de escalabilidade manual. O escalonamento manual é uma alternativa ao escalonamento automático, especialmente se você quiser fazer alterações de capacidade únicas.

Depois de escalar manualmente seu grupo, o Amazon EC2 Auto Scaling retoma as atividades normais de escalabilidade automática com base nas políticas de escalabilidade e ações programadas que você definiu. Para grupos com o aquecimento de instâncias padrão ativado, todas as novas instâncias passam por um período de aquecimento antes de começarem a contribuir com as métricas usadas para o escalonamento automático. Esse período de aquecimento ajuda a estabilizar o grupo na nova capacidade. Para ter mais informações, consulte Definir o aquecimento padrão da instância para um grupo do Auto Scaling.

Às vezes, talvez você queira desativar temporariamente as políticas de escalabilidade e as ações agendadas antes de escalar manualmente um grupo. Isso evita que surjam conflitos entre as ações manuais de escalonamento e as atividades de escalonamento automatizadas. Para ter mais informações, consulte Desative as atividades de escalabilidade.

Conteúdo

- Alterar a capacidade desejada de um grupo do Auto Scaling existente
- Encerrar uma instância no seu grupo do Auto Scaling (AWS CLI)

Alterar a capacidade desejada de um grupo do Auto Scaling existente

Quando você altera a capacidade desejada do seu grupo de Auto Scaling, o Amazon EC2 Auto Scaling gerencia o processo de lançamento e encerramento de instâncias para atingir o novo tamanho desejado.

Console

Para alterar o tamanho de seu grupo do Auto Scaling

1. Abra o console do Amazon EC2 em https://console.aws.amazon.com/ec2/ e escolha 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 exibido na parte inferior da página.
3. Na guia Detalhes, escolha Detalhes do grupo, Editar.
4. Para Capacidade desejada, aumente ou diminua a capacidade desejada. Por exemplo, para aumentar o tamanho do grupo em um, se o valor atual for1, insira2.

   Se o novo valor para a capacidade desejada for maior que a capacidade mínima desejada e a capacidade máxima desejada, a capacidade máxima desejada será automaticamente aumentada para o novo valor de capacidade desejada.
5. Quando terminar, escolha Atualizar.

Verifique se o tamanho do grupo que você especificou resultou na mesma quantidade de instâncias sendo executadas. Por exemplo, se você aumentou o tamanho do grupo em um, verifique se seu grupo de Auto Scaling iniciou uma instância adicional.

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

1. Na guia Atividade, em Histórico de atividades, você pode ver o progresso das atividades associadas ao grupo Auto Scaling. A coluna Status mostra o status atual de sua instância. Enquanto sua instância está ativando, a coluna de status mostra `Not yet in service`. O status muda para `Successful` depois que a instância é ativada. Você também pode usar o ícone de atualização para ver o status atual da sua instância. Para ter mais informações, consulte Verificar uma ação de escalabilidade para um grupo do Auto Scaling.
2. Na guia Gerenciamento de instâncias, em Instâncias, você pode ver o status da instância. Demora um pouco para iniciar uma instância.

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

AWS CLI

O exemplo a seguir pressupõe que você criou um grupo do Auto Scaling com um tamanho mínimo de 1 e um tamanho máximo de 5. Portanto, o grupo atualmente tem uma instância em execução.

Para alterar o tamanho de seu grupo do Auto Scaling

Use o set-desired-capacitycomando para alterar o tamanho do seu grupo de Auto Scaling, conforme mostrado no exemplo a seguir.

```
aws autoscaling set-desired-capacity --auto-scaling-group-name my-asg \
  --desired-capacity 2
```

Se você optar por cumprir o período de desaquecimento padrão para seu grupo do Auto Scaling, especifique a opção `--honor-cooldown`, conforme mostrado no exemplo a seguir. Para ter mais informações, consulte Desaquecimento de escalabilidade para o Amazon EC2 Auto Scaling.

```
aws autoscaling set-desired-capacity --auto-scaling-group-name my-asg \
  --desired-capacity 2 --honor-cooldown
```

Para verificar o tamanho de seu grupo do Auto Scaling

Use o describe-auto-scaling-groupscomando para confirmar que o tamanho do seu grupo de Auto Scaling foi alterado, como no exemplo a seguir.

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

Veja a seguir um exemplo de saída, que fornece detalhes sobre o grupo e as instâncias lançadas.

```
{
    "AutoScalingGroups": [
        {
```

```
            "AutoScalingGroupName": "my-asg",
            "AutoScalingGroupARN": "arn",
            "LaunchTemplate": {
                "LaunchTemplateName": "my-launch-template",
                "Version": "1",
                "LaunchTemplateId": "lt-050555ad16a3f9c7f"
            },
            "MinSize": 1,
            "MaxSize": 5,
            "DesiredCapacity": 2,
            "DefaultCooldown": 300,
            "AvailabilityZones": [
                "us-west-2a"
            ],
            "LoadBalancerNames": [],
            "TargetGroupARNs": [],
            "HealthCheckType": "EC2",
            "HealthCheckGracePeriod": 300,
            "Instances": [
                {
                    "ProtectedFromScaleIn": false,
                    "AvailabilityZone": "us-west-2a",
                    "LaunchTemplate": {
                        "LaunchTemplateName": "my-launch-template",
                        "Version": "1",
                        "LaunchTemplateId": "lt-050555ad16a3f9c7f"
                    },
                    "InstanceId": "i-05b4f7d5be44822a6",
                    "InstanceType": "t3.micro",
                    "HealthStatus": "Healthy",
                    "LifecycleState": "Pending"
                },
                {
                    "ProtectedFromScaleIn": false,
                    "AvailabilityZone": "us-west-2a",
                    "LaunchTemplate": {
                        "LaunchTemplateName": "my-launch-template",
                        "Version": "1",
                        "LaunchTemplateId": "lt-050555ad16a3f9c7f"
                    },
                    "InstanceId": "i-0c20ac468fa3049e8",
                    "InstanceType": "t3.micro",
                    "HealthStatus": "Healthy",
                    "LifecycleState": "InService"
```

```
                }
            ],
            "CreatedTime": "2019-03-18T23:30:42.611Z",
            "SuspendedProcesses": [],
            "VPCZoneIdentifier": "subnet-c87f2be0",
            "EnabledMetrics": [],
            "Tags": [],
            "TerminationPolicies": [
                "Default"
            ],
            "NewInstancesProtectedFromScaleIn": false,
            "ServiceLinkedRoleARN": "arn",
            "TrafficSources": []
        }
    ]
}
```

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

Encerrar uma instância no seu grupo do Auto Scaling (AWS CLI)

Há momentos em que talvez você queira manualmente reduzir a escala horizontalmente em seu grupo do Auto Scaling, mas queira e encerrar uma instância específica. Você pode escalar manualmente em seu grupo de Auto Scaling usando o comando terminate-instance-in-auto-scaling-group e especificando o ID da instância que você deseja encerrar e a `--should-decrement-desired-capacity` opção, conforme mostrado no exemplo a seguir.

```
aws autoscaling terminate-instance-in-auto-scaling-group \
  --instance-id i-026e4c9f62c3e448c --should-decrement-desired-capacity
```

Veja a seguir um exemplo de saída, que fornece detalhes sobre a atividade de escalabilidade.

```
{
    "Activities": [
        {
            "ActivityId": "b8d62b03-10d8-9df4-7377-e464ab6bd0cb",
            "AutoScalingGroupName": "my-asg",
            "Description": "Terminating EC2 instance: i-026e4c9f62c3e448c",
            "Cause": "At 2023-09-23T06:39:59Z instance i-026e4c9f62c3e448c was taken
 out of service in response to a user request, shrinking the capacity from 1 to 0.",
```

```
            "StartTime": "2023-09-23T06:39:59.015000+00:00",
            "StatusCode": "InProgress",
            "Progress": 0,
            "Details": "{\"Subnet ID\":\"subnet-6194ea3b\",\"Availability Zone\":\"us-
west-2c\"}"
        }
    ]
}
```

Esta opção não está disponível no console. No entanto, você pode usar a página Instâncias do console do Amazon EC2 para encerrar uma instância em seu grupo de Auto Scaling. Quando você faz isso, o Amazon EC2 Auto Scaling detecta que a instância não está mais em execução e a substitui automaticamente como parte do processo de verificação de integridade. Depois de encerrar a instância, leva um ou dois minutos para que uma nova instância seja executada. Para obter informações sobre como encerrar uma instância, consulte Encerrar uma instância no Guia do usuário do Amazon EC2 para instâncias Linux.

Se você encerrar instâncias em seu grupo e isso causar uma distribuição desigual entre as zonas de disponibilidade, o Amazon EC2 Auto Scaling reequilibrará o grupo para restabelecer uma distribuição uniforme, a menos que você suspenda o processo. `AZRebalance` Para ter mais informações, consulte Suspender e retomar os processos do Amazon EC2 Auto Scaling.

Escalabilidade programada para o Amazon EC2 Auto Scaling

Com o escalonamento programado, você pode configurar o escalonamento automático para seu aplicativo com base em mudanças de carga previsíveis. Você cria ações programadas que aumentam ou diminuem a capacidade desejada do seu grupo em horários específicos.

Por exemplo, você experimenta um padrão regular de tráfego semanal em que a carga aumenta no meio da semana e diminui no final da semana. Você pode configurar um cronograma de escalabilidade no Amazon EC2 Auto Scaling que se alinhe a esse padrão:

- Na manhã de quarta-feira, uma ação programada aumenta a capacidade aumentando a capacidade desejada previamente definida do grupo Auto Scaling.
- Na sexta-feira à noite, outra ação programada diminui a capacidade ao diminuir a capacidade desejada previamente definida do grupo Auto Scaling.

Essas ações de escalabilidade programadas permitem otimizar os custos e a performance. Seu aplicativo tem capacidade suficiente para lidar com o pico de tráfego no meio da semana, mas não provisiona demais a capacidade desnecessária em outros momentos.

Você pode usar o escalonamento programado e as políticas de escalabilidade em conjunto para obter os benefícios de ambas as abordagens de escalabilidade. Após a execução de uma ação de escalabilidade programada, a política de escalabilidade pode continuar a tomar decisões sobre a necessidade de escalar ainda mais a capacidade. Isso ajuda a garantir que você tenha capacidade suficiente para lidar com a carga de sua aplicação. Embora sua aplicação seja escalada para atender à demanda, a capacidade atual deve estar dentro das capacidades mínima e máxima definidas pela ação agendada.

Conteúdo

- Como a escalabilidade programada funciona
- Programações recorrentes
- Fuso horário
- Considerações
- Criar uma ação programada
- Exibir detalhes da ação agendada
- Verificar as atividades de escalabilidade
- Excluir uma ação programada
- Limitações

Como a escalabilidade programada funciona

Para usar a escalabilidade programada, crie ações programadas que instruam o Amazon EC2 Auto Scaling a realizar atividades de escalabilidade em horários específicos. Ao criar uma ação programada, você especifica o grupo Auto Scaling, quando a atividade de escalabilidade deve ocorrer, a nova capacidade desejada e, opcionalmente, uma nova capacidade mínima e uma nova capacidade máxima. É possível criar ações programadas para escalar uma única vez ou de forma programada.

No momento especificado, o Amazon EC2 Auto Scaling escala com base nos novos valores de capacidade, comparando a capacidade atual com a capacidade desejada especificada.

- Se a capacidade atual for menor do que a capacidade desejada especificada, o Amazon EC2 Auto Scaling expande ou adiciona instâncias à capacidade desejada especificada.
- Se a capacidade atual for maior do que a capacidade desejada especificada, o Amazon EC2 Auto Scaling expande ou remove instâncias até a capacidade desejada especificada.

Uma ação programada define a capacidade desejada, mínima e máxima do grupo na data e hora especificadas. Você pode criar uma ação programada para somente uma dessas capacidades por vez, por exemplo, a capacidade desejada. No entanto, há alguns casos em que você deve incluir a capacidade mínima e máxima para garantir que a capacidade desejada especificada na ação não esteja fora desses limites.

Programações recorrentes

Para criar uma agenda recorrente usando o AWS CLI ou um SDK, especifique uma expressão cron e um fuso horário para descrever quando essa ação agendada deve ocorrer novamente. Opcionalmente, você pode especificar uma data e hora para a hora de início, a hora de término ou ambas.

Para criar uma agenda recorrente usando o AWS Management Console, especifique o padrão de recorrência, o fuso horário, a hora de início e a hora de término opcional da ação agendada. Todas as opções de padrão de recorrência são baseadas em expressões do cron. Alternativamente, você pode escrever sua própria expressão do cron personalizada.

A expressão do cron consiste em cinco campos separados por espaços: [Minuto] [Hora] [Dia_do_mês] [Mês_do_ano] [Dia_da_semana]. Por exemplo, a expressão do cron `30 6 * * 2` configura uma ação programada que se repete todas as terças-feiras às 6h30. O asterisco é usado como um curinga para corresponder a todos os valores de um campo. Para obter outros exemplos de expressões do cron, consultehttps://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.
- A hora de início e a hora de término devem ser definidas em UTC quando você usa o AWS CLI ou um SDK.

Fuso horário

Por padrão, as programações recorrentes definidas por você estão no fuso horário UTC (Tempo Universal Coordenado). É possível alterar o fuso para corresponder a seu fuso horário local ou a um fuso horário de outra parte da rede. Se você especificar um o fuso horário que siga o horário de verão, ele se ajustará automaticamente ao horário de verão (DST).

Os valores válidos são os nomes canônicos dos fusos horários do banco de dados de fusos horários da Internet Assigned Numbers Authority (IANA). Por exemplo, o horário do Leste dos EUA é canonicamente identificado como. `America/New_York` Para obter mais informações, consulte https://www.iana.org/time-zones.

Fusos horários baseados em localização, como ajuste `America/New_York` automático para o horário de verão. No entanto, um fuso horário baseado em UTC, como `Etc/UTC`, é uma hora absoluta e não se ajustará para o horário de verão.

Por exemplo, você tem uma programação recorrente cujo fuso horário é `America/New_York`. A primeira ação de escalabilidade acontece no fuso horário `America/New_York`, antes do horário de verão ser iniciado. A próxima ação de escalabilidade acontece no fuso horário `America/New_York`, depois do horário de verão ser iniciado. A primeira ação começa às 8:00 UTC-5 na hora local, enquanto a segunda vez começa às 8:00 UTC-4 no horário local.

Se você criar uma ação agendada usando o AWS Management Console e especificar um fuso horário que observe o horário de verão, tanto a programação recorrente quanto os horários de início e término se ajustarão automaticamente para o horário de verão.

Considerações

Ao criar uma ação programada, lembre-se do seguinte:

- A ordem de execução das ações programadas é garantida no mesmo grupo, mas não das ações programadas entre grupos.
- Uma ação agendada geralmente é executada em segundos. No entanto, a ação pode ser atrasada em até dois minutos da hora de início programada. Como as ações programadas em um grupo do Auto Scaling são executadas na ordem em que são especificadas, as ações com horas de início programadas próximas umas das outras podem demorar mais para serem executadas.
- Você pode desativar temporariamente a escalabilidade programada para um grupo do Auto Scaling, suspendendo o processo `ScheduledActions`. Isso ajuda você a impedir que

ações programadas fiquem ativas sem precisar excluí-las. Em seguida, você pode retomar a escalabilidade programada quando quiser usá-la novamente. Para ter mais informações, consulte Suspender e retomar os processos do Amazon EC2 Auto Scaling.

- Depois de criar uma ação programada, você pode atualizar qualquer uma de suas configurações, exceto o nome.

Criar uma ação programada

Para criar uma ação agendada para seu grupo de Auto Scaling, use um dos seguintes métodos:

Console

Para criar 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), escolha Create scheduled action (Criar ação programada).
4. Insira um Name (Nome), para a ação programada.
5. Em Capacidade desejada, Mín., Máx., escolha a nova capacidade desejada do grupo e os novos limites de tamanho mínimo e máximo. A capacidade desejada deve ser maior ou igual ao tamanho mínimo do grupo e menor ou igual ao tamanho máximo do grupo.
6. Em Recurrence (Recorrência), selecione uma das opções disponíveis.
   - Se você quiser escalar em uma programação recorrente, escolha com que frequência o Amazon EC2 Auto Scaling deve executar a ação programada.
     - Se você escolher uma opção que começa com Every (A cada), a expressão Cron será criada para você.
     - Se você escolher Cron, insira uma expressão do cron que especifique quando executar a ação, em UTC.
   - Se você quiser escalar apenas uma vez, escolha Once (Uma vez).
7. Em Time zone (Fuso horário), escolha um fuso horário. O padrão é `Etc/UTC`.

Todos os fusos horários listados são do banco de dados de fuso horário da IANA. Para obter mais informações, consulte https://en.wikipedia.org/wiki/List_of_tz_database_time_zones.

8. Defina uma data e hora para Specific start time (Horário de início específico).

   - Se você escolher uma programação recorrente, o horário inicial definirá quando a primeira ação programada na série recorrente será executada.
   - Se você escolheu Once (Uma vez) como recorrência, o horário inicial define a data e a hora para a ação programada ser executada.

9. (Opcional) Para programações recorrentes, você pode especificar uma hora final escolhendo Set End Time (Definir horário de término) e, em seguida, escolher uma data e hora para End by (Encerrar em).
10. Escolha Criar. O console exibe as ações programadas para o grupo do Auto Scaling.

AWS CLI

Para criar uma ação agendada, você pode usar um dos seguintes comandos de exemplo. Substitua cada *espaço reservado para entrada do usuário* por suas próprias informações.

Exemplo: para escalar apenas uma vez

Use o seguinte comando put-scheduled-update-group-action com as `--desired-capacity` opções `--start-time "YYYY-MM-DDThh:mm:ssZ"` e.

```
aws autoscaling put-scheduled-update-group-action --scheduled-action-name my-one-time-action \
  --auto-scaling-group-name my-asg --start-time "2021-03-31T08:00:00Z" --desired-capacity 3
```

Exemplo: Para agendar o escalonamento em uma programação recorrente

Use o seguinte comando put-scheduled-update-group-action com as `--desired-capacity` opções `--recurrence "cron expression"` e.

```
aws autoscaling put-scheduled-update-group-action --scheduled-action-name my-recurring-action \
  --auto-scaling-group-name my-asg --recurrence "0 9 * * *" --desired-capacity 3
```

Por padrão, o Amazon EC2 Auto Scaling executa a programação de recorrência especificada com base no fuso horário UTC. Para especificar um fuso horário diferente, inclua a `--time-zone` opção e o nome do fuso horário da IANA, como no exemplo a seguir.

```
--time-zone "America/New_York"
```

Para obter mais informações, consulte https://en.wikipedia.org/wiki/List_of_tz_database_time_zones.

Exibir detalhes da ação agendada

Para ver detalhes das próximas ações agendadas para seu grupo de Auto Scaling, use um dos seguintes métodos:

Console

Para ver os detalhes da ação agendada

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 Escala automática, na seção Ações agendadas, você pode ver as próximas ações agendadas.

Observe que o console mostra os valores da hora de início e hora de término em seu horário local com o deslocamento 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 valorde Time zone (Fuso horário) mostra seu fuso horário solicitado, por exemplo, `America/New_York`.

AWS CLI

Use o seguinte comando describe-scheduled-actions:

```
aws autoscaling describe-scheduled-actions --auto-scaling-group-name my-asg
```

Se houver êxito, o comando gerará uma saída semelhante à seguinte.

```
{
```

```
  "ScheduledUpdateGroupActions": [
    {
      "AutoScalingGroupName": "my-asg",
      "ScheduledActionName": "my-recurring-action",
      "Recurrence": "30 0 1 1,6,12 *",
      "ScheduledActionARN": "arn:aws:autoscaling:us-
west-2:123456789012:scheduledUpdateGroupAction:8e86b655-b2e6-4410-8f29-
b4f094d6871c:autoScalingGroupName/my-asg:scheduledActionName/my-recurring-action",
      "StartTime": "2020-12-01T00:30:00Z",
      "Time": "2020-12-01T00:30:00Z",
      "MinSize": 1,
      "MaxSize": 6,
      "DesiredCapacity": 4
    }
  ]
}
```

Verificar as atividades de escalabilidade

Para verificar as atividades de escalabilidade associadas à escalabilidade programada, consulteVerificar uma ação de escalabilidade para um grupo do Auto Scaling.

Excluir uma ação programada

Para excluir uma ação agendada, use um dos seguintes métodos:

Console

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 Ações, Excluir.
5. Quando a confirmação for solicitada, escolha Sim, excluir.

AWS CLI

Use o seguinte comando delete-scheduled-action:

```
aws autoscaling delete-scheduled-action --auto-scaling-group-name my-asg \
  --scheduled-action-name my-recurring-action
```

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.

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.

É altamente recomendável que você use políticas de escalabilidade de rastreamento de metas e escolha uma métrica que mude inversamente proporcional a uma alteração na capacidade do

seu grupo de Auto Scaling. Portanto, se você dobrar o tamanho do seu grupo de Auto Scaling, a métrica diminuirá em 50%. Isso permite que os dados de métricas acionem com precisão eventos de escalabilidade proporcionais. Estão incluídas métricas como utilização média da CPU ou contagem média de solicitações por alvo.

Com o rastreamento de metas, seu grupo de Auto Scaling é dimensionado 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.

As políticas de rastreamento de metas também eliminam a necessidade de definir manualmente CloudWatch alarmes e ajustes de escala. O Amazon EC2 Auto Scaling lida com isso automaticamente com base na meta que você definiu.

Conteúdo

- Como funcionam as políticas de escalabilidade dinâmica
- Várias políticas de escalabilidade dinâmica
- Políticas de escalabilidade com monitoramento do objetivo para o Amazon EC2 Auto Scaling
- Políticas de escalabilidade simples e em etapas do Amazon EC2 Auto Scaling
- Desaquecimento de escalabilidade para o Amazon EC2 Auto Scaling
- Escalabilidade baseada no Amazon SQS
- Verificar uma ação de escalabilidade para um grupo do Auto Scaling
- Desabilitar uma política de escalabilidade para um grupo do Auto Scaling
- Excluir uma política de escalabilidade
- Exemplo de políticas de escalabilidade para a AWS Command Line Interface (AWS CLI)

Como funcionam as políticas de escalabilidade dinâmica

Uma política de escalabilidade dinâmica instrui o Amazon EC2 Auto Scaling a rastrear CloudWatch uma métrica específica e define a ação a ser tomada quando CloudWatch o 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 maneiras: usando as mesmas unidades que você escolheu ao definir a capacidade desejada em termos de instâncias ou usando unidades de capacidade (se os pesos das instâncias forem aplicados).

- 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 pesos de instância. Nesse caso, o Amazon EC2 Auto Scaling pode aumentar a escala na horizontal acima do limite de tamanho máximo, mas somente até o peso máximo da instância. Sua intenção é chegar o mais próximo possível da nova capacidade desejada, mas ainda seguir as estratégias de alocação especificadas para o grupo. As estratégias de alocação determinam quais tipos de instância serão executados. Os pesos determinam quantas com unidades de capacidade cada instância contribui para a capacidade desejada do grupo com base no seu tipo de instância.

- Exemplo 3: um grupo do Auto Scaling tem uma capacidade máxima de 12, uma capacidade atual de 10 e uma política de escalabilidade dinâmica que adiciona 5 unidades de capacidade. Os tipos de instância têm um dos três pesos atribuídos: 1, 4 ou 6. Ao invocar a política, o Amazon EC2 Auto Scaling opta por iniciar um tipo de instância com um peso de 6 com base na estratégia de alocação. O resultado desse evento de expansão é um grupo com uma capacidade desejada de 12 e uma capacidade atual de 16.

Várias políticas de escalabilidade dinâmica

Na maioria dos casos, uma política de escalabilidade com monitoramento do objetivo é suficiente para configurar o grupo do Auto Scaling para aumentar e reduzir a escala na horizontal automaticamente. Uma política de escalabilidade com monitoramento do objetivo permite que você selecione um resultado desejado e faça com que o grupo do Auto Scaling adicione e remova instâncias conforme necessário para atingir o resultado.

Para uma configuração de escalabilidade avançada, seu grupo do Auto Scaling pode ter mais de uma política de escalabilidade. Por exemplo, você pode definir uma ou mais políticas de escalabilidade de rastreamento de destino, uma ou mais políticas de escalabilidade em etapas, ou ambas. Isso fornece maior flexibilidade para abranger vários cenários.

Para ilustrar como várias políticas de escalabilidade dinâmica trabalham em conjunto, considere uma aplicação que use um grupo do Auto Scaling e uma fila do Amazon SQS para enviar solicitações a uma única instância do EC2. Para ajudar a garantir que a aplicação seja executada em níveis ideais, há duas políticas que controlam quando o grupo do Auto Scaling deve ter a escala aumentada na horizontal. Uma é uma política de rastreamento de destino que usa uma métrica personalizada para adicionar e remover capacidade com base no número de mensagens do SQS na fila. A outra é uma política de escalabilidade por etapas que usa a CloudWatch `CPUUtilization` métrica da Amazon para adicionar capacidade quando a instância excede 90% de utilização por um período de tempo especificado.

Quando há várias políticas em vigor ao mesmo tempo, há uma chance de que cada política instrua o grupo do Auto Scaling a ter a escala na horizontal aumentada (ou reduzida) ao mesmo tempo. Por exemplo, é possível que a `CPUUtilization` métrica aumente e ultrapasse o limite do CloudWatch alarme ao mesmo tempo em que a métrica personalizada do SQS aumente e ultrapasse o limite do alarme métrico personalizado.

Quando ocorrem essas situações, o Amazon EC2 Auto Scaling escolhe a política que fornece a maior capacidade para aumentar e para reduzir a escala na horizontal. Por exemplo, suponha que a política `CPUUtilization` execute uma instância, enquanto a política da fila do SQS executa duas instâncias. Se os critérios de aumento de escala na horizontal das duas políticas forem atendidos ao mesmo tempo, o Amazon EC2 Auto Scaling dará precedência à política da fila do SQS. Isso resulta na execução de duas instâncias no grupo do Auto Scaling.

A abordagem de dar precedência à política que fornece a maior capacidade se aplica mesmo quando as políticas usam critérios diferentes para aumentar. Por exemplo, se uma política terminar três instâncias, outra política diminuir o número de instâncias em 25%, e o grupo tiver oito instâncias

no momento da redução da escala na horizontal, o Amazon EC2 Auto Scaling dará precedência à política que fornece o maior número de instâncias para o grupo. Isso faz com que o grupo do Auto Scaling termine duas instâncias (25% de 8 = 2). A intenção é evitar que o Amazon EC2 Auto Scaling remova instâncias demais.

No entanto, recomendamos cautela ao usar políticas de escalabilidade de rastreamento de destino com políticas de escalabilidade de etapas, pois conflitos entre essas políticas podem causar um comportamento indesejável. Por exemplo, se a política de escalabilidade de etapas iniciar uma atividade de redução antes que a política de rastreamento de destino esteja pronta para ser reduzida, a atividade de redução não será bloqueada. Após a conclusão da atividade de redução, a política de rastreamento de destino poderá instruir o grupo a expandir novamente.

Políticas de escalabilidade com monitoramento do objetivo para o Amazon EC2 Auto Scaling

Uma política de escalabilidade de rastreamento de metas dimensiona automaticamente a capacidade do seu grupo de Auto Scaling com base em um valor métrico alvo. Isso permite que a aplicação mantenha uma performance ideal e uma eficiência de custos sem a necessidade de intervenção manual.

Com o rastreamento de destinos, você seleciona uma métrica e um valor de destino para representar a utilização média ideal ou o nível de throughput para a aplicação. O Amazon EC2 Auto Scaling cria e gerencia CloudWatch os alarmes que invocam eventos de escalabilidade quando a métrica se desvia da meta. Por exemplo, isso é semelhante à forma como um termostato mantém uma temperatura alvo.

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 de Auto Scaling se expandirá ou aumentará a capacidade quando a CPU exceder 50% para lidar com o aumento da carga. Ele aumentará ou diminuirá a capacidade quando a CPU cair abaixo de 50% para otimizar os custos durante períodos de baixa utilização.

Tópicos

- Várias políticas de escalabilidade de monitoramento de objetivo

- Escolher métricas
- Definir valor de objetivo
- Defina o tempo de aquecimento da instância
- Considerações
- Criar uma política de dimensionamento com monitoramento do objetivo
- Crie uma política de escalabilidade de rastreamento de destino para Amazon EC2 Auto Scaling usando matemática em métricas

Várias políticas de escalabilidade de monitoramento de objetivo

Para ajudar a otimizar o desempenho de escalonamento, você pode usar várias políticas de escalabilidade com monitoramento de objetivo juntas desde que cada uma delas use uma métrica diferente. Por exemplo, utilização e throughput podem promover influência cruzada. Sempre que uma dessas métricas muda, geralmente isso significa que outras métricas também serão afetadas. Portanto, o uso de várias métricas fornece informações adicionais sobre a carga sob a qual seu grupo de Auto Scaling está. Isso pode ajudar o Amazon EC2 Auto Scaling a tomar decisões mais informadas ao determinar quanta capacidade adicionar ao seu grupo.

A intenção do Amazon EC2 Auto Scaling é sempre priorizar a disponibilidade. Ele expandirá o grupo de Auto Scaling se alguma das políticas de rastreamento de alvos estiver pronta para ser expandida. Ele será ampliado somente se todas as políticas de rastreamento de alvos (com a parte de expansão ativada) estiverem prontas para serem ampliadas.

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 destino com um tipo de métrica predefinida, você escolhe uma métrica da lista de métricas predefinidas a seguir.

- `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 destino podem ser encontradas no tópico Listar as métricas CloudWatch disponíveis para suas instâncias no Guia do usuário do Amazon EC2 para instâncias Linux e as métricas do seu Application Load Balancer no tópico Listar CloudWatch as métricas disponíveis para suas instâncias 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. Você deve usar o AWS CLI ou um SDK para criar uma política de rastreamento de metas com uma especificação métrica personalizada. Para obter um exemplo que especifica uma especificação métrica personalizada para uma política de escalabilidade de rastreamento de metas usando o AWS CLI, consulte. Exemplo de políticas de escalabilidade para a AWS Command Line Interface (AWS CLI)

Lembre-se do seguinte ao escolher uma métrica:

- Recomendamos que você somente use as métricas que estão disponíveis em intervalos de um minuto para ajudar a escalar mais rapidamente em resposta a alterações na 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 habilitar o monitoramento detalhado, consulteConfigurar monitoramento para instâncias do Auto Scaling.
- Nem todas as métricas personalizadas funcionam para rastreamento de destino. A métrica deve ser de utilização válida e descrever o quão ocupada uma instância está. O valor da métrica deve aumentar e diminuir em proporção ao número das instâncias no grupo do Auto Scaling. Isso é para que os dados da métrica possam ser usados para expandir ou reduzir o número de instâncias. Por exemplo, a utilização da CPU de um grupo do Auto Scaling funcionará (ou seja, a métrica

`CPUUtilization` do Amazon EC2 com a dimensão da métrica `AutoScalingGroupName`) se a carga no grupo do Auto Scaling for distribuída entre as instâncias.

- As métricas a seguir não funcionam para rastreamento de destino:
  - O número de solicitações recebidas pelo balanceador de carga voltadas para o grupo do Auto Scaling (ou seja, a métrica `RequestCount` do Elastic Load Balancing). O número de solicitações recebidas pelo balanceador de carga não é alterado com base na utilização do grupo do Auto Scaling.
  - A latência da solicitação do balanceador de carga (ou seja, a métrica `Latency` do Elastic Load Balancing). A latência da solicitação pode aumentar com base no aumento da utilização, mas não necessariamente muda de forma proporcional.
  - A CloudWatch métrica de fila do Amazon SQS. `ApproximateNumberOfMessagesVisible` O número de mensagens em uma fila pode não mudar proporcionalmente ao tamanho do grupo do Auto Scaling que processa mensagens da fila. Contudo, uma métrica personalizada que meça o número de mensagens na fila por instância do EC2 no grupo do Auto Scaling pode funcionar. Para ter mais informações, consulte Escalabilidade baseada no Amazon SQS.
- Para usar a métrica `ALBRequestCountPerTarget`, é necessário especificar o parâmetro `ResourceLabel` a fim de identificar o grupo de destino do balanceador de carga que está associado à métrica. Para obter um exemplo que especifica o `ResourceLabel` parâmetro para uma política de escalabilidade de rastreamento de metas usando o AWS CLI, consulte. Exemplo de políticas de escalabilidade para a AWS Command Line Interface (AWS CLI)
- Quando uma métrica emite valores reais de 0 para CloudWatch (por exemplo,`ALBRequestCountPerTarget`), um grupo de Auto Scaling pode escalar até 0 quando não há tráfego para seu aplicativo por um período prolongado. A capacidade mínima do grupo deve estar definida como 0 para que seu grupo do Auto Scaling reduza a escala horizontalmente para 0 quando não houver solicitação roteada para ele.
- Em vez de publicar novas métricas para usar em sua política de escalabilidade, é possível usar a matemática métrica para combinar métricas existentes. Para ter mais informações, consulte Crie uma política de escalabilidade de rastreamento de destino para Amazon EC2 Auto Scaling usando matemática em métricas.

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.

Defina 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 contabilizada nas métricas agregadas da instância EC2 do grupo Auto Scaling.

Enquanto as instâncias estão no período de aquecimento, suas políticas de escalabilidade só se expandem se o valor métrico das instâncias que não estão se aquecendo for maior do que a meta de utilização da política.

Se o grupo voltar a aumentar a escala na horizontal, as instâncias que ainda estão se aquecendo serão contadas como parte da capacidade desejada para a próxima ação de aumento da escala na horizontal. A intenção é expandir de forma contínua (mas não excessivamente).

Enquanto a atividade de aumentar a escala na horizontal estiver em andamento, todas as atividades de reduzir a escala na horizontal iniciadas por políticas de escalabilidade serão bloqueadas até que as instâncias terminem de aquecer. Quando as instâncias terminarem de se aquecer, se ocorrer um evento de reduzir a escala horizontalmente, todas as instâncias atualmente em processo de encerramento serão contabilizadas na capacidade atual do grupo ao calcular a nova capacidade desejada. Portanto, não removemos mais instâncias do que o necessário do grupo do Auto Scaling.

Valor padrão

Se nenhum valor for definido, a política de escalabilidade usará o valor padrão, que é o valor 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 de instância padrão para facilitar a atualização de todas as políticas de escalabilidade quando o horário de aquecimento mudar.

Considerações

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

- Não crie, edite ou exclua os CloudWatch alarmes usados com uma política de escalabilidade de rastreamento de metas. O Amazon EC2 Auto Scaling cria e gerencia CloudWatch os alarmes associados às suas políticas de escalabilidade de rastreamento de destino 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 faltarem pontos de dados na métrica, isso fará com que o estado do CloudWatch alarme mude para`INSUFFICIENT_DATA`. Quando isso acontece, o Amazon EC2 Auto Scaling não poderá escalar seu grupo até que novos pontos de dados sejam encontrados.
- A matemática métrica pode ser útil se a métrica for intencionalmente relatada de maneira esparsa. Por exemplo, para usar os valores mais recentes, use a função `FILL(m1,REPEAT)`, na qual `m1` é a métrica.
- É possível ver lacunas entre o valor de destino e os pontos de dados de métrica reais. Isso ocorre porque agimos de maneira conservadora arredondando para cima ou para baixo, ao determinarmos quantas instâncias adicionar ou remover. Isso evita a adição de um número insuficiente de instâncias ou remova muitas instâncias. No entanto, para grupos do Auto Scaling menores, com um número menor de instâncias, a utilização do grupo pode parecer distante do valor do objetivo. Por exemplo, vamos supor que você defina um valor de objetivo de 50% para a utilização da CPU, e o seu grupo do Auto Scaling exceda o objetivo. Podemos determinar que a adição de 1,5 instância diminuirá a utilização da CPU em cerca de 50%. Como não é possível adicionar 1,5 instância, arredondamos para cima e adicionamos duas instâncias. Isso pode diminuir a utilização da CPU para um valor abaixo de 50%, mas garante que sua aplicação tenha recursos suficientes para oferecer suporte a ele. Da mesma forma, se determinarmos que 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.

Criar uma política de dimensionamento com monitoramento do objetivo

Para criar uma política de escalabilidade de rastreamento de metas para seu grupo de Auto Scaling, use um dos métodos a seguir.

Antes de começar, confirme se sua métrica preferida está disponível em intervalos de 1 minuto (em comparação com o intervalo padrão de 5 minutos das métricas do Amazon EC2).

Console

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

1. Abra o console do Amazon EC2 em https://console.aws.amazon.com/ec2/ e escolha 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 Escalabilidade, especifique o intervalo no qual você deseja escalar atualizando a capacidade mínima desejada e a capacidade máxima desejada. Essas duas configurações permitem que seu grupo do Auto Scaling seja escalado dinamicamente. Para ter mais informações, consulte Definir limites de escalabilidade para seu grupo do Auto Scaling.
5. Em Escalabilidade automática, escolha Política de escalabilidade com rastreamento do destino.
6. Para definir a política, faça o seguinte:
   a. Especifique um nome para a política.
   b. Escolha uma métrica para o Tipo de métrica.

Se tiver escolhido Application Load Balancer request count per target (Contagem de solicitações do Application Load Balancer por destino), escolha um grupo de destino em Target group (Grupo de destino).

c. Especifique um Target value (Valor de destino) para a métrica.

d. (Opcional) Para aquecimento da instância, atualize o valor do aquecimento da instância conforme necessário.

e. (Opcional) Selecione Disable scale in to create only a scale-out policy (Desabilitar redução para criar somente uma política de expansão). Isso permite que você crie uma política de redução separada de um tipo diferente, se desejado.

7. Prossiga para criar o grupo do Auto Scaling. Sua política de escalabilidade será criada depois que o grupo do Auto Scaling for criado.

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

1. Abra o console do Amazon EC2 em https://console.aws.amazon.com/ec2/ e escolha 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. Verificar se os limites de escalabilidade estão definidos adequadamente. Por exemplo, se sua capacidade desejada já estiver no máximo, especifique um novo máximo para aumentar a escala horizontalmente. Para ter mais informações, consulte Definir limites de escalabilidade para seu grupo do Auto Scaling.
4. Na guia Automatic scaling (Escalabilidade automática), em Dynamic scaling policies (Políticas dinâmicas de escalabilidade), selecione Create dynamic scaling policy (Criar política dinâmica de escalabilidade).
5. Para definir a política, faça o seguinte:

   a. Em Tipo de política, mantenha o padrão de Escalabilidade de rastreamento de destino.

   b. Especifique um nome para a política.

   c. Escolha uma métrica para o Tipo de métrica. É possível escolher apenas um tipo de métrica. Para usar mais de uma métrica, crie várias políticas.

Se você escolheu Application Load Balancer request count per target (Contagem de solicitações do balanceador de carga da aplicação por destino), escolha um grupo de destino em Target group (Grupo de destino).

d. Especifique um Target value (Valor de destino) para a métrica.

e. (Opcional) Para aquecimento da instância, atualize o valor do aquecimento da instância conforme necessário.

f. (Opcional) Selecione Disable scale in to create only a scale-out policy (Desabilitar redução para criar somente uma política de expansão). Isso permite que você crie uma política de redução separada de um tipo diferente, se desejado.

6. Escolha Criar.

AWS CLI

Para criar uma política de escalabilidade de rastreamento de metas, você pode usar o exemplo a seguir para ajudá-lo a começar. Substitua cada *espaço reservado para entrada do usuário* por suas próprias informações.

> **Note**
>
> Para obter mais exemplos, consulte Exemplo de políticas de escalabilidade para a AWS Command Line Interface (AWS CLI).

Para criar uma política de escalabilidade com rastreamento do destino (AWS CLI)

1. Use o `cat` comando a seguir para armazenar um valor alvo para sua política de escalabilidade e uma especificação métrica predefinida em um arquivo JSON nomeado `config.json` em seu diretório inicial. Veja a seguir um exemplo de configuração de rastreamento de metas que mantém a utilização média da CPU em 50%.

```
$ cat ~/config.json
{
  "TargetValue": 50.0,
  "PredefinedMetricSpecification":
    {
      "PredefinedMetricType": "ASGAverageCPUUtilization"
    }
```

```
}
```

Para obter mais informações, consulte a PredefinedMetricSpecificationReferência da API Auto Scaling do Amazon EC2.

2. Use o put-scaling-policycomando, junto com o `config.json` arquivo que você criou na etapa anterior, para criar sua política de escalabilidade.

```
aws autoscaling put-scaling-policy --policy-name cpu50-target-tracking-scaling-policy \
  --auto-scaling-group-name my-asg --policy-type TargetTrackingScaling \
  --target-tracking-configuration file://config.json
```

Se for bem-sucedido, esse comando retornará os ARNs e os nomes dos dois CloudWatch alarmes criados em seu nome.

```
{
    "PolicyARN": "arn:aws:autoscaling:us-
west-2:123456789012:scalingPolicy:228f02c2-c665-4bfd-
aaac-8b04080bea3c:autoScalingGroupName/my-asg:policyName/cpu50-target-tracking-
scaling-policy",
    "Alarms": [
        {
            "AlarmARN": "arn:aws:cloudwatch:us-
west-2:123456789012:alarm:TargetTracking-my-asg-AlarmHigh-
fc0e4183-23ac-497e-9992-691c9980c38e",
            "AlarmName": "TargetTracking-my-asg-AlarmHigh-
fc0e4183-23ac-497e-9992-691c9980c38e"
        },
        {
            "AlarmARN": "arn:aws:cloudwatch:us-
west-2:123456789012:alarm:TargetTracking-my-asg-AlarmLow-61a39305-ed0c-47af-
bd9e-471a352ee1a2",
            "AlarmName": "TargetTracking-my-asg-AlarmLow-61a39305-ed0c-47af-
bd9e-471a352ee1a2"
        }
    ]
}
```

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

Usando a matemática métrica, você pode consultar várias CloudWatch métricas e usar expressões matemáticas para criar novas séries temporais com base nessas métricas. Você pode visualizar as séries temporais resultantes no CloudWatch console e adicioná-las aos painéis. Para obter mais informações sobre matemática métrica, consulte Usando matemática métrica no Guia CloudWatch do usuário da Amazon.

As considerações a seguir se aplicam a expressões matemática em métricas:

- Você pode consultar qualquer CloudWatch métrica disponível. Cada métrica corresponde a uma combinação exclusiva de nome de métrica, espaço nominal e zero ou mais dimensões.
- Você pode usar qualquer operador aritmético (+ - */^), função estatística (como AVG ou SUM) ou outra função compatível. CloudWatch
- Você pode usar as métricas e os resultados de outras expressões matemáticas nas fórmulas da expressão matemática.
- Qualquer expressão usada em uma especificação de métrica deve eventualmente retornar uma única série temporal.
- Você pode verificar se uma expressão matemática métrica é válida usando o CloudWatch console ou a CloudWatch GetMetricDataAPI.

> **Note**
>
> Você pode criar uma política de escalabilidade de rastreamento de metas usando matemática métrica somente se usar o AWS CLI ou um SDK. Esse recurso ainda não está disponível no console AWS CloudFormation e.

Exemplo: lista de pendências da fila do Amazon SQS por instância

Para calcular a lista de pendências da fila do Amazon SQS por instância, use o número aproximado de mensagens disponíveis para recuperação da fila e divida esse número pela capacidade de execução do grupo do, que corresponde ao número de instâncias no estado`InService`. Para ter mais informações, consulte Escalabilidade baseada no Amazon SQS.

A lógica da expressão é a seguinte:

```
sum of (number of messages in the queue)/(number of InService instances)
```

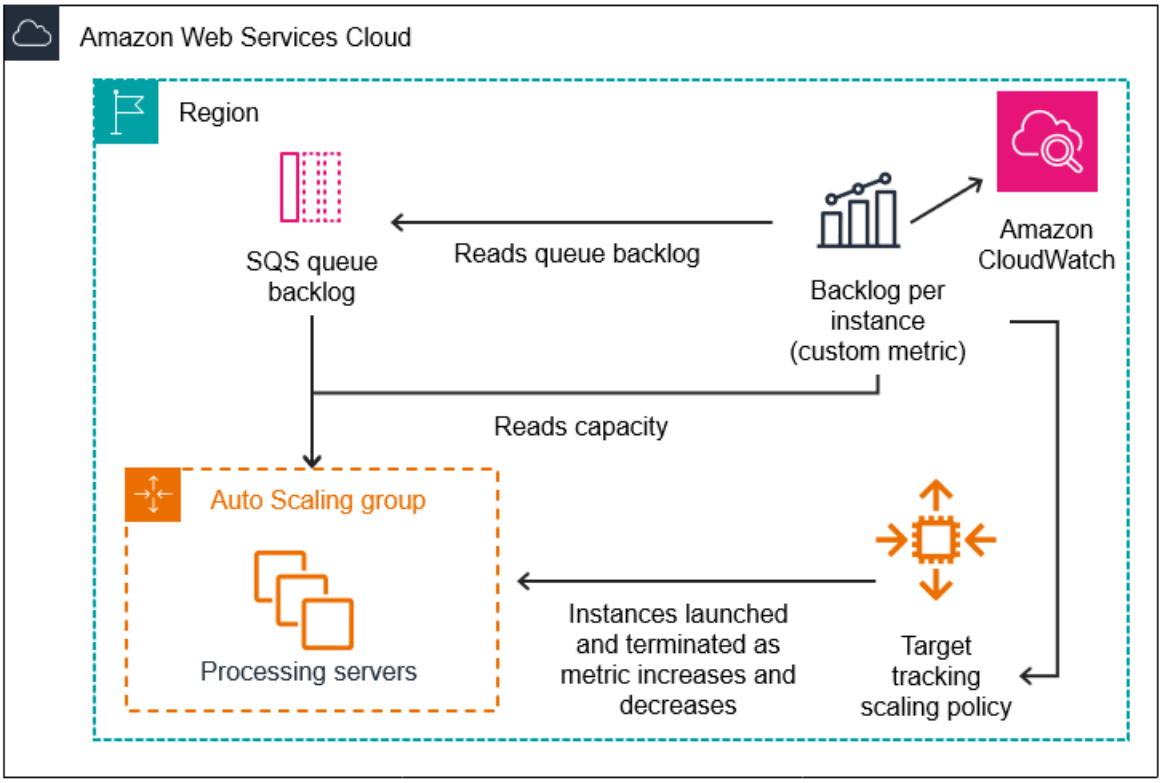
Então, suas informações CloudWatch métricas são as seguintes.

ID	CloudWatch métrica	Estatística	Período
m1	ApproximateNumberOfMessagesVisible	Soma	1 minuto
m2	GroupInServiceInstances	Média	1 minuto

O ID e a expressão matemáticos da métrica são os seguintes:

ID	Expressão
e1	`(m1)/(m2)`

O diagrama a seguir ilustra a arquitetura dessa métrica:



Para usar essa matemática em métricas na criação de uma política de escalabilidade com monitoramento de destino (AWS CLI)

1. Armazene a expressão matemática em métricas como parte de uma especificação de métrica personalizada em um arquivo JSON denominado `config.json`.

   Use o exemplo a seguir como auxílio para começar. Substitua cada *espaço reservado para entrada do usuário* por suas próprias informações.

```
{
    "CustomizedMetricSpecification": {
        "Metrics": [
            {
                "Label": "Get the queue size (the number of messages waiting to be
 processed)",
                "Id": "m1",
                "MetricStat": {
                    "Metric": {
                        "MetricName": "ApproximateNumberOfMessagesVisible",
                        "Namespace": "AWS/SQS",
                        "Dimensions": [
                            {
                                "Name": "QueueName",
                                "Value": "my-queue"
                            }
                        ]
                    },
                    "Stat": "Sum"
                },
                "ReturnData": false
            },
            {
                "Label": "Get the group size (the number of InService instances)",
                "Id": "m2",
                "MetricStat": {
                    "Metric": {
                        "MetricName": "GroupInServiceInstances",
                        "Namespace": "AWS/AutoScaling",
                        "Dimensions": [
                            {
                                "Name": "AutoScalingGroupName",
                                "Value": "my-asg"
                            }
```

```
                    ]
                },
                "Stat": "Average"
            },
            "ReturnData": false
        },
        {
            "Label": "Calculate the backlog per instance",
            "Id": "e1",
            "Expression": "m1 / m2",
            "ReturnData": true
        }
    ]
  },
  "TargetValue": 100
}
```

Para obter mais informações, consulte a TargetTrackingConfigurationReferência da API 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 o nome exato da métrica, o namespace e as dimensões (se aplicável) de uma CloudWatch métrica com o AWS CLI, consulte list-metrics.

2. Para criar essa política, execute o put-scaling-policycomando usando o arquivo JSON como entrada, conforme demonstrado no exemplo a seguir.

```
aws autoscaling put-scaling-policy --policy-name sqs-backlog-target-tracking-scaling-policy \
  --auto-scaling-group-name my-asg --policy-type TargetTrackingScaling \
  --target-tracking-configuration file://config.json
```

Se for bem-sucedido, esse comando retornará o Amazon Resource Name (ARN) da política e os ARNs dos dois CloudWatch alarmes criados em seu nome.


```
{
    "PolicyARN": "arn:aws:autoscaling:us-
west-2:123456789012:scalingPolicy:228f02c2-c665-4bfd-
aaac-8b04080bea3c:autoScalingGroupName/my-asg:policyName/sqs-backlog-target-
tracking-scaling-policy",
    "Alarms": [
        {
            "AlarmARN": "arn:aws:cloudwatch:us-
west-2:123456789012:alarm:TargetTracking-my-asg-AlarmHigh-
fc0e4183-23ac-497e-9992-691c9980c38e",
            "AlarmName": "TargetTracking-my-asg-AlarmHigh-
fc0e4183-23ac-497e-9992-691c9980c38e"
        },
        {
            "AlarmARN": "arn:aws:cloudwatch:us-
west-2:123456789012:alarm:TargetTracking-my-asg-AlarmLow-61a39305-ed0c-47af-
bd9e-471a352ee1a2",
            "AlarmName": "TargetTracking-my-asg-AlarmLow-61a39305-ed0c-47af-
bd9e-471a352ee1a2"
        }
    ]
}
```


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

Políticas de escalabilidade simples e em etapas do Amazon EC2 Auto Scaling

O escalonamento por etapas e as políticas de escalabilidade simples escalam a capacidade do seu grupo de Auto Scaling em incrementos predefinidos com base em alarmes. CloudWatch É possível definir políticas de escalabilidade separadas para lidar com o aumento horizontal da escala (aumento

da capacidade) e com a redução horizontal da escala (diminuição da capacidade) quando um limite de alarme é violado.

Com o escalonamento por etapas e o escalonamento simples, você cria e gerencia os CloudWatch alarmes que invocam o processo de escalabilidade. Quando um alarme é violado, o Amazon EC2 Auto Scaling inicia a política de escalabilidade associada a esse alarme.

É altamente recomendável que você use políticas de escalabilidade de rastreamento de metas para escalar métricas como a utilização média da CPU ou a contagem média de solicitações por alvo. 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 ter mais informações, consulte Políticas de escalabilidade de rastreamento de destino.

Conteúdo

- Funcionamento das políticas de escalabilidade em etapas
- Ajustes em etapas para escalabilidade em etapas
- Tipos de ajuste da escalabilidade
- Aquecimento da instância
- Considerações
- Crie uma política de escalonamento por etapas para expansão horizontal
- Crie uma política de escalonamento por etapas para escalar em
- Políticas de escalabilidade simples

Funcionamento das políticas de escalabilidade em etapas

Para usar o escalonamento por etapas, primeiro você cria um CloudWatch alarme que monitora uma métrica para seu grupo de Auto Scaling. Defina a métrica, o valor limite e o número de períodos de avaliação que determinam uma violação de alarme. Em seguida, crie uma política de escalonamento de etapas que defina como escalar seu grupo quando o limite de alarme for violado.

Adicione os ajustes de etapas na política. É possível definir diferentes ajustes de etapas com base na dimensão da violação do alarme. Por exemplo: .

- Expanda em 10 instâncias se a métrica de alarme atingir 60%

- Expanda em 30 instâncias se a métrica de alarme atingir 75 por cento
- Expanda em 40 instâncias se a métrica de alarme atingir 85%

Quando o limite de alarme for violado para o número especificado de períodos de avaliação, o Amazon EC2 Auto Scaling aplicará os ajustes de etapas definidos na política. Os ajustes podem continuar para violações de alarmes adicionais até que o estado do alarme retorne a 0K.

Cada instância tem um período de aquecimento para evitar que as atividades de escalabilidade sejam muito reativas às mudanças que ocorrem em curtos períodos de tempo. Opcionalmente, você pode configurar o período de aquecimento para sua política de escalabilidade. No entanto, recomendamos usar o aquecimento padrão da instância para facilitar a atualização de todas as políticas de escalabilidade quando o horário de aquecimento mudar. Para ter mais informações, consulte Definir o aquecimento padrão da instância para um grupo do Auto Scaling.

As políticas de escalabilidade simples são semelhantes às políticas de escalabilidade por etapas, exceto que se baseiam em um único ajuste de escalabilidade, com um período de espera entre cada atividade de escalabilidade. Para ter mais informações, consulte Políticas de escalabilidade simples.

Ajustes em etapas para escalabilidade em etapas

Ao criar uma política de escalabilidade em etapas, especifique um ou mais ajustes em etapas que dimensionem automaticamente o número de instâncias de forma dinâmica com base no tamanho da violação do alarme. Cada ajuste em etapas especifica o seguinte:

- Um limite inferior para o valor da métrica
- Um limite superior para o valor da métrica
- O valor de acordo com o qual dimensionar com base no tipo de ajuste de dimensionamento

CloudWatch agrega pontos de dados métricos com base na estatística da métrica associada ao seu CloudWatch alarme. Quando o alarme é violado, a política de dimensionamento apropriada é invocada. O Amazon EC2 Auto Scaling aplica o tipo de agregação aos pontos CloudWatch de dados métricos mais recentes (em oposição aos dados métricos brutos). Ele compara esse valor de métrica agregada com os limites superior e inferior definidos pelo ajustes em etapa para determinar qual deles deve ser executado.

Você especifica os limites superior e inferior em relação ao limite de ruptura. Por exemplo, digamos que você tenha criado um CloudWatch alarme e uma política de expansão para quando a métrica estiver acima de 50%. Em seguida, você criou um segundo alarme e uma política para reduzir a

escala horizontalmente em momentos em que a métrica está abaixo de 50%. Você fez um conjunto de ajustes de etapas com um tipo de ajuste `PercentChangeInCapacity` (ou porcentagem do grupo no console) para cada política:

Exemplo: ajustes em etapas para política de expansão

Limite inferior	Limite superior	Ajuste
0	10	0
10	20	10
20	nulo	30

Exemplo: ajustes em etapas para política de redução

Limite inferior	Limite superior	Ajuste
-10	0	0
-20	-10	-10
nulo	-20	-30

Isso cria a seguinte configuração de escalabilidade.

```
Metric value

-infinity          30%    40%          60%     70%             infinity
-------------------------------------------------------------------------
          -30%      | -10% | Unchanged  | +10%  |       +30%
-------------------------------------------------------------------------
```

Agora, digamos que você use essa configuração de escalabilidade em um grupo de Auto Scaling que tenha uma capacidade atual e uma capacidade desejada de 10. Os pontos a seguir resumem o comportamento da configuração de escalabilidade em relação às capacidades desejada e atual do grupo:

- A capacidade atual e desejada será mantida enquanto o valor agregado da métrica for maior que 40 e menor que 60.

- Se o valor da métrica chegar a 60, a capacidade desejada do grupo aumenta em 1 instância, para 11 instâncias, com base no segundo ajuste em etapas da política de expansão (adicionar 10% de 10 instâncias). Depois que a nova instância estiver em execução e o tempo de aquecimento especificado expirar, a capacidade atual do grupo aumentará para 11 instâncias. Se o valor da métrica subir para 70 mesmo após esse aumento na capacidade, a capacidade desejada do grupo aumentará em outras 3 instâncias, para 14 instâncias. Isso se baseia no ajuste da terceira etapa da política de expansão (adicione 30% de 11 instâncias, 3,3 instâncias, arredondadas para 3 instâncias).
- Se o valor da métrica chegar a 40, a capacidade desejada do grupo será reduzida em 1 instância, para 13 instâncias, com base no segundo ajuste em etapas da política de redução (removerá 10% das 14 instâncias, 1,4 instâncias, arredondadas para 1 instância). Se o valor da métrica cair para 30 mesmo após essa diminuição na capacidade, a capacidade desejada do grupo diminuirá em outras 3 instâncias, para 10 instâncias. Isso se baseia no ajuste da terceira etapa da política de expansão (remova 30% de 13 instâncias, 3,9 instâncias, arredondadas para 3 instâncias).

Ao especificar os ajustes em etapas para sua política de escalabilidade, observe o seguinte:

- Se você usar o AWS Management Console, você especifica os limites superior e inferior como valores absolutos. Se você usa o AWS CLI ou um SDK, especifica os limites superior e inferior em relação ao limite de violação.
- Os intervalos de seus ajustes em etapas não podem se sobrepor ou ter uma lacuna.
- Somente um ajuste em etapas pode ter um limite inferior nulo (infinito negativo). Se um ajuste em etapas tiver um limite inferior negativo, não deverá haver um ajuste em etapas com um limite inferior nulo.
- Somente um ajuste em etapas pode ter um limite superior nulo (infinito positivo). Se um ajuste em etapas tiver um limite superior positivo, deverá haver um ajuste em etapas com um limite superior nulo.
- Os limites inferior e superior não podem ser nulos no mesmo ajuste em etapas.
- Se o valor da métrica estiver acima do limite de violação, o limite inferior será inclusivo e o limite superior será exclusivo. Se o valor da métrica estiver abaixo do limite de violação, o limite inferior será exclusivo e o limite superior será inclusivo.

Tipos de ajuste da escalabilidade

É possível definir uma política de escalabilidade que execute a ação de escalabilidade ideal, com base no tipo de ajuste de escalabilidade escolhido. É possível especificar o tipo de ajuste como um percentual da capacidade atual do seu grupo do Auto Scaling ou em unidades de capacidade. Normalmente, uma unidade de capacidade significa uma instância, a menos que você esteja usando o recurso de pesos de instância.

O Amazon EC2 Auto Scaling oferece suporte aos seguintes tipos de ajuste de escalabilidade simples e em etapa:

- `ChangeInCapacity`: aumentar ou diminuir a capacidade atual do grupo no valor especificado. Um valor de ajuste positivo aumenta a capacidade e um valor negativo diminui a capacidade. Por exemplo: se a capacidade atual do grupo for 3 e o ajuste for 5, quando essa política for executada, adicionaremos 5 unidades de capacidade à capacidade, para um total de 8 unidades de capacidade.
- `ExactCapacity`: alterar a capacidade atual do grupo para o valor especificado. Especifique um valor não negativo com esse tipo de ajuste. Exemplo: se a capacidade atual do grupo for 3 instâncias e o ajuste for 5, quando essa política for executada, alteraremos a capacidade para 5 unidades de capacidade.
- `PercentChangeInCapacity`: aumentar ou diminuir a capacidade atual do grupo no percentual especificado. Um valor positivo aumenta a capacidade e um valor negativo diminui a capacidade. Por exemplo: se a capacidade atual for 10 e o ajuste for 10%, quando essa política for executada, adicionaremos 1 unidade de capacidade à capacidade, para um total de 11 unidades de capacidade.

> **Note**
>
> Se o valor resultante não for um inteiro, o arredondamento é feito da seguinte forma:
>
> - Valores maiores que 1 serão arredondados para baixo. Por exemplo, `12.7` será arredondado para `12`.
> - Os valores entre 0 e 1 serão arredondados para 1. Por exemplo, `.67` será arredondado para `1`.
> - Os valores entre 0 e -1 serão arredondados para -1. Por exemplo, `-.58` será arredondado para `-1`.

- Os valores menores que -1 serão arredondado para cima. Por exemplo, `-6.67` será arredondado para `-6`.

Com `PercentChangeInCapacity`, também é possível especificar o número mínimo de instâncias a serem dimensionadas usando o parâmetro `MinAdjustmentMagnitude`. Por exemplo, vamos supor que você crie uma política que adiciona 25% e especifique um incremento mínimo de 2 instâncias. Se você tiver um grupo do Auto Scaling com 4 instâncias e a política de escalabilidade for executada, 25% de 4 será 1 instância. No entanto, como você especificou um incremento mínimo de 2, serão adicionadas 2 instâncias.

Quando você usa pesos de instância, o efeito de definir o `MinAdjustmentMagnitude` parâmetro para um valor diferente de zero muda. 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 usar pesos de instância, lembre-se de que a capacidade atual do seu grupo de Auto Scaling pode 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 contabilizada nas métricas agregadas da instância EC2 do grupo Auto Scaling.

Enquanto as instâncias estão no período de aquecimento, suas políticas de escalabilidade 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.

Se o grupo voltar a aumentar a escala na horizontal, as instâncias que ainda estão se aquecendo serão contadas como parte da capacidade desejada para a próxima ação de aumento da escala na horizontal. Portanto, várias violações de alarme que caem no intervalo do mesmo ajuste em etapas

resultam em uma única ação de escalabilidade. A intenção é expandir de forma contínua (mas não excessivamente).

Por exemplo, vamos supor que você cria uma política com duas etapas. A primeira etapa adiciona 10 por cento quando a métrica chega a 60, e a segunda etapa adiciona 30 por cento quando a métrica chega a 70 por cento. Seu grupo do Auto Scaling tem uma capacidade desejada e atual de 10. A capacidade atual e desejada não altera enquanto o valor agregado da métrica for menor que 60. Suponha que a métrica chegue a 60, então 1 instância é adicionada (10 por cento de 10 instâncias). Em seguida, a métrica chega a 62 enquanto a nova instância ainda está se aquecendo. A política de escalabilidade calcula a nova capacidade desejada com base na capacidade atual, que ainda é 10. No entanto, a capacidade desejada do grupo já aumentou para 11 instâncias, portanto, a política de escalabilidade não aumenta mais a capacidade desejada. Se a métrica chegar a 70 enquanto a nova instância ainda está em processo de aquecimento, deveremos adicionar 3 instâncias (30% de 10 instâncias). No entanto, como a capacidade desejada do grupo já é 11, adicionaremos apenas 2 instâncias, para uma nova capacidade desejada de 13 instâncias.

Enquanto a atividade de aumentar a escala na horizontal estiver em andamento, todas as atividades de reduzir a escala na horizontal iniciadas por políticas de escalabilidade serão bloqueadas até que as instâncias terminem de aquecer. Quando as instâncias terminarem de se aquecer, se ocorrer um evento de reduzir a escala horizontalmente, todas as instâncias atualmente em processo de encerramento serão contabilizadas na capacidade atual do grupo ao calcular a nova capacidade desejada. Portanto, não removemos mais instâncias do que o necessário do grupo do Auto Scaling. Por exemplo, enquanto uma instância já estiver sendo encerrada, se um alarme estiver em violação no intervalo do mesmo ajuste de etapa que diminuiu a capacidade desejada em 1, nenhuma ação de escalabilidade será realizada.

Valor padrão

Se nenhum valor for definido, a política de escalabilidade usará o valor padrão, que é o valor 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.

Considerações

As considerações a seguir são aplicáveis ao trabalhar com políticas de escalabilidade simples e em etapas:

- Avalie se é possível prever os ajustes em etapas na aplicação com precisão suficiente para usar a escalabilidade em etapas. Se a métrica de escalabilidade aumentar ou diminuir proporcionalmente

à capacidade do destino dimensionável, recomendamos que você use uma política de escalabilidade de rastreamento do objetivo. Você ainda tem a opção de usar a escalabilidade em etapas como política adicional para uma configuração mais avançada. Por exemplo, é possível configurar uma resposta mais agressiva quando a utilização atinge determinado nível.

- Para evitar oscilações, certifique-se de escolher uma margem adequada entre os limites de redução e aumento da escala. Oscilação é um ciclo infinito de aumento e redução de escala horizontal. Ou seja, se o sistema adotar alguma ação de escalabilidade, o valor da métrica mudaria e iniciaria outra ação de escalabilidade na direção inversa.

Crie uma política de escalonamento por etapas para expansão horizontal

Para criar uma política de escalabilidade por etapas para expansão horizontal para seu grupo de Auto Scaling, use um dos seguintes métodos:

Console

Etapa 1: criar um CloudWatch alarme para o limite máximo métrico

1. Abra o CloudWatch console em https://console.aws.amazon.com/cloudwatch/.
2. Se necessário, altere a região da . Na barra de navegação, selecione a região na qual o grupo do Auto Scaling reside.
3. No painel de navegação, escolha Alarms, All alarms (Alarmes, Todos os alarmes) e Create alarm (Criar alarme).
4. Escolha Selecionar métrica.
5. Na guia All metrics (Todas as métricas), escolha EC2, By Auto Scaling Group (Por grupo do Auto Scaling) e insira o nome do grupo do Auto Scaling no campo de pesquisa. Depois, selecione `CPUUtilization` e escolha Selecionar métrica. A página Especificar métrica e condições será exibida, mostrando um gráfico e outras informações sobre a métrica.
6. Em Period (Período), escolha o período de avaliação para o alarme, por exemplo, 1 minuto. Ao avaliar o alarme, todos os períodos são agregados em um único ponto de dados.

   > **Note**
   >
   > Um período mais curto cria um alarme mais sensível.

7. Em Condições, faça o seguinte:

- Em Tipo de limite, escolha Estático.
- Em Whenever **`CPUUtilization`** is, especifique se você deseja que o valor da métrica seja maior ou maior ou igual ao limite para violar o 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.

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 CloudWatch do usuário da Amazon.

9. Escolha Próximo.

   A página Configure actions (Configurar ações) é exibida.

10. Em Notification (Notificação), selecione um tópico do Amazon SNS para notificar quando o alarme estiver no estado `ALARM`, `OK` ou `INSUFFICIENT_DATA`.

    Para que o alarme envie várias notificações para o mesmo estado de alarme ou para diferentes estados de alarme, escolha Add notification (Adicionar notificação).

    Para que o alarme não envie notificações, escolha Remove (Remover).

11. Você pode deixar vazias as outras seções da página Configure actions (Configurar ações). Deixar as outras seções vazias cria um alarme sem associá-lo a uma política de escalabilidade. Em seguida, você pode associar o alarme a uma política de escalabilidade do console do Amazon EC2 Auto Scaling.

12. Escolha Próximo.

13. Insira um nome (por exemplo, `Step-Scaling-AlarmHigh-AddCapacity`) e, opcionalmente, uma descrição para o alarme e escolha Próximo.
14. Selecione Criar alarme.

Use o procedimento a seguir para continuar de onde parou depois de criar o CloudWatch alarme.

Etapa 2: criar uma política de escalonamento por etapas para expansão horizontal

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. Verificar se os limites de escalabilidade estão definidos adequadamente. Por exemplo, se sua capacidade desejada já estiver no máximo, especifique um novo máximo para aumentar a escala horizontalmente. Para ter mais informações, consulte Definir limites de escalabilidade para seu grupo do Auto Scaling.
4. Na guia Automatic scaling (Escalabilidade automática), em Dynamic scaling policies (Políticas dinâmicas de escalabilidade), selecione Create dynamic scaling policy (Criar política dinâmica de escalabilidade).
5. Em Tipo de política, escolha Escalabilidade de etapas e, em seguida, especifique um nome para a política.
6. Para CloudWatch alarme, escolha seu alarme. Se você ainda não criou um alarme, escolha Criar um CloudWatch alarme e conclua as etapas 4 a 14 no procedimento anterior para criar um alarme.
7. Especifique a alteração no tamanho do grupo atual que essa política fará quando executada usando Take the action (Executar a ação). É possível adicionar um número específico de instâncias ou uma porcentagem do tamanho do grupo existente ou definir o grupo para um tamanho exato.

   Por exemplo, para criar uma política de expansão que aumente a capacidade do grupo em 30%, escolhaAdd, insira `30` no próximo campo e escolha. `percent of group` Por padrão, o limite inferior desse ajuste em etapas é o limite do alarme, e o limite superior é positivo (+) infinito.

8. Para adicionar outra etapa, escolha Add step (Adicionar etapa) e defina o valor de acordo com o qual dimensionar e os limites inferior e superior da etapa em relação ao limite do alarme.
9. Para definir um número mínimo de instâncias a serem dimensionadas, atualize o campo número em Add capacity units in increments of at least (Adicionar unidades de capacidade em incrementos de pelo menos) 1 capacity units (unidades de capacidade).
10. (Opcional) Para aquecimento da instância, atualize o valor do aquecimento da instância conforme necessário.
11. Escolha Criar.

AWS CLI

Para criar uma política de escalabilidade por etapas para expansão horizontal (aumentar a capacidade), você pode usar os seguintes exemplos de comandos. Substitua cada *espaço reservado para entrada do usuário* por suas próprias informações.

Ao usar o AWS CLI, primeiro você cria uma política de escalabilidade por etapas que fornece instruções ao Amazon EC2 Auto Scaling sobre como escalar quando o valor de uma métrica está aumentando. Em seguida, você cria o alarme identificando a métrica a ser observada, definindo o limite máximo métrico e outros detalhes para os alarmes e associando o alarme à política de escalabilidade.

Etapa 1: criar uma política para expansão horizontal

Use o put-scaling-policycomando a seguir para criar uma política de escalabilidade de etapas chamada`my-step-scale-out-policy`, com um tipo de ajuste `PercentChangeInCapacity` que aumente a capacidade do grupo com base nos seguintes ajustes de etapa (supondo um limite de CloudWatch alarme de 60%):

- Aumentar a contagem de instâncias em 10 por cento quando o valor da métrica for maior que ou igual a 60 por cento, mas menor que 75 por cento
- Aumentar a contagem de instâncias em 20 por cento quando o valor da métrica for maior que ou igual a 75 por cento, mas menor que 85 por cento
- Aumentar a contagem de instâncias em 30 por cento quando o valor da métrica for maior ou igual 85 por cento

```
aws autoscaling put-scaling-policy \
```

```
  --auto-scaling-group-name my-asg  \
  --policy-name my-step-scale-out-policy \
  --policy-type StepScaling \
  --adjustment-type PercentChangeInCapacity \
  --metric-aggregation-type Average \
  --step-adjustments
 MetricIntervalLowerBound=0.0,MetricIntervalUpperBound=15.0,ScalingAdjustment=10 \

 MetricIntervalLowerBound=15.0,MetricIntervalUpperBound=25.0,ScalingAdjustment=20 \
                     MetricIntervalLowerBound=25.0,ScalingAdjustment=30 \
  --min-adjustment-magnitude 1
```

Anote o nome de recurso da Amazon (ARN) da política. Você precisa dele para criar um CloudWatch alarme para a política.

```
{
    "PolicyARN":
 "arn:aws:autoscaling:region:123456789012:scalingPolicy:4ee9e543-86b5-4121-b53b-
aa4c23b5bbcc:autoScalingGroupName/my-asg:policyName/my-step-scale-in-policy
}
```

Etapa 2: criar um CloudWatch alarme para o limite máximo métrico

Use o CloudWatch put-metric-alarmcomando a seguir para criar um alarme que aumente o tamanho do grupo de Auto Scaling com base em um valor limite médio de CPU de 60% por pelo menos dois períodos consecutivos de avaliação de dois minutos. Para usar sua própria métrica personalizada, especifique o nome em `--metric-name` e o namespace em `--namespace`.

```
aws cloudwatch put-metric-alarm --alarm-name Step-Scaling-AlarmHigh-AddCapacity \
  --metric-name CPUUtilization --namespace AWS/EC2 --statistic Average \
  --period 120 --evaluation-periods 2 --threshold 60 \
  --comparison-operator GreaterThanOrEqualToThreshold \
  --dimensions "Name=AutoScalingGroupName,Value=my-asg" \
  --alarm-actions PolicyARN
```

Crie uma política de escalonamento por etapas para escalar em

Para criar uma política de escalonamento de etapas para escalar seu grupo de Auto Scaling, use um dos seguintes métodos:

Console

Etapa 1: criar um CloudWatch alarme para o limite métrico baixo

1. Abra o CloudWatch console em https://console.aws.amazon.com/cloudwatch/.
2. Se necessário, altere a região da . Na barra de navegação, selecione a região na qual o grupo do Auto Scaling reside.
3. No painel de navegação, escolha Alarms, All alarms (Alarmes, Todos os alarmes) e Create alarm (Criar alarme).
4. Escolha Selecionar métrica.
5. Na guia All metrics (Todas as métricas), escolha EC2, By Auto Scaling Group (Por grupo do Auto Scaling) e insira o nome do grupo do Auto Scaling no campo de pesquisa. Depois, selecione `CPUUtilization` e escolha Selecionar métrica. A página Especificar métrica e condições será exibida, mostrando um gráfico e outras informações sobre a métrica.
6. Em Period (Período), escolha o período de avaliação para o alarme, por exemplo, 1 minuto. Ao avaliar o alarme, todos os períodos são agregados em um único ponto de dados.

   > **Note**
   >
   > Um período mais curto cria um alarme mais sensível.

7. Em Condições, faça o seguinte:

   - Em Tipo de limite, escolha Estático.
   - Em Whenever **`CPUUtilization`** is, especifique se você deseja que o valor da métrica seja menor ou menor ou igual ao limite para violar o alarme. Em than (que), insira o valor do limite desejado de violação de alarme.

     > **Important**
     >
     > Para um alarme a ser usado com uma política para reduzir a escala horizontalmente (alarme inferior), certifique-se de não escolher um valor maior que ou igual ao limite.

8. Em Configuração adicional, faça o seguinte:

   - Em Datapoints to alarm (Pontos de dados para alarme), insira o número de pontos de dados (períodos de avaliação) durante os quais o valor da métrica deverá atender às

condições de limite para o alarme. Por exemplo, com dois períodos consecutivos de 5 minutos, o estado de alarme levaria 10 minutos para ser invocado.

- Em Tratamento de dados ausentes, escolha Tratar dados ausentes como inválidos (limite de violação). Para obter mais informações, consulte Configurando como CloudWatch os alarmes tratam os dados perdidos no Guia CloudWatch do usuário da Amazon.

9. Escolha Próximo.

   A página Configure actions (Configurar ações) é exibida.

10. Em Notification (Notificação), selecione um tópico do Amazon SNS para notificar quando o alarme estiver no estado `ALARM`, `OK` ou `INSUFFICIENT_DATA`.

    Para que o alarme envie várias notificações para o mesmo estado de alarme ou para diferentes estados de alarme, escolha Add notification (Adicionar notificação).

    Para que o alarme não envie notificações, escolha Remove (Remover).

11. Você pode deixar vazias as outras seções da página Configure actions (Configurar ações). Deixar as outras seções vazias cria um alarme sem associá-lo a uma política de escalabilidade. Em seguida, você pode associar o alarme a uma política de escalabilidade do console do Amazon EC2 Auto Scaling.
12. Escolha Próximo.
13. Insira um nome (por exemplo, `Step-Scaling-AlarmLow-RemoveCapacity`) e, opcionalmente, uma descrição para o alarme e escolha Próximo.
14. Selecione Criar alarme.

Use o procedimento a seguir para continuar de onde parou depois de criar o CloudWatch alarme.

Etapa 2: criar uma política de escalonamento por etapas para escalar em

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. Verificar se os limites de escalabilidade estão definidos adequadamente. Por exemplo, se a capacidade desejada do seu grupo já estiver no mínimo, você precisará especificar um

novo mínimo para poder escalar. Para ter mais informações, consulte Definir limites de escalabilidade para seu grupo do Auto Scaling.

4. Na guia Automatic scaling (Escalabilidade automática), em Dynamic scaling policies (Políticas dinâmicas de escalabilidade), selecione Create dynamic scaling policy (Criar política dinâmica de escalabilidade).
5. Em Tipo de política, escolha Escalabilidade de etapas e, em seguida, especifique um nome para a política.
6. Para CloudWatch alarme, escolha seu alarme. Se você ainda não criou um alarme, escolha Criar um CloudWatch alarme e conclua as etapas 4 a 14 no procedimento anterior para criar um alarme.
7. Especifique a alteração no tamanho do grupo atual que essa política fará quando executada usando Take the action (Executar a ação). É possível remover um número específico de instâncias ou uma porcentagem do tamanho do grupo existente ou definir o grupo para um tamanho exato.

   Por exemplo, para criar uma política de escalabilidade que diminua a capacidade do grupo em duas instâncias, escolha`Remove`, insira 2 no próximo campo 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.
8. Para adicionar outra etapa, escolha Add step (Adicionar etapa) e defina o valor de acordo com o qual dimensionar e os limites inferior e superior da etapa em relação ao limite do alarme.
9. Escolha Criar.

AWS CLI

Para criar uma política de escalabilidade por etapas para escalar (diminuir a capacidade), você pode usar os seguintes exemplos de comandos. Substitua cada *`espaço reservado para entrada do usuário`* por suas próprias informações.

Ao usar o AWS CLI, primeiro você cria uma política de escalabilidade por etapas que fornece instruções ao Amazon EC2 Auto Scaling sobre como escalar quando o valor de uma métrica está diminuindo. Em seguida, você cria o alarme identificando a métrica a ser observada, definindo o limite mínimo métrico e outros detalhes para os alarmes e associando o alarme à política de escalabilidade.

Etapa 1: criar uma política para escalar em

Use o put-scaling-policycomando a seguir para criar uma política de escalabilidade de etapas chamadamy-step-scale-in-policy, com um tipo de ajuste ChangeInCapacity que diminui a capacidade do grupo em 2 instâncias quando o CloudWatch alarme associado ultrapassa o valor mínimo do limite métrico.

```
aws autoscaling put-scaling-policy \
  --auto-scaling-group-name my-asg  \
  --policy-name my-step-scale-in-policy \
  --policy-type StepScaling \
  --adjustment-type ChangeInCapacity \
  --step-adjustments MetricIntervalUpperBound=0.0,ScalingAdjustment=-2
```

Anote o nome de recurso da Amazon (ARN) da política. Você precisa dele para criar o CloudWatch alarme para a política.

```
{
    "PolicyARN": "arn:aws:autoscaling:region:123456789012:scalingPolicy:ac542982-cbeb-4294-891c-a5a941dfa787:autoScalingGroupName/my-asg:policyName/my-step-scale-out-policy
}
```

Etapa 2: criar um CloudWatch alarme para o limite métrico baixo

Use o CloudWatch put-metric-alarmcomando a seguir para criar um alarme que diminua o tamanho do grupo de Auto Scaling com base no valor limite médio de CPU de 40 por cento por pelo menos dois períodos de avaliação consecutivos de dois minutos. Para usar sua própria métrica personalizada, especifique o nome em `--metric-name` e o namespace em `--namespace`.

```
aws cloudwatch put-metric-alarm --alarm-name Step-Scaling-AlarmLow-RemoveCapacity \
  --metric-name CPUUtilization --namespace AWS/EC2 --statistic Average \
  --period 120 --evaluation-periods 2 --threshold 40 \
  --comparison-operator LessThanOrEqualToThreshold \
  --dimensions "Name=AutoScalingGroupName,Value=my-asg" \
  --alarm-actions PolicyARN
```

Políticas de escalabilidade simples

Os exemplos a seguir mostram como você pode usar os comandos da CLI para criar políticas de escalabilidade simples. Eles permanecem neste documento como referência para qualquer cliente

que queira usá-los, mas recomendamos que você use políticas de rastreamento de metas ou escalonamento de etapas em vez disso.

Assim como as políticas de escalabilidade por etapas, as políticas de escalabilidade simples exigem que você crie CloudWatch alarmes para suas políticas de escalabilidade. Nas políticas que você cria, você também deve definir se deseja adicionar ou remover instâncias e quantas, ou definir o grupo com um tamanho exato.

Uma das principais diferenças entre políticas de escalabilidade de etapas e políticas de escalabilidade simples são os ajustes de etapas que você obtém com as políticas de escalabilidade de etapas. Com o escalonamento de etapas, você pode fazer alterações maiores ou menores no tamanho do grupo com base nos ajustes de etapas que você especificar.

Uma política de escalabilidade simples também deve aguardar a conclusão de uma atividade de escalonamento em andamento ou a substituição da verificação de integridade e o término de um período de espera antes de responder a alarmes adicionais. Por outro lado, com o escalonamento por etapas, a política continua respondendo a alarmes adicionais, mesmo quando uma atividade de escalonamento ou substituição da verificação de integridade está em andamento. Isso significa que o Amazon EC2 Auto Scaling avalia todas as violações de alarme à medida que recebe as mensagens de alarme. Por isso, recomendamos que você use políticas de escalabilidade por etapas, mesmo que tenha apenas um único ajuste de escalabilidade.

O Amazon EC2 Auto Scaling originalmente oferecia suporte apenas a políticas de escalabilidade simples. Se você criou sua política de escalabilidade antes da introdução das políticas de rastreamento de metas e escalabilidade de etapas, sua política será tratada como uma política de escalabilidade simples.

Crie uma política de escalabilidade simples para expansão horizontal

Use o put-scaling-policycomando 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 ultrapassa o valor máximo do limite métrico.

```
aws autoscaling put-scaling-policy --policy-name my-simple-scale-out-policy \
  --auto-scaling-group-name my-asg --scaling-adjustment 30 \
  --adjustment-type PercentChangeInCapacity
```

Anote o nome de recurso da Amazon (ARN) da política. Você precisa dele para criar o CloudWatch alarme para a política.

Crie uma política de escalabilidade simples para escalar em

Use o put-scaling-policycomando a seguir para criar uma política de escalabilidade simples chamadamy-simple-scale-in-policy, com um tipo de ajuste ChangeInCapacity que diminui a capacidade do grupo em uma instância quando o CloudWatch alarme associado viola o valor mínimo do limite métrico.

```
aws autoscaling put-scaling-policy --policy-name my-simple-scale-in-policy \
  --auto-scaling-group-name my-asg --scaling-adjustment -1 \
  --adjustment-type ChangeInCapacity --cooldown 180
```

Anote o nome de recurso da Amazon (ARN) da política. Você precisa dele para criar o CloudWatch alarme para a política.

Desaquecimento de escalabilidade para o Amazon EC2 Auto Scaling

> **Important**
>
> Como prática recomendada, recomendamos não usar políticas de escalabilidade simples e desaquecimento de escalabilidade. Uma política de escalabilidade com rastreamento do destino ou uma política de escalabilidade em etapas é melhor para a performance da escalabilidade. Para uma política de escalabilidade que altera o tamanho do grupo do Auto Scaling proporcionalmente à medida que o valor da métrica de escalabilidade diminui ou aumenta, recomendamos o monitoramento do objetivo em escalabilidade simples ou escalabilidade em etapas.

Ao criar políticas de escalabilidade simples para seu grupo do Auto Scaling, recomendamos que você configure o esfriamento de escalabilidade ao mesmo tempo.

Após iniciar ou terminar instâncias, o grupo do Auto Scaling espera o período de desaquecimento encerrar antes que qualquer outra ação de escalabilidade iniciada por políticas de escalabilidade simples possa ser iniciada. A intenção do período de esfriamento é impedir que o grupo do Auto Scaling inicie ou encerre outras instâncias antes que os efeitos de atividades de escalabilidade anteriores sejam visíveis.

Suponha, por exemplo, que uma política de escalabilidade simples para utilização da CPU recomende iniciar duas instâncias. O Amazon EC2 Auto Scaling inicia duas instâncias e pausa

as ações de escalabilidade até o período de desaquecimento terminar. Quando o período de desaquecimento terminar, será possível retomar todas as ações de escalabilidade iniciadas por políticas de escalabilidade simples. Se a utilização da CPU violar o limite alto do alarme novamente, o grupo do Auto Scaling aumentará a escala na horizontal novamente, e o período de desaquecimento entrará em vigor novamente. Porém, se duas instâncias forem suficientes para diminuir o valor da métrica, o grupo permanecerá no tamanho atual.

Conteúdo

- Considerações
- hooks do ciclo de vida podem causar mais atrasos
- Alterar o período de desaquecimento padrão
- Definir um período de desaquecimento para políticas de escalabilidade simples específicas

Considerações

As considerações a seguir se aplicam ao trabalhar com políticas de escalabilidade simples e desaquecimentos de escalabilidade:

- As políticas de monitoramento do objetivo e escalabilidade de etapas podem iniciar uma ação de aumento da escala na horizontal imediatamente sem esperar que o período de desaquecimento termine. Em vez disso, sempre que seu grupo de Auto Scaling inicia instâncias, as instâncias individuais têm um período de aquecimento. Para ter mais informações, consulte Definir o aquecimento padrão da instância para um grupo do Auto Scaling.
- Quando uma ação programada começa no horário programado, ela pode acionar uma ação de escalabilidade imediatamente, sem esperar que o período de desaquecimento termine.
- Se uma instância se tornar não íntegra, o Amazon EC2 Auto Scaling não aguardará o fim do período de desaquecimento para substituir a instância não íntegra.
- Quando várias instâncias são iniciadas ou terminadas, o período de desaquecimento (o desaquecimento padrão ou o desaquecimento específico da política de escalabilidade) entra em vigor quando a última instância conclui seu início ou término.
- Quando o grupo do Auto Scaling é escalado manualmente, o padrão é não aguardar o desaquecimento terminar. No entanto, você pode ignorar esse comportamento e respeitar o tempo de recarga padrão ao usar o AWS CLI ou um SDK para escalar manualmente.
- Por padrão, o Elastic Load Balancing aguarda 300 segundos para concluir o processo de cancelamento do registro (descarga da conexão). Se o grupo estiver atrás de um balanceador de

carga do Elastic Load Balancing, ele aguardará que as instâncias de encerramento cancelem o registro antes de iniciar o período de desaquecimento.

hooks do ciclo de vida podem causar mais atrasos

Caso um gancho do ciclo de vida seja invocado, o período de desaquecimento começará após a conclusão da ação do ciclo de vida ou após o período do tempo limite terminar. Por exemplo, considere um grupo do Auto Scaling com um gancho do ciclo de vida para iniciar a instância. Quando a aplicação passa por um aumento na demanda, o grupo executa uma instância para adicionar capacidade. Como há um gancho do ciclo de vida, a instância é colocada em estado de espera, e as ações de escalabilidade causadas por políticas de escalabilidade simples são pausadas. Quando a instância entra no estado `InService`, o período de desaquecimento é iniciado. Quando o período de desaquecimento termina, atividades de políticas de escalabilidade simples são retomadas.

Quando o Elastic Load Balancing está ativado, para fins de escalabilidade, o período de espera começa quando a instância selecionada para encerramento inicia a drenagem da conexão (atraso no cancelamento do registro). O período de resfriamento não espera que a drenagem da conexão termine ou que o gancho do ciclo de vida conclua sua ação. 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 cooldown padrão, use o describe-auto-scaling-groupscomando.

Definir um período de desaquecimento para políticas de escalabilidade simples específicas

Por padrão, todas as políticas de escalabilidade simples usam o período de desaquecimento padrão definido para o grupo do Auto Scaling. Para configurar um período de desaquecimento para poíticas de escalabilidade simples específicas, use o parâmetro de desaquecimento opcional ao criar ou atualizar a política. Quando um período de desaquecimento é especificado para uma política, ele substitui o desaquecimento padrão.

Um uso comum para um período de desaquecimento específico de política de escalabilidade é com uma política de redução da escala na horizontal. Como essa política termina instâncias, o Amazon EC2 Auto Scaling precisa de menos tempo para determinar se deve terminar instâncias adicionais. Encerrar instâncias deve ser uma operação muito mais rápida do que iniciar instâncias. O desaquecimento padrão de 300 segundos é, portanto, muito longo. Nesse caso, um período de desaquecimento específico de política de escalabilidade com um valor inferior para política de redução da escala na horizontal pode ajudar a diminuir custos, permitindo que o grupo reduza a escala na horizontal mais rapidamente.

Para criar ou atualizar políticas de escalabilidade simples no console, escolha a guia Automatic scaling (Escalabilidade automática) depois de criar o grupo. Para criar ou atualizar políticas de escalabilidade simples usando o AWS CLI, use o put-scaling-policycomando. Para ter mais informações, consulte Políticas de escalabilidade simples e em etapas.

Escalabilidade baseada no Amazon SQS

> **Important**
>
> As informações e etapas a seguir mostram como calcular o backlog de filas do Amazon SQS por instância usando o atributo `ApproximateNumberOfMessages` queue antes de publicá-

lo como uma métrica personalizada para. CloudWatch No entanto, é possível economizar o custo e o esforço investidos na publicação de sua própria métrica usando a matemática em métricas. Para ter mais informações, consulte Crie uma política de escalabilidade de rastreamento de destino para Amazon EC2 Auto Scaling usando matemática em métricas.

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.

Conteúdo

- Usar o monitoramento do objetivo com a métrica correta
- Limitações e pré-requisitos
- Configurar escalabilidade baseada no Amazon SQS
- Amazon SQS e proteção contra redução de escala na horizontal de instâncias

Usar o monitoramento do objetivo com a métrica correta

Se você usar uma política de escalabilidade com monitoramento de objetivo baseada em uma métrica de fila do Amazon SQS personalizada, a escalabilidade dinâmica poderá se ajustar à curva de demanda da aplicação de forma mais eficaz. Para obter mais informações sobre como escolher métricas para rastreamento de destino, consulte Escolher métricas.

O problema com o uso de uma métrica do CloudWatch Amazon SQS, como `ApproximateNumberOfMessagesVisible` para rastreamento de metas, é que o número de mensagens na fila pode não mudar proporcionalmente ao tamanho do grupo de Auto Scaling que processa as mensagens da fila. Isso ocorre porque número de mensagens na fila do SQS não define exclusivamente o número de instâncias necessário. O número de instâncias no grupo do Auto Scaling pode ser determinado por vários fatores, incluindo o tempo necessário para processar uma mensagem e a quantidade de latência (atraso na fila) aceitável.

A solução é usar uma métrica backlog por instância com o valor de destino sendo o backlog aceitável por instância a ser mantido. Você pode calcular esses números da seguinte maneira:

- Backlog por instância: para calcular o backlog por instância, comece com o atributo da fila `ApproximateNumberOfMessages` para determinar o comprimento da fila do SQS (número de mensagens disponíveis para recuperação da fila). Divida esse número pela capacidade de execução da frota, que para um grupo do Auto Scaling é o número de instâncias no estado `InService`, para obter o backlog por instância.
- Backlog aceitável por instância: para calcular o valor de destino, primeiro determine o que a aplicação pode aceitar em termos de latência. Depois, pegue o valor de latência aceitável e divida-o pelo tempo médio que uma instância do EC2 leva para processar uma mensagem.

Como exemplo, digamos que você tenha um grupo do Auto Scaling com 10 instâncias e o número de mensagens visíveis na fila (`ApproximateNumberOfMessages`) seja de 1.500. Se o tempo médio de processamento for de 0,1 segundo para cada mensagem e a latência mais longa aceitável for de 10 segundos, o backlog aceitável por instância será 10/0,1, o que equivale a 100 mensagens. Isso significa que 100 é o valor de destino da sua política de rastreamento de destino. O evento de aumento horizontal da escala ocorrerá quando o backlog por instância atingir o valor desejado. Como o backlog por instância já está em 150 mensagens (1.500 mensagens/10 instâncias), o grupo passa por um aumento da escala na horizontal com 5 instâncias para manter a proporção em relação ao valor do objetivo.

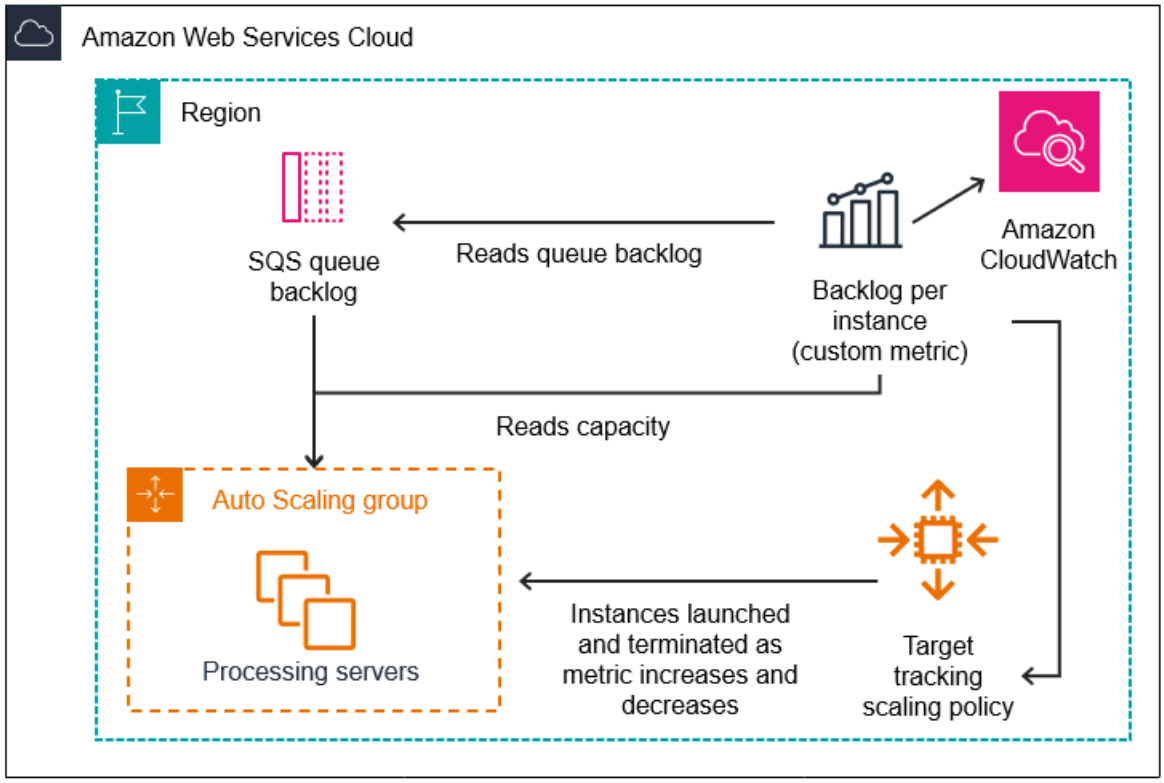
Os procedimentos a seguir demonstram como publicar a métrica personalizada e criar a política de escalabilidade com monitoramento do objetivo que configura o grupo do Auto Scaling para escalar com base nesses cálculos.

> **Important**
>
> Para reduzir custos, lembre-se de usar a matemática em métricas. Para ter mais informações, consulte Crie uma política de escalabilidade de rastreamento de destino para Amazon EC2 Auto Scaling usando matemática em métricas.

Existem três partes principais nessa configuração:

- Um grupo do Auto Scaling para gerenciar instâncias do EC2 para fins de processamento de mensagens de uma fila do SQS.
- Uma métrica personalizada para enviar à Amazon CloudWatch que mede o número de mensagens na fila por instância do EC2 no grupo Auto Scaling.
- Uma política de rastreamento de metas que configura seu grupo de Auto Scaling para escalar com base na métrica personalizada e em um valor alvo definido. CloudWatch os alarmes invocam a política de escalabilidade.

O diagrama a seguir ilustra a arquitetura dessa configuração.



Limitações 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 no. CloudWatch Em seguida, você pode monitorar sua métrica com AWS Management Console o.
- 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 o AWS CLI ou um SDK para especificar uma métrica personalizada para sua política de escalabilidade.

As seções a seguir orientam você a usar o AWS CLI para as tarefas que você precisa realizar. Por exemplo, para obter dados métricos que reflitam 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
- Etapa 2: Criar uma política de escalabilidade com monitoramento do objetivo
- Etapa 3: Testar sua política de escalabilidade

Etapa 1: criar uma métrica CloudWatch personalizada

Uma métrica personalizada é definida usando um nome de métrica e um namespace de sua escolha. Namespaces para métricas personalizadas não podem começar com `AWS/`. Para obter mais informações sobre a publicação de métricas personalizadas, consulte o tópico Publicar métricas personalizadas no Guia CloudWatch do usuário da Amazon.

Siga este procedimento para criar a métrica personalizada lendo primeiro as informações da sua AWS conta. Depois, calcule a métrica de backlog por instância, conforme recomendado em uma seção anterior. Por fim, publique esse número com uma CloudWatch granularidade de 1 minuto. Sempre que possível, é altamente recomendável que você escale as métricas com uma

granularidade de um minuto para garantir uma resposta mais rápida às alterações na carga do sistema.

Para criar uma métrica CloudWatch personalizada (AWS CLI)

1. Use o get-queue-attributescomando 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 describe-auto-scaling-groupscomando para obter a capacidade de execução do grupo, que é o número de instâncias no estado do `InService` ciclo de vida. Esse comando retorna as instâncias de um grupo do Auto Scaling juntamente com seu estado de ciclo de vida.

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

3. Calcule o backlog por instância dividindo o número aproximado de mensagens disponíveis para recuperação da fila pela capacidade de execução do grupo.
4. Crie um script que seja executado a cada minuto para recuperar o valor do backlog por instância e publicá-lo em uma métrica CloudWatch personalizada. Ao publicar uma métrica personalizada, você especifica o nome da métrica, o espaço nominal, a unidade, o valor e zero ou mais dimensões. Uma dimensão consiste em um nome e um valor de dimensão.

   Para publicar sua métrica personalizada, substitua os valores do espaço reservado em *itálico* pelo nome da métrica preferida, pelo valor da métrica, por um namespace (desde que não comece com "AWS“) e pelas dimensões (opcional) e execute o comando a seguir. 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 CloudWatch. A métrica é visível no CloudWatch console. Você pode acessá-lo fazendo login AWS Management Console e navegando até a CloudWatch página. Depois, visualize a métrica navegando até a página de métricas ou procurando-a usando a caixa de pesquisa. Para obter

informações sobre métricas de visualização, consulte Visualizar métricas disponíveis no Guia CloudWatch do usuário da Amazon.

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

A métrica que você criou agora pode ser adicionada a uma política de escalabilidade com rastreamento de destino.

Para criar uma política de escalabilidade com rastreamento do destino (AWS CLI)

1. Use o comando `cat` a seguir para especificar um valor de destino para sua política de escalabilidade e uma especificação de métrica personalizada em um arquivo JSON chamado `config.json` em seu diretório inicial. Substitua cada *espaço reservado para entrada do usuário* por suas próprias informações. Para o `TargetValue`, calcule a métrica backlog aceitável por instância e insira-a aqui. Para calcular esse número, decida um valor de latência normal e divida-o pelo tempo médio necessário para processar uma mensagem, conforme descrito em uma seção anterior.

   Se você não especificou nenhuma dimensão para a métrica criada na etapa 1, não inclua nenhuma dimensão na especificação métrica personalizada.

   ```
   $ cat ~/config.json
   {
      "TargetValue":100,
      "CustomizedMetricSpecification":{
         "MetricName":"MyBacklogPerInstance",
         "Namespace":"MyNamespace",
         "Dimensions":[
            {
               "Name":"MyOptionalMetricDimensionName",
               "Value":"MyOptionalMetricDimensionValue"
            }
         ],
         "Statistic":"Average",
         "Unit":"None"
      }
   }
   ```

2. Use o put-scaling-policycomando, junto com o `config.json` arquivo que você criou na etapa anterior, para criar sua política de escalabilidade.

```
aws autoscaling put-scaling-policy --policy-name sqs100-target-tracking-scaling-
policy \
  --auto-scaling-group-name my-asg --policy-type TargetTrackingScaling \
  --target-tracking-configuration file://~/config.json
```

Isso cria dois alarmes: um para escalabilidade e outro para dimensionamento. Ele também retorna o Amazon Resource Name (ARN) da política registrada CloudWatch, que CloudWatch usa para invocar a escalabilidade sempre que o limite métrico é violado.

Etapa 3: Testar sua política de escalabilidade

Depois que a configuração estiver concluída, verifique se a sua política de escalabilidade está funcionando. É possível testá-la aumentando o número de mensagens na fila do SQS e verificando se o grupo do Auto Scaling iniciou uma instância do EC2 adicional. Também é possível testá-la diminuindo o número de mensagens na fila do SQS e verificando se o grupo do Auto Scaling terminou uma instância do EC2.

Para testar a função de expansão

1. Siga as etapas em Criar uma fila padrão do Amazon SQS e enviar uma mensagem ou Criar uma fila FIFO do Amazon SQS e enviar uma mensagem para adicionar mensagens à sua fila. Certifique-se de que você aumentou o número de mensagens na fila para que a métrica backlog por instância exceda o valor de destino.

   Pode levar alguns minutos para que as alterações invoquem o alarme.
2. Use o describe-auto-scaling-groupscomando para verificar se o grupo iniciou uma instância.

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

Para testar a função de redução

1. Siga as etapas em Receber e excluir uma mensagem (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 describe-auto-scaling-groupscomando 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 ter mais informações, consulte Crie seus aplicativos no Amazon EC2 Auto Scaling para lidar com o encerramento de instâncias sem problemas.

Verificar uma ação de escalabilidade para um grupo do Auto Scaling

Na seção Amazon EC2 Auto Scaling do console do Amazon EC2, o Activity history (Histórico de atividades) para um grupo do Auto Scaling permite exibir o status atual de uma ação de escalabilidade que esteja em andamento. Quando a ação de escalabilidade estiver concluída, você poderá ver se ela foi bem-sucedida ou não. Isso é particularmente útil quando você está criando grupos do Auto Scaling ou adicionando condições de escalabilidade a grupos existentes.

Quando você adiciona uma etapa do monitoramento do objetivo, ou política de escalabilidade simples ao grupo do Auto Scaling, o Amazon EC2 Auto Scaling começa imediatamente a avaliar a política em relação à métrica. O alarme da métrica passa para o estado ALARM (ALARME) quando

a métrica viola o limite em um determinado número de períodos de avaliação. Isso significa que uma política de escalabilidade pode resultar em uma ação de escalabilidade logo após sua criação. Depois de o Amazon EC2 Auto Scaling alterar a capacidade desejada em resposta a uma política de escalabilidade, é possível verificar a ação de escalabilidade em sua conta. Se deseja receber uma notificação por email do Amazon EC2 Auto Scaling informando sobre uma ação de escalabilidade, siga as instruções em Opções de notificação do Amazon SNS para o Amazon EC2 Auto Scaling.

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

Para visualizar as ações de escalabilidade do seu grupo do Auto Scaling (console)

1. Abra o console do Amazon EC2 em https://console.aws.amazon.com/ec2/ e escolha grupos do Auto Scaling no painel de navegação.
2. Na barra de navegação na parte superior da tela, escolha a região onde está o seu grupo do Auto Scaling.
3. Marque a caixa de seleção ao lado do grupo do Auto Scaling.

   Um painel dividido é aberto na parte inferior da página.
4. Na guia Activity (Atividade), em Activity history (Histórico de atividades), a coluna Status (Status) mostra se o seu grupo do Auto Scaling iniciou ou encerrou instâncias com êxito, ou se a ação de escalabilidade ainda está em andamento.
5. (Opcional) Se houver muitas ações de escalabilidade, você poderá escolher o ícone > na borda superior do histórico de atividades para ver a próxima página de ações de escalabilidade.
6. Na guia Instance management (Gerenciamento de instâncias), em Instances (Instâncias), a coluna Lifecycle (Ciclo de vida) contém o estado das suas instâncias. Após a instância iniciar e todos os ganchos do ciclo de vida terminarem, seu estado 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, um exemplo de saída.

As ações de escalabilidade são ordenadas por horário de início. As atividades ainda em andamento são descritas primeiro.

```
{
  "Activities": [
    {
      "ActivityId": "5e3a1f47-2309-415c-bfd8-35aa06300799",
      "AutoScalingGroupName": "my-asg",
      "Description": "Terminating EC2 instance: i-06c4794c2499af1df",
      "Cause": "At 2020-02-11T18:34:10Z a monitor alarm TargetTracking-my-asg-AlarmLow-
b9376cab-18a7-4385-920c-dfa3f7783f82 in state ALARM triggered policy my-target-
tracking-policy changing the desired capacity from 3 to 2.  At 2020-02-11T18:34:31Z
 an instance was taken out of service in response to a difference between desired and
 actual capacity, shrinking the capacity from 3 to 2.  At 2020-02-11T18:34:31Z instance
 i-06c4794c2499af1df was selected for termination.",
      "StartTime": "2020-02-11T18:34:31.268Z",
      "EndTime": "2020-02-11T18:34:53Z",
      "StatusCode": "Successful",
      "Progress": 100,
      "Details": "{\"Subnet ID\":\"subnet-5ea0c127\",\"Availability Zone\":\"us-west-2a
\"...}",
      "AutoScalingGroupARN": "arn"
    },
...
  ]
}
```

Para obter uma descrição dos campos na saída, consulte Atividade na Referência da API do Amazon EC2 Auto Scaling.

Para obter ajuda para recuperar as atividades de dimensionamento para um grupo excluído e obter informações sobre os tipos de erros que você pode encontrar e como tratá-los, consulte Solucionar problemas do Amazon EC2 Auto Scaling.

Desabilitar uma política de escalabilidade para um grupo do Auto Scaling

Este tópico descreve como desabilitar temporariamente uma política de escalabilidade para que ela não inicie alterações no número de instâncias no grupo do Auto Scaling. Quando você desabilita uma política de escalabilidade, os detalhes de configuração são preservados, para que seja possível habilitar novamente e rapidamente a política. Isso é mais fácil do que excluir temporariamente uma política quando ela não é necessária e recriá-la mais tarde.

Quando uma política de escalabilidade é desabilitada, o grupo do Auto Scaling não sofre aumento ou redução de escala na horizontal para os alarmes de métrica que são violados enquanto a política de escalabilidade está desabilitada. No entanto, as ações de escalabilidade ainda em andamento não são interrompidas.

Observe que as políticas de escalabilidade desabilitadas ainda são contabilizadas em relação às suas cotas para o número de políticas de escalabilidade que podem ser adicionadas a um grupo do Auto Scaling.

Para desabilitar uma política de escalabilidade (console)

1. Abra o console do Amazon EC2 em https://console.aws.amazon.com/ec2/ e escolha 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 put-scaling-policycomando com a `--no-enabled` opção a seguir. 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 put-scaling-policycomando com a `--enabled` opção a seguir. 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, 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 você também precise excluir os CloudWatch alarmes. A exclusão de uma política de escalabilidade de rastreamento de metas também exclui todos os alarmes associados. CloudWatch A exclusão de uma política de escalonamento de etapas ou de uma política de escalabilidade simples exclui a ação de alarme subjacente, mas não exclui o CloudWatch alarme, mesmo que não tenha mais uma ação associada.

Para excluir uma política de escalabilidade (console)

1. Abra o console do Amazon EC2 em https://console.aws.amazon.com/ec2/ e escolha 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 Sim, excluir.
6. (Opcional) Se você excluiu uma política de escalabilidade por etapas ou uma política de escalabilidade simples, faça o seguinte para excluir o CloudWatch alarme associado à política. É possível ignorar essas subetapas para manter o alarme para uso futuro.

    a. Abra o CloudWatch console em https://console.aws.amazon.com/cloudwatch/.
    b. No painel de navegação, escolha Alarms (Alarmes).
    c. Escolha o alarme (por exemplo, `Step-Scaling-AlarmHigh-AddCapacity`) e escolha Action (Ação) e Delete (Excluir).
    d. Quando a confirmação for solicitada, escolha Excluir.

Para obter as políticas de escalabilidade para um grupo do Auto Scaling (AWS CLI)

Antes de excluir uma política de escalabilidade, use o seguinte comando describe-policies para ver quais políticas de escalabilidade foram criadas para o grupo do Auto Scaling. Você pode usar a saída ao excluir a política e os CloudWatch alarmes.

```
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, um exemplo de saída.

```
[
    {
        "AutoScalingGroupName": "my-asg",
```

```
        "PolicyName": "cpu50-target-tracking-scaling-policy",
        "PolicyARN": "PolicyARN",
        "PolicyType": "TargetTrackingScaling",
        "StepAdjustments": [],
        "Alarms": [
            {
                "AlarmARN": "arn:aws:cloudwatch:us-
west-2:123456789012:alarm:TargetTracking-my-asg-AlarmHigh-
fc0e4183-23ac-497e-9992-691c9980c38e",
                "AlarmName": "TargetTracking-my-asg-AlarmHigh-
fc0e4183-23ac-497e-9992-691c9980c38e"
            },
            {
                "AlarmARN": "arn:aws:cloudwatch:us-
west-2:123456789012:alarm:TargetTracking-my-asg-AlarmLow-61a39305-ed0c-47af-
bd9e-471a352ee1a2",
                "AlarmName": "TargetTracking-my-asg-AlarmLow-61a39305-ed0c-47af-
bd9e-471a352ee1a2"
            }
        ],
        "TargetTrackingConfiguration": {
            "PredefinedMetricSpecification": {
                "PredefinedMetricType": "ASGAverageCPUUtilization"
            },
            "TargetValue": 50.0,
            "DisableScaleIn": false
        },
        "Enabled": true
    }
]
```

Para excluir a política de dimensionamento (AWS CLI)

Use o comando delete-scaling-policy.

```
aws autoscaling delete-policy --auto-scaling-group-name my-asg \
  --policy-name cpu50-target-tracking-scaling-policy
```

Para excluir seu CloudWatch alarme (AWS CLI)

Para políticas de escalabilidade escalonadas e simples, use o comando delete-alarms para excluir o CloudWatch alarme associado à política. Você pode ignorar essa etapa para manter o alarme para uso futuro. É possível excluir um ou mais alarmes por vez. Por exemplo, use o comando a

seguir para excluir os alarmes `Step-Scaling-AlarmHigh-AddCapacity` e `Step-Scaling-AlarmLow-RemoveCapacity`.

```
aws cloudwatch delete-alarms --alarm-name Step-Scaling-AlarmHigh-AddCapacity Step-Scaling-AlarmLow-RemoveCapacity
```

Exemplo de políticas de escalabilidade para a AWS Command Line Interface (AWS CLI)

Você pode criar políticas de escalabilidade para o Amazon EC2 Auto Scaling por meio AWS Management Console do,, ou SDKs. AWS CLI

Os exemplos a seguir mostram como você pode criar políticas de escalabilidade para o Amazon EC2 Auto Scaling com o comando. AWS CLI put-scaling-policy Substitua cada *espaço reservado para entrada do usuário* por suas próprias informações.

Para começar a escrever políticas de escalabilidade usando o AWS CLI, consulte os exercícios introdutórios em e. Políticas de escalabilidade de rastreamento de destino Políticas de escalabilidade simples e em etapas

Exemplo 1: como aplicar uma política de escalabilidade com monitoramento do objetivo com uma especificação de métrica predefinida

```
aws autoscaling put-scaling-policy --policy-name cpu50-target-tracking-scaling-policy \
  --auto-scaling-group-name my-asg --policy-type TargetTrackingScaling \
  --target-tracking-configuration file://config.json
{
  "TargetValue": 50.0,
  "PredefinedMetricSpecification": {
    "PredefinedMetricType": "ASGAverageCPUUtilization"
  }
}
```

Para obter mais informações, consulte a PredefinedMetricSpecificationReferência da API Auto Scaling do Amazon EC2.

> **Note**
>
> Se o arquivo não estiver no diretório atual, digite o caminho completo para o arquivo. Para obter mais informações sobre a leitura de valores de AWS CLI parâmetros de um arquivo,

consulte Carregamento de AWS CLI parâmetros de um arquivo no Guia AWS Command Line Interface do usuário.

Exemplo 2: como aplicar uma política de escalabilidade com monitoramento do objetivo com uma especificação de métrica personalizada

```
aws autoscaling put-scaling-policy --policy-name sqs100-target-tracking-scaling-policy
 \
  --auto-scaling-group-name my-asg --policy-type TargetTrackingScaling \
  --target-tracking-configuration file://config.json
{
  "TargetValue": 100.0,
  "CustomizedMetricSpecification": {
    "MetricName": "MyBacklogPerInstance",
    "Namespace": "MyNamespace",
    "Dimensions": [{
      "Name": "MyOptionalMetricDimensionName",
      "Value": "MyOptionalMetricDimensionValue"
    }],
    "Statistic": "Average",
    "Unit": "None"
  }
}
```

Para obter mais informações, consulte a CustomizedMetricSpecificationReferência da API Auto Scaling do Amazon EC2.

Exemplo 3: como aplicar uma política de escalabilidade com monitoramento do objetivo somente para expansão

```
aws autoscaling put-scaling-policy --policy-name alb1000-target-tracking-scaling-policy
 \
  --auto-scaling-group-name my-asg --policy-type TargetTrackingScaling \
  --target-tracking-configuration file://config.json
{
  "TargetValue": 1000.0,
  "PredefinedMetricSpecification": {
    "PredefinedMetricType": "ALBRequestCountPerTarget",
    "ResourceLabel": "app/my-alb/778d41231b141a0f/targetgroup/my-alb-target-
group/943f017f100becff"
  },
```

```
  "DisableScaleIn": true
}
```

Exemplo 4: como aplicar uma política de escalabilidade em etapas para expansão

```
aws autoscaling put-scaling-policy \
  --auto-scaling-group-name my-asg  \
  --policy-name my-step-scale-out-policy \
  --policy-type StepScaling \
  --adjustment-type PercentChangeInCapacity \
  --metric-aggregation-type Average \
  --step-adjustments
 MetricIntervalLowerBound=10.0,MetricIntervalUpperBound=20.0,ScalingAdjustment=10 \

 MetricIntervalLowerBound=20.0,MetricIntervalUpperBound=30.0,ScalingAdjustment=20 \
                     MetricIntervalLowerBound=30.0,ScalingAdjustment=30 \
  --min-adjustment-magnitude 1
```

Anote o nome de recurso da Amazon (ARN) da política. Você precisa do ARN ao criar o CloudWatch alarme.

Exemplo 5: como aplicar uma política de escalabilidade em etapas para redução

```
aws autoscaling put-scaling-policy \
  --auto-scaling-group-name my-asg  \
  --policy-name my-step-scale-in-policy \
  --policy-type StepScaling \
  --adjustment-type ChangeInCapacity \
  --step-adjustments MetricIntervalUpperBound=0.0,ScalingAdjustment=-2
```

Anote o nome de recurso da Amazon (ARN) da política. Você precisa do ARN ao criar o CloudWatch alarme.

Exemplo 6: como aplicar uma política de escalabilidade simples para expansão

```
aws autoscaling put-scaling-policy --policy-name my-simple-scale-out-policy \
  --auto-scaling-group-name my-asg --scaling-adjustment 30 \
  --adjustment-type PercentChangeInCapacity --min-adjustment-magnitude 2
```

Anote o nome de recurso da Amazon (ARN) da política. Você precisa do ARN ao criar o CloudWatch alarme.

Exemplo 7: como aplicar uma política de escalabilidade simples para redução

```
aws autoscaling put-scaling-policy --policy-name my-simple-scale-in-policy \
  --auto-scaling-group-name my-asg --scaling-adjustment -1 \
  --adjustment-type ChangeInCapacity --cooldown 180
```

Anote o nome de recurso da Amazon (ARN) da política. Você precisa do ARN ao criar o CloudWatch alarme.

Escala preditiva para o Amazon EC2 Auto Scaling

O escalonamento preditivo funciona analisando dados históricos de carga para detectar padrões diários ou semanais nos fluxos de tráfego. Ele usa essas informações para prever as necessidades futuras de capacidade para que o Amazon EC2 Auto Scaling possa aumentar proativamente a capacidade do seu grupo de Auto Scaling de acordo com a carga prevista.

A escalabilidade preditiva é adequada para situações em que há:

- Tráfego cíclico, como alta utilização de recursos durante o horário comercial e baixa utilização de recursos durante a noite e nos fins de semana
- Padrões on-and-off de carga de trabalho recorrentes, como processamento em lote, testes ou análise periódica de dados
- Aplicações que demoram muito para inicializar, causando um impacto de latência considerável na performance da aplicação durante eventos de aumento da escala na horizontal

Em geral, se houver padrões regulares de aumento de tráfego e aplicações que demoram muito para inicializar, considere o uso da escalabilidade preditiva. A escalabilidade preditiva pode ajudar você a expandir mais rapidamente, lançando a capacidade antes da carga prevista, em comparação com o uso apenas da escalabilidade dinâmica, que é reativa por natureza. O escalonamento preditivo também pode potencialmente economizar dinheiro em sua fatura do EC2, ajudando você a evitar a necessidade de provisionar demais a capacidade.

Por exemplo, considere uma aplicação com elevado índice de utilização durante o horário comercial e baixo uso durante a noite. No início de cada dia útil, a escalabilidade preditiva pode adicionar capacidade antes do primeiro fluxo de tráfego. Isso ajuda sua aplicação a manter alta disponibilidade e performance ao passar de um período de menor utilização para um período de maior utilização. Você não precisa esperar que a escalabilidade dinâmica reaja à mudança de tráfego. Você

também não precisa gastar tempo revisando os padrões de carga da aplicação e tentando alocar a quantidade certa de capacidade usando a escalabilidade programada.

Tópicos

- Como a escalabilidade preditiva funciona
- Crie uma política de escalabilidade preditiva
- Avaliar as políticas de escalabilidade preditiva
- Substituir valores de previsão usando ações programadas
- Configurações avançadas de política de escalabilidade preditiva usando métricas personalizadas

Como a escalabilidade preditiva funciona

Este tópico explica como a escalabilidade preditiva funciona e descreve o que considerar ao criar uma política de escalabilidade preditiva.

Tópicos

- Como funciona
- Limite máximo de capacidade
- Considerações
- Regiões compatíveis

Como funciona

Para usar a escala preditiva, crie uma política de escalabilidade preditiva que especifique a CloudWatch métrica a ser monitorada e analisada. Para que a escala preditiva comece a prever valores futuros, essa métrica deve ter pelo menos 24 horas de dados.

Depois de criar a política, a escalabilidade preditiva começa a analisar os dados métricos dos últimos 14 dias para identificar padrões. Ele usa essa análise para gerar uma previsão horária dos requisitos de capacidade para as próximas 48 horas. A previsão é atualizada a cada 6 horas usando os CloudWatch dados mais recentes. À medida que novos dados chegam, a escala preditiva é capaz de melhorar continuamente a precisão das previsões futuras.

Quando você ativa a escala preditiva pela primeira vez, ela é executada somente no modo de previsão. Nesse modo, ele gera previsões de capacidade, mas na verdade não escala seu grupo de Auto Scaling com base nessas previsões. Isso permite avaliar a precisão e a

adequação da previsão. Você pode visualizar os dados de previsão usando a operação da `GetPredictiveScalingForecast` API ou AWS Management Console o.

Depois de analisar os dados de previsão e decidir começar a escalar com base nesses dados, mude a política de escalabilidade para o modo de previsão e escala. Neste modo:

- Se a previsão espera um aumento na carga, o Amazon EC2 Auto Scaling aumentará a capacidade com a escalabilidade horizontal.
- Se a previsão esperar uma diminuição na carga, ela não será ampliada para remover a capacidade. Se quiser remover a capacidade que não é mais necessária, você deve criar políticas de escalabilidade dinâmica.

Por padrão, o Amazon EC2 Auto Scaling escala seu grupo de Auto Scaling no início de cada hora com base na previsão daquela hora. Opcionalmente, você pode especificar um horário de início anterior usando a `SchedulingBufferTime` propriedade na operação da `PutScalingPolicy` API ou a configuração de instâncias de pré-lançamento no. AWS Management Console Isso faz com que o Amazon EC2 Auto Scaling lance novas instâncias antes da demanda prevista, dando a elas tempo para inicializar e se preparar para lidar com o tráfego.

Para oferecer suporte ao lançamento de novas instâncias antes da demanda prevista, é altamente recomendável que você ative o aquecimento de instâncias padrão para seu grupo de Auto Scaling. Isso especifica um período após uma atividade de escalabilidade horizontal durante o qual o Amazon EC2 Auto Scaling não será escalado, mesmo que as políticas de escalabilidade dinâmica indiquem que a capacidade deve ser reduzida. Isso ajuda você a garantir que as instâncias recém-lançadas tenham tempo suficiente para começar a atender ao aumento do tráfego antes de serem consideradas para operações de expansão. Para ter mais informações, consulte Definir o aquecimento padrão da instância para um grupo do Auto Scaling.

Limite máximo de capacidade

Os grupos do Auto Scaling têm uma configuração de capacidade máxima que limita o número máximo de instâncias do EC2 que podem ser executadas para o grupo. Por padrão, quando as políticas de escalabilidade são definidas, elas não podem aumentar a capacidade acima da capacidade máxima.

Como alternativa, você pode permitir que a capacidade máxima do grupo seja aumentada automaticamente se a capacidade prevista se aproximar ou exceder a capacidade máxima do grupo Auto Scaling. Para ativar esse comportamento, use as `MaxCapacityBuffer` propriedades

`MaxCapacityBreachBehavior` e na operação da `PutScalingPolicy` API ou a configuração de comportamento de capacidade máxima no AWS Management Console.

> **Warning**
>
> Tenha cuidado ao permitir que a capacidade máxima seja aumentada automaticamente. Isso pode fazer com que mais instâncias sejam lançadas do que o pretendido se o aumento da capacidade máxima não for monitorado e gerenciado. A capacidade máxima aumentada então se torna a nova capacidade máxima normal para o grupo Auto Scaling até que você a atualize manualmente. A capacidade máxima não diminui automaticamente de volta ao máximo original.

Considerações

- Confirme se a escalabilidade preditiva é adequada para sua workload. Uma workload será uma boa opção para o uso da escalabilidade preditiva se ela apresentar padrões de carga recorrentes específicos do dia da semana ou da hora do dia. Para verificar isso, configure políticas de escalabilidade preditiva no modo somente previsão e consulte as recomendações do console. O Amazon EC2 Auto Scaling fornece recomendações com base em observações sobre a performance potencial da política. Avalie a previsão e as recomendações antes de permitir que a escalabilidade preditiva escale ativamente sua aplicação.
- A escalabilidade preditiva precisa de pelo menos 24 horas de dados históricos para começar a previsão. No entanto, as previsões serão mais eficazes se os dados históricos abrangerem duas semanas completas. Se você atualizar sua aplicação criando um novo do grupo do Auto Scaling e excluindo o antigo, o novo grupo do Auto Scaling precisará de 24 horas de dados históricos de carga antes que a escalabilidade preditiva possa começar a gerar previsões novamente. É possível usar métricas personalizadas para agregar métricas em grupos do Auto Scaling novos e antigos. Senão, talvez seja necessário esperar alguns dias para obter uma previsão mais precisa.
- Escolha uma métrica de carga que represente com precisão a carga total do seu aplicativo e seja o aspecto do seu aplicativo que é mais importante escalar.
- Usar escalabilidade dinâmica com escalabilidade preditiva ajuda você a acompanhar de perto a curva de demanda do seu aplicativo, aumentando a escala durante períodos de baixo tráfego e aumentando a escala quando o tráfego é maior do que o esperado. Quando várias políticas de escalabilidade estão ativas, cada política determina a capacidade desejada de forma independente e a capacidade desejada é definida como a capacidade máxima entre essas. Por exemplo, se 10 instâncias forem necessárias para permanecer na utilização-alvo em uma política de

escalabilidade com monitoramento do objetivo e 8 instâncias forem necessárias para permanecer na utilização-alvo em uma política de dimensionamento preditiva, a capacidade desejada do grupo será definida como 10. Se você não conhece o escalonamento dinâmico, recomendamos o uso de políticas de escalabilidade de rastreamento de metas. Para ter mais informações, consulte Escalabilidade dinâmica para o Amazon EC2 Auto Scaling.

- Um pressuposto básico da escalabilidade preditivo é que o grupo do Auto Scaling é homogêneo e todas as instâncias têm capacidade igual. Se isso não for verdade para seu grupo, a capacidade prevista pode ser imprecisa. Portanto, tenha cuidado ao criar políticas de escalabilidade preditiva para grupos mistos de instâncias, pois podem ser provisionadas instâncias de diferentes tipos com capacidade desigual. 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.

Regiões compatíveis

O Amazon EC2 Auto Scaling oferece suporte a políticas de escalabilidade preditiva no Regiões da AWS seguinte: Leste dos EUA (Norte da Virgínia), Leste dos EUA (Ohio), Oeste dos EUA (Norte da Califórnia), África (Cidade do Cabo), Canadá (Central), UE (Frankfurt), UE (Irlanda), UE (Londres), UE (Milão), UE (Paris), UE (Estocolmo), Ásia-Pacífico (Hong Kong), Ásia-Pacífico (Jacarta), Ásia-Pacífico (Mumbai), Ásia-Pacífico (Osaka), Ásia-Pacífico (Tóquio), Ásia-Pacífico (Cingapura), Ásia-Pacífico (Seul), 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).

Crie uma política de escalabilidade preditiva

Os procedimentos a seguir ajudam você a criar uma política de escalabilidade preditiva usando o AWS Management Console ou. AWS CLI

Se o grupo do Auto Scaling for novo, ele deverá fornecer pelo menos 24 horas de dados antes que o Amazon EC2 Auto Scaling possa gerar uma previsão para ele.

Conteúdo

- Criar uma política de escalabilidade preditiva (console)
- Criar uma política de escalabilidade preditiva (AWS CLI)

Criar uma política de escalabilidade preditiva (console)

Se esta é a primeira vez que você cria uma política de escalabilidade preditiva, recomendamos usar o console para criar várias políticas de escalabilidade preditiva no modo somente previsão. Isso permite testar os efeitos potenciais de diferentes métricas e valores-alvo. Você pode criar várias políticas de escalabilidade preditiva para cada grupo do Auto Scaling, mas somente uma das políticas pode ser usada para a escalabilidade ativa.

Criar uma política de escalação preditiva no console (métricas predefinidas)

Siga o procedimento a seguir para criar uma política de escalação preditiva usando métricas predefinidas (CPU, E/S da rede ou contagem de solicitações do Application Load Balancer por destino). A maneira mais fácil de criar uma política de escalação preditiva é usar métricas predefinidas. Se você preferir usar métricas personalizadas, consulte Criar uma política de escalação preditiva no console (métricas personalizadas).

Para criar uma política de escalabilidade preditiva

1. Abra o console do Amazon EC2 em https://console.aws.amazon.com/ec2/ e escolha 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. As métricas personalizadas podem incluir outras métricas fornecidas por CloudWatch ou nas quais você publica CloudWatch. Para usar a contagem de solicitações de CPU, E/S de rede ou Application Load Balancer por destino, consulte Criar uma política de escalação preditiva no console (métricas predefinidas).

Para criar uma política de escalação preditiva usando métricas personalizadas, você deve fazer o seguinte:

- Você deve fornecer as consultas brutas que permitem que o Amazon EC2 Auto Scaling interaja com as métricas inseridas. CloudWatch Para ter mais informações, consulte Configurações avançadas de política de escalabilidade preditiva usando métricas personalizadas. Para ter certeza de que o Amazon EC2 Auto Scaling pode extrair os dados CloudWatch métricos, confirme se cada consulta está retornando pontos de dados. Confirme isso usando o CloudWatch console ou a operação CloudWatch GetMetricDatada API.

  > **Note**
  >
  > Fornecemos exemplos de cargas úteis JSON no editor de JSON no console do Amazon EC2 Auto Scaling. Esses exemplos fornecem uma referência para os pares de valores-

chave necessários para adicionar outras CloudWatch métricas fornecidas por AWS ou nas quais você publicou anteriormente. CloudWatch Você pode usá-las como ponto de partida e depois personalizá-las de acordo com as suas necessidades.

- Se você usar qualquer matemática de métricas, deverá estruturar manualmente o JSON para adequá-lo ao seu cenário específico. Para ter mais informações, consulte Usar expressões de matemática métrica. Antes de usar matemática de métricas em sua política, confirme se as consultas de métricas baseadas em expressões matemáticas de métricas são válidas e retornam uma única série temporal. Confirme isso usando o CloudWatch console ou a operação CloudWatch GetMetricDatada API.

Se você cometer um erro em uma consulta fornecendo dados incorretos, como o nome errado do grupo do Auto Scaling, a previsão não terá nenhum dado. Para solucionar problemas de métricas personalizadas, consulte Considerações e solução de problemas.

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 agregar os dados em uma única série temporal.

7. Em Target utilization (Utilização-alvo), insira o valor-alvo que o Amazon EC2 Auto Scaling deveria manter. O Amazon EC2 Auto Scaling aumentará a escala na horizontal até que a utilização média seja igual à utilização-alvo ou até atingir o número máximo de instâncias especificado.
8. (Opcional) Em Pre-launch instances (Iniciar instâncias previamente), escolha com que antecedência você deseja que suas instâncias sejam iniciadas antes que a previsão solicite o aumento de carga.
9. (Opcional) Em Max capacity behavior (Comportamento na capacidade máxima), escolha se será permitido que o Amazon EC2 Auto Scaling aumente a escala horizontalmente além da capacidade máxima do grupo quando a capacidade prevista exceder o máximo definido. A ativação dessa configuração permite aumentar a escala horizontalmente nos períodos em que estão previstos picos de tráfego.
10. (Opcional) Em Buffer maximum capacity above the forecasted capacity (Capacidade máxima do buffer acima da capacidade prevista), escolha a quantidade de capacidade adicional a ser usada quando a capacidade prevista estiver próxima de ou exceder a capacidade máxima. O valor é especificado como um percentual em relação à capacidade de prevista. Por exemplo, se o buffer é 10, isso significa um buffer de 10%. Portanto, se a capacidade prevista for 50 e a capacidade máxima for 40, a capacidade máxima real será 55.

    Se esse valor for definido como 0, o Amazon EC2 Auto Scaling poderá escalar a capacidade acima da capacidade máxima para igualar, mas não exceder, a capacidade prevista.
11. Selecione Create predictive scaling policy (Criar política de escalabilidade preditiva).

Criar uma política de escalabilidade preditiva (AWS CLI)

Use o AWS CLI seguinte para configurar políticas de escalabilidade preditiva para seu grupo de Auto Scaling. Substitua cada *espaço reservado para entrada do usuário* por suas próprias informações.

Para obter mais informações sobre as CloudWatch métricas que você pode especificar, consulte PredictiveScalingMetricSpecificationa Referência da API do Amazon EC2 Auto Scaling.

Exemplo 1: Uma política de escalabilidade preditiva que cria previsões, mas não implementa a escalabilidade

O exemplo a seguir mostra uma configuração de política completa que usa métricas de utilização da CPU para escalabilidade preditiva com uma utilização-alvo de `40`. O modo `ForecastOnly` é usado por padrão, a menos que você especifique explicitamente qual modo usar. Salve esta configuração em um arquivo chamado `config.json`.

```
{
    "MetricSpecifications": [
        {
            "TargetValue": 40,
            "PredefinedMetricPairSpecification": {
                "PredefinedMetricType": "ASGCPUUtilization"
            }
        }
    ]
}
```

Para criar a política na linha de comando, execute o put-scaling-policycomando com o arquivo de configuração especificado, conforme demonstrado no exemplo a seguir.

```
aws autoscaling put-scaling-policy --policy-name cpu40-predictive-scaling-policy \
  --auto-scaling-group-name my-asg --policy-type PredictiveScaling \
  --predictive-scaling-configuration file://config.json
```

Se bem-sucedido, esse comando gerará o nome do recurso da Amazon (ARN) da política.

```
{
  "PolicyARN": "arn:aws:autoscaling:region:account-id:scalingPolicy:2f4f5048-d8a8-4d14-
b13a-d1905620f345:autoScalingGroupName/my-asg:policyName/cpu40-predictive-scaling-
policy",
```

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

Exemplo 2: Uma política de escalabilidade preditiva que cria previsões e implementa a escalabilidade

Para uma política que permite que o Amazon EC2 Auto Scaling preveja e implemente a escalabilidade, adicione a propriedade `Mode` com um valor `ForecastAndScale`. O exemplo a seguir mostra uma configuração de política que usa métricas de número de solicitações do Application Load Balancer. A utilização-alvo é `1000` e a escalabilidade preditiva é definida no modo `ForecastAndScale`.

```
{
    "MetricSpecifications": [
        {
            "TargetValue": 1000,
            "PredefinedMetricPairSpecification": {
                "PredefinedMetricType": "ALBRequestCount",
                "ResourceLabel": "app/my-alb/778d41231b141a0f/targetgroup/my-alb-
target-group/943f017f100becff"
            }
        }
    ],
    "Mode": "ForecastAndScale"
}
```

Para criar essa política, execute o put-scaling-policycomando com o arquivo de configuração especificado, conforme demonstrado no exemplo a seguir.

```
aws autoscaling put-scaling-policy --policy-name alb1000-predictive-scaling-policy \
  --auto-scaling-group-name my-asg --policy-type PredictiveScaling \
  --predictive-scaling-configuration file://config.json
```

Se bem-sucedido, esse comando gerará o nome do recurso da Amazon (ARN) da política.

```
{
  "PolicyARN": "arn:aws:autoscaling:region:account-
id:scalingPolicy:19556d63-7914-4997-8c81-d27ca5241386:autoScalingGroupName/my-
asg:policyName/alb1000-predictive-scaling-policy",
  "Alarms": []
}
```

Exemplo 3: Uma política de escalabilidade preditiva que pode escalar acima da capacidade máxima

O exemplo a seguir mostra como criar uma política que poderá escalar além do limite máximo de tamanho do grupo quando você precisar que ele lide com uma carga maior do que o normal. Por padrão, o Amazon EC2 Auto Scaling não aumenta a capacidade do EC2 além da capacidade máxima definida. No entanto, pode ser útil deixá-lo ir além com um pouco mais de capacidade para evitar problemas de performance ou disponibilidade.

Para fornecer espaço para o Amazon EC2 Auto Scaling provisionar capacidade adicional quando a capacidade estiver no tamanho máximo do grupo ou muito próxima a ele, especifique as propriedades `MaxCapacityBreachBehavior` e `MaxCapacityBuffer`, conforme mostrado no exemplo a seguir. É necessário especificar `MaxCapacityBreachBehavior` com um valor de `IncreaseMaxCapacity`. O número máximo de instâncias que seu grupo pode ter depende do valor de `MaxCapacityBuffer`.

```
{
    "MetricSpecifications": [
        {
            "TargetValue": 70,
            "PredefinedMetricPairSpecification": {
                "PredefinedMetricType": "ASGCPUUtilization"
            }
        }
    ],
    "MaxCapacityBreachBehavior": "IncreaseMaxCapacity",
    "MaxCapacityBuffer": 10
}
```

Neste exemplo, a política é configurada para usar um buffer de 10% (`"MaxCapacityBuffer": 10`). Assim, se a capacidade prevista for 50 e a capacidade máxima for 40, a capacidade máxima efetiva será 55. Uma política que pudesse escalar a capacidade acima da capacidade máxima para igualar, mas não exceder, a capacidade prevista teria um buffer de 0 (`"MaxCapacityBuffer": 0`).

Para criar essa política, execute o put-scaling-policycomando com o arquivo de configuração especificado, conforme demonstrado no exemplo a seguir.

```
aws autoscaling put-scaling-policy --policy-name cpu70-predictive-scaling-policy \
  --auto-scaling-group-name my-asg --policy-type PredictiveScaling \
  --predictive-scaling-configuration file://config.json
```

Se bem-sucedido, esse comando gerará o nome do recurso da Amazon (ARN) da política.

```
{
  "PolicyARN": "arn:aws:autoscaling:region:account-id:scalingPolicy:d02ef525-8651-4314-
bf14-888331ebd04f:autoScalingGroupName/my-asg:policyName/cpu70-predictive-scaling-
policy",
  "Alarms": []
}
```

Avaliar as políticas de escalabilidade preditiva

Antes de usar uma política de escalabilidade preditiva para escalar seu grupo do Auto Scaling, analise as recomendações e outros dados da política no console do Amazon EC2 Auto Scaling. Isso é importante porque você não quer que uma política de escalabilidade preditiva escale sua capacidade real até saber que suas previsões estão corretas.

Se o grupo do Auto Scaling for novo, permita que o Amazon EC2 Auto Scaling tenha 24 horas para criar a primeira previsão.

Ao criar uma previsão, o Amazon EC2 Auto Scaling usa dados históricos. Se o grupo do Auto Scaling ainda não tiver muitos dados históricos recentes, o Amazon EC2 Auto Scaling poderá preencher temporariamente a previsão com agregados criados com base nos agregados históricos disponíveis atualmente. As previsões são preenchidas até duas semanas anteriores à data de criação da política.

Conteúdo

- Visualizar recomendações de escalabilidade preditiva
- Analisar grafos de monitoramento de escalabilidade preditiva
- Monitore métricas de escalabilidade preditiva com CloudWatch

Visualizar recomendações de escalabilidade preditiva

Para realizar uma análise eficaz, o Amazon EC2 Auto Scaling deve ter pelo menos duas políticas de escalabilidade preditiva para comparação. (Porém, ainda é possível analisar as conclusões de uma única política.) Ao criar várias políticas, é possível avaliar uma política que usa uma métrica em relação a uma política que usa outra métrica. Também é possível avaliar o impacto de diferentes combinações de valores de destino e métricas. Depois que as políticas de escalabilidade preditiva são criadas, o Amazon EC2 Auto Scaling imediatamente começa a avaliar qual política faria um trabalho melhor ao escalar seu grupo.

Para visualizar as recomendações no console do Amazon EC2 Auto Scaling

1. Abra o console do Amazon EC2 em https://console.aws.amazon.com/ec2/ e escolha 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.
4. (Opcional) Para selecionar o período desejado para os resultados da recomendação, escolha o valor de sua preferência no menu suspenso Período de avaliação: 2 dias, 1 semana, 2 semanas, 4 semanas, 6 semanas ou 8 semanas. Por padrão, o período de avaliação são as duas últimas semanas. Um período de avaliação mais longo oferece mais pontos de dados para os resultados da recomendação. Porém, adicionar mais pontos de dados pode não melhorar os resultados, se seus padrões de carga tiverem sido alterados, como após um período de demanda excepcional. Nesse caso, é possível obter uma recomendação mais focada analisando dados mais recentes.

> **Note**
>
> As recomendações são geradas somente para políticas que estão no modo Somente previsão. O recurso de recomendações funciona melhor quando uma política está no modo Somente previsão durante o período de avaliação. Se você iniciar uma política no modo de Prever e escalar e alterná-la para o modo Somente previsão posteriormente, é provável que as conclusões dessa política tenham desvios. Isso ocorre porque a política já contribuiu em favor da capacidade real.

Analisar grafos de monitoramento de escalabilidade preditiva

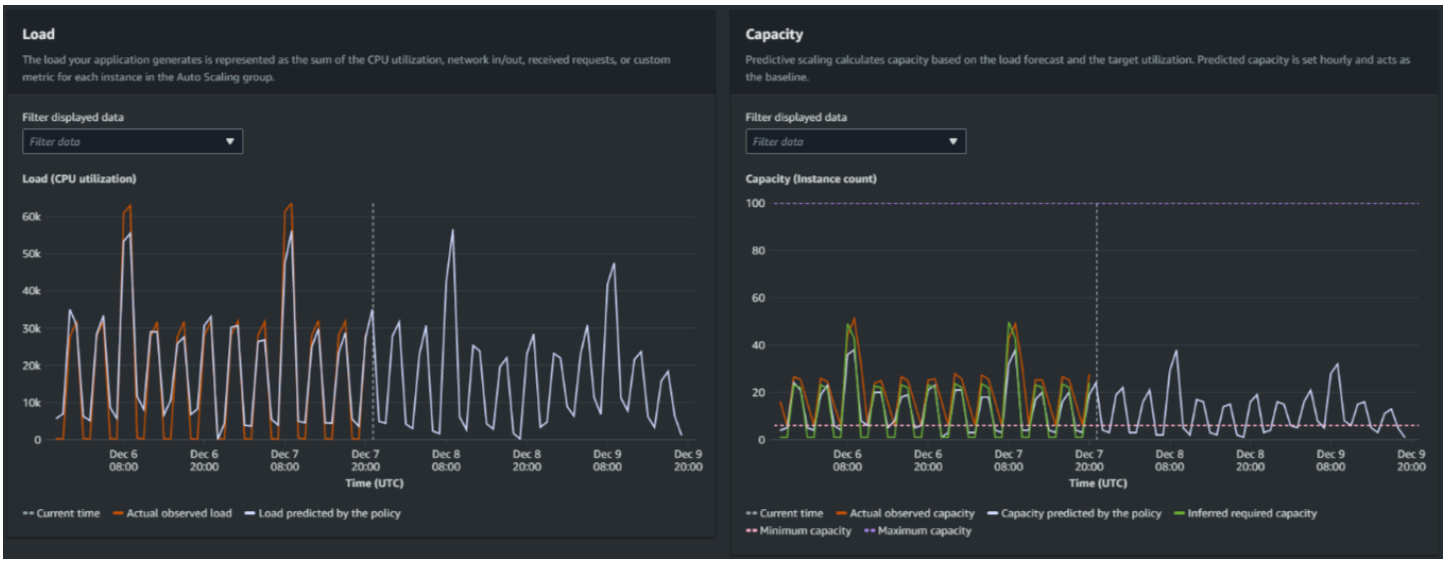
No console do Amazon EC2 Auto Scaling, é possível analisar a previsão dos dias, semanas ou meses anteriores para visualizar a performance da política ao longo do tempo. Você também pode usar essas informações para avaliar a precisão das previsões ao decidir se permitirá que uma política escale a capacidade real.

Para analisar grafos de monitoramento de escalabilidade preditiva no console do Amazon EC2 Auto Scaling

1. Escolha uma política na lista Políticas de escalabilidade preditiva.
2. Na seção Monitorar, você pode visualizar as previsões passadas e futuras de carga e de capacidade da política em relação aos valores reais. O grafo Carga exibe a previsão de carga e os valores reais para a métrica de carga escolhida. O grafo Capacidade exibe o número de instâncias previstas pela política. Também inclui o número real de instâncias iniciadas. A linha vertical separa os valores históricos das previsões futuras. Esses grafos ficam disponíveis logo após a criação da política.
3. (Opcional) Para alterar a quantidade de dados históricos exibidos no gráfico, escolha o valor de sua preferência no menu suspenso Período de avaliação na parte superior da página. O período de avaliação não transforma os dados desta página de maneira alguma. Ele altera apenas a quantidade de dados históricos exibidos.

A imagem a seguir exibe os grafos Carga e Capacidade quando as previsões foram aplicadas várias vezes. A escalabilidade preditiva prevê a carga com base nos dados históricos de carga. A carga que sua aplicação gera é representada como a soma da utilização da CPU, entrada/saída da rede, solicitações recebidas ou métrica personalizada para cada instância no grupo do Auto Scaling. A

escalabilidade preditiva calcula as necessidades futuras de capacidade com base na previsão de carga e na utilização desejada que você deseja alcançar para a métrica de escalabilidade.



Compare dados no grafo Carga

Cada linha horizontal representa um conjunto diferente de pontos de dados relatados em intervalos de uma hora:

1. A Carga real observada usa a estatística SUM da métrica de carga escolhida para exibir a carga horária total 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.

Monitore métricas de escalabilidade preditiva com CloudWatch

Dependendo de suas necessidades, talvez você prefira acessar dados de monitoramento para escalabilidade preditiva da Amazon CloudWatch em vez do console Amazon EC2 Auto Scaling. Após criar uma política de escalabilidade preditiva, a política coletará dados que são usados para prever sua carga e capacidade futuras. Depois que esses dados são coletados, eles são armazenados automaticamente em CloudWatch intervalos regulares. Em seguida, você pode usar CloudWatch para visualizar o desempenho da política ao longo do tempo. Você também pode criar CloudWatch alarmes para notificá-lo quando os indicadores de desempenho mudarem além dos limites definidos em CloudWatch.

Tópicos

- Visualizar dados históricos de previsão
- Criar métricas de precisão usando matemática métrica

Visualizar dados históricos de previsão

Você pode visualizar os dados de previsão de carga e capacidade para uma política de escalabilidade preditiva em. CloudWatch Isso pode ser útil ao visualizar previsões em relação a outras CloudWatch métricas em um único gráfico. O recurso também pode ajudar ao visualizar um intervalo mais amplo de tempo para que você possa ver tendências ao longo do tempo. É possível acessar até 15 meses de métricas históricas a fim de obter uma melhor visão do desempenho da política.

Para ter mais informações, consulte Métricas e dimensões de escalabilidade preditiva.

Para visualizar dados históricos de previsão usando o CloudWatch console

1. Abra o CloudWatch console em https://console.aws.amazon.com/cloudwatch/.
2. No painel de navegação, escolha Metrics (Métricas) e, em seguida, All metrics (Todas as métricas).

3. Selecione o namespace da métrica Auto Scaling (Escalabilidade automática).
4. Escolha uma das seguintes opções para visualizar a previsão de carga ou as métricas de previsão de capacidade:

   - Previsões de carga de escalabilidade preditiva
   - Previsões de capacidade de escalabilidade preditiva
5. No campo de pesquisa, insira o nome da política de escalabilidade preditiva ou o nome do grupo do Auto Scaling e pressione a tecla Enter para filtrar os resultados.
6. Para criar um gráfico de uma métrica, marque a caixa de seleção ao lado da métrica. Para alterar o nome do gráfico, escolha o ícone de lápis. Para alterar o período, selecione um dos valores predefinidos ou escolha custom (personalizado). Para obter mais informações, consulte Representação gráfica de uma métrica no Guia do CloudWatch usuário da Amazon.
7. Para alterar a estatística, escolha a guia Métricas em gráfico. Escolha o cabeçalho de coluna ou um valor individual e, em seguida, escolha uma estatística diferente. Embora você possa escolher qualquer estatística para cada métrica, nem todas as estatísticas são úteis para PredictiveScalingLoadForecastas PredictiveScalingCapacityForecastmétricas. Por exemplo, as estatísticas 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 Ações, Adicionar ao painel.

Criar métricas de precisão usando matemática métrica

Com a matemática métrica, você pode consultar várias CloudWatch métricas e usar expressões matemáticas para criar novas séries temporais com base nessas métricas. Você pode visualizar as séries temporais resultantes no CloudWatch console e adicioná-las aos painéis. Para obter mais informações sobre matemática métrica, consulte Usando matemática métrica no Guia CloudWatch do usuário da Amazon.

Usando matemática métrica, você pode representar graficamente de diferentes maneiras os dados que o Amazon EC2 Auto Scaling gera sobre escalabilidade preditiva. Isso ajuda a monitorar o desempenho da política ao longo do tempo e a entender se é possível melhorar sua combinação de métricas.

Por exemplo, você pode usar uma expressão de matemática métrica para monitorar o mean absolute percentage error (MAPE – Erro percentual absoluto médio). A métrica MAPE ajuda a monitorar a diferença entre os valores previstos e os valores efetivos observados durante uma determinada janela de previsão. Mudanças no valor de MAPE podem indicar se o desempenho da política está se degradando ao longo do tempo conforme a natureza 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). `m1`é a CloudWatch métrica que contém os valores reais de carga (supondo que a utilização da CPU seja a métrica de carga originalmente especificada para a política nomeada`my-`

`predictive-scaling-policy`). `m2`é a CloudWatch métrica que contém os valores de carga previstos. A sintaxe matemática para a métrica MAPE é a seguinte:

média de (abs ((efetivo - previsto)/(efetivo)))

Visualizar suas métricas de precisão e definir alarmes

Para visualizar os dados métricos de precisão, selecione a guia Métricas no CloudWatch console. Nele, é possível representar graficamente os dados. Para obter mais informações, consulte Adicionar uma expressão matemática a um CloudWatch gráfico no Guia do CloudWatch usuário da Amazon.

Na seção Metrics (Métricas), você também pode criar um alarme com base em uma métrica que esteja monitorando. Enquanto estiver na guia Graphed metrics (Métricas representadas em gráficos), selecione o ícone Create alarm (Criar alarme) na coluna Actions (Ações). O ícone Create alarm (Criar alarme) é representado como um pequeno sino. Para obter mais informações e opções de notificação, consulte Criação de um CloudWatch alarme com base em uma expressão matemática métrica e Notificação de usuários sobre alterações de alarme no Guia do CloudWatch usuário da Amazon.

Como alternativa, você pode usar GetMetricDatae PutMetricAlarmrealizar cálculos usando matemática métrica e criar alarmes com base na saída.

Substituir valores de previsão usando ações programadas

Às vezes, você pode ter informações adicionais sobre seus futuros requisitos de aplicações que o cálculo de previsão não pode levar em conta. Por exemplo, os cálculos de previsão podem subestimar a capacidade necessária para um evento de marketing futuro. Você pode usar ações programadas para substituir temporariamente a previsão durante períodos futuros. As ações programadas podem ser executadas de forma recorrente ou em uma data e hora específicas quando houver flutuações de demanda únicas.

Por exemplo, você pode criar uma ação programada com uma capacidade mínima maior do que a prevista. Em 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.

Conteúdo

- Etapa 1: (Opcional) Analisar dados de séries temporais
- Etapa 2: Criar duas ações programadas

Etapa 1: (Opcional) Analisar dados de séries temporais

Comece analisando os dados de séries temporais de previsão. Essa é uma etapa opcional, mas é útil quando você deseja entender os detalhes da previsão.

1. Recuperar a previsão

   Após a criação da previsão, é possível consultar um período específico na previsão. O objetivo da consulta é obter uma visão completa dos dados de séries temporais para um período específico.

   Sua consulta pode incluir até dois dias de dados de previsão futura. Se você usa a escalabilidade preditiva há algum tempo, também pode acessar seus dados de previsão anteriores. No entanto, a duração máxima de tempo entre as horas inicial e final é de 30 dias.

   Para obter a previsão usando o get-predictive-scaling-forecast AWS CLI comando, forneça os seguintes parâmetros no comando:

   - Insira o nome do grupo do Auto Scaling no parâmetro `--auto-scaling-group-name`.
   - Insira o nome da política no parâmetro `--policy-name`.
   - Insira a hora de início no parâmetro `--start-time` para retornar apenas os dados de previsão para depois ou no horário especificado.
   - Insira a hora de término no parâmetro `--end-time` para retornar apenas os dados de previsão para antes do horário especificado.

```
aws autoscaling get-predictive-scaling-forecast --auto-scaling-group-name my-asg \
    --policy-name cpu40-predictive-scaling-policy \
    --start-time "2021-05-19T17:00:00Z" \
    --end-time "2021-05-19T23:00:00Z"
```

   Se bem-sucedido, o comando retornará uma resposta semelhante à seguinte.

```
{
```

```
    "LoadForecast": [
        {
            "Timestamps": [
                "2021-05-19T17:00:00+00:00",
                "2021-05-19T18:00:00+00:00",
                "2021-05-19T19:00:00+00:00",
                "2021-05-19T20:00:00+00:00",
                "2021-05-19T21:00:00+00:00",
                "2021-05-19T22:00:00+00:00",
                "2021-05-19T23:00:00+00:00"
            ],
            "Values": [
                153.0655799339254,
                128.8288551285919,
                107.1179447150675,
                197.3601844551528,
                626.4039934516954,
                596.9441277518481,
                677.9675713779869
            ],
            "MetricSpecification": {
                "TargetValue": 40.0,
                "PredefinedMetricPairSpecification": {
                    "PredefinedMetricType": "ASGCPUUtilization"
                }
            }
        }
    ],
    "CapacityForecast": {
        "Timestamps": [
            "2021-05-19T17:00:00+00:00",
            "2021-05-19T18:00:00+00:00",
            "2021-05-19T19:00:00+00:00",
            "2021-05-19T20:00:00+00:00",
            "2021-05-19T21:00:00+00:00",
            "2021-05-19T22:00:00+00:00",
            "2021-05-19T23:00:00+00:00"
        ],
        "Values": [
            2.0,
            2.0,
            2.0,
            2.0,
            4.0,
```

```
            4.0,
            4.0
        ]
    },
    "UpdateTime": "2021-05-19T01:52:50.118000+00:00"
}
```

A resposta inclui duas previsões: `LoadForecast` e `CapacityForecast`. `LoadForecast` mostra a previsão de carga horária. `CapacityForecast` mostra os valores de previsão para a capacidade que é necessária em uma base horária para lidar com a carga prevista enquanto mantém um `TargetValue` de 40,0 (40% de utilização média da CPU).

2. Identificar o período-alvo

   Identifique a hora ou horas em que a flutuação de demanda única deverá ocorrer. Lembre-se de que as datas e os horários mostrados na previsão estão em UTC.

Etapa 2: Criar duas ações programadas

Em seguida, crie duas ações programadas para um período específico em que sua aplicação terá uma carga maior do que a prevista. Por exemplo, se você tiver um evento de marketing que irá direcionar o tráfego para seu site por um período limitado, poderá programar uma ação única para atualizar a capacidade mínima quando ele começar. Em seguida, agende outra ação para retornar a capacidade mínima para a configuração original quando o evento terminar.

Para criar duas ações programadas para eventos únicos (console)

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

   O console exibe as ações programadas para o grupo do Auto Scaling.

6. Configure uma segunda ação programada para retornar a capacidade mínima para a configuração original no final do evento. A escalabilidade preditiva pode escalar a capacidade somente quando o valor definido para Min (Mínima) é menor que os valores da previsão.

Para criar duas ações programadas para eventos únicos (AWS CLI)

Para usar o AWS CLI para criar as ações agendadas, use o comando put-scheduled-update-group-action.

Por exemplo, vamos definir uma programação que mantenha uma capacidade mínima de três instâncias em 19 de maio às 17h por oito horas. Os comandos a seguir mostram como implementar esse cenário.

O primeiro comando put-scheduled-update-group-action instrui o Amazon EC2 Auto Scaling a atualizar a capacidade mínima do grupo de Auto Scaling especificado às 17h UTC de 19 de maio de 2021.

```
aws autoscaling put-scheduled-update-group-action --scheduled-action-name my-event-start \
  --auto-scaling-group-name my-asg --start-time "2021-05-19T17:00:00Z" --minimum-capacity 3
```

O segundo comando instrui o Amazon EC2 Auto Scaling a definir a capacidade mínima do grupo como um à 1h da manhã UTC em 20 de maio de 2021.

```
aws autoscaling put-scheduled-update-group-action --scheduled-action-name my-event-end \
  --auto-scaling-group-name my-asg --start-time "2021-05-20T01:00:00Z" --minimum-capacity 1
```

Após você adicionar essas ações programadas ao grupo do Auto Scaling, o Amazon EC2 Auto Scaling fará o seguinte:

- Às 17h UTC em 19 de maio de 2021, a primeira ação programada é executada. Se o grupo tiver menos de três instâncias, ele será expandido para três instâncias. Durante esse período e nas próximas oito horas, o Amazon EC2 Auto Scaling poderá continuar a aumentar a escala na horizontal se a capacidade prevista for maior do que a capacidade real ou se houver uma política de escalabilidade dinâmica em vigor.
- À 1h da manhã UTC em 20 de maio de 2021, a segunda ação programada é executada. Isso retorna a capacidade mínima para sua configuração original no final do evento.

Escalabilidade com base em programações recorrentes

Para substituir a previsão para o mesmo período de tempo todas as semanas, crie duas ações programadas e forneça a lógica de hora e data usando uma expressão cron.

A expressão cron consiste em cinco campos separados por espaços: [Minute] [Hour] [Day_of_Month] [Month_of_Year] [Day_of_Week]. Os campos podem conter quaisquer valores permitidos, incluindo caracteres especiais.

Por exemplo, esta expressão cron executa a ação todas as terças-feiras às 6h30. O asterisco é usado como um curinga para corresponder a todos os valores de um campo.

```
30 6 * * 2
```

Consulte também

Para obter mais informações sobre como criar, listar, editar e excluir ações programadas, consulte Escalabilidade programada para o Amazon EC2 Auto Scaling.

Configurações avançadas de política de escalabilidade preditiva usando métricas personalizadas

Em uma política de escalabilidade preditiva é possível usar métricas predefinidas ou personalizadas. Métricas personalizadas são úteis quando as métricas predefinidas (CPU, E/S da rede e contagem de solicitações do Application Load Balancer) não descrevem suficientemente a carga da aplicação.

Ao criar uma política de escalabilidade preditiva com métricas personalizadas, você pode especificar outras CloudWatch métricas fornecidas por AWS, ou você pode especificar métricas que você

mesmo define e publica. Você também pode usar a matemática métrica para agregar e transformar métricas existentes em uma nova série temporal que AWS não é rastreada automaticamente. A combinação de valores em seus dados, por exemplo, calculando novas somas ou médias, é chamada de agregação. Os dados resultantes são chamados de um agregado.

A seção a seguir contém as práticas recomendados e exemplos de como sstruturar o JSON para a política.

Conteúdo

- Práticas recomendadas
- Pré-requisitos
- Estruture o JSON para métricas personalizadas
- Considerações e solução de problemas
- Limitações

Práticas recomendadas

As seguintes práticas recomendadas podem ajudar no uso mais eficaz de métricas personalizadas:

- Para a especificação da métrica de carga, a métrica mais útil é uma métrica que represente a carga em um grupo do Auto Scaling como um todo, independentemente da capacidade do grupo.
- Para a especificação da métrica de escalabilidade, a métrica mais útil para escalar é throughput ou utilização média por métrica de instância.
- A métrica de escalabilidade deve ser inversamente proporcional à capacidade. Ou seja, se o número de instâncias no grupo do Auto Scaling aumentar, a métrica de escalabilidade deve diminuir aproximadamente na mesma proporção. Para garantir que a escalabilidade preditiva se comporte conforme o esperado, a métrica de carga e a métrica de escalabilidade também devem se correlacionar fortemente entre si.
- A utilização visada deve corresponder ao tipo de métrica de escalabilidade. Para uma configuração de política que use a utilização da CPU, essa é uma porcentagem visada. Para uma configuração de política que use throughput, como o número de solicitações ou mensagens, esse é o número visado de solicitações ou mensagens por instância durante qualquer intervalo de um minuto.
- Se essas recomendações não forem seguidas, provavelmente os valores futuros previstos da série temporal estarão incorretos. Para validar se os dados estão corretos, você pode visualizar os valores previstos no console do Amazon EC2 Auto Scaling. Como alternativa, depois de criar sua

política de escalabilidade preditiva, inspecione os `CapacityForecast` objetos `LoadForecast` retornados por uma chamada para a API. GetPredictiveScalingForecast

- Recomendamos a configuração da escalabilidade preditiva no modo apenas previsão para avaliar a previsão antes que a escalabilidade preditiva comece a modificar ativamente a capacidade.

Pré-requisitos

Para adicionar métricas personalizadas à política de escalação preditiva, você deve ter as permissões `cloudwatch:GetMetricData`.

Para especificar suas próprias métricas em vez das métricas AWS fornecidas, você deve primeiro publicar suas métricas em CloudWatch. Para obter mais informações, consulte Publicação de métricas personalizadas no Guia CloudWatch do usuário da Amazon.

Se publicar suas próprias métricas, certifique-se de publicar os pontos de dados com uma frequência mínima de cinco minutos. O Amazon EC2 Auto Scaling recupera os pontos CloudWatch de dados com base na duração do período necessário. Por exemplo, a especificação da métrica de carga usa métricas horárias para medir a carga em seu aplicativo. CloudWatch usa seus dados métricos publicados para fornecer um único valor de dados para qualquer período de uma hora, agregando todos os pontos de dados com registros de data e hora que se enquadram em cada período de uma hora.

Estruture o JSON para métricas personalizadas

A seção a seguir contém exemplos de como configurar a escala preditiva para consultar dados. CloudWatch Há dois métodos diferentes de configurar essa opção, e o método escolhido afeta qual será o formato usado para estruturar JSON para a política de escalação preditiva. Quando você usa matemática de métricas, o formato do JSON varia ainda mais com base na matemática de métrica que está sendo aplicada.

1. Para criar uma política que obtenha dados diretamente de outras CloudWatch métricas fornecidas AWS ou nas quais você publica CloudWatch, consulteExemplo de política de escalação preditiva com métricas personalizadas de carga e de dimensionamento (AWS CLI).
2. Para criar uma política que possa consultar várias CloudWatch métricas e usar expressões matemáticas para criar novas séries temporais com base nessas métricas, consulteUsar expressões de matemática métrica.

Exemplo de política de escalação preditiva com métricas personalizadas de carga e de dimensionamento (AWS CLI)

Para criar uma política de escalabilidade preditiva com métricas personalizadas de carga e escalabilidade com o. AWS CLI, armazene os argumentos para `--predictive-scaling-configuration` em um arquivo JSON chamado. `config.json`

Você começa a adicionar métricas personalizadas substituindo os valores substituíveis no exemplo a seguir por suas métricas e sua meta de utilização.

```
{
  "MetricSpecifications": [
    {
      "TargetValue": 50,
      "CustomizedScalingMetricSpecification": {
        "MetricDataQueries": [
          {
            "Id": "scaling_metric",
            "MetricStat": {
              "Metric": {
                "MetricName": "MyUtilizationMetric",
                "Namespace": "MyNameSpace",
                "Dimensions": [
                  {
                    "Name": "MyOptionalMetricDimensionName",
                    "Value": "MyOptionalMetricDimensionValue"
                  }
                ]
              },
              "Stat": "Average"
            }
          }
        ]
      },
      "CustomizedLoadMetricSpecification": {
        "MetricDataQueries": [
          {
            "Id": "load_metric",
            "MetricStat": {
              "Metric": {
                "MetricName": "MyLoadMetric",
                "Namespace": "MyNameSpace",
                "Dimensions": [
```

```
                {
                  "Name": "MyOptionalMetricDimensionName",
                  "Value": "MyOptionalMetricDimensionValue"
                }
              ]
            },
            "Stat": "Sum"
          }
        }
      ]
    }
  }
]
}
```

Para obter mais informações, consulte a MetricDataQueryReferência da API Auto Scaling do Amazon EC2.

> **Note**
>
> Veja a seguir alguns recursos adicionais que podem ajudá-lo a encontrar nomes de métricas, namespaces, dimensões e estatísticas para CloudWatch métricas:
>
> - Para obter informações sobre as métricas disponíveis para AWS serviços, consulte AWS serviços que publicam CloudWatch métricas no Guia CloudWatch do usuário da Amazon.
> - Para obter o nome exato da métrica, o namespace e as dimensões (se aplicável) de uma CloudWatch métrica com o AWS CLI, consulte list-metrics.

Para criar essa política, execute o put-scaling-policycomando usando o arquivo JSON como entrada, conforme demonstrado no exemplo a seguir.

```
aws autoscaling put-scaling-policy --policy-name my-predictive-scaling-policy \
  --auto-scaling-group-name my-asg --policy-type PredictiveScaling \
  --predictive-scaling-configuration file://config.json
```

Se bem-sucedido, esse comando gerará o nome do recurso da Amazon (ARN) da política.

```
{
  "PolicyARN": "arn:aws:autoscaling:region:account-id:scalingPolicy:2f4f5048-d8a8-4d14-
b13a-d1905620f345:autoScalingGroupName/my-asg:policyName/my-predictive-scaling-policy",
```

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

Usar expressões de matemática métrica

A seção a seguir fornece informações e exemplos de políticas de escalação preditiva que mostram como você pode usar a matemática de métricas em sua política.

Conteúdo

- Noções básicas de matemática métrica
- Exemplo de política de escalação preditiva que combina métricas por meio da matemática de métricas (AWS CLI)
- Exemplo de política de escalação preditiva para usar em um cenário de implantação azul/verde (AWS CLI)

Noções básicas de matemática métrica

Se tudo o que você quer fazer é agregar dados métricos existentes, a matemática CloudWatch métrica poupa o esforço e o custo de publicar outra métrica no. CloudWatch Você pode usar qualquer métrica que AWS forneça e também pode usar métricas que você define como parte de seus aplicativos. Por exemplo, talvez você queira calcular a lista de pendências da fila do Amazon SQS por instância. Você pode fazer isso usando o número aproximado de mensagens disponíveis para recuperação da fila e dividindo esse número pela capacidade de execução do grupo do Auto Scaling.

Para obter mais informações, consulte Usando matemática métrica no Guia CloudWatch do usuário da Amazon.

Se você optar por usar uma expressão matemática métrica em sua política de escalabilidade preditiva, considere os seguintes pontos:

- As operações matemáticas métricas usam os pontos de dados da combinação exclusiva de nome da métrica, namespace e pares de métricas de chaves-valor da dimensão.
- Você pode usar qualquer operador aritmético (+ - */^), função estatística (como AVG ou SUM) ou outra função compatível. CloudWatch
- Você pode usar as métricas e os resultados de outras expressões matemáticas nas fórmulas da expressão matemática.

- Suas expressões matemáticas métricas podem ser compostas de agregações diferentes. No entanto, uma prática recomendada para o resultado final da agregação é usar `Average` para a métrica de escalabilidade e `Sum` para a métrica de carga.
- Qualquer expressão usada em uma especificação de métrica deve eventualmente retornar uma única série temporal.

Para usar matemática métrica, faça o seguinte:

- Escolha uma ou mais CloudWatch métricas. Em seguida, crie a expressão. Para obter mais informações, consulte Usando matemática métrica no Guia CloudWatch do usuário da Amazon.
- Verifique se a expressão matemática métrica é válida usando o CloudWatch console ou a CloudWatch GetMetricDataAPI.

Exemplo de política de escalação preditiva que combina métricas por meio da matemática de métricas (AWS CLI)

Às vezes, ao invés de especificar a métrica diretamente, talvez seja necessário processar seus dados de alguma forma, primeiramente. Por exemplo, você pode ter uma aplicação que extrai o trabalho de uma fila do Amazon SQS e talvez queira usar o número de itens na fila como critério para escalabilidade preditiva. O número de mensagens na fila não define unicamente o número necessário de instâncias. Portanto, é necessário mais trabalho para criar uma métrica que possa ser usada para calcular a lista de pendências por instância. Para ter mais informações, consulte Escalabilidade baseada no Amazon SQS.

Veja a seguir um exemplo de política de escalabilidade preditiva para esse cenário. Ele especifica métricas de escalabilidade e carga baseadas na métrica `ApproximateNumberOfMessagesVisible` do Amazon SQS, que é o número de mensagens disponíveis para recuperação da fila. Ele também usa a métrica `GroupInServiceInstances` do Amazon EC2 Auto Scaling e uma expressão matemática para calcular a lista de pendências por instância para a métrica de escalabilidade.

```
aws autoscaling put-scaling-policy --policy-name my-sqs-custom-metrics-policy \
  --auto-scaling-group-name my-asg --policy-type PredictiveScaling \
  --predictive-scaling-configuration file://config.json
{
  "MetricSpecifications": [
    {
      "TargetValue": 100,
```

```
      "CustomizedScalingMetricSpecification": {
        "MetricDataQueries": [
          {
            "Label": "Get the queue size (the number of messages waiting to be
processed)",
            "Id": "queue_size",
            "MetricStat": {
              "Metric": {
                "MetricName": "ApproximateNumberOfMessagesVisible",
                "Namespace": "AWS/SQS",
                "Dimensions": [
                  {
                    "Name": "QueueName",
                    "Value": "my-queue"
                  }
                ]
              },
              "Stat": "Sum"
            },
            "ReturnData": false
          },
          {
            "Label": "Get the group size (the number of running instances)",
            "Id": "running_capacity",
            "MetricStat": {
              "Metric": {
                "MetricName": "GroupInServiceInstances",
                "Namespace": "AWS/AutoScaling",
                "Dimensions": [
                  {
                    "Name": "AutoScalingGroupName",
                    "Value": "my-asg"
                  }
                ]
              },
              "Stat": "Sum"
            },
            "ReturnData": false
          },
          {
            "Label": "Calculate the backlog per instance",
            "Id": "scaling_metric",
            "Expression": "queue_size / running_capacity",
            "ReturnData": true
```

```
          }
        ]
      },
      "CustomizedLoadMetricSpecification": {
        "MetricDataQueries": [
          {
            "Id": "load_metric",
            "MetricStat": {
              "Metric": {
                "MetricName": "ApproximateNumberOfMessagesVisible",
                "Namespace": "AWS/SQS",
                "Dimensions": [
                  {
                    "Name": "QueueName",
                    "Value": "my-queue"
                  }
                ],
              },
              "Stat": "Sum"
            },
            "ReturnData": true
          }
        ]
      }
    }
  ]
}
```

O exemplo retorna o ARN da política.

```
{
  "PolicyARN": "arn:aws:autoscaling:region:account-id:scalingPolicy:2f4f5048-d8a8-4d14-
b13a-d1905620f345:autoScalingGroupName/my-asg:policyName/my-sqs-custom-metrics-policy",
  "Alarms": []
}
```

Exemplo de política de 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 Usando políticas de escalabilidade preditiva do EC2 Auto Scaling com implantações azul/verdes no blog de computação. AWS

O exemplo de política a seguir mostra como isso pode ser feito. Neste exemplo, a política usa a métrica `CPUUtilization` emitida pelo Amazon EC2. Ela também usa a métrica `GroupInServiceInstances` do Amazon EC2 Auto Scaling e uma expressão matemática para calcular o valor da métrica de escalabilidade por instância. Ela também especifica uma especificação de métrica de capacidade para obter a métrica `GroupInServiceInstances`.

A expressão de pesquisa encontra o `CPUUtilization` de instâncias em vários grupos do Auto Scaling com base nos critérios de pesquisa especificados. Se, posteriormente, você criar um novo grupo do Auto Scaling que corresponda aos mesmos critérios de pesquisa, o `CPUUtilization` das instâncias no novo grupo do Auto Scaling são incluídas automaticamente.

```
aws autoscaling put-scaling-policy --policy-name my-blue-green-predictive-scaling-
policy \
  --auto-scaling-group-name my-asg --policy-type PredictiveScaling \
  --predictive-scaling-configuration file://config.json
{
  "MetricSpecifications": [
    {
      "TargetValue": 25,
      "CustomizedScalingMetricSpecification": {
        "MetricDataQueries": [
          {
```

```
            "Id": "load_sum",
            "Expression": "SUM(SEARCH('{AWS/EC2,AutoScalingGroupName} MetricName=
\"CPUUtilization\" ASG-myapp', 'Sum', 300))",
            "ReturnData": false
          },
          {
            "Id": "capacity_sum",
            "Expression": "SUM(SEARCH('{AWS/AutoScaling,AutoScalingGroupName}
 MetricName=\"GroupInServiceInstances\" ASG-myapp', 'Average', 300))",
            "ReturnData": false
          },
          {
            "Id": "weighted_average",
            "Expression": "load_sum / capacity_sum",
            "ReturnData": true
          }
        ]
      },
      "CustomizedLoadMetricSpecification": {
        "MetricDataQueries": [
          {
            "Id": "load_sum",
            "Expression": "SUM(SEARCH('{AWS/EC2,AutoScalingGroupName} MetricName=
\"CPUUtilization\" ASG-myapp', 'Sum', 3600))"
          }
        ]
      },
      "CustomizedCapacityMetricSpecification": {
        "MetricDataQueries": [
          {
            "Id": "capacity_sum",
            "Expression": "SUM(SEARCH('{AWS/AutoScaling,AutoScalingGroupName}
 MetricName=\"GroupInServiceInstances\" ASG-myapp', 'Average', 300))"
          }
        ]
      }
    }
  ]
}
```

O exemplo retorna o ARN da política.

```
{
```

```
  "PolicyARN": "arn:aws:autoscaling:region:account-id:scalingPolicy:2f4f5048-d8a8-4d14-
b13a-d1905620f345:autoScalingGroupName/my-asg:policyName/my-blue-green-predictive-
scaling-policy",
  "Alarms": []
}
```

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 pesquisa, consulte a sintaxe CloudWatch da expressão de pesquisa no Guia CloudWatch do usuário da Amazon.
- Se você não validou uma expressão com antecedência, o put-scaling-policycomando a valida ao criar sua política de escalabilidade. No entanto, existe a possibilidade de que esse comando não identifique a causa exata dos erros detectados. Para corrigir os problemas, solucione os erros que você recebe em uma resposta de uma solicitação ao get-metric-datacomando. Você também pode solucionar o problema da expressão no CloudWatch console.
- Quando você visualiza os gráficos de Carga e de Capacidade no console, o gráfico da Capacidade pode não mostrar nenhum dado. Para garantir que os gráficos tenham dados completos, certifique-se de habilitar consistentemente métricas de grupo para seus grupos do Auto Scaling. Para ter mais informações, consulte Ativar métricas do grupo do Auto Scaling (console).
- A especificação da métrica de capacidade só é útil para implantações azul/verde quando você tem aplicações que são executadas em diferentes grupos do Auto Scaling ao longo de suas vidas úteis. Essa métrica personalizada permite que você forneça a capacidade total de vários grupos do Auto Scaling. A escalabilidade preditiva usa isso para mostrar dados históricos nos gráficos de Capacidade no console.
- 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.

Controlar quais instâncias do Auto Scaling serão terminadas durante uma redução de escala na horizontal

O Amazon EC2 Auto Scaling usa políticas de encerramento para decidir a ordem de encerramento de instâncias. Você pode usar uma política predefinida ou criar uma política personalizada para atender aos seus requisitos específicos. Ao usar uma política personalizada ou uma proteção escalável de instâncias, você também pode impedir que seu grupo de Auto Scaling encerre instâncias que ainda não estão prontas para serem encerradas.

Conteúdo

- Quando o Amazon EC2 Auto Scaling usa políticas de rescisão
- Configurar políticas de rescisão para o Amazon EC2 Auto Scaling
- Criar uma política de término personalizada com o Lambda
- Usar proteção de redução na escala na horizontal de instâncias
- Crie seus aplicativos no Amazon EC2 Auto Scaling para lidar com o encerramento de instâncias sem problemas

Quando o Amazon EC2 Auto Scaling usa políticas de rescisão

As seções a seguir descrevem os cenários em que o Amazon EC2 Auto Scaling usa políticas de término.

Conteúdo

- Eventos de redução de escala na horizontal
- Atualização de instância
- Rebalanceamento de zona de disponibilidade

Eventos de redução de escala na horizontal

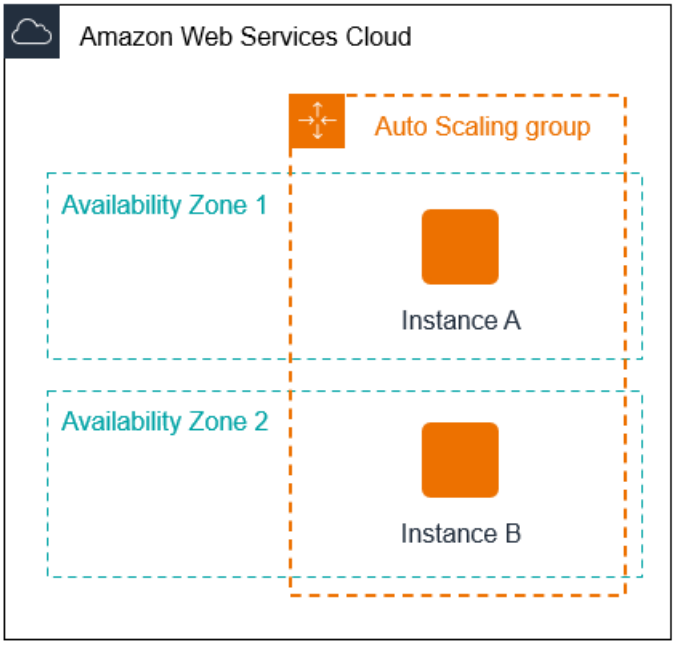
Um evento de redução de escala na horizontal ocorre quando há um novo valor para a capacidade desejada de um grupo do Auto Scaling que é menor do que a capacidade atual do grupo.

Eventos de redução de escala na horizontal ocorrem nos seguintes casos:

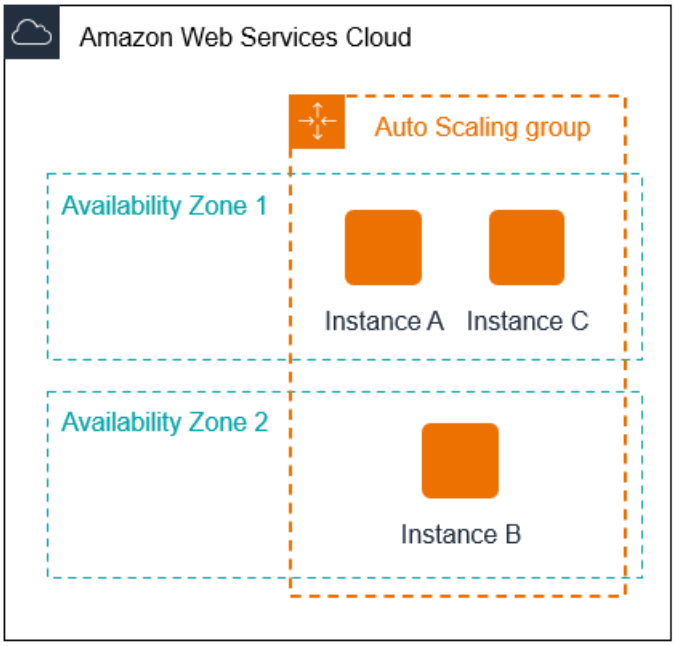
- Ao usar políticas de escalabilidade dinâmica e o tamanho do grupo diminui como resultado de alterações no valor de uma métrica
- Ao usar a escalabilidade programada e o tamanho do grupo diminui como resultado de uma ação programada
- Quando você reduz o tamanho do grupo manualmente

O exemplo a seguir mostra como as políticas de término funcionam quando há um evento de redução de capacidade na horizontal.

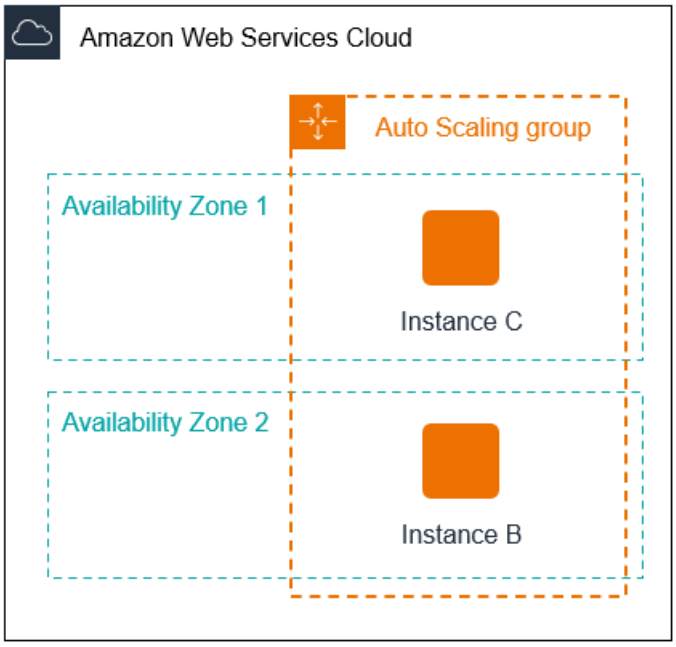
1. O grupo do Auto Scaling deste exemplo tem um tipo de instância, duas zonas de disponibilidade e uma capacidade desejada de duas instâncias. Ele também tem uma política de escalabilidade dinâmica que adiciona e remove instâncias quando a utilização de recursos aumenta ou diminui. As duas instâncias 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 seleciona a zona de disponibilidade com duas instâncias e encerra a instância que foi executada a partir de uma configuração de execução, de um modelo de execução diferente ou da versão mais antiga do modelo de execução atual. Se as instâncias foram lançadas a partir do mesmo modelo e versão de execução, o Amazon EC2 Auto Scaling seleciona a instância que está mais próxima da próxima hora de cobrança e a encerra.



Atualização de instância

Você pode iniciar uma atualização de instância para atualizar as instâncias em seu grupo de 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.

Configurar políticas de rescisão para o Amazon EC2 Auto Scaling

Uma política de rescisão fornece os critérios que o Amazon EC2 Auto Scaling segue para encerrar instâncias em uma ordem específica.

Por padrão, o Amazon EC2 Auto Scaling usa uma política de encerramento projetada para encerrar instâncias que estão usando primeiro configurações desatualizadas. Você pode alterar a política de encerramento para controlar quais instâncias são mais importantes para serem encerradas primeiro.

Quando o Amazon EC2 Auto Scaling encerra instâncias, ele tenta manter o equilíbrio entre as zonas de disponibilidade que estão habilitadas para seu grupo de Auto Scaling. A manutenção do equilíbrio zonal tem precedência sobre a política de rescisão. Se uma zona de disponibilidade tiver mais instâncias do que outras, o Amazon EC2 Auto Scaling aplica primeiro a política de rescisão à zona desequilibrada. Se as zonas de disponibilidade estiverem equilibradas, ela aplicará a política de rescisão em todas as zonas.

Tópicos

- Como funciona a política de rescisão padrão
- Política de término padrão e grupos de instâncias mistas
- Políticas de rescisão predefinidas
- Alterar a política de rescisão de um grupo do Auto Scaling

Como funciona a política de rescisão padrão

Quando o Amazon EC2 Auto Scaling precisa encerrar uma instância, ele primeiro identifica qual zona (ou zonas) de disponibilidade tem mais instâncias e pelo menos uma instância que não está protegida da escalabilidade. Em seguida, avalia as instâncias desprotegidas dentro da zona de disponibilidade identificada da seguinte forma:

Instâncias que usam configurações desatualizadas

- Para grupos que usam um modelo de execução — determine se alguma das instâncias usa configurações desatualizadas, priorizando nesta ordem:
  1. Primeiro, verifique as instâncias lançadas com uma configuração de execução.

2. Em seguida, verifique se há instâncias lançadas usando um modelo de execução diferente em vez do modelo de execução atual.
3. Por fim, verifique as instâncias usando a versão mais antiga do modelo de lançamento atual.

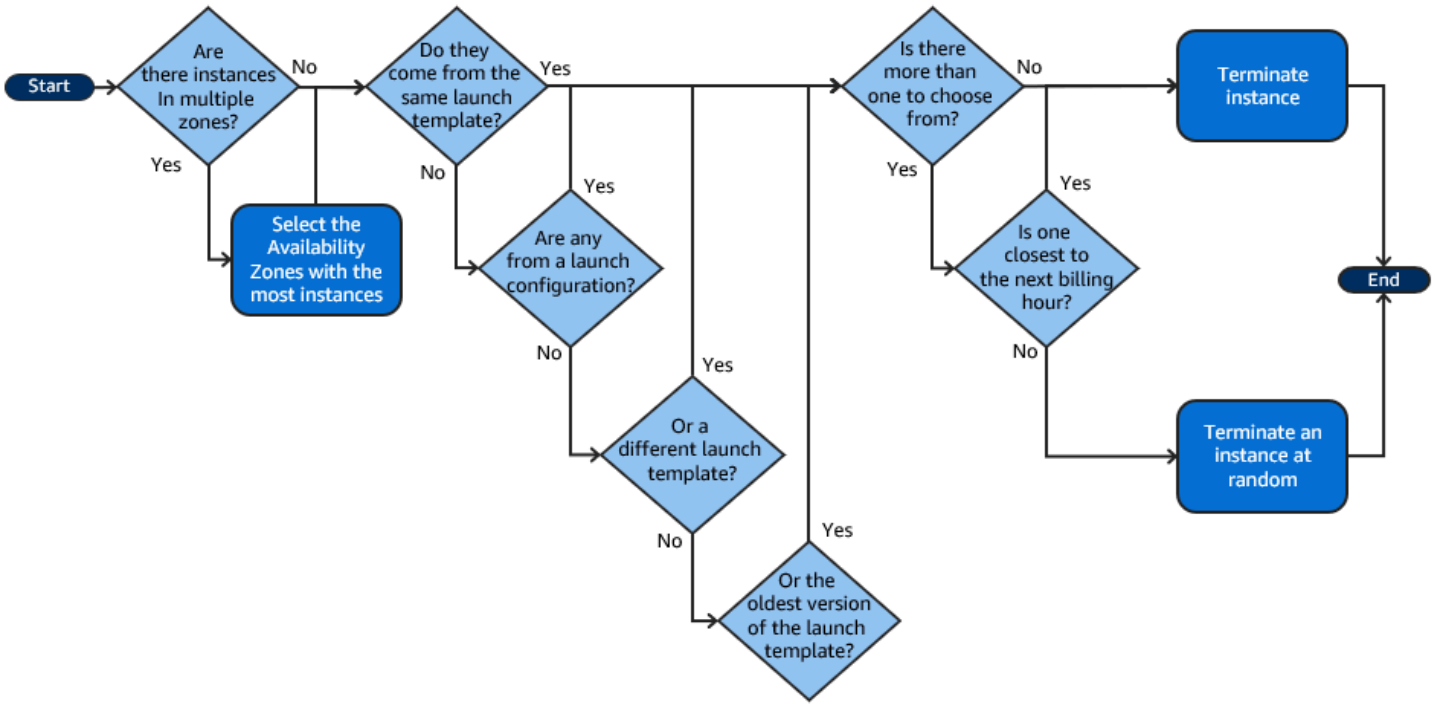
- Para grupos que usam uma configuração de execução — determine se alguma das instâncias usa a configuração de execução mais antiga.

Se nenhuma instância com configurações desatualizadas for encontrada ou se houver várias instâncias para escolher, o Amazon EC2 Auto Scaling considera os próximos critérios de instâncias que se aproximam da próxima hora de cobrança.

Instâncias que se aproximam da próxima hora de cobrança

Determine se alguma das instâncias que atendem aos critérios anteriores está mais próxima da próxima hora de cobrança. Se várias instâncias estiverem igualmente próximas, encerre uma aleatoriamente. Isso ajuda você a maximizar o uso de suas instâncias que são cobradas por hora. No entanto, a maior parte do uso do EC2 agora é cobrada por segundo, portanto, essa otimização oferece menos benefícios. Para obter mais informações, consulte Definição de preço do Amazon EC2.

O diagrama de fluxo a seguir ilustra como a política de rescisão padrão funciona para grupos que usam um modelo de lançamento.



Política de término padrão e grupos de instâncias mistas

O Amazon EC2 Auto Scaling aplica critérios adicionais ao encerrar instâncias em grupos de instâncias mistas.

Quando o Amazon EC2 Auto Scaling precisa encerrar uma instância, ele primeiro identifica qual opção de compra (spot ou sob demanda) deve ser encerrada com base nas configurações do grupo. Isso garante que o grupo tenha uma tendência em direção à proporção especificada de instâncias spot e sob demanda ao longo do tempo.

Em seguida, aplica a política de rescisão de forma independente dentro de cada zona de disponibilidade. Ele determina qual instância spot ou sob demanda em qual zona de disponibilidade encerrar para manter as zonas de disponibilidade equilibradas. A mesma lógica se aplica a um grupo misto de instâncias com pesos definidos para os tipos de instância.

Em cada zona, a política de encerramento padrão funciona da seguinte forma para determinar qual instância desprotegida dentro da opção de compra identificada pode ser encerrada:

1. Determine se alguma das instâncias pode ser encerrada para melhorar o alinhamento com a estratégia de alocação especificada para o grupo Auto Scaling. Se nenhuma instância for identificada para otimização ou se houver várias instâncias para escolher, a avaliação continuará.
2. Determine se alguma das instâncias usa configurações desatualizadas, priorizando nesta ordem:
   a. Primeiro, verifique as instâncias lançadas com uma configuração de execução.
   b. Em seguida, verifique se há instâncias lançadas usando um modelo de execução diferente em vez do modelo de execução atual.
   c. Por fim, verifique as instâncias usando a versão mais antiga do modelo de lançamento atual.

   Se nenhuma instância com configurações desatualizadas for encontrada ou se houver várias instâncias para escolher, a avaliação continuará.
3. Determine se alguma das instâncias está mais próxima da próxima hora de cobrança. Se várias instâncias estiverem igualmente próximas, escolha uma aleatoriamente.

Políticas de rescisão predefinidas

Você escolhe entre as seguintes políticas de rescisão predefinidas:

- **`Default`**— Encerrar instâncias de acordo com a política de rescisão padrão.

- **`AllocationStrategy`**— Encerre as instâncias no grupo Auto Scaling para alinhar as instâncias restantes à estratégia de alocação do tipo de instância que está sendo encerrada (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`**— Encerre instâncias que tenham o modelo de lançamento 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`**— Encerre instâncias que tenham 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`**— Encerre as instâncias que estão mais próximas da próxima hora de cobrança. Essa política ajuda a maximizar o uso de suas instâncias que têm uma taxa por hora.
- **`NewestInstance`**— Encerre 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`**— Encerre a instância mais antiga do grupo. Essa opção é útil quando você está atualizando as instâncias no grupo do Auto Scaling para um novo tipo de instância do EC2. Você pode substituir instâncias do tipo antigo gradualmente por instâncias do tipo novo.

> **Note**
>
> O Amazon EC2 Auto Scaling sempre equilibra as instâncias entre as zonas de disponibilidade primeiro, independentemente da política de término usada. Como resultado, você pode encontrar situações em que algumas instâncias mais recentes são terminadas antes de instâncias mais antigas. Por exemplo, quando há uma zona de disponibilidade adicionada mais recentemente ou quando uma zona de disponibilidade tiver mais instâncias que as outras zonas de disponibilidade que sejam usadas pelo grupo.

Alterar a política de rescisão de um grupo do Auto Scaling

Para alterar a política de rescisão do seu grupo de Auto Scaling, use um dos métodos a seguir.

Console

Você não pode alterar a política de rescisão ao criar inicialmente um grupo de Auto Scaling no console do Amazon EC2 Auto Scaling. A política de término padrão é usada automaticamente. Depois que seu grupo de Auto Scaling for criado, você poderá substituir a política padrão por uma política de rescisão diferente ou por várias políticas de rescisão listadas na ordem em que devem ser aplicadas.

Para alterar a política de rescisão de 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 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 ter mais informações, consulte Criar uma política de término personalizada com o Lambda.

   > **Note**
   >
   > Ao usar várias políticas, a ordem delas devem ser definida corretamente:
   >
   > - Se você usar a política Default (Padrão), coloque-a em último lugar na lista.
   > - Se você usar uma Custom termination policy (Política personalizada de encerramento), ela deve ser a primeira política na lista.

5. Escolha Atualizar.

AWS CLI

A política de término padrão é usada automaticamente, a menos que uma política diferente seja especificada.

Para alterar a política de rescisão de um grupo do Auto Scaling

Use um dos seguintes comandos:

- create-auto-scaling-group
- update-auto-scaling-group

Você pode usar as políticas de término individualmente ou combiná-las em uma lista de políticas. Por exemplo, use o comando a seguir para atualizar um grupo do Auto Scaling a fim de usar primeiro a política `OldestLaunchConfiguration` e, depois disso, usar a política `ClosestToNextInstanceHour`.

```
aws autoscaling update-auto-scaling-group --auto-scaling-group-name my-asg --
termination-policies "OldestLaunchConfiguration" "ClosestToNextInstanceHour"
```

Se você usar a política de encerramento `Default`, coloque-a no final da lista de políticas de encerramento. Por exemplo, `--termination-policies "OldestLaunchConfiguration" "Default"`.

Para usar uma política de rescisão personalizada, você deve primeiro criar sua política de rescisão usando AWS Lambda. Para especificar a função do Lambda a ser usada como política de término, torne-a a primeira na lista de políticas de término. Por exemplo, `--termination-policies "arn:aws:lambda:us-west-2:123456789012:function:HelloFunction:prod" "OldestLaunchConfiguration"`. Para ter mais informações, consulte Criar uma política de término personalizada com o Lambda.

Criar uma política de término personalizada com o Lambda

O Amazon EC2 Auto Scaling usa políticas de término para priorizar quais instâncias serão terminadas primeiro ao diminuir o tamanho do seu grupo do Auto Scaling (referido como redução de

escala na horizontal). O grupo do Auto Scaling usa uma política de término padrão, mas você pode, opcionalmente, escolher ou criar suas próprias políticas de término. Para obter mais informações sobre como escolher uma política de término predefinida, consulte Configurar políticas de rescisão para o Amazon EC2 Auto Scaling.

Neste tópico, você aprenderá como criar uma política de término personalizada usando uma função do AWS Lambda que o Amazon EC2 Auto Scaling chama em resposta a determinados eventos. A função do Lambda que você cria processa as informações nos dados de entrada enviados pelo Amazon EC2 Auto Scaling e devolve uma lista de instâncias que estão prontas para término.

Uma política de término personalizada fornece melhor controle sobre quais instâncias são terminadas e quando. Por exemplo, quando seu grupo do Auto Scaling sofre redução de escala na horizontal, o Amazon EC2 Auto Scaling não pode determinar se há workloads em execução que não devem ser interrompidas. Com uma função do Lambda, você pode validar a solicitação de término e aguardar até que a workload seja concluída antes de retornar o ID da instância ao Amazon EC2 Auto Scaling para término.

Conteúdo

- Dados de entrada
- Dados de resposta
- Considerações
- Criar a função do Lambda
- Limitações

Dados de entrada

O Amazon EC2 Auto Scaling gera uma carga útil JSON para eventos de redução de escala na horizontal e também faz isso quando as instâncias estão prestes a ser terminadas como resultado da duração máxima da instância ou dos recursos de atualização da instância. Ele também gera uma carga JSON para os eventos de redução de escala na horizontal que podem ser iniciados ao rebalancear seu grupo nas zonas de disponibilidade.

Essa carga contém informações sobre a capacidade que o Amazon EC2 Auto Scaling precisa terminar, uma lista de instâncias sugeridas para término e o evento que iniciou o término.

Esta é uma carga útil de exemplo:

```
{
```

```
  "AutoScalingGroupARN": "arn:aws:autoscaling:us-east-1:<account-
id>:autoScalingGroup:d4738357-2d40-4038-ae7e-b00ae0227003:autoScalingGroupName/my-asg",
  "AutoScalingGroupName": "my-asg",
  "CapacityToTerminate": [
    {
      "AvailabilityZone": "us-east-1b",
      "Capacity": 2,
      "InstanceMarketOption": "on-demand"
    },
    {
      "AvailabilityZone": "us-east-1b",
      "Capacity": 1,
      "InstanceMarketOption": "spot"
    },
    {
      "AvailabilityZone": "us-east-1c",
      "Capacity": 3,
      "InstanceMarketOption": "on-demand"
    }
  ],
  "Instances": [
    {
      "AvailabilityZone": "us-east-1b",
      "InstanceId": "i-0056faf8da3e1f75d",
      "InstanceType": "t2.nano",
      "InstanceMarketOption": "on-demand"
    },
    {
      "AvailabilityZone": "us-east-1c",
      "InstanceId": "i-02e1c69383a3ed501",
      "InstanceType": "t2.nano",
      "InstanceMarketOption": "on-demand"
    },
    {
      "AvailabilityZone": "us-east-1c",
      "InstanceId": "i-036bc44b6092c01c7",
      "InstanceType": "t2.nano",
      "InstanceMarketOption": "on-demand"
    },
    ...
  ],
  "Cause": "SCALE_IN"
}
```

A carga útil inclui o nome do grupo do Auto Scaling, seu nome do recurso da Amazon (ARN) e os seguintes elementos:

- `CapacityToTerminate` descreve o quanto da sua capacidade spot ou sob demanda está definida para ser terminada em uma determinada zona de disponibilidade.
- `Instances` representa as instâncias que o Amazon EC2 Auto Scaling sugere para término com base nas informações em `CapacityToTerminate`.
- `Cause` descreve o evento que acionou o término `SCALE_IN`, `INSTANCE_REFRESH`, `MAX_INSTANCE_LIFETIME` ou `REBALANCE`.

As informações a seguir descrevem os fatores mais significativos em como o Amazon EC2 Auto Scaling gera as `Instances` nos dados de entrada:

- A manutenção do equilíbrio entre as zonas de disponibilidade tem precedência quando uma instância está sendo terminada devido a eventos de aumento de escala na horizontal e términos baseados na atualização de instância. Dessa forma, se uma zona de disponibilidade tiver mais instâncias que as outras que são usadas pelo grupo, os dados de entrada contêm instâncias qualificáveis para término somente a partir da zona de disponibilidade desbalanceada. Se as zonas de disponibilidade usadas pelo grupo forem balanceadas, os dados de entrada conterão instâncias de todas as zonas de disponibilidade do grupo.
- Ao usar uma política de instâncias mistas, a manutenção das suas capacidades spot e sob demanda em equilíbrio com base nos percentuais desejados para cada opção de compra também tem precedência. Primeiro, identificamos qual dos dois tipos (spot ou sob demanda) deve ser terminado. Em seguida, identificamos quais instâncias (dentro da opção de compra identificada) em que zonas de disponibilidade serão terminadas que resultarão no maior equilíbrio das zonas de disponibilidade.

Dados de resposta

Os dados de entrada e os dados de resposta trabalham juntos para restringir a lista de instâncias a serem terminadas.

Com a entrada dada, a resposta de sua função do Lambda deve se parecer com o exemplo a seguir:

```
{
  "InstanceIDs": [
    "i-02e1c69383a3ed501",
```

```
    "i-036bc44b6092c01c7",
    ...
  ]
}
```

Os `InstanceIDs` na resposta representam as instâncias que estão prontas para sere terminadas.

Como alternativa, você pode devolver um conjunto diferente de instâncias que estão prontas para serem terminadas, o que substitui as instâncias nos dados de entrada. Se nenhuma instância estiver pronta para ser terminada quando sua função do Lambda for chamada, você também pode optar por não devolver nenhuma instância.

Quando não houver nenhuma instância pronta para encerramento, a resposta de sua função do Lambda deverá se parecer com o exemplo a seguir:

```
{
  "InstanceIDs": [ ]
}
```

Considerações

Observe as seguintes considerações ao usar uma política de término personalizada:

- Devolver uma instância primeiro nos dados de resposta não garante seu término. Se mais do que o número necessário de instâncias for devolvido quando sua função do Lambda for chamada, o Amazon EC2 Auto Scaling avaliará cada instância em relação às outras políticas de término especificadas para seu grupo do Auto Scaling. Quando há várias diretivas de término, ele tenta aplicar a próxima diretiva de término na lista e, se houver mais instâncias do que as necessárias para término, ele passa para a próxima diretiva de término, e assim por diante. Se nenhuma outra política de término for especificada, a política de término padrão será usada para determinar quais instâncias serão terminadas.
- Se nenhuma instância for devolvida ou se sua função do Lambda expirar, o Amazon EC2 Auto Scaling aguardará um curto período de tempo antes de chamar sua função novamente. Para qualquer evento de redução de escala na horizontal, ele continua tentando desde que a capacidade desejada do grupo seja menor que sua capacidade atual. Por exemplo, términos baseados em atualização, ele continua tentando por uma hora. Depois disso, se continuar a falhar ao terminar quaisquer instâncias, a operação de atualização da instância falhará. Com a duração máxima da instância, o Amazon EC2 Auto Scaling continua tentando terminar a instância identificada como excedendo sua vida útil máxima.

- Como sua função é repetida continuamente, certifique-se de testar e corrigir quaisquer erros permanentes em seu código antes de usar uma função do Lambda como uma política de término personalizada.
- Se você substituir os dados de entrada com sua própria lista de instâncias a serem terminadas, e o término dessas instâncias deixar as zonas de disponibilidade fora de equilíbrio, o Amazon EC2 Auto Scaling reequilibrará gradualmente a distribuição de capacidade entre zonas de disponibilidade. Primeiro, ele invoca sua função do Lambda para ver se existem instâncias que estão prontas para serem terminadas para que ele possa determinar se deseja iniciar o rebalanceamento. Se houver instâncias prontas para serem terminadas, ele iniciará novas instâncias primeiro. Quando as instâncias terminam de ser iniciadas, elas detectam que a capacidade atual do grupo é maior do que a capacidade desejada e iniciam um evento de redução de escala na horizontal.
- Uma política de encerramento personalizada não afeta sua capacidade de também usar proteção para reduzir a escala horizontalmente para evitar que determinadas instâncias sejam encerradas. Para ter mais informações, consulte Usar proteção de redução na escala na horizontal de instâncias.

Criar a função do Lambda

Comece criando a função do Lambda, para que você possa especificar seu nome do recurso da Amazon (ARN) nas políticas de término do seu grupo do Auto Scaling.

Para criar uma função do Lambda (console)

1. Abra a página Functions (Funções) no console do Lambda.
2. Na barra de navegação na parte superior da tela, escolha a mesma região usada ao criar o grupo do Auto Scaling.
3. Escolha Create function (Criar função) e Author from scratch (Criar desde o início).
4. Em Basic information (Informações básicas), para Function name (Nome da função), insira um nome para a função.
5. Escolha a opção Criar função. Você é retornado ao código e configuração da função.
6. Com sua função ainda aberta no console, em Function code (Código da função), cole seu código no editor.
7. Escolha Implantar.

8. Opcionalmente, crie uma versão publicada da função do Lambda escolhendo a guia Versions (Versões), e depois, Publish new version (Publicar nova versão). Para saber mais sobre controle de versões no Lambda, consulte Versões de função do Lambda no Guia do desenvolvedor do AWS Lambda .
9. Se você optou por publicar uma versão, escolha a guia Aliases caso deseje associar um alias a essa versão da função do Lambda. Para saber mais sobre aliases no Lambda, consulte Aliases de função do Lambda no Guia do desenvolvedor do AWS Lambda .
10. Em seguida, escolha a guia Configuration (Configuração) e, depois, Permissions (Permissões).
11. Role para baixo até Resource-based policy (Política baseada em recurso) e, em seguida, escolha Add permissions (Adicionar permissões). Uma política baseada em recurso é usada para conceder permissões para invocar sua função no principal que é especificado na política. Neste caso, o principal será 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. Para Ação, escolha lambda: InvokeFunction.
    d. Em Statement ID (ID da instrução), insira um ID de instrução exclusivo, como **`AllowInvokeByAutoScaling`**.
    e. Selecione 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 ter mais informações, consulte Alterar a política de rescisão de um grupo do Auto Scaling.

> **Note**
>
> Para exemplos que você pode usar como referência para desenvolver sua função Lambda, consulte o GitHub repositório do Amazon EC2 Auto Scaling.

Limitações

- Você só pode especificar uma função do Lambda nas políticas de término para um grupo do Auto Scaling. Se houver várias políticas de término especificadas, a função do Lambda deve ser especificada primeiro.
- Você pode referenciar sua função do Lambda usando um ARN não qualificado (sem um sufixo) ou um ARN qualificado que tenha uma versão ou um alias como sufixo. Se um ARN não qualificado for usado (por exemplo, `function:my-function`), sua política baseada em recurso deve ser criada na versão não publicada da sua função. Se um ARN qualificado for usado (por exemplo, `function:my-function:1` ou `function:my-function:prod`), sua política baseada em recurso deve ser criada na versão publicada específica da sua função.
- Você não pode usar um ARN qualificado com o sufixo `$LATEST`. Se você tentar adicionar uma política de término personalizada que se refira a um ARN qualificado com o sufixo `$LATEST`, isso resultará em um erro.
- O número de instâncias fornecidas nos dados de entrada é limitado a 30.000 instâncias. Se houver mais de 30.000 instâncias que possam ser terminadas, os dados de entrada incluirão `"HasMoreInstances": true` para indicar que o número máximo de instâncias é devolvido.
- O tempo máximo de execução para sua função do Lambda é de dois segundos (2000 milissegundos). Como prática recomendada, 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.
- Se o tempo de execução exceder o limite de 2 segundos, qualquer ação de expansão ficará suspensa até que o tempo de execução fique abaixo desse limite. Para funções do Lambda com tempos de execução consistentemente mais longos, encontre uma maneira de reduzir o tempo de execução, como armazenar em cache os resultados onde eles possam ser recuperados durante as invocações subsequentes do Lambda.

Usar proteção de redução na escala na horizontal de instâncias

A proteção escalável de instâncias permite que você controle quais instâncias o Amazon EC2 Auto Scaling pode encerrar. Um caso de uso comum desse recurso é escalar cargas de trabalho baseadas em contêineres. Para ter mais informações, consulte Crie seus aplicativos no Amazon EC2 Auto Scaling para lidar com o encerramento de instâncias sem problemas.

Por padrão, a proteção de escalabilidade de instâncias é desativada quando você cria um grupo de Auto Scaling. Isso significa que o Amazon EC2 Auto Scaling pode encerrar qualquer instância no grupo.

É possível proteger as instâncias assim que elas são iniciadas ao habilitar a configuração de proteção contra redução de escala na horizontal de instâncias no seu grupo do Auto Scaling. A proteção de redução de instâncias começa quando o estado da instância é `InService`. Em seguida, para controlar quais instâncias podem ser encerradas, desabilite a configuração de proteção escalável em instâncias individuais dentro do grupo do Auto Scaling. Ao fazer isso, você pode continuar protegendo determinadas instâncias contra encerramentos indesejados.

Tópicos

- Considerações
- Altere a proteção de escalabilidade para um grupo de Auto Scaling
- Alterar a proteção escalável de uma instância

Considerações

Veja a seguir algumas considerações ao usar a proteção escalável de instâncias:

- 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 histórico de atividades do grupo Auto Scaling inclui a seguinte mensagem se todas as instâncias em um grupo de Auto Scaling estiverem protegidas da escalabilidade quando ocorrer um evento de expansão: `Could not scale to desired capacity because all remaining instances are protected from scale-in.`
- Se você desvincular uma instância protegida contra redução de escala na horizontal, sua configuração de proteção de redução de instâncias será perdida. Quando a instância é associada ao grupo novamente, ela herda a configuração de proteção de redução de instâncias atual do grupo. Quando o Amazon EC2 Auto Scaling executa uma instância ou move uma instância de um 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.
- A proteção contra redução de escala na horizontal de instâncias não protege as instâncias do Auto Scaling contra o seguinte:

- Substituição da verificação se a instância não passar nas verificações de integridade. Para ter mais informações, consulte Verificações de integridade para instâncias em um grupo do Auto Scaling.
- Interrupções de instâncias spot Uma instância spot é encerrada quando a capacidade não está mais disponível ou o preço spot excede seu preço máximo.
- Uma reserva de bloco de capacidade termina. O Amazon EC2 recupera as instâncias do Capacity Block mesmo que elas estejam protegidas da escalabilidade.
- Encerramento manual por meio do `terminate-instance-in-auto-scaling-group` comando. Para ter mais informações, consulte Encerrar uma instância no seu grupo do Auto Scaling (AWS CLI).
- Encerramento manual por meio do console do Amazon EC2, comandos da CLI e operações de API. Para proteger as instâncias do Auto Scaling contra término manual, habilite a proteção contra término do Amazon EC2. (Isso não impede que o Amazon EC2 Auto Scaling encerre instâncias ou encerre manualmente por meio do comando.) `terminate-instance-in-auto-scaling-group` Para obter informações sobre como ativar a proteção contra rescisão do Amazon EC2 em um modelo de lançamento, consulte. Criar um modelo de execução usando configurações avançadas

Altere a proteção de escalabilidade para um grupo de Auto Scaling

É possível habilitar ou desabilitar a configuração de proteção contra redução de escala na horizontal de instâncias para um grupo do Auto Scaling. Quando você a habilita, todas as novas instâncias lançadas pelo grupo terão a proteção de escalabilidade de instância ativada.

Ativar ou desativar essa configuração para um grupo de Auto Scaling não afeta as instâncias existentes.

Console

Para habilitar a proteção escalável para um novo grupo de Auto Scaling

Ao criar o grupo Auto Scaling, na página Configurar tamanho do grupo e políticas de escalabilidade, em Proteção de escalabilidade de instância, marque a caixa de seleção Habilitar proteção de escalabilidade de instância.

Para ativar ou desativar a proteção de expansão para 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. 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 de escalabilidade de instância, marque ou desmarque a caixa de seleção Ativar proteção de escalabilidade de instância para ativar ou desativar essa opção conforme necessário.
5. Escolha Atualizar.

AWS CLI

Para habilitar a proteção escalável para um novo grupo de Auto Scaling

Use o create-auto-scaling-groupcomando a seguir para ativar a proteção de escalabilidade da instância.

```
aws autoscaling create-auto-scaling-group --auto-scaling-group-name my-asg --new-instances-protected-from-scale-in ...
```

Para habilitar a proteção escalável para um grupo existente

Use o update-auto-scaling-groupcomando a seguir para ativar a proteção de escalabilidade da instância para o grupo de Auto Scaling especificado.

```
aws autoscaling update-auto-scaling-group --auto-scaling-group-name my-asg --new-instances-protected-from-scale-in
```

Para desativar a proteção de expansão para um grupo existente

Use o seguinte comando para desabilitar a proteção de redução de instâncias para o grupo especificado.

```
aws autoscaling update-auto-scaling-group --auto-scaling-group-name my-asg --no-new-instances-protected-from-scale-in
```

Alterar a proteção escalável de uma instância

Por padrão, uma instância obtém sua configuração de proteção contra redução de escala na horizontal de instâncias de seu grupo do Auto Scaling. No entanto, você pode ativar ou desativar a proteção de escalabilidade de instâncias para instâncias individuais após a inicialização.

Console

Para ativar ou desativar a proteção de expansão para uma 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.
3. Na guia Instance management (Gerenciamento de instâncias), em Instances (Instâncias), selecione uma instância.
4. Para habilitar a proteção de redução de instâncias, escolha Actions (Ações) e Set scale-in protection (Definir proteção de redução). Quando solicitado, escolha Set scale-in protection (Definir proteção de redução).
5. Para desabilitar a proteção de redução de instâncias, escolha Actions (Ações) e Remove scale-in protection (Remover proteção de redução). Quando solicitado, escolha Remove Scale In Protection (Remover proteção de redução).

AWS CLI

Para habilitar a proteção escalável para uma instância

Use o set-instance-protectioncomando a seguir para ativar a proteção de escalabilidade da instância para a instância especificada.

```
aws autoscaling set-instance-protection --instance-ids i-5f2e8a0d --auto-scaling-
group-name my-asg --protected-from-scale-in
```

Para desativar a proteção de expansão para uma instância

Use o seguinte comando para desabilitar a proteção de redução para a instância especificada,

```
aws autoscaling set-instance-protection --instance-ids i-5f2e8a0d --auto-scaling-
group-name my-asg --no-protected-from-scale-in
```

> **Note**
>
> Lembre-se de que a proteção escalável de instâncias não garante que as instâncias não sejam encerradas no caso de um erro humano, por exemplo, se alguém encerrar manualmente uma instância usando o console do Amazon EC2 ou. AWS CLI Para proteger sua instância contra término acidental, use a proteção contra término do Amazon EC2. No entanto, mesmo com a proteção contra término e a proteção de aumento de escala na horizontal de instâncias habilitadas, os dados salvos no armazenamento da instância podem ser perdidos se uma verificação de integridade determinar que uma instância não está íntegra ou se o próprio grupo for excluído acidentalmente. Como em qualquer ambiente, uma prática recomendada é fazer backup de seus dados com frequência ou sempre que for apropriado para seus requisitos de continuidade de negócios.

Crie seus aplicativos no Amazon EC2 Auto Scaling para lidar com o encerramento de instâncias sem problemas

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 redução horizontal da escala.

Por exemplo, suponha que você tenha uma fila do Amazon SQS que coleta mensagens recebidas para trabalhos de longa execução. Quando uma nova mensagem chega, uma instância no grupo do 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 do Auto Scaling. À medida que o número de mensagens diminui, as instâncias existentes são automaticamente encerradas. 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 criar seu aplicativo para responder a esse cenário.

Você pode usar os seguintes atributos para evitar que seu grupo do Auto Scaling encerre instâncias que ainda não estão prontas para serem encerradas ou encerre instâncias muito rapidamente para que concluam os trabalhos atribuídos. Todos esses três atributos podem ser usados em combinação ou separadamente.

Conteúdo

- Proteção de redução horizontal de escala de instâncias
- Política de encerramento personalizada
- hook do ciclo de vida de encerramento

> **Important**
>
> Ao criar seus aplicativos no Amazon EC2 Auto Scaling para lidar com o encerramento de instâncias sem problemas, lembre-se desses pontos.
>
> - Se uma instância não estiver íntegra, o Amazon EC2 Auto Scaling a substituirá independentemente do atributo que você usar (a menos que você suspenda o processo `ReplaceUnhealthy`). Você pode usar um hook do ciclo de vida para permitir que o aplicativo seja desligado normalmente ou copie quaisquer dados que você precise recuperar antes que a instância seja encerrada.
> - Não é garantido que um hook 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 encerra a instância.

Proteção de redução horizontal de escala de instâncias

Você pode usar a proteção de redução horizontal de escala de instâncias em muitas situações em que o encerramento de instâncias é uma ação crítica que deve ser negada por padrão e permitida apenas explicitamente para instâncias específicas. Por exemplo, ao executar workloads 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 criaram integrações com proteção de redução horizontal de escala de instâncias em seus produtos.

Você pode habilitar a proteção de redução horizontal de escala no grupo do Auto Scaling para aplicar a proteção de redução horizontal de escala às instâncias quando elas são 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 ambiente de gerenciamento centralizado que gerencia se uma instância pode ser encerrada ou não, ou das próprias instâncias. No entanto, uma grande frota pode enfrentar problemas de controle de utilização se um grande número de instâncias estiver alternando continuamente sua proteção de redução horizontal de escala.

Para ter mais informações, consulte Usar proteção de redução na escala na horizontal de instâncias.

Política de encerramento personalizada

Assim como a proteção contra redução redução horizontal de escala de instâncias, uma política de encerramento personalizada ajuda a impedir que seu grupo do Auto Scaling encerre instâncias específicas.

Por padrão, o grupo do Auto Scaling usa uma política de encerramento padrão para determinar quais instâncias ele encerra primeiro. Se você quiser ter mais controle sobre quais instâncias serão encerradas primeiro, você pode implementar sua própria política de encerramento personalizada usando uma função do Lambda. O Amazon EC2 Auto Scaling chama a função sempre que precisa decidir qual instância deve ser encerrada. 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 se souber quando uma instância é suficientemente redundante ou subutilizada para que possa ser encerrada. Para oferecer suporte a isso, você precisa implementar seu aplicativo com um ambiente de gerenciamento que monitore o workload em todo o grupo. Dessa forma, se uma instância ainda estiver processando trabalhos, a função do Lambda sabe que não deve incluí-la.

Para ter mais informações, consulte Criar uma política de término personalizada com o Lambda.

hook do ciclo de vida de encerramento

Um hook do ciclo de vida de encerramento prolonga a vida útil de uma instância que já foi selecionada para encerramento. Ele fornece tempo extra para concluir todas as mensagens ou solicitações atualmente atribuídas à instância, ou para salvar o avanço e transferir o trabalho para outra instância.

Para muitos workloads, um hook do ciclo de vida pode ser suficiente para encerrar normalmente um aplicativo em uma instância selecionada para encerramento. Esta é a melhor abordagem e não pode ser usada para evitar o encerramento em caso de falha.

Para usar um hook do 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
No interior da instância	O serviço de metadados de instância (IMDS) é um endpoint seguro que você pode pesquisar o status de uma instância diretamente da instância. Se os metadados voltarem com,`Terminated` sua instância está programada para ser encerrada.	Aplicativos em que você deve realizar uma ação na instância antes que ela seja encerrada.	Recuperar o estado de destino do ciclo de vida
Fora da instância	Quando uma instância é encerrada, uma notificação de evento é gerada. Você pode criar regras usando Amazon EventBridge, 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 destino de notificação

Para usar um hook do ciclo de vida, você também precisa saber quando sua instância está pronta para ser totalmente encerrada. O Amazon EC2 Auto Scaling não instruirá o Amazon EC2 a encerrar a instância até receber CompleteLifecycleActionuma chamada ou até que o tempo limite termine, o que ocorrer primeiro.

Por padrão, uma instância pode continuar em execução por uma hora (tempo limite de pulsação) devido a um hook 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 RecordLifecycleActionHeartbeatchamadas de API.

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

Suspender e retomar os processos do Amazon EC2 Auto Scaling

Este tópico descreve como suspender e, em seguida, retomar um ou mais dos processos do seu grupo de Auto Scaling para desativar temporariamente determinadas operações.

Suspender processos pode ser útil quando você precisa investigar ou solucionar um problema sem interferência de políticas de escalabilidade ou ações programadas. Isso também ajuda a evitar que o Amazon EC2 Auto Scaling marque instâncias não íntegras e as substitua enquanto você faz alterações em seu grupo de Auto Scaling.

Tópicos

- Tipos de processos
- Considerações
- Suspender processos
- Processos de currículo
- Como os processos suspensos afetam outros processos

> **Note**
>
> Além de suspensões que você inicia, o Amazon EC2 Auto Scaling também pode suspender processos dos grupos do Auto Scaling que falharem repetidamente ao iniciar instâncias. Isso é conhecido como suspensão administrativa. Uma suspensão administrativa se aplica mais comumente a grupos do Auto Scaling que estão tentando iniciar instâncias por mais de 24 horas, mas não tiveram êxito. Você pode retomar os processos que foram suspensos pelo Amazon EC2 Auto Scaling por motivos administrativos.

Tipos de processos

O recurso suspender-retomar é compatível com os seguintes processos:

- `Launch`— Adiciona instâncias ao grupo Auto Scaling quando o grupo se expande ou quando o Amazon EC2 Auto Scaling opta por iniciar instâncias por outros motivos, como quando adiciona instâncias a um pool aquecido.
- `Terminate`— Remove instâncias do grupo Auto Scaling quando o grupo se expande ou quando o Amazon EC2 Auto Scaling opta por encerrar instâncias por outros motivos, como quando uma instância é encerrada por exceder sua duração máxima de vida útil ou por falhar em uma verificação de saúde.
- `AddToLoadBalancer`— Adiciona instâncias ao grupo-alvo do balanceador de carga vinculado ou ao Classic Load Balancer quando elas são iniciadas. Para ter mais informações, consulte Usar o Elastic Load Balancing para distribuir tráfego entre as instâncias no grupo do Auto Scaling.
- `AlarmNotification`— Aceita notificações de CloudWatch alarmes associados a políticas de escalabilidade dinâmica. Para ter mais informações, consulte Escalabilidade dinâmica para o Amazon EC2 Auto Scaling.
- `AZRebalance`— Equilibra o número de instâncias do EC2 no grupo uniformemente em todas as zonas de disponibilidade especificadas quando o grupo fica desequilibrado, por exemplo, quando uma zona de disponibilidade anteriormente indisponível retorna a um estado saudável. Para ter mais informações, consulte Atividades de rebalanceamento.
- `HealthCheck`— Verifica a integridade das instâncias e marca uma instância como não íntegra se o Amazon EC2 ou o Elastic Load Balancing informarem 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 ter mais informações, consulte Verificações de integridade para instâncias em um grupo do Auto Scaling.
- `InstanceRefresh`— Encerra e substitui instâncias usando o recurso de atualização de instâncias. Para ter mais informações, consulte Use uma atualização de instância para atualizar instâncias em um grupo de Auto Scaling.
- `ReplaceUnhealthy`— Encerra instâncias marcadas como não íntegras e, em seguida, cria novas instâncias para substituí-las. Para ter mais informações, consulte Verificações de integridade para instâncias em um grupo do Auto Scaling.
- `ScheduledActions`— Executa as ações de escalabilidade programadas que você cria ou que são criadas para você quando você cria um plano de AWS Auto Scaling escalabilidade e ativa a

escala preditiva. Para ter mais informações, consulte Escalabilidade programada para o Amazon EC2 Auto Scaling.

Considerações

Antes de suspender processos, considere o seguinte:

- A suspensão `AlarmNotification` permite que você interrompa temporariamente as políticas de rastreamento, etapas e escalabilidade simples de metas 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.
- Você pode optar por suspender `ReplaceUnhealthy` os processos `HealthCheck` e reinicializar as instâncias sem que o Amazon EC2 Auto Scaling encerre as instâncias com base em suas verificações de saúde. No entanto, se você precisar do Amazon EC2 Auto Scaling para continuar realizando verificações de saúde nas instâncias restantes, use o recurso de espera em vez disso. Para ter mais informações, consulte Remover temporiamente instâncias do grupo do Auto Scaling.
- Se você suspender os processos `Launch` e `Terminate`, ou `AZRebalance`, e então fizer alterações em seu grupo do Auto Scaling, p. ex., desvinculando instâncias ou alterando as zonas de disponibilidade especificadas, seu grupo poderá ficar desbalanceado entre as zonas de disponibilidade. Se isso acontecer, depois que você retomar os processos suspensos, o Amazon EC2 Auto Scaling redistribuirá gradualmente as instâncias de modo uniforme entre as zonas de disponibilidade.
- Se você suspender o `Terminate` processo, ainda poderá forçar o encerramento das instâncias usando o delete-auto-scaling-groupcomando com a opção force delete.
- A suspensão do `Terminate` processo se aplica somente às instâncias que estão atualmente no `InService` estado. Isso não impede o encerramento de instâncias em outros estados, como`Pending`, ou instâncias que não sejam retomadas adequadamente do modo de espera.
- O `RemoveFromLoadBalancerLowPriority` processo pode ser ignorado quando está presente em chamadas para descrever grupos de Auto Scaling usando os SDKs AWS CLI ou. Esse processo está defasado e é retido somente para fins de compatibilidade retroativa.

Suspender processos

Para suspender um processo para um grupo de Auto Scaling, use um dos seguintes métodos:

Console

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 Atualizar.

AWS CLI

Use o seguinte comando suspend-processes para suspender processos individuais.

```
aws autoscaling suspend-processes --auto-scaling-group-name my-asg --scaling-
processes HealthCheck ReplaceUnhealthy
```

Para suspender todos os processos, omita a opção `--scaling-processes` como indicado abaixo.

```
aws autoscaling suspend-processes --auto-scaling-group-name my-asg
```

Processos de currículo

Para retomar um processo suspenso para um grupo do Auto Scaling, use um dos seguintes métodos:

Console

Para retomar um processo suspenso

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 suspenso.
5. Escolha Atualizar.

AWS CLI

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
```

Como os processos suspensos afetam outros processos

As seções a seguir descrevem o que acontece quando processos diferentes são suspensos individualmente.

Tópicos

- Launchestá suspenso
- Terminateestá suspenso
- AddToLoadBalancerestá suspenso
- AlarmNotificationestá suspenso
- AZRebalanceestá suspenso
- HealthCheckestá suspenso
- InstanceRefreshestá suspenso
- ReplaceUnhealthyestá suspenso
- ScheduledActionsestá suspenso
- Considerações adicionais

Launchestá suspenso

- `AlarmNotification` ainda está ativo, mas seu grupo do Auto Scaling não pode iniciar atividades de aumento da escala na horizontal para alarmes que estejam violados.
- `ScheduledActions` está ativo, mas seu grupo do Auto Scaling não pode iniciar atividades de aumento da escala na horizontal para nenhuma ação de alarme que ocorra.
- `AZRebalance` deixa de rebalancear o grupo.
- `ReplaceUnhealthy` continua a encerrar instâncias não íntegras, mas não inicia instâncias substitutas. Quando você retomar o processo`Launch`, o Amazon EC2 Auto Scaling substituirá imediatamente todas as instâncias que ele encerrou enquanto `Launch` estava suspenso.
- `InstanceRefresh` não substitui as instâncias.

Terminateestá 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.

AddToLoadBalancerestá suspenso

- O Amazon EC2 Auto Scaling executa as instâncias, mas não as adiciona ao grupo de destino do balanceador de carga ou ao Classic Load Balancer. Quando você retomar o processo `AddToLoadBalancer`, ele retomará a adição de instâncias ao balanceador de carga quando elas

forem iniciadas. No entanto, ele não adicionará as instâncias que foram iniciadas enquanto esse processo estava suspenso. Você deve registrar essas instâncias manualmente.

`AlarmNotification`está suspenso

- O Amazon EC2 Auto Scaling não invoca políticas de escalabilidade quando CloudWatch um limite de alarme é violado. Quando você retomar o `AlarmNotification`, o Amazon EC2 Auto Scaling levará em consideração as políticas com limites de alarme que estejam sendo violados no momento.

`AZRebalance`está suspenso

- Seu grupo do Amazon EC2 Auto Scaling não tenta redistribuir instâncias após determinados eventos. No entanto, se ocorrer um evento de expansão ou de redução, o processo de escalabilidade ainda tentará balancear as zonas de disponibilidade. Por exemplo, durante 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`.

`HealthCheck`está suspenso

- O Amazon EC2 Auto Scaling interrompe a marcação de instâncias com problemas de integridade como resultado das verificações de integridade do EC2 e do Elastic Load Balancing. Suas verificações personalizadas de integridade continuam funcionando corretamente. Depois que você suspender `HealthCheck`, se precisar, defina manualmente o estado de integridade das instâncias no seu grupo e faça com que o `ReplaceUnhealthy` as substitua.

`InstanceRefresh`está suspenso

- O Amazon EC2 Auto Scaling interrompe a substituição de instâncias como resultado de uma atualização de instância. Se houver uma atualização de instância em andamento, isso pausará a operação sem cancelá-la.

`ReplaceUnhealthy`está suspenso

- O Amazon EC2 Auto Scaling interrompe a substituição de instâncias que estão marcadas como não íntegras. As instâncias que falharem nas verificações de integridade do EC2 ou do Elastic Load Balancing ainda serão marcadas como não íntegras. Assim que você retomar o processo `ReplaceUnhealthy`, o Amazon EC2 Auto Scaling substituirá as instâncias que foram marcadas como não íntegras enquanto esse processo estava suspenso. O processo `ReplaceUnhealthy` chama `Terminate` primeiro e depois `Launch`.

`ScheduledActions`está suspenso

- O Amazon EC2 Auto Scaling não executa ações de escalabilidade que estejam programadas para execução durante o período de suspensão. Quando você retomar o `ScheduledActions`, o Amazon EC2 Auto Scaling considerará apenas ações programadas cuja programação ainda não tenha expirado.

Considerações adicionais

Além disso, quando `Launch` ou `Terminate` estiverem suspensos, os seguintes recursos podem não funcionar corretamente:

- Vida útil máxima da instância — Quando `Launch` ou `Terminate` são suspensas, o recurso de vida útil máxima da instância não pode substituir nenhuma instância.
- Interrupções de instâncias spot — Se `Terminate` estiver suspensa e seu grupo de 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.
- Rebalanceamento de capacidade — Se `Terminate` estiver suspenso e você usar o rebalanceamento de capacidade para lidar com interrupções de instâncias 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.

- Anexação e desvinculação de instâncias — Quando `Launch` e `Terminate` são suspensas, você pode desanexar instâncias que estão anexadas ao seu grupo de Auto Scaling, mas enquanto estiverem `Launch` suspensas, você não poderá anexar novas instâncias ao grupo.
- Instâncias em espera — Quando `Launch` e `Terminate` estão suspensas, você pode colocar uma instância no `Standby` estado, mas enquanto `Launch` estiver suspensa, você não pode devolver uma instância no `Standby` estado ao serviço.

Monitore seus grupos do Auto Scaling

O monitoramento é uma parte importante para manter a confiabilidade, a disponibilidade e o desempenho do Amazon EC2 Auto Scaling e de suas Nuvem AWS soluções. AWS fornece as seguintes ferramentas de monitoramento para observar o Amazon EC2 Auto Scaling, relatar quando algo está errado e tomar ações automáticas quando apropriado:

Verificações de integridade

O Amazon EC2 Auto Scaling executa verificações de integridade nas instâncias do seu grupo do Auto Scaling. Se uma instância não passar na verificação de integridade, ela será marcada como não íntegra e será encerrada enquanto o Amazon EC2 Auto Scaling inicia uma nova instância para substituí-la. Para obter mais informações, consulte Verificações de integridade para instâncias em um grupo do Auto Scaling.

AWS Health Dashboard

O AWS Health Dashboard exibe informações e também fornece notificações que são invocadas por alterações no funcionamento dos AWS recursos. As informações são apresentadas de duas formas: em um painel que mostra eventos recentes e futuros organizados por categoria e em um log de eventos completo que mostra todos os eventos dos últimos 90 dias. Para obter mais informações, consulte AWS Health Dashboard notificações para Amazon EC2 Auto Scaling.

CloudTrail

Com o,AWS CloudTrail você pode rastrear as chamadas feitas para a API do Amazon EC2 Auto Scaling por ou em nome do seu clienteConta da AWS. O CloudTrail armazena as informações em arquivos de log no bucket do Amazon S3 que você especificar. Você pode usar esses arquivos de log para monitorar a atividade de seus grupos do Auto Scaling. Os logs incluem quais solicitações foram feitas, os endereços IP de onde as solicitações vieram, quem fez a solicitação, quando a solicitação foi feita e assim por diante. Para obter mais informações, consulte Registre chamadas da API do Amazon EC2 Auto Scaling com AWS CloudTrail.

> **Coleta de logs nas instâncias do Amazon EC2**
>
> Você pode usar o CloudWatch para coletar logs dos sistemas operacionais para suas instâncias do EC2. Para obter mais informações, consulte Coletar métricas e logs de instâncias do Amazon EC2 e servidores locais com o agente do CloudWatch e Visualizar dados de log enviados ao CloudWatch Logs no Guia do usuário do Amazon CloudWatch.

Para obter informações sobre outros serviços da AWS que podem ajudar você a registrar em log e coletar dados sobre suas workloads, consulte o Guia de registro em log e monitoramento para proprietários de aplicações na Orientação prescritiva da AWS.

Amazon CloudWatch

O Amazon CloudWatch ajuda você a analisar logs, além de monitorar em tempo real as métricas dos seus recursos e aplicações hospedadas na AWS. 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 o uso do serviço de monitorar as métricas dos grupos do Auto Scaling e instâncias, consulte Monitorar métricas do CloudWatch para grupos e instâncias do Auto Scaling.

O CloudWatch também rastreia métricas de uso de AWS API para Amazon EC2 Auto Scaling. Você pode usar essas métricas para configurar alarmes que alertem quando o volume de chamadas da API violar um limite definido por você. Para obter mais informações, consulte Métricas de uso da AWS no Guia do usuário do Amazon CloudWatch.

AWS Compute Optimizer

O Compute Optimizer informa as recomendações de instâncias do Amazon EC2 que podem ajudar a decidir se deseja passar para um novo tipo de instância. Ele analisa se o tipo de instância do grupo do Auto Scaling está em condições ideais e gera recomendações de otimização para reduzir o custo e melhorar o desempenho de suas workloads.. Para obter mais informações, consulte Use AWS Compute Optimizer para obter recomendações sobre o tipo de instância para um grupo de Auto Scaling.

Amazon EventBridge

O Amazon EventBridge é um serviço de barramento de eventos sem servidor que facilita a conexão de aplicações a dados de diversas origens. O EventBridge fornece um fluxo de dados em tempo real de suas próprias aplicações, de aplicações de software como serviço (SaaS) e de serviços da AWS e roteia esses dados para destinos como o Lambda. Isso permite monitorar eventos que ocorrem em serviços e criar arquiteturas orientadas a eventos. Para obter mais informações, consulte Use EventBridge para lidar com eventos do Auto Scaling.

AWS Security Hub

Use AWS Security Hub para monitorar o uso do Amazon EC2 Auto Scaling já que ele se refere às melhores práticas de segurança. O Security Hub usa controles de segurança detetives para avaliar configurações de recursos e padrões de segurança que ajudam você a cumprir vários frameworks 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 no 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 Opções de notificação do Amazon SNS para o Amazon EC2 Auto Scaling.

Verificações de integridade para instâncias em um grupo do Auto Scaling

O Amazon EC2 Auto Scaling monitora continuamente o status de saúde das instâncias em um grupo de Auto Scaling para manter a capacidade desejada.

Todas as instâncias em um grupo de Auto Scaling começam 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. Ele pode receber notificações de várias fontes quando uma instância não é íntegra e precisa ser substituída. Essas fontes incluem:

- Amazon EC2
- Elastic Load Balancing
- VPC Lattice
- Verificações de saúde personalizadas que você define

Quando o Amazon EC2 Auto Scaling determina que `InService` uma instância não está íntegra, ele a substitui por uma nova instância para manter a capacidade desejada do grupo. A nova instância é iniciada usando as configurações atuais do grupo do Auto Scaling e seu modelo de execução associado ou configuração de execução.

Instâncias não íntegras também podem ocorrer quando uma instância é encerrada inesperadamente, como devido à interrupção de uma instância spot ou ao encerramento manual por um usuário.

Novamente, o Amazon EC2 Auto Scaling iniciará automaticamente uma instância substituta nesses casos para manter a capacidade desejada.

Conteúdo

- Sobre as verificações de integridade do seu grupo do Auto Scaling
- Veja o motivo das falhas na verificação de integridade
- Definir um período de carência da verificação de integridade para um grupo do Auto Scaling

Sobre as verificações de integridade do seu grupo do Auto Scaling

Este tópico dá uma visão geral dos tipos de verificação de integridade padrão e disponíveis e descreve como eles funcionam.

Conteúdo

- Tipo de verificação de integridade
- Verificações de integridade do Amazon EC2
- Verificações de integridade do Elastic Load Balancing
- Verificações de integridade do VPC Lattice
- Como o Amazon EC2 Auto Scaling minimiza o tempo de inatividade
- Considerações sobre a verificação de integridade
- Verificações de integridade personalizadas
- Recursos relacionados

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	• Verifica se a instância está em execução • Verifica se há problemas subjacentes de hardware ou software capazes de prejudicar a instância

Health check type (Tipo de verificação de integridade)	O que ele verifica
	Esse é o tipo padrão de verificação de integridade para um grupo do Auto Scaling.
Verificações de integridade do Elastic Load Balancing	• Verifica se o balanceador de carga relata a instância como íntegra, confirmando se a instância está disponível para processar solicitações. Para executar esse tipo de verificação de integridade, você deve habilitá-lo para seu grupo do Auto Scaling.
Verificações de integridade do VPC Lattice	• Verifica se o VPC Lattice relata a instância como íntegra, confirmando se a instância está disponível para lidar com solicitações Para executar esse tipo de verificação de integridade, você deve habilitá-lo para seu grupo do Auto Scaling.
Verificações de integridade personalizadas	• Verifica se há outros problemas que possam indicar problemas de integridade da instância de acordo com suas verificações de integridade personalizadas

Verificações de integridade do Amazon EC2

Depois que uma instância é executada, ela é anexada ao grupo do Auto Scaling e entra no estado`InService`. Para obter mais informações sobre os diferentes status do ciclo de vida de instâncias em um grupo do Auto Scaling, consulte Ciclo de vida das instâncias do Amazon EC2 Auto Scaling.

O Amazon EC2 Auto Scaling verifica periodicamente o status de integridade de todas as instâncias no grupo do Auto Scaling para garantir que elas estejam em execução e em boas condições.

Verificações do status

O Amazon EC2 Auto Scaling usa os resultados das verificações de status da instância do Amazon EC2 para determinar o status de integridade de uma instância. Se a instância estiver em qualquer

estado do Amazon EC2 diferente de `running` ou se o status para as verificações de status mudar para `impaired`, o Amazon EC2 Auto Scaling considerará a instância como não íntegra e a substituirá. Isso inclui quando a instância tenha qualquer um dos seguintes estados:

- `stopping`
- `stopped`
- `shutting-down`
- `terminated`

As verificações de status do Amazon EC2 não exigem nenhuma configuração especial e estão sempre habilitadas. Para obter mais informações, consulte Tipos de verificações de status no Manual do usuário do Amazon EC2 para instâncias do Linux.

> **Important**
>
> O Amazon EC2 Auto Scaling permite que essas verificações de status falhem ocasionalmente sem realizar qualquer ação. Quando uma verificação de status falha, o Amazon EC2 Auto Scaling espera alguns minutos AWS para corrigir o problema. Ele não marca imediatamente uma instância `Unhealthy` quando seu status para as 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. Neste caso, marca imediatamente a instância `Unhealthy` e a substitui.

Eventos agendados

O Amazon EC2 pode agendar ocasionalmente eventos em suas instâncias que serão executados após um carimbo de data/hora específico. Para obter mais informações, consulte Eventos programados para sua instância no Guia do usuário do Amazon EC2 para instâncias do Linux.

Se uma de suas instâncias for afetada por um evento programado, o Amazon EC2 Auto Scaling considerará a instância como não íntegra e a substituirá. A instância não começa a ser encerrada até que a data e a hora especificadas no carimbo de data/hora sejam atingidas.

Verificações de 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. Depois de fazer isso, quando o Elastic Load Balancing relata uma instância registrada como`Unhealthy`, o Amazon EC2 Auto Scaling marca a instância `Unhealthy` na próxima verificação de integridade periódica e a substitui.

Se a drenagem da conexão (atraso de cancelamento de registro) estiver habilitada para seu balanceador de carga, o Amazon EC2 Auto Scaling aguardará que as solicitações em andamento sejam concluídas ou que o tempo limite máximo expire antes de terminar instâncias não íntegras.

Para saber como habilitar verificações de integridade do Elastic Load Balancing para seu grupo de Auto Scaling, consulte Anexe um balanceador de carga do Elastic Load Balancing ao seu grupo de Auto Scaling.

> **Note**
>
> Quando você habilita as verificações de integridade do Elastic Load Balancing para um grupo, o Amazon EC2 Auto Scaling pode substituir instâncias que o Elastic Load Balancing

relata como não íntegras, mas somente depois que o balanceador de carga estiver no `InService` estado. Para ter mais informações, consulte Verificar o status do anexo de seu balanceador de carga.

Verificações de integridade do VPC Lattice

Por padrão, o Amazon EC2 Auto Scaling ignora os resultados das verificações de integridade da VPC Lattice. Opcionalmente, você pode ativar essas verificações de integridade para seu grupo do Auto Scaling. Depois de fazer isso, quando o VPC Lattice relata uma instância registrada como`Unhealthy`, o Amazon EC2 Auto Scaling marca a instância `Unhealthy` na próxima verificação de integridade periódica e a substitui. O processo de registrar instâncias e, em seguida, verificar sua integridade funciona da mesma forma que as verificações de integridade do Elastic Load Balancing.

Para saber como habilitar verificações de integridade do VPC Lattice para seu grupo de Auto Scaling, consulte Anexe um grupo de destino VPC Lattice ao seu grupo do Auto Scaling.

> **Note**
>
> Quando você habilita verificações de integridade do VPC Lattice para um grupo, o Amazon EC2 Auto Scaling pode substituir instâncias que o VPC Lattice relata como não íntegras, mas somente depois que o grupo de destino estiver no estado `InService`. Para ter mais informações, consulte Verifique o status do anexo do grupo de destino do VPC Lattice.

Como o Amazon EC2 Auto Scaling minimiza o tempo de inatividade

Por padrão, as substituições de verificações 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 `Unhealthy` marcada com set-instance-healtho comando), ele a substituirá imediatamente. No entanto, se houver outras instâncias não íntegras, o Amazon EC2 Auto Scaling usará a abordagem a seguir para se recuperar de falhas. Esta abordagem minimiza qualquer tempo de inatividade que possa ocorrer devido a problemas temporários ou verificações de integridade mal configuradas.

- Se houver uma ação de escalabilidade em andamento e seu grupo do Auto Scaling estiver abaixo da capacidade desejada em 10% ou mais, o Amazon EC2 Auto Scaling aguarda a atividade de escalabilidade em andamento antes de substituir as instâncias não íntegras.
- Ao aumentar a escala horizontalmente, o Amazon EC2 Auto Scaling aguarda que as instâncias passem por uma verificação de integridade inicial. Ele também aguarda que o aquecimento padrão de instância seja concluído para garantir que as novas instâncias estejam prontas.
- Depois que as instâncias terminarem de aquecer e o grupo tiver aumentado para acima de 90% da capacidade desejada, o Amazon EC2 Auto Scaling substituirá as instâncias não íntegras da seguinte maneira:
  - O Amazon EC2 Auto Scaling substitui apenas até 10% da capacidade desejada do grupo por vez. Ele faz isso até que todas as instâncias não íntegras sejam substituídas.
  - Ao substituir instâncias, ele espera que as novas instâncias passem por uma verificação de integridade inicial. Ele também aguarda que o aquecimento padrão de instância seja concluído antes de continuar.

> **Note**
>
> Se o tamanho de um grupo do Auto Scaling for suficientemente pequeno para que o valor resultante de 10% seja menor que um, o Amazon EC2 Auto Scaling substituirá cada uma das instâncias não íntegras de cada vez. Isso pode resultar em tempo de inatividade para o grupo.
> 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 balanceador de carga estiver no estado`InService`, o Amazon EC2 Auto Scaling poderá marcar menos instâncias não íntegras de cada vez. Isso pode resultar em muito menos instâncias substituídas por vez do que os 10% aplicados em outros cenários. Isso fornece tempo para resolver o problema sem que o Amazon EC2 Auto Scaling encerre automaticamente todo o grupo.

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 ter mais informações, consulte Ganchos do ciclo de vida do Amazon EC2 Auto Scaling.
- O Amazon EC2 Auto Scaling não fornece um modo de remover as verificações de status e eventos programados do Amazon EC2 das verificações de integridade. Se você não quiser que as instâncias sejam substituídas, recomendamos suspender o processo `ReplaceUnhealthy` e `HealthCheck` para grupos do Auto Scaling individuais. Para ter mais informações, consulte Suspender e retomar os processos do Amazon EC2 Auto Scaling.
- Para redefinir manualmente o status de integridade de uma instância não `Healthy` íntegra, tente usar o set-instance-healthcomando. Se você receber um erro, provavelmente a instância já está encerrando. Geralmente, redefinir o status de integridade de uma instância `Healthy` com o set-instance-healthcomando só é útil nos casos em que o `ReplaceUnhealthy` processo ou o `Terminate` processo estão suspensos.
- O Amazon EC2 Auto Scaling não executa verificações de integridade em instâncias que estão no estado `Standby`. Para ter mais informações, consulte Remover temporiamente instâncias do grupo do Auto Scaling.
- Quando a instância é encerrada, qualquer endereço IP elástico é dissociado e não é automaticamente associado à nova instância. É necessário associar manualmente os endereços IP elásticos à nova instância ou fazer isso automaticamente com uma solução baseada em gancho do ciclo de vida. Para obter mais informações, consulte Endereços 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 Associar um volume do Amazon EBS a uma instância no Guia do usuário do Amazon EBS.

Verificações de integridade personalizadas

Opcionalmente, você pode executar tarefas personalizadas de detecção de integridade nas instâncias do seu grupo do Auto Scaling e definir o status de integridade de uma instância como `Unhealthy` caso a tarefa falhe. 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-healthcomando a seguir para definir o status de integridade da instância especificada como**`Unhealthy`**.

```
aws autoscaling set-instance-health --instance-id i-1234567890abcdef0 --health-
status Unhealthy
```

Por padrão, esse comando respeita o período de carência da verificação de integridade. Porém, é possível substituir esse comportamento e não respeitar o período de carência incluindo a opção`--no-should-respect-grace-period`.

Use o describe-auto-scaling-groupscomando a seguir para verificar se o status de integridade da instância é`Unhealthy`.

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

O exemplo a seguir é de uma resposta que mostra que o status de integridade da instância é `Unhealthy` e que a instância está sendo encerrada.

```
{
    "AutoScalingGroups": [
        {
            ....
            "Instances": [
                {
                    "ProtectedFromScaleIn": false,
                    "AvailabilityZone": "us-west-2a",
                    "LaunchTemplate": {
                        "LaunchTemplateName": "my-launch-template",
                        "Version": "1",
                        "LaunchTemplateId": "lt-1234567890abcdef0"
                    },
                    "InstanceId": "i-1234567890abcdef0",
                    "InstanceType": "t2.micro",
                    "HealthStatus": "Unhealthy",
                    "LifecycleState": "Terminating"
                },
```

```
                ...
            ]
        }
    ]
}
```

Recursos relacionados

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. 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 ter mais informações, consulte Visualizar o status e o motivo de falhas da verificação de integridade.

Veja o motivo das falhas na verificação de integridade

Usando o procedimento a seguir, você pode exibir informações sobre todas as instâncias substituídas devido a uma verificação de integridade.

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`.

Para obter informações sobre os tipos de erros que você pode encontrar e como tratá-los, consulte Solucionar problemas com as verificações de integridade do Amazon EC2 Auto Scaling.

> **Note**
>
> Por padrão, o Amazon EC2 Auto Scaling cria uma nova atividade de escalabilidade para encerrar a instância não íntegra e, em seguida, finaliza-a. Enquanto a instância estiver sendo encerrada, outra atividade de escalonamento iniciará uma nova instância.
> Você pode alterar esse comportamento para executar uma nova instância primeiro usando uma política de manutenção de instâncias. Com uma política de manutenção de instâncias, você pode estabelecer limites para um grupo do Auto Scaling quando há eventos que levam à substituição de instâncias, e seu grupo do Auto Scaling só pode substituir instâncias dentro dessa faixa de limite. Porém, como o Amazon EC2 Auto Scaling encerra imediatamente as instâncias que não estão mais em execução, o limite inferior da política de manutenção de sua instância pode ser excedido se uma instância for encerrada inesperadamente ou se você parar ou reinicializar manualmente uma instância. Para ter mais informações, consulte Políticas de manutenção de instância.

Definir um período de carência da verificação de integridade para um grupo do Auto Scaling

Quando uma verificação de integridade do Amazon EC2 Auto Scaling determina que uma `InService` instância não está íntegra, ela a substitui por uma nova instância. O período de carência da verificação de integridade especifica o tempo mínimo (em segundos) que uma nova instância será mantida em serviço antes de ser terminada, caso não esteja íntegra.

Um exemplo de caso de uso pode ser a exigência de que o Amazon EC2 Auto Scaling evite realizar ações se as verificações de integridade do Elastic Load Balancing falharem porque uma instância ainda está sendo inicializada. As verificações de integridade do Elastic Load Balancing são executadas em paralelo, começando quando a instância é registrada no balanceador de carga. O período de carência evita que o Amazon EC2 Auto Scaling marque suas instâncias recém-iniciadas `Unhealthy` e as encerre desnecessariamente se elas não passarem imediatamente nessas verificações de integridade após entrarem no `InService` estado.

Por padrão, o período de carência da verificação de integridade é de 300 segundos quando você cria um grupo do Auto Scaling. Seu valor padrão é 0 segundos quando você cria um grupo de Auto

Scaling usando o AWS CLI ou um SDK. O valor 0 desativa um período de carência da verificação de integridade.

Definir um valor muito alto reduz a eficácia das verificações de integridade do Amazon EC2 Auto Scaling. Se você usar ganchos do ciclo de vida para iniciar a instância, poderá definir o valor do período de carência da verificação de integridade como 0. Com ganchos 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 ter mais informações, consulte Ganchos do ciclo de vida do Amazon EC2 Auto Scaling.

O período de carência se aplica às seguintes instâncias:

- Instâncias recém-lançadas
- Instâncias que são colocadas de volta em serviço após estarem em espera
- Instâncias que você anexa manualmente ao grupo

> **Important**
>
> Durante o período de carência da verificação de integridade, se o Amazon EC2 Auto Scaling detectar que uma instância não está mais no estado Amazon EC2, `running` ele marcará imediatamente a instância `Unhealthy` e a substituirá. Por exemplo, se você interromper uma instância em um grupo do Auto Scaling, ela será marcada `Unhealthy` e substituída.

Definir um período de carência da verificação de integridade para um grupo

Definir um período de carência da verificação de integridade para grupos do Auto Scaling existentes.

Console

Para modificar o período de carência da verificação de integridade para um grupo novo (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-groupcomando. O exemplo a seguir configura o período de carência da verificação de integridade com um valor de *60* segundos para um novo grupo do Auto Scaling denominado *my-asg*.

```
aws autoscaling create-auto-scaling-group --auto-scaling-group-name my-asg --health-check-grace-period 60 ...
```

Console

Para modificar o período de carência da verificação de integridade para um grupo existente (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 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 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-groupcomando. O exemplo a seguir configura o período de carência da verificação de integridade com um valor de *120* segundos para um grupo do Auto Scaling existente denominado *my-asg*.

```
aws autoscaling update-auto-scaling-group --auto-scaling-group-name my-asg --health-check-grace-period 120
```

> **Note**
>
> É altamente recomendável também definir o tempo padrão de aquecimento da instância para seu grupo de Auto Scaling. Para ter mais informações, consulte Definir o aquecimento padrão da instância para um grupo do Auto Scaling.

AWS Health Dashboard notificações para Amazon EC2 Auto Scaling

Você AWS Health Dashboard fornece suporte para notificações provenientes do Amazon EC2 Auto Scaling. Essas notificações proporcionam conhecimento e orientação de remediação para problemas de performance de recursos ou disponibilidade que podem afetar suas aplicações. Somente eventos específicos para modelos de execução e grupos de segurança ausentes estão disponíveis no momento.

Isso AWS Health Dashboard faz parte do AWS Health serviço. Ele não requer configuração e pode ser visualizado por qualquer usuário autenticado em sua conta. Para obter mais informações, consulte Introdução ao seu AWS Health painel.

Se você receber uma mensagem semelhante às seguintes, ela deverá ser tratada como um alarme para executar uma ação.

Exemplo: a escala do grupo do Auto Scaling não está aumentando na horizontal devido a um grupo de segurança ausente

```
 Hello,

 At 2020-01-11 04:00 UTC, we detected an issue with your Auto Scaling group [ARN] in
Conta da AWS 123456789012.
```

```
 A security group associated with this Auto Scaling group cannot be found. Each time
a
 scale out operation is performed, it will be prevented until you make a change that
 fixes the issue.

 We recommend that you review and update your Auto Scaling group configuration to
change
 the launch template or launch configuration that depends on the unavailable security
 group.

 Sincerely,
 Amazon Web Services
```

Exemplo: a escala do grupo do Auto Scaling não está aumentando na horizontal devido a um modelo de execução ausente

```
 Hello,

 At 2021-05-11 04:00 UTC, we detected an issue with your Auto Scaling group [ARN] in
Conta da AWS 123456789012.

 The launch template associated with this Auto Scaling group cannot be found. Each
time
 a scale out operation is performed, it will be prevented until you make a change
that
 fixes the issue.

 We recommend that you review and update your Auto Scaling group configuration and
 specify an existing launch template to use.

 Sincerely,
 Amazon Web Services
```

Monitorar métricas do CloudWatch para grupos e instâncias do Auto Scaling

As métricas são um conceito fundamental do Amazon CloudWatch. Uma métrica representa um conjunto de pontos de dados ordenados ao longo do tempo 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 suas instâncias e grupos do Auto Scaling diretamente do Amazon CloudWatch em vez de usar o console do Amazon EC2 Auto Scaling.

Para obter mais informações, consulte o Guia do usuário do Amazon CloudWatch.

Índice

- Visualizar grafos de monitoramento no console do Amazon EC2 Auto Scaling
- Métricas do Amazon CloudWatch para o Amazon EC2 Auto Scaling
- Configurar monitoramento para instâncias do Auto Scaling

Visualizar grafos de monitoramento no console do Amazon EC2 Auto Scaling

Na seção do Amazon EC2 Auto Scaling do console do Amazon EC2, você pode monitorar o progresso minuto a minuto de grupos do Auto Scaling individuais, usando métricas do CloudWatch.

É possível monitorar os seguintes tipos de métricas:

- Métricas do Auto Scaling – as métricas de Auto Scaling são ativadas somente quando você as habilita. Para obter mais informações, consulte Ativar métricas do grupo do Auto Scaling (console). Quando as métricas do Auto Scaling estão habilitadas, os gráficos de monitoramento mostram dados publicados em granularidade de um minuto para métricas de Auto Scaling.
- Métricas do EC2 – as métricas da instância do Amazon EC2 estão sempre habilitadas. Se o monitoramento detalhado estiver habilitado, os gráficos de monitoramento mostrarão dados publicados em granularidade de um minuto para métricas de instância. Para obter mais informações, consulte Configurar monitoramento para instâncias do Auto Scaling.

Para visualizar gráficos de monitoramento usando o console do Amazon EC2 Auto Scaling

1. Abra o console do Amazon EC2 em https://console.aws.amazon.com/ec2/ e escolha 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.
- Selecione View logs (Visualizar logs) na ferramenta de menu para exibir fluxos de log associados (se houver) no console do CloudWatch.
- Para visualizar um gráfico no CloudWatch, selecione View in metrics (Visualizar em métricas) na ferramenta de menu. Isso o levará para a página do CloudWatch 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).

Na seção Auto Scaling, as métricas do gráfico incluem as seguintes métricas. Essas métricas fornecem medições que podem ser indicadores de um problema potencial, como número de instâncias de terminação ou número de instâncias pendentes. Você pode encontrar definições para essas métricas em Métricas do Amazon CloudWatch para o Amazon EC2 Auto Scaling.

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`
Instâncias em encerramento	`GroupTerminatingInstances`
Total de instâncias	`GroupTotalInstances`

Na seção EC2, você pode encontrar as seguintes métricas de gráfico com base nas métricas de performance cruciais para suas instâncias do Amazon EC2. Essas métricas do EC2 são um agregado de métricas de todas as instâncias do grupo. Você pode encontrar as definições para essas métricas em Listar as métricas disponíveis do CloudWatch para as instâncias no Guia do usuário do Amazon EC2 para instâncias do Linux.

Nome de exibição	Nome da métrica do CloudWatch
Utilização da CPU	`CPUUtilization`

Nome de exibição	Nome da métrica do CloudWatch
Leituras de disco	`DiskReadBytes`
Operações de leitura de disco	`DiskReadOps`
Gravações de disco	`DiskWriteBytes`
Operações de gravação de disco	`DiskWriteOps`
Entrada de rede	`NetworkIn`
Saída de rede	`NetworkOut`
Falha na verificação de status (qualquer)	`StatusCheckFailed`
Falha na verificação de status (instância)	`StatusCheckFailed_Instance`
Falha na verificação de status (sistema)	`StatusCheckFailed_System`

Além disso, algumas métricas estão disponíveis para casos de uso específicos nas métricas do gráfico do Auto Scaling.

As especificações dos gráficos a seguir são úteis se o seu grupo usa pesos que definem quantas unidades de instância estão disponíveis para a capacidade desejada do grupo. Você pode encontrar definições para essas métricas em Métricas do Amazon CloudWatch para o Amazon EC2 Auto Scaling.

Nome de exibição	Nome da métrica do CloudWatch
Unidades de capacidade em serviço	`GroupInServiceCapacity`
Unidades de capacidade pendentes	`GroupPendingCapacity`

Nome de exibição	Nome da métrica do CloudWatch
Unidades de capacidade em espera	`GroupStandbyCapacity`
Unidades de capacidade em encerramento	`GroupTerminatingCapacity`
Total de unidades de capacidade	`GroupTotalCapacity`

As métricas a seguir são úteis se o seu grupo usa o recurso de grupo de aquecimento. Você pode encontrar definições para essas métricas em Métricas do Amazon CloudWatch para o Amazon EC2 Auto Scaling.

Nome de exibição	Nome da métrica do CloudWatch
Tamanho mínimo do grupo de aquecimentoWarm Pool Minimum Size	`WarmPoolMinSize`
Capacidade desejada do grupo de aquecimento	`WarmPoolDesiredCapacity`
Unidades de capacidade pendentes no grupo de aquecimento	`WarmPoolPendingCapacity`
Unidades de capacidade em encerramento no grupo de aquecimento	`WarmPoolTerminatingCapacity`
Unidades de capacidade aquecidas no grupo de aquecimento	`WarmPoolWarmedCapacity`

Nome de exibição	Nome da métrica do CloudWatch
Total de unidades de capacidade executadas no grupo de aquecimento	`WarmPoolTotalCapacity`
Capacidade desejada do grupo e do grupo de aquecimento	`GroupAndWarmPoolDesiredCapacity`
Total de unidades de capacidade executadas no grupo e no grupo de aquecimento	`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 Guia do usuário do Amazon CloudWatch.

Métricas do Amazon CloudWatch 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 dependem das métricas de grupo habilitadas e de quais métricas de grupo você habilitou. As métricas do grupo estão disponíveis a uma granularidade de um minuto sem custo adicional, mas é necessário habilitá-las.

Ao habilitar métricas do grupo do Auto Scaling, o Amazon EC2 Auto Scaling envia os dados amostrados ao CloudWatch a cada minuto com base no melhor esforço. Em casos raros de interrupções de serviço no CloudWatch, os dados não são preenchidos retroativamente para completar lacunas no histórico de métricas de grupo.

Índice

- Métricas do grupo do Auto Scaling
- Dimensões para métricas do grupo do Auto Scaling
- Métricas e dimensões de escalabilidade preditiva
- Ativar métricas do grupo do Auto Scaling (console)
- Habilitar métricas do grupo do Auto Scaling (AWS CLI)

Métricas do grupo do Auto Scaling

Com essas métricas, você obtém visibilidade quase contínua sobre o histórico de seu grupo do Auto Scaling, como alterações no tamanho do grupo ao longo do tempo.

Métrica	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 você configura um grupo misto de instâncias para medir a capacidade desejada em unidades diferentes, por exemplo, atribuindo pesos com base na contagem de vCPU de cada tipo de instância, as métricas contam o número de unidades usadas pelo seu grupo do Auto Scaling. Se você não configurou um grupo de instâncias mistas para medir a capacidade desejada em unidades diferentes, as métricas seguintes são preenchidas, mas são iguais às métricas definidas na tabela anterior. Para obter mais informações, consulte Visão geral da configuração.

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. 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.

Métrica	Descrição
`WarmPoolMinSize`	O tamanho mínimo do grupo de alta atividade.

Métrica	Descrição
	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.

Métrica	Descrição
`WarmPoolWarmedCapacity`	A quantidade de capacidade disponível para se inserir no grupo do Auto Scaling durante a redução da escala. Essa métrica não inclui instâncias pendentes ou sendo encerradas. Critérios de relatório: relatado se a coleta de métricas estiver habilitada.
`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 Monitore métricas de escalabilidade preditiva com CloudWatch.

Métrica	Descrição	Dimensões
`PredictiveScalingLoadForecast`	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`
`PredictiveScalingCapacityForecast`	A quantidade 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. As estatísticas`Average`, `Minimum` e `Maximum` são úteis, mas a estatística `Sum` não. Critérios de relatório: reportado após a criação da previsão inicial.	`AutoScalingGroupName`, `PolicyName`
`PredictiveScalingMetricPairCorrelation`	A correlação entre a métrica de escalabilidade e a média por instância da métrica de carga. A escalabilidade preditiva pressupõe alta correlação. Então, se você observar um valor baixo para essa métrica, é melhor não usar um par de métricas.	`AutoScalingGroupName`, `PolicyName`, `PairIndex`

Métrica	Descrição	Dimensões
	As estatísticas `Average`, `Minimum` e `Maximum` são úteis, mas a estatística `Sum` não. Critérios de relatório: reportado após a criação da previsão inicial.	

> **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 habilitar métricas do grupo do Auto Scaling

Habilite uma ou mais métricas de grupo usando o comando enable-metrics-collection. Por exemplo, o comando a seguir habilita uma única métrica para o grupo do Auto Scaling especificado.

```
aws autoscaling enable-metrics-collection --auto-scaling-group-name my-asg \
  --metrics GroupDesiredCapacity --granularity "1Minute"
```

Se você omitir a opção `--metrics`, todas as métricas serão habilitadas.

```
aws autoscaling enable-metrics-collection --auto-scaling-group-name my-asg \
  --granularity "1Minute"
```

Para desabilitar métricas do grupo do Auto Scaling

Use o comando disable-metrics-collection para desabilitar 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 Preços do Amazon CloudWatch e Monitoramento de instâncias usando o CloudWatch no Manual do usuário do Amazon EC2 para instâncias do 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)
- Habilitar o monitoramento detalhado (AWS CLI)
- Alternar entre monitoramento básico e detalhado
- Coletar métricas adicionais usando o atendente do CloudWatch

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

Ao criar o modelo de execução usando o AWS Management Console, escolha Enable (Habilitar) para a opção Detailed CloudWatch monitoring (Monitoramento detalhado do CloudWatch) na seção Advanced details (Detalhes avançados). Caso contrário, o monitoramento básico será habilitado. Para obter mais informações, consulte Criar um modelo de execução usando configurações avançadas.

Para habilitar o monitoramento detalhado em uma configuração de execução

Ao criar a configuração de execução usando o AWS Management Console, na seção Additional configuration (Configuração adicional), selecione Enable EC2 instance detailed monitoring within CloudWatch (Habilitar monitoramento detalhado da instância do EC2 no CloudWatch). Caso contrário, o monitoramento básico será habilitado. Para obter mais informações, consulte Criar uma configuração de execução.

Habilitar o monitoramento detalhado (AWS CLI)

Por padrão, o monitoramento básico é habilitado quando você cria um modelo de execução usando a AWS CLI. O monitoramento detalhado é habilitado por padrão quando você cria uma configuração de execução usando a AWS CLI ou um SDK.

Para habilitar o monitoramento detalhado em um modelo de execução

Para modelos de execução, use o comando create-launch-template e envie um arquivo JSON que contenha as informações para criar o modelo de execução. Defina o parâmetro de monitoramento como `"Monitoring":{"Enabled":true}` para habilitar o monitoramento detalhado ou `"Monitoring":{"Enabled":false}` para habilitar o monitoramento básico.

Para habilitar o monitoramento detalhado em uma configuração de execução

Para as configurações de execução, use o comando create-launch-configuration com a opção `--instance-monitoring`. Defina essa opção como `true` para habilitar o monitoramento detalhado ou `false` para habilitar o monitoramento básico.

```
--instance-monitoring Enabled=true
```

Alternar entre monitoramento básico e detalhado

Para alterar o tipo de monitoramento habilitado em novas instâncias do EC2, atualize o modelo de execução ou o grupo do Auto Scaling para usar um novo modelo ou uma nova configuração de execução. As instâncias existentes continuam a usar o tipo de monitoramento habilitado anteriormente. Para atualizar todas as instâncias, termine-as para que elas sejam substituídas por seu grupo do Auto Scaling ou atualize as instâncias individualmente usando monitor-instances e unmonitor-instances.

> **Note**
>
> Com os recursos de tempo de vida máximo e atualização de instância e de atualização da instância, também é possível substituir todas as instâncias no grupo do Auto Scaling para iniciar novas instâncias que usem as novas configurações. Para obter mais informações, consulte Recicle as instâncias em seu grupo do Auto Scaling.

Ao alternar entre monitoramento básico e detalhado:

Se houver alarmes do CloudWatch associados às políticas de escalabilidade em etapas ou políticas de escalabilidade simples no seu grupo do Auto Scaling, use o comando put-metric-alarm 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.

Coletar métricas adicionais usando o atendente do CloudWatch

Para coletar métricas no nível do sistema operacional, como memória disponível e memória utilizada, você deve instalar o agente do CloudWatch. Podem ser cobrados taxas adicionais. É possível

usar um agente do CloudWatch para coletar métricas do sistema e arquivos de log das instâncias do Amazon EC2. Para obter mais informações, consulte Métricas coletadas pelo atendente do CloudWatch no Guia do usuário do Amazon CloudWatch.

Registre chamadas da API do Amazon EC2 Auto Scaling com AWS CloudTrail

O Amazon EC2 Auto Scaling é integrado AWS CloudTrail com um serviço que fornece um registro das ações realizadas por um usuário, função ou serviço usando o Amazon EC2 Auto Scaling. CloudTrail captura 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 de CloudTrail eventos para um bucket do Amazon S3, incluindo eventos para o Amazon EC2 Auto Scaling. Se você não configurar uma trilha, ainda poderá ver os eventos mais recentes no CloudTrail console no Histórico de eventos. Usando as informações coletadas por CloudTrail, você pode determinar a solicitação que foi feita ao Amazon EC2 Auto Scaling, o endereço IP a partir do qual a solicitação foi feita, quem fez a solicitação, quando ela foi feita e detalhes adicionais.

Para saber mais sobre isso CloudTrail, consulte o Guia AWS CloudTrail do usuário.

Informações do Amazon EC2 Auto Scaling em CloudTrail

CloudTrail é ativado na 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 CloudTrail um evento junto com outros eventos da Amazon Web Services no histórico de eventos. Você pode visualizar, pesquisar e baixar os eventos recentes em sua conta da Amazon Web Services. Para obter mais informações, consulte Visualização de eventos com histórico de CloudTrail eventos.

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. Uma trilha permite CloudTrail entregar arquivos de log para 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 de eventos coletados nos CloudTrail registros. Para mais informações, consulte:

- Visão geral da criação de uma trilha
- CloudTrail serviços e integrações suportados
- Configurando notificações do Amazon SNS para CloudTrail
- Recebendo arquivos de CloudTrail log de várias regiões e Recebendo arquivos de CloudTrail log de várias contas

Todas as ações do Amazon EC2 Auto Scaling são registradas e CloudTrail documentadas na Amazon EC2 Auto Scaling API Reference. Por exemplo, chamadas para as UpdateAutoScalingGroupações CreateLaunchConfigurationDescribeAutoScalingGroup, e geram entradas nos arquivos de CloudTrail log.

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 root ou AWS Identity and Access Management (IAM).
- Se a solicitação foi feita com credenciais de segurança temporárias de um perfil ou de um usuário federado.
- Se a solicitação foi feita por outro serviço da .

Para obter mais informações, consulte o CloudTrail`userIdentity`elemento.

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 para um bucket do Amazon S3 que você especificar. CloudTrail os arquivos de log 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 da solicitação e assim por diante. CloudTrail os arquivos de log não são um rastreamento de pilha ordenado das chamadas públicas de API, portanto, eles não aparecem em nenhuma ordem específica.

O exemplo a seguir mostra uma entrada de CloudTrail registro que demonstra a CreateLaunchConfiguration ação.

```
{
    "eventVersion": "1.05",
```

```
    "userIdentity": {
        "type": "Root",
        "principalId": "123456789012",
        "arn": "arn:aws:iam::123456789012:root",
        "accountId": "123456789012",
        "accessKeyId": "AKIAIOSFODNN7EXAMPLE",
        "sessionContext": {
            "attributes": {
                "mfaAuthenticated": "false",
                "creationDate": "2018-08-21T17:05:42Z"
            }
        }
    },
    "eventTime": "2018-08-21T17:07:49Z",
    "eventSource": "autoscaling.amazonaws.com",
    "eventName": "CreateLaunchConfiguration",
    "awsRegion": "us-west-2",
    "sourceIPAddress": "192.0.2.0",
    "userAgent": "Coral/Jakarta",
    "requestParameters": {
        "ebsOptimized": false,
        "instanceMonitoring": {
            "enabled": false
        },
        "instanceType": "t2.micro",
        "keyName": "EC2-key-pair-oregon",
        "blockDeviceMappings": [
            {
                "deviceName": "/dev/xvda",
                "ebs": {
                    "deleteOnTermination": true,
                    "volumeSize": 8,
                    "snapshotId": "snap-01676e0a2c3c7de9e",
                    "volumeType": "gp2"
                }
            }
        ],
        "launchConfigurationName": "launch_configuration_1",
        "imageId": "ami-6cd6f714d79675a5",
        "securityGroups": [
            "sg-00c429965fd921483"
        ]
    },
    "responseElements": null,
```

```
    "requestID": "0737e2ea-fb2d-11e3-bfd8-99133058e7bb",
    "eventID": "3fcfb182-98f8-4744-bd45-b38835ab61cb",
    "eventType": "AwsApiCall",
    "recipientAccountId": "123456789012"
}
```

Recursos relacionados

Com o CloudWatch Logs, você pode monitorar e receber alertas para eventos específicos capturados pelo CloudTrail. Os eventos enviados para o CloudWatch Logs são aqueles configurados para serem registrados por sua trilha, portanto, certifique-se de ter configurado sua trilha ou trilhas para registrar os tipos de eventos que você está interessado em monitorar. CloudWatch Os registros podem monitorar as informações nos arquivos de log e notificá-lo quando determinados limites forem atingidos. É possível também arquivar seus dados de log em armazenamento resiliente. Para obter mais informações, consulte o Guia do usuário do Amazon CloudWatch Logs e o tópico Monitoramento de arquivos de CloudTrail log com o Amazon CloudWatch Logs no Guia AWS CloudTrail do usuário.

Opções de notificação do Amazon SNS para o Amazon EC2 Auto Scaling

Você pode configurar seu grupo de Auto Scaling para notificá-lo sobre eventos importantes que afetam seu aplicativo. Com as notificações, você também pode eliminar a pesquisa e não encontrará o `RequestLimitExceeded` erro que às vezes resulta da pesquisa.

Há duas maneiras de receber notificações sobre o Amazon EC2 Auto Scaling:

- Amazon Simple Notification Service — O Amazon SNS pode notificá-lo quando seu grupo de Auto Scaling inicia ou encerra instâncias. Só é possível ativar e desativar notificações do Amazon SNS. Para ter mais informações, consulte Amazon SNS e Amazon EC2 Auto Scaling.
- Amazon EventBridge — EventBridge fornece notificações mais avançadas, orientadas por eventos, de acordo com critérios específicos e enviadas para uma variedade de destinos, incluindo o Amazon SNS. EventBridge também pode monitorar uma variedade maior de eventos do Auto Scaling para um monitoramento mais preciso. Para ter mais informações, consulte Use EventBridge para lidar com eventos do Auto Scaling.

Você também pode realizar uma ação personalizada quando uma instância entra em um estado pendente durante a inicialização ou o encerramento usando ganchos e serviços de ciclo de vida, como Amazon EventBridge SNS e Amazon SQS. Os ganchos de ciclo de vida também podem fornecer tempo extra para que uma nova instância conclua um script especificado nos dados do usuário antes que o Amazon EC2 Auto Scaling adicione a instância ao grupo. Para ter mais informações, consulte Ganchos do ciclo de vida do Amazon EC2 Auto Scaling.

Amazon SNS e Amazon EC2 Auto Scaling

Esta seção mostra como usar o Amazon SNS para monitorar quando seu grupo de Auto Scaling inicia ou encerra instâncias.

Por exemplo, se você configurar o grupo do Auto Scaling para usar o tipo de notificação `autoscaling: EC2_INSTANCE_TERMINATE` e seu grupo do Auto Scaling terminar uma instância, ele enviará uma notificação por e-mail. Esse e-mail contém os detalhes da instância encerrada, como o ID da instância e o motivo pelo qual a instância foi encerrada.

Observe que, à medida que o Amazon EC2 Auto Scaling adiciona ou remove instâncias do grupo, as notificações sobre essas alterações são enviadas para você, com uma notificação enviada por instância. No entanto, a entrega dessas notificações é feita com base no melhor esforço, e suas instâncias ainda podem falhar após a notificação inicial, por exemplo, se uma verificação de saúde posterior falhar. Portanto, mesmo que o Amazon EC2 Auto Scaling notifique você no início, uma instância ainda pode falhar posteriormente. Observe que você pode configurar quanto tempo depois de iniciar uma instância o Amazon EC2 Auto Scaling espera antes de realizar a primeira verificação de saúde. Para ter mais informações, consulte Definir um período de carência da verificação de integridade para um grupo do Auto Scaling.

Para obter mais informações sobre o Amazon SNS em geral, consulte o Guia do desenvolvedor do Amazon Simple Notification Service.

Conteúdo

- Notificações do SNS
- Configurar notificações do Amazon SNS para o Amazon EC2 Auto Scaling
  - Crie um tópico do Amazon SNS
  - Assinar o tópico do Amazon SNS
  - Confirmar sua assinatura do Amazon SNS
  - Configurar o grupo do Auto Scaling para enviar notificações

- Testar a notificação
- Excluir a configuração da notificação
- Política de chaves para um tópico do Amazon SNS criptografado

Notificações do SNS

O Amazon EC2 Auto Scaling oferece suporte ao envio de notificações do Amazon SNS quando os seguintes eventos ocorrem.

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.

Crie um tópico do Amazon SNS

Um tópico do SNS é um ponto de acesso lógico, um canal de comunicação que seu grupo do Auto Scaling usa para enviar notificações. Você cria um tópico especificando um nome para o tópico.

Quando você cria o nome de um tópico, ele deve atender aos seguintes requisitos:

- Ter entre 1 e 256 caracteres
- Conter letras maiúsculas e minúsculas ASCIIs, números, sublinhados ou hífens

Para obter mais informações, consulte Criação de um tópico do Amazon SNS no Guia do desenvolvedor do Amazon Simple Notification Service.

Assinar o tópico do Amazon SNS

Para receber as notificações que seu grupo do Auto Scaling envia ao tópico, você deve assinar um endpoint para o tópico. Neste procedimento, em Endpoint, especifique o endereço de e-mail no qual você deseja receber as notificações do Amazon EC2 Auto Scaling.

Para obter instruções, consulte Assinatura de um tópico do Amazon SNS no Guia do desenvolvedor do Amazon Simple Notification Service.

Confirmar sua assinatura do Amazon SNS

O Amazon SNS envia um e-mail de confirmação para o endereço de e-mail especificado na etapa anterior.

Certifique-se de abrir o e-mail em AWS Notifications (Notificações) e escolher o link para confirmar a assinatura antes de prosseguir para a próxima etapa.

Você receberá uma mensagem de confirmação de. AWS O Amazon SNS agora está configurado para receber notificações e enviar a notificação como um e-mail para o endereço de e-mail que você especificou.

Configurar o grupo do Auto Scaling para enviar notificações

Você pode configurar seu grupo do Auto Scaling para enviar notificações para o Amazon SNS quando um evento de escalabilidade ocorre, como lançamento ou término de instâncias. O Amazon SNS envia uma notificação com informações sobre as instâncias para o endereço de e-mail que você especificou.

Para configurar notificações do Amazon SNS para o seu grupo do Auto Scaling (console)

1. Abra o console do Amazon EC2 em https://console.aws.amazon.com/ec2/ e escolha 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 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 Atualizar.
6. Depois de alguns minutos, você receberá uma notificação para o evento. Se não for necessário ter a instância adicional executada para este teste, será possível reduzir Desired capacity (Capacidade desejada) em 1. Depois de alguns minutos, você receberá uma notificação para o evento.

Excluir a configuração da notificação

Você poderá excluir sua configuração de notificação do Amazon EC2 Auto Scaling se ela não estiver mais sendo usada.

Para excluir a configuração de notificação do Amazon EC2 Auto Scaling (console)

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

O tópico do Amazon SNS que você especificar pode ser criptografado com uma chave gerenciada pelo cliente criada com o AWS Key Management Service. Para dar permissão ao Amazon EC2 Auto Scaling para publicar em tópicos criptografados, você deve primeiro criar sua chave KMS e depois adicionar a seguinte declaração à política da chave KMS. Substitua o ARN de exemplo pelo ARN da função apropriada vinculada ao serviço que tem permissão de acesso à chave. Para obter mais informações, consulte Configurar AWS KMS permissões no Guia do desenvolvedor do Amazon Simple Notification Service.

Neste exemplo, a declaração de política dá à função vinculada ao serviço denominada AWSServiceRoleForAutoScalingpermissões para usar a chave gerenciada pelo cliente. Para saber mais sobre a função vinculada ao serviço do Amazon EC2 Auto Scaling, consulte Funções vinculadas ao serviço do Amazon EC2 Auto Scaling.

```
{
  "Sid": "Allow service-linked role use of the customer managed key",
  "Effect": "Allow",
  "Principal": {
```

```
    "AWS": "arn:aws:iam::123456789012:role/aws-service-role/autoscaling.amazonaws.com/
AWSServiceRoleForAutoScaling"
  },
  "Action": [
    "kms:GenerateDataKey*",
    "kms:Decrypt"
  ],
  "Resource": "*"
}
```

As chaves de condição de `aws:SourceArn` e `aws:SourceAccount` não são suportadas em políticas de chaves que permitem que o Amazon EC2 Auto Scaling publique em tópicos criptografados.

AWS serviços integrados ao Amazon EC2 Auto Scaling

O Amazon EC2 Auto Scaling pode ser integrado a outros serviços. 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
- Use reservas de capacidade sob demanda para reservar capacidade em zonas de disponibilidade específicas
- Crie grupos de Auto Scaling a partir da linha de comando usando AWS CloudShell
- Criar um grupo do Auto Scaling com AWS CloudFormation
- Use AWS Compute Optimizer para obter recomendações sobre o tipo de instância para um grupo de Auto Scaling
- Usar o Elastic Load Balancing para distribuir tráfego entre as instâncias no grupo do Auto Scaling
- Rotear o tráfego para o grupo do Auto Scaling com um grupo de destino do VPC Lattice
- Use EventBridge para lidar com eventos do Auto Scaling
- Fornecer conectividade de rede para suas instâncias do Auto Scaling usando a Amazon VPC

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.

Conteúdo

- Visão geral
- Comportamento de rebalanceamento de capacidade
- Considerações
- Habilitar o rebalanceamento de capacidade (console)
- Habilitar o rebalanceamento de capacidade (AWS CLI)
- Recursos relacionados
- Limitações

Visão geral

Para usar o rebalanceamento de capacidade com seu grupo do Auto Scaling, as etapas básicas são:

1. Configure seu grupo do Auto Scaling para usar vários tipos de instância e zonas de disponibilidade. Dessa forma, o Amazon EC2 Auto Scaling pode analisar a capacidade disponível para instâncias spot em cada zona de disponibilidade. Para ter mais informações, consulte Grupos do Auto Scaling com vários tipos de instâncias e opções de compra.
2. Adicione hooks do ciclo de vida conforme necessário para realizar um desligamento suave do seu aplicativo dentro das instâncias que recebem a notificação de rebalanceamento. Para ter mais informações, consulte Ganchos do ciclo de vida do Amazon EC2 Auto Scaling.

   A seguir estão alguns motivos pelos quais você pode usar um hook do ciclo de vida:

   - Para o encerramento suave de operadores do Amazon SQS
   - Para concluir o cancelamento do registro do sistema de nomes de domínio (DNS)

- Para extrair logs do sistema ou do aplicativo e carregá-los no Amazon Simple Storage Service (Amazon S3)

3. Desenvolva uma ação personalizada para o hook do ciclo de vida. Para invocar sua ação personalizada o mais rápido possível, você precisa saber quando uma instância está pronta para ser encerrada. Descubra isso detectando o estado do ciclo de vida da instância.
   - Para invocar uma ação fora da instância, escreva uma EventBridge regra e automatize a ação a ser tomada quando um padrão de evento corresponder à regra.
   - Para invocar uma ação dentro da instância, configure a instância para executar um script de desligamento e recuperar o estado do ciclo de vida por meio dos metadados da instância.

   É fundamental projetar a ação personalizada para ser concluída em menos de dois minutos. Isso garante que haja tempo suficiente para concluir as tarefas antes do encerramento da instância.

Depois de concluir essas etapas, será possível começar a usar o rebalanceamento de capacidade.

Comportamento de rebalanceamento de capacidade

Com o rebalanceamento de capacidade, o Amazon EC2 Auto Scaling se comporta da seguinte maneira quando uma instância recebe uma recomendação de rebalanceamento:

- Quando a nova instância spot é iniciada, o Amazon EC2 Auto Scaling aguarda até que a nova instância passe na verificação de integridade antes de encerrar a instância anterior. Ao substituir mais de uma instância, o término de cada instância anterior é iniciado depois que a nova instância foi iniciada e aprovada na verificação de integridade.
- Como o Amazon EC2 Auto Scaling tenta iniciar novas instâncias antes de terminar as anteriores, estar na capacidade máxima especificada ou próximo a ela pode impedir ou parar completamente as atividades de rebalanceamento. Para evitar esse problema, o Amazon EC2 Auto Scaling pode exceder temporariamente o tamanho máximo do grupo em até 10% da capacidade desejada.
- Se você não adicionou um hook do ciclo de vida ao grupo do Auto Scaling, o Amazon EC2 Auto Scaling começará a encerrar as instâncias anteriores assim que as novas instâncias passarem na verificação de integridade.
- Se você adicionou um hook do ciclo de vida, isso aumenta o tempo necessário até começarmos a encerrar as instâncias anteriores pelo valor de tempo limite especificado para o hook do ciclo de vida.
- Se você estiver usando políticas de escalabilidade ou escalabilidade agendada, as atividades de escalabilidade serão executadas em paralelo. Se uma ação de escalabilidade estiver em

andamento e seu grupo do Auto Scaling estiver abaixo da nova capacidade desejada, o Amazon EC2 Auto Scaling aumentará a escala na horizontal antes de terminar as instâncias anteriores.

Se não houver capacidade para seus tipos de instância em uma zona de disponibilidade, o Amazon EC2 Auto Scaling continuará tentando executar instâncias spot em outras zonas de disponibilidade habilitadas até que ela seja bem-sucedida.

Na pior das hipóteses, se as novas instâncias falharem na inicialização ou se suas verificações de integridade falharem, o Amazon EC2 Auto Scaling continuará tentando reiniciá-las. Enquanto ele tenta iniciar novas instâncias, as anteriores serão eventualmente interrompidas e encerradas à força com um aviso de interrupção de dois minutos.

Considerações

Considere o seguinte ao usar o reequilíbrio de capacidade:

Projete seu aplicativo para ser tolerante a interrupções pontuais

Seu aplicativo deve ser capaz de lidar com alterações dinâmicas no número de instâncias e com a possibilidade de uma instância spot ser interrompida antecipadamente. Por exemplo, se o seu grupo do Auto Scaling estiver atrás de um balanceador de carga do Elastic Load Balancing, o Amazon EC2 Auto Scaling aguardará o cancelamento do registro da instância no balanceador de carga antes de chamar o hook do ciclo de vida. Se o tempo para cancelar o registro da instância e concluir a ação do ciclo de vida demorar muito, a instância poderá ser interrompida enquanto o Amazon EC2 Auto Scaling aguarda a conclusão da ação do ciclo de vida antes de encerrar a instância.

Nem sempre é possível para o Amazon EC2 enviar o sinal de recomendação de rebalanceamento antes do aviso de interrupção da instância spot de dois minutos. Às vezes, o sinal de recomendação de reequilíbrio chega ao mesmo tempo que o aviso de interrupção de dois minutos. Quando isso acontece, o Amazon EC2 Auto Scaling chama o hook do ciclo de vida e tenta executar uma nova instância spot imediatamente.

Evitar um risco elevado de interrupção das instâncias spot substitutas

Suas instâncias spot de substituição poderão correr um risco elevado de interrupção se você usar a estratégia de alocação`lowest-price`. Isso ocorre porque lançamos instâncias no pool de menor preço que tem capacidade disponível naquele momento, mesmo que suas instâncias spot substitutas possam ser interrompidas logo após serem lançadas. Para evitar um alto risco de interrupção, recomendamos fortemente que não use a estratégia de alocação`lowest-price`.

Em vez disso, recomendamos a estratégia de alocação`price-capacity-optimized`. Essa estratégia executa instâncias spot de substituição em grupos spot com menor probabilidade de interrupção e com o preço mais baixo possível. Portanto, é menos provável que sejam interrompidos em um futuro próximo.

O Amazon EC2 Auto Scaling só lançará uma nova instância se a disponibilidade for igual ou melhor

Um dos objetivos do rebalanceamento de capacidade é melhorar a disponibilidade de uma instância spot. Se uma instância spot existente receber uma recomendação de rebalanceamento, o Amazon EC2 Auto Scaling só iniciará uma nova instância se a nova instância fornecer a mesma ou melhor disponibilidade do que a instância existente. Se o risco de interrupção de uma nova instância for pior do que a instância existente, o Amazon EC2 Auto Scaling não iniciará uma nova instância. No entanto, o Amazon EC2 Auto Scaling continuará avaliando os pools de capacidade Spot com base nas informações fornecidas pelo serviço Amazon EC2 Spot e iniciará uma nova instância se a disponibilidade melhorar.

Há uma chance de que sua instância existente seja interrompida sem que o Amazon EC2 Auto Scaling inicie proativamente uma nova instância. Quando isso acontece, o Amazon EC2 Auto Scaling tenta executar uma nova instância assim que recebe o aviso de interrupção da instância spot. Isso acontece independentemente de a nova instância ter um alto risco de interrupção.

O Rebalanceamento da capacidade não aumenta a taxa de interrupção de instâncias Spot

Quando o Rebalanceamento da capacidade é habilitado, ele não aumenta a Taxa de interrupção de instâncias Spot (o número de instâncias Spot que são recuperadas quando o Amazon EC2 precisa novamente de capacidade). Porém, se o Rebalanceamento da capacidade detectar que uma instância está sob risco de interrupção, o Amazon EC2 Auto Scaling tentará iniciar imediatamente uma nova instância. Portanto, mais instâncias poderão ser substituídas do que se você esperasse que o Amazon EC2 Auto Scaling iniciasse uma nova instância após a interrupção da instância em risco.

Embora você possa substituir mais instâncias com o Rebalanceamento de capacidade habilitado, você se beneficia por ser proativo em vez de reativo. Isso lhe dá mais tempo para agir antes que suas instâncias sejam interrompidas. Com um Aviso de interrupção de instâncias Spot, normalmente você só tem até dois minutos para encerrar sua instância sem problemas. Com o Rebalanceamento da capacidade iniciando uma nova instância com antecedência, os processos existentes têm maiores chances de serem concluídos na instância em risco. Além disso, você pode iniciar os procedimentos de desligamento da instância, impedir que novos trabalhos sejam agendados na instância em risco e preparar a instância recém-executada para assumir o controle

do aplicativo. Com a substituição proativa no reequilíbrio de capacidade, você se beneficia de uma continuidade tranquila.

O exemplo teórico a seguir demonstra os riscos e benefícios do uso do Rebalanceamento de Capacidade:

- 14h – Uma recomendação de rebalanceamento é recebida para a instância A. O Amazon EC2 Auto Scaling tenta imediatamente iniciar a instância de substituição B, dando a você tempo para iniciar seus procedimentos de encerramento.
- 14h30 – Uma recomendação de rebalanceamento é recebida para a instância B, que é substituída pela instância C. Isso lhe dá tempo para iniciar seus procedimentos de desligamento.
- 14h32 – Se o reequilíbrio de capacidade não estiver ativado e se um aviso de interrupção da instância spot tiver sido recebido às 14h32, para a instância A, você terá apenas dois minutos para agir. Porém, a instância A teria continuado em execução até esse momento.

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. Para Etapa 1: escolher modelo de execução ou configuração, insira um nome para o grupo do Auto Scaling, escolha um modelo de execução e escolha Próximo para prosseguir para a próxima etapa.
4. Para a Etapa 2: escolha as opções de execução da instância, em Requisitos do tipo de instância, escolha as configurações para criar um grupo misto de instâncias. Isso inclui os tipos de instância que ele pode lançar, as opções de compra de instâncias e as estratégias de alocação para instâncias sob demanda e spot. Por padrão, essas configurações não estão definidas. Para configurá-las, é necessário selecionar Override launch template (Substituir modelo de execução). Para obter mais informações sobre como criar um grupo de instâncias mistas, consulte Grupos do Auto Scaling com vários tipos de instâncias e opções de compra.

5. Em Rede, escolha as opções desejadas. Verifique se as sub-redes que você deseja utilizar se encontram em diferentes zonas de disponibilidade.
6. Na seção Estratégias de alocação, escolha uma estratégia de alocação spot. Ative ou desative o Rebalanceamento de capacidade marcando ou desmarcando a caixa de seleção em Rebalanceamento de capacidade. Você só vê essa opção quando solicita que uma porcentagem do grupo do Auto Scaling seja iniciada como instâncias spot na seção Opções de compra de instância.
7. Crie o grupo do Auto Scaling.
8. (Opcional) Adicione hooks do ciclo de vida conforme necessário. Para ter mais informações, consulte Adicionar ganchos do ciclo de vida.

Para ativar ou desativar o reequilíbrio de capacidade para um grupo existente do Auto Scaling

1. Abra o console do Amazon EC2 em https://console.aws.amazon.com/ec2/ e escolha 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. Na seção Estratégias de alocação, ative ou desative o rebalanceamento de capacidade marcando ou desmarcando a caixa de seleção em Rebalanceamento de capacidade.
5. Escolha Atualizar.

Habilitar o rebalanceamento de capacidade (AWS CLI)

Os exemplos a seguir mostram como usar o AWS CLI para ativar e desativar o rebalanceamento de capacidade.

Use o update-auto-scaling-groupcomando create-auto-scaling-groupou com o seguinte parâmetro:

- `--capacity-rebalance/--no-capacity-rebalance`— Valor booleano que indica se o rebalanceamento de capacidade está ativado.

Antes de chamar o create-auto-scaling-groupcomando, você precisa do nome de um modelo de execução configurado para uso com um grupo de Auto Scaling. Para ter mais informações, consulte Criar um modelo de execução para um grupo do Auto Scaling.

> **Note**
>
> Os procedimentos a seguir mostram como usar um arquivo de configuração formatado em JSON ou YAML. Se você usar a AWS CLI versão 1, deverá especificar um arquivo de configuração formatado em JSON. Se você usar a AWS CLI versão 2, poderá especificar um arquivo de configuração formatado em YAML ou JSON.

JSON

Para criar e configurar um novo grupo do Auto Scaling

- Use o create-auto-scaling-groupcomando a seguir para criar um novo grupo de Auto Scaling e ativar o rebalanceamento de capacidade. Este comando faz referência a um arquivo JSON como único parâmetro para seu grupo de Auto Scaling.

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

Se você ainda não possui um arquivo de configuração da CLI que especifique uma política de instâncias mistas, crie um.

Adicione a entrada a seguir ao objeto JSON de nível superior no arquivo de configuração.

```
{
    "CapacityRebalance": true
}
```

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

```
{
    "AutoScalingGroupName": "my-asg",
    "DesiredCapacity": 12,
    "MinSize": 12,
    "MaxSize": 15,
    "CapacityRebalance": true,
    "MixedInstancesPolicy": {
        "InstancesDistribution": {
            "OnDemandBaseCapacity": 0,
            "OnDemandPercentageAboveBaseCapacity": 25,
            "SpotAllocationStrategy": "price-capacity-optimized"
```

```
        },
        "LaunchTemplate": {
            "LaunchTemplateSpecification": {
                "LaunchTemplateName": "my-launch-template",
                "Version": "$Default"
            },
            "Overrides": [
                {
                    "InstanceType": "c5.large"
                },
                {
                    "InstanceType": "c5a.large"
                },
                {
                    "InstanceType": "m5.large"
                },
                {
                    "InstanceType": "m5a.large"
                },
                {
                    "InstanceType": "c4.large"
                },
                {
                    "InstanceType": "m4.large"
                },
                {
                    "InstanceType": "c3.large"
                },
                {
                    "InstanceType": "m3.large"
                }
            ]
        }
    },
    "TargetGroupARNs": "arn:aws:elasticloadbalancing:us-
west-2:123456789012:targetgroup/my-alb-target-group/943f017f100becff",
    "VPCZoneIdentifier": "subnet-5ea0c127,subnet-6194ea3b,subnet-c934b782"
}
```

YAML

Para criar e configurar um novo grupo do Auto Scaling

- Use o create-auto-scaling-groupcomando a seguir para criar um novo grupo de Auto Scaling e ativar o rebalanceamento de capacidade. Este comando faz referência a um arquivo YAML como único parâmetro para seu grupo do Auto Scaling.

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

Adicione a linha a seguir ao arquivo de configuração formatado em YAML.

```
CapacityRebalance: true
```

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

```
---
AutoScalingGroupName: my-asg
DesiredCapacity: 12
MinSize: 12
MaxSize: 15
CapacityRebalance: true
MixedInstancesPolicy:
  InstancesDistribution:
    OnDemandBaseCapacity: 0
    OnDemandPercentageAboveBaseCapacity: 25
    SpotAllocationStrategy: price-capacity-optimized
  LaunchTemplate:
    LaunchTemplateSpecification:
      LaunchTemplateName: my-launch-template
      Version: $Default
    Overrides:
    - InstanceType: c5.large
    - InstanceType: c5a.large
    - InstanceType: m5.large
    - InstanceType: m5a.large
    - InstanceType: c4.large
    - InstanceType: m4.large
    - InstanceType: c3.large
    - InstanceType: m3.large
TargetGroupARNs:
```

```
- arn:aws:elasticloadbalancing:us-west-2:123456789012:targetgroup/my-alb-target-
group/943f017f100becff
VPCZoneIdentifier: subnet-5ea0c127,subnet-6194ea3b,subnet-c934b782
```

Para habilitar o rebalanceamento de capacidade para um grupo do Auto Scaling existente

- Use o update-auto-scaling-groupcomando a seguir para ativar o rebalanceamento de capacidade.

```
aws autoscaling update-auto-scaling-group --auto-scaling-group-name my-asg \
  --capacity-rebalance
```

Para verificar se o rebalanceamento de capacidade está habilitado para um grupo do Auto Scaling

- Use o describe-auto-scaling-groupscomando a seguir para verificar se o rebalanceamento de capacidade está ativado e para ver os detalhes.

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

A seguir, uma exemplo de resposta.

```
{
    "AutoScalingGroups": [
        {
            "AutoScalingGroupName": "my-asg",
            "AutoScalingGroupARN": "arn",
            ...
            "CapacityRebalance": true
        }
    ]
}
```

Para desabilitar o rebalanceamento de capacidade

Use o update-auto-scaling-groupcomando com a `--no-capacity-rebalance` opção de desativar o rebalanceamento de capacidade.

```
aws autoscaling update-auto-scaling-group --auto-scaling-group-name my-asg \
```

```
  --no-capacity-rebalance
```

Recursos relacionados

Para obter mais informações sobre o rebalanceamento de capacidade, consulte Gerenciar proativamente o ciclo de vida da instância spot usando o novo recurso de rebalanceamento de capacidade para o EC2 Auto Scaling no blog Compute. AWS

Para obter mais informações sobre as recomendações de rebalanceamento de instâncias do EC2, consulte Recomendações de rebalanceamento de instâncias do EC2 no Guia do usuário do Amazon EC2 para instâncias do Linux.

Para saber mais sobre hooks do ciclo de vida, consulte os recursos a seguir.

- Tutorial: Configurar um gancho do ciclo de vida que invoca uma função do Lambda(usando EventBridge)
- Tutorial: configurar dados do usuário para recuperar o estado de destino do ciclo de vida por meio de metadados de instância

Limitações

- O Amazon EC2 Auto Scaling poderá substituir a instância que recebe a notificação de rebalanceamento somente se a instância não estiver protegida contra redução vertical. No entanto, a proteção escalável não impede o encerramento de uma interrupção Spot. Para obter mais informações, consulte Usar proteção de redução na escala na horizontal de instâncias.
- O suporte para reequilíbrio de capacidade está disponível em todas as áreas comerciais Regiões da AWS onde o Amazon EC2 Auto Scaling está disponível, exceto na região do Oriente Médio (Emirados Árabes Unidos).

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 as Reservas de Capacidade, crie a reserva de capacidade na 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 estiver,`open` todas as novas instâncias e as instâncias existentes que tiverem atributos correspondentes serão 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 do Auto Scaling que executa instâncias sob demanda em Reservas de Capacidade `targeted`. Isto dá a você maior 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 número de instâncias.
2. Reservas de capacidade de grupo usando AWS Resource Groups.
3. Criar um grupo do Auto Scaling com um modelo de execução direcionado ao grupo de recursos, usando as mesmas zonas de disponibilidade das reservas de capacidade.

Conteúdo

- Etapa 1: criar as Reservas de Capacidade
- Etapa 2: criar um grupo de Reserva de Capacidade
- Etapa 3: criar um modelo de execução
- Etapa 4: criar um grupo do Auto Scaling
- Recursos relacionados

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 do Auto Scaling será implantado.

> **Note**
>
> Você só pode criar reservas `targeted` ao criar as Reservas de Capacidade pela primeira vez.

Console

Para criar sua reserva 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. Na página Criar uma Reserva de capacidade, atente para as seguintes configurações na seção Detalhes da instância. 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, selecione o tipo de instância a ser executada na capacidade reservada.
   b. Para Platforma, selecione o sistema operacional das suas instâncias.
   c. Para Zona de disponibilidade, escolha a primeira zona de disponibilidade na qual você deseja reservar capacidade.
   d. Para Capacidade total, escolha o número de instâncias de que você precisa. Calcule o número total de instâncias necessárias para seu grupo do Auto Scaling dividido pelo número de zonas de disponibilidade que você planeja usar.
4. Em detalhes de Reserva de Capacidade, para encerramento de Reserva de Capacidade, selecione uma das seguintes opções:
   - Em um horário específico — cancele a reserva de capacidade automaticamente na data e hora especificadas.
   - Manualmente — reserve a capacidade até cancelá-la explicitamente.
5. Em Elegibilidade da instância, escolha Direcionada: somente instâncias que têm como destino a Reserva de Capacidade.
6. (Opcional) Para Tags, especifique as tags a serem associadas à Reserva de Capacidade.
7. Escolha Criar.

8. Anote o ID da recém-criada Reserva de Capacidade. Você precisa dele para configurar o grupo de Reserva de Capacidade.

Repita este procedimento para cada zona de disponibilidade que você deseja habilitar para seu grupo do Auto Scaling, alterando somente o valor da opção Zona de disponibilidade.

AWS CLI

Para criar suas Reservas de Capacidade

Use o create-capacity-reservationcomando a seguir 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 recém-criada Reserva de Capacidade. Você precisa dele para configurar o grupo de Reserva de Capacidade.

Repita este comando para cada zona de disponibilidade que você deseja habilitar para seu grupo do Auto Scaling, alterando somente o valor da opção`--availability-zone`.

Etapa 2: criar um grupo de Reserva de Capacidade

Ao terminar de criar as reservas de capacidade, você poderá agrupá-las usando o serviço AWS Resource Groups. AWS O Resource Groups oferece suporte a vários tipos diferentes de grupos para diferentes usos. O Amazon EC2 usa um grupo de propósito especial, 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 no Guia do Usuário de Grupos de Recruso AWS .

Para criar um grupo de reserva de capacidade usando o AWS CLI

Use o comando create-group 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"]}]}'
```

A seguir, uma exemplo de resposta.

```
{
    "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"
                        ]
                    }
                ]
            }
        ],
        "Status": "UPDATE_COMPLETE"
    }
}
```

Anote o ARN do novo grupo de recursos. Você precisa configurar o modelo de execução para seu grupo do Auto Scaling.

Para associar suas Reservas de Capacidade ao grupo recém-criado usando o AWS CLI

Use o comando group-resources a seguir para associar as Reservas de Capacidade ao grupo de Reserva de Capacidade recém-criado. Para a opção,`--resource-arns` especifique as Reservas de Capacidade usando seus ARNs. Estruture os ARNs usando a região relevante, seu ID de conta e os IDs de reserva que você anotou anteriormente. Neste exemplo, as reservas com IDs *cr-1234567890abcdef1* e *cr-54321abcdef567890* serão agrupadas 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
```

A seguir, uma exemplo de resposta.

```
{
   "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 a AWS Referência da API dos Grupos de Recurso.

Etapa 3: criar um modelo de execução

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 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. Criar o modelo de execução. Selecione um AMI e o tipo de instância que corresponde às Reservas de Capacidade que está planejando usar, 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. Amplie os Detalhes avançados e faça o seguinte:
    a. Em Reserva de Capacidade, escolha Destino por grupo.
    b. Em Reserva de Capacidade - Destino por grupo, escolha o grupo de Reservas de Capacidade que você criou na seção anterior e, em seguida, escolha Salvar.
7. Escolha 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 create-launch-templatecomando a seguir para criar um modelo de execução que especifica que a Reserva de Capacidade tem como alvo um grupo de recursos específico. Substitua o valor da amostra por `--launch-template-name`. Substitua *`c5.xlarge`* pelo tipo de instância que

você usou na Reserva de Capacidade e *ami-0123456789EXAMPLE* pelo ID da AMI que você deseja usar. Substitua *arn:aws:resource-groups:region:account-id:group/my-cr-group* pelo o ARN do grupo de recurso que você criou no início da etapa 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" }
            }
        }'
```

A seguir, uma exemplo de resposta.

```
{
    "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 do Auto Scaling

Console

Crie seu grupo do Auto Scaling como você costuma fazer, mas ao escolher suas sub-redes VPC, escolha uma sub-rede de cada zona de disponibilidade que corresponda às Reservas de Capacidade `targeted` que você criou. Em seguida, quando seu grupo do 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 atingida, lançaremos qualquer coisa além da capacidade reservada como capacidade normal sob demanda.

Para criar um grupo do Auto Scaling simples

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 que você usou ao criar o modelo de lançamento.
3. Selecione Criar um grupo do Auto Scaling.
4. Na página Choose launch template or configuration (Escolher 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 Escolha as opções de execução da instância, pule a seção Requisitos do tipo de instância para usar o tipo de instância do EC2 especificado no modelo de execução.
8. Em Network (Rede), para VPC, escolha uma VPC. O grupo do Auto Scaling deve ser criado na mesma VPC do grupo de segurança especificado no modelo de execução. Se você não especificou um grupo de segurança para seu modelo de execução, você pode selecionar qualquer VPC que tenha sub-redes nas mesmas zonas de disponibilidade que suas Reservas de Capacidade.
9. Para zonas de disponibilidade e sub-redes, selecione sub-redes de cada zona de disponibilidade que você deseja incluir, com base nas zonas de disponibilidade em que suas Reservas de Capacidade se encontram.
10. Escolha Next (Próximo) duas vezes.
11. Na página Configurar tamanho do grupo e políticas de escalabilidade, em 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 ter mais informações, consulte Definir limites de escalabilidade para seu grupo do Auto Scaling.
12. Escolha Skip to review (Ir para revisão).
13. Na página Review (Revisão), escolha Create Auto Scaling group (Criar grupo do Auto Scaling).

AWS CLI

Para criar um grupo do Auto Scaling simples

Use o create-auto-scaling-groupcomando a seguir e especifique o nome e a versão do seu modelo de lançamento como o valor da `--launch-template` opção. Substitua os valores de amostra por,`--auto-scaling-group-name`,`--min-size` `--max-size` e `--vpc-zone-identifier`.

Para a opção,`--availability-zones` especifique as zonas de disponibilidade para as quais você criou reservas de capacidade. Por exemplo, se suas reservas de capacidade especificarem as zonas de disponibilidade `us-east-1a` e,`us-east-1b` então você deverá criar seu grupo do Auto Scaling nas mesmas zonas. Em seguida, quando seu grupo do 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 atingida, 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 o AWS CloudFormation modelo no seguinte GitHub repositório de AWS exemplos: https://github.com/aws-samples/ aws-auto-scaling-backed - by-on-demand-capacity -reservations/.

Os tópicos relacionados a seguir podem ser úteis à medida que você aprende sobre Reservas de Capacidade.

- On-Demand Capacity Reservations
  - Crie uma reserva de capacidade no Guia do usuário do Amazon EC2 para instâncias do Linux
  - Reservas de capacidade sob demanda, no Guia do usuário do Amazon EC2 para instâncias do Linux

  - Escolha um grupo de reservas de capacidade sob demanda do Amazon EC2 no blog de operações e migrações na AWS nuvem
- Blocos de capacidade (Reservas de Capacidade com uma duração definida)
  - Blocos de capacidade para ML no Guia do Usuário do Amazon EC2 para Instâncias do Linux
  - Use blocos de capacidade para cargas de trabalho de aprendizado de máquina

Crie grupos de Auto Scaling a partir da linha de comando usando AWS CloudShell

Se houver suporte Regiões da AWS, você pode executar AWS CLI comandos usando AWS CloudShell um shell pré-autenticado baseado em navegador que é iniciado diretamente do. AWS Management Console Você pode executar AWS CLI comandos em serviços usando seu shell preferido (shell Bash ou Z). PowerShell

Você pode iniciar a AWS CloudShell partir do AWS Management Console usando um dos dois métodos a seguir:

- Escolha o AWS CloudShell ícone na barra de navegação do console. Ele está à direita da caixa de pesquisa.
- Use a caixa de pesquisa na barra de navegação do console para pesquisar CloudShelle, em seguida, escolha a CloudShellopção.

Quando AWS CloudShell é iniciado em uma nova janela do navegador pela primeira vez, um painel de boas-vindas é exibido e lista os principais recursos. Depois de fechar esse painel, as atualizações de status serão fornecidas enquanto o shell configura e encaminha suas credenciais do console. Quando o prompt de comando for exibido, o shell estará pronto para interação.

Para obter mais informações sobre esse serviço, consulte o Manual do usuário do AWS CloudShell.

Criar um grupo do Auto Scaling com AWS CloudFormation

O Amazon EC2 Auto Scaling é integrado AWS CloudFormation com um serviço que ajuda você a modelar e configurar AWS seus recursos para que você possa gastar menos tempo criando e gerenciando seus recursos e infraestrutura. Você cria um modelo que descreve todos os AWS recursos desejados (como grupos do Auto Scaling) e AWS CloudFormation provisiona e configura esses recursos para você.

Ao usar AWS CloudFormation, você pode reutilizar seu modelo para configurar seus recursos do Amazon EC2 Auto Scaling de forma consistente e repetida. Descreva seus recursos uma vez e, em seguida, provisione os mesmos recursos repetidamente em várias Contas da AWS regiões.

Amazon EC2 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 em suas AWS CloudFormation pilhas. Se você não estiver familiarizado com JSON ou YAML, você pode usar o AWS CloudFormation Designer para ajudá-lo a começar a usar modelos. AWS CloudFormation Para obter mais informações, consulte O que é AWS CloudFormation Designer? no Guia do AWS CloudFormation usuário.

Para começar a criar seus próprios modelos de pilha para o Amazon EC2 Auto Scaling, realize as tarefas a seguir:

- Crie um modelo de lançamento usando AWS::EC2::LaunchTemplateo.
-

Para ver um passo a passo que mostra como implantar um grupo do Auto Scaling por trás de um Application Load Balancer, consulte Demonstração: criar uma aplicação escalável com balanceamento de carga no Guia do usuário do AWS CloudFormation .

Você pode encontrar outros exemplos úteis de trechos de modelos que criam grupos de Auto Scaling e recursos relacionados nas seguintes seções do Guia do AWS CloudFormation Usuário:

- Referência de tipo de recurso Amazon EC2 Auto Scaling Referência de tipo de recurso Amazon Scaling
- Configure os recursos do Amazon EC2 Auto Scaling com AWS CloudFormation

Saiba mais sobre AWS CloudFormation

Para saber mais sobre isso AWS CloudFormation, consulte os seguintes recursos:

- AWS CloudFormation
- AWS CloudFormation Guia do usuário
- AWS CloudFormation API Reference

- AWS CloudFormation Guia do usuário da interface de linha de comando

Use AWS Compute Optimizer para obter recomendações sobre o tipo de instância para um grupo de Auto Scaling

AWS fornece recomendações de instâncias do Amazon EC2 para ajudá-lo a melhorar o desempenho, economizar dinheiro ou ambos, usando recursos desenvolvidos por. AWS Compute OptimizerÉ 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 Conceitos básicos do AWS Compute Optimizer no Manual do usuário do AWS Compute Optimizer .

Conteúdo

- Limitações
- Descobertas
- Exibir recomendações
- Considerações para avaliação das recomendações

Limitações

O Compute Optimizer gera recomendações para instâncias em grupos do Auto Scaling configurados para iniciar e executar os tipos de instância M, C, R, T e X. No entanto, ele não gera recomendações para tipos de instância -g alimentados por processadores AWS Graviton2 (por exemplo, C6g) e para tipos de instância -n que têm maior desempenho de largura de banda de rede (por exemplo, M5n).

Os grupos do Auto Scaling também devem ser configurados para executar um único tipo de instância (ou seja, nenhum tipo de instância mista), não devem ter uma política de escalabilidade anexada a eles e ter os mesmos valores para a capacidade desejada, mínima e máxima (ou seja, um grupo do Auto Scaling com um número fixo de instâncias). O Compute Optimizer gera recomendações para instâncias em grupos do Auto Scaling que atendam todos esses requisitos de configuração.

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.

Exibir recomendações

Depois de optar pelo Compute Optimizer, é possível visualizar as descobertas e as recomendações que ele gera para seus grupos do Auto Scaling. Caso tenho realizado a opção recentemente, as recomendações poderão não estar disponíveis durante até 12 horas.

Como visualizar as recomendações geradas para um grupo do Auto Scaling

1. Abra o console do Compute Optimizer em https://console.aws.amazon.com/compute-optimizer/.

   A página Dashboard (Painel) é aberta.
2. Escolha View recommendations for all Auto Scaling groups (Visualizar recomendações para todos os grupos de Auto Scaling).
3. Selecione seu grupo do Auto Scaling.

4. Escolha View detail (Visualizar detalhes).

   A visualização muda para exibir até três recomendações de instância diferentes em uma visualização pré-configurada, com base nas configurações de tabela padrão. Ele também fornece dados CloudWatch métricos recentes (utilização média da CPU, média de entrada e média de saída de rede) para o grupo Auto Scaling.

Determine se deseja usar uma das recomendações. Decida se deseja realizar a otimização para melhorar a performance, reduzir custos ou ambos.

Para alterar o tipo de instância no grupo do Auto Scaling, atualize o modelo de execução ou o grupo do Auto Scaling para usar uma nova configuração de execução. As instâncias existentes continuam a usar a configuração anterior. Para atualizar as instâncias existentes, termine-as para que elas sejam substituídas pelo grupo do Auto Scaling, ou permita que a escalabilidade automática substitua gradualmente as instâncias mais antigas por instâncias mais novas com base em suas políticas de término.

> **Note**
>
> Com os recursos de tempo de vida máximo e atualização de instância, você também pode substituir instâncias existentes no grupo do Auto Scaling para iniciar novas instâncias que usem o modelo de execução ou a configuração de execução. Para obter mais informações, consulte Substituir instâncias do Auto Scaling com base na vida útil máxima da instância e Use uma atualização de instância para atualizar instâncias em um grupo de Auto Scaling.

Considerações para avaliação das recomendações

Antes de passar para um novo tipo de instância, considere o seguinte:

- As recomendações não preveem seu uso. As recomendações são baseadas em seu histórico de uso durante os últimos 14 dias. Escolha um tipo de instância que atenda às suas necessidades de uso futuras.
- Concentre-se nas métricas gráficas para determinar se o uso real é menor do que a capacidade da instância. Você também pode visualizar dados métricos (média, pico, percentil) CloudWatch para avaliar melhor suas recomendações de instância do EC2. Por exemplo, observe como as métricas de porcentagem da CPU mudam durante o dia e se há picos que precisem ser acomodados. Para

obter mais informações, consulte Visualização das métricas disponíveis no Guia CloudWatch do usuário da Amazon.

- O Compute Optimizer pode fornecer recomendações para instâncias expansíveis, que são as instâncias T3, T3a e T2. Se você 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 do

Além dos tópicos desta página, consulte os seguintes recursos:

- Tipos de instância do Amazon EC2
- AWS Compute Optimizer Guia do usuário

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. Isso registra o grupo com o balanceador de carga, o qual atua como um ponto único de contato para todo o tráfego da Web de entrada para seu grupo do Auto Scaling.

Quando você usa o Elastic Load Balancing com seu grupo do Auto Scaling, não é necessário registrar suas instâncias do EC2 no balanceador de carga. As instâncias iniciadas pelo grupo do Auto Scaling serão automaticamente registradas no balanceador de carga. Da mesma forma, as instâncias que são terminadas pelo grupo do Auto Scaling terão o registro cancelado automaticamente no balanceador de carga.

Depois de anexar um load balancer ao grupo do Auto Scaling, você poderá configurar o grupo do Auto Scaling para usar métricas do Elastic Load Balancing (como a contagem de solicitações do Application Load Balancer por destino) para dimensionar o número de instâncias no grupo conforme a demanda flutua.

Opcionalmente, você pode adicionar verificações de integridade do Elastic Load Balancing ao seu grupo do Auto Scaling para que o Amazon EC2 Auto Scaling possa identificar e substituir instâncias não íntegras com base nessas verificações de integridade adicionais. Caso contrário, você pode criar um CloudWatch alarme que o notifique se a contagem de hosts saudáveis do grupo-alvo for menor do que a permitida.

Conteúdo

- Tipos de Elastic Load Balancing
- Prepare-se para conectar um balanceador de carga do Elastic Load Balancing ao seu grupo de Auto Scaling
- Anexe um balanceador de carga do Elastic Load Balancing ao seu grupo de Auto Scaling
- Configurar um Application Load Balancer ou Network Load Balancer pelo console do Amazon EC2 Auto Scaling
- Verificar o status do anexo de seu balanceador de carga
- Adicionar e remover zonas de disponibilidade
- Exemplos de como trabalhar com o Elastic Load Balancing com o AWS Command Line Interface ()AWS CLI

Tipos de Elastic Load Balancing

O Elastic Load Balancing oferece quatro tipos de balanceadores de carga que podem ser usados com seu grupo do Auto Scaling: balanceadores de carga de aplicação, balanceadores de carga de rede, balanceadores de carga de gateway e balanceadores de carga clássicos

Há uma diferença fundamental em como os tipos de balanceadores de carga são configurados. Com os balanceadores de carga de aplicação, balanceadores de carga de rede e balanceadores de carga de gateway, as instâncias são registradas como destinos com um grupo de destino, e o tráfego deve ser roteado para o grupo de destino. Com balanceadores de carga clássicos, as instâncias são registradas diretamente no balanceador de carga.

Application Load Balancer

Roteia e faz balanceamento de carga na camada da aplicação (HTTP/HTTPS) e é compatível com roteamento baseado em caminho. Um Application Load Balancer pode rotear solicitações para portas em um ou mais destinos registrados, como instâncias do EC2, na sua nuvem privada virtual (VPC).

Network Load Balancer

Roteia e promove o balanceamento de carga na camada de transporte (camada 4 do TCP/UDP) com base nas informações de endereço extraídas do cabeçalho da camada 4. Os balanceadores de carga de rede podem lidar com picos de tráfego, reter o IP de origem do cliente e usar um IP fixo para a vida útil do balanceador de carga.

Balanceador de carga de gateway

Distribui o tráfego para uma frota de instâncias de dispositivos. Fornece escalabilidade, disponibilidade e simplicidade para dispositivos virtuais de terceiros, como firewalls, sistemas de detecção e prevenção de intrusões e outros dispositivos. Os balanceadores de carga de gateway funcionam com dispositivos virtuais compatíveis com o protocolo GENEVE. Integração técnica adicional é necessária, portanto, certifique-se de consultar o manual do usuário antes de escolher um balanceador de carga de gateway.

Classic Load Balancer

Roteia e faz balanceamento de carga na camada de transporte (TCP/SSL) ou na camada de aplicação (HTTP/HTTPS).

Para obter uma compreensão mais profunda dos diferentes tipos de balanceadores de carga disponíveis, consulte os seguintes recursos:

- O que é Elastic Load Balancing?
- O que é um Application Load Balancer?
- O que é um Network Load Balancer?
- O que é um balanceador de carga de gateway?
- O que é um Classic Load Balancer?

Prepare-se para conectar um balanceador de carga do Elastic Load Balancing ao seu grupo de Auto Scaling

Antes de conectar um balanceador de carga do Elastic Load Balancing ao seu grupo de Auto Scaling, você deve preencher os seguintes pré-requisitos:

- Você já deve ter criado o balanceador de carga e o grupo-alvo usados para rotear o tráfego para seu grupo de Auto Scaling.

  Há duas maneiras de criar o balanceador de carga e o grupo-alvo:

  - Usando o Elastic Load Balancing — Siga os procedimentos na documentação do Elastic Load Balancing para criar e configurar o balanceador de carga e o grupo-alvo antes de criar o grupo Auto Scaling. Ignore a etapa para registrar suas instâncias do Amazon EC2. O Amazon EC2 Auto Scaling cuida automaticamente do registro (e cancelamento do registro) de instâncias quando você anexa um grupo-alvo ao seu grupo de Auto Scaling. Para obter mais informações, consulte Conceitos básicos do Elastic Load Balancing no Manual do usuário do Elastic Load Balancing.
  - Usando o Amazon EC2 Auto Scaling — Crie, configure e conecte o balanceador de carga e o grupo-alvo com uma configuração básica do console do Amazon EC2 Auto Scaling. Para ter mais informações, consulte Configurar um Application Load Balancer ou Network Load Balancer pelo console do Amazon EC2 Auto Scaling.
- Antes de criar um balanceador de carga, saiba o tipo de balanceador de carga que você precisa. Para ter mais informações, consulte Tipos de Elastic Load Balancing.
- O balanceador de carga e seu grupo-alvo devem estar na mesma Conta da AWS VPC e região do seu grupo de Auto Scaling.

- O grupo de destino deve especificar um tipo de destino `instance`. Não é possível especificar um tipo de destino `ip` ao usar um grupo do Auto Scaling.
- Se o modelo de execução do seu grupo de Auto Scaling não contiver o grupo de segurança correto para permitir o tráfego de entrada necessário do balanceador de carga, 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. As instâncias existentes não são atualizadas com as novas configurações quando o modelo de execução é modificado. Para atualizar as instâncias existentes, você pode iniciar uma atualização da instância para substituir as instâncias. Para ter mais informações, consulte Use uma atualização de instância para atualizar instâncias em um grupo de Auto Scaling.
- Os grupos de segurança no modelo de lançamento também devem permitir o acesso do balanceador de carga na porta correta para que o Elastic Load Balancing realize suas verificações de integridade.
- Ao implantar dispositivos virtuais por trás de um Gateway Load Balancer, a Amazon Machine Image (AMI) no modelo de lançamento deve especificar a ID de uma AMI que suporte o protocolo GENEVE para permitir que o grupo Auto Scaling troque tráfego com um Gateway Load Balancer. Além disso, os grupos de segurança no modelo de lançamento devem permitir o tráfego UDP na porta 6081.

> **Tip**
>
> Se você tiver scripts de bootstrap que levam um tempo para serem concluídos, você pode adicionar opcionalmente um hook 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. No entanto, você pode adicionar um gancho de ciclo de vida após a criação do grupo. Para ter mais informações, consulte Ganchos do ciclo de vida do Amazon EC2 Auto Scaling.

Configurar verificações de saúde para alvos

Você pode configurar verificações de saúde para seus destinos registrados com um balanceador de carga do Elastic Load Balancing para garantir que eles sejam capazes de lidar com o tráfego

adequadamente. As etapas específicas variam de acordo com o tipo de balanceador de carga que você está usando. Para obter mais informações, consulte os seguintes recursos do :

- Application Load Balancer — Consulte Verificações de saúde para seus grupos-alvo no Guia do usuário do Application Load Balancers.
- Network Load Balancer — Consulte Verificações de saúde para seus grupos-alvo no Guia do usuário para Network Load Balancers.
- Gateway Load Balancer — Consulte Verificações de saúde para seus grupos-alvo no Guia do usuário de Gateway Load Balancers.
- Classic Load Balancer — Consulte Configurar verificações de saúde para seu Classic Load Balancer no Guia do usuário para Classic Load Balancers.

Por padrão, o Amazon EC2 Auto Scaling não considera uma instância não íntegra e a substitui se ela falhar nas verificações de saúde do Elastic Load Balancing. As verificações de integridade padrão para um grupo do Auto Scaling são somente verificações de integridade do EC2. Para ter mais informações, consulte Verificações de integridade para instâncias em um grupo do Auto Scaling.

Para permitir que o Amazon EC2 Auto Scaling substitua instâncias consideradas insalubres pelo Elastic Load Balancing, você pode configurar seu grupo de Auto Scaling para usar as verificações de saúde do Elastic Load Balancing. Ao fazer isso, o Amazon EC2 Auto Scaling considera a instância insalubre se ela falhar nas verificações de saúde do EC2 ou nas verificações de saúde do Elastic Load Balancing. Se você anexar vários grupos de destino do balanceador de carga ou balanceadores de carga clássicos ao grupo, todos eles deverão informar que a instância é íntegra para que ela seja considerada íntegra. Se um deles relatar uma instância como não íntegra, o grupo do Auto Scaling substituirá a instância, mesmo que outros a relatem como íntegra.

Para obter informações sobre como habilitar essas verificações de saúde para seu grupo de Auto Scaling, consulte. Anexe um balanceador de carga do Elastic Load Balancing ao seu grupo de Auto Scaling

> **Note**
>
> Para garantir que essas verificações de saúde comecem o mais rápido possível, certifique-se de que o período de carência da verificação de saúde do seu grupo não esteja definido como alto demais, 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 ter mais

informações, consulte Definir um período de carência da verificação de integridade para um grupo do Auto Scaling.

Anexe um balanceador de carga do Elastic Load Balancing ao seu grupo de Auto Scaling

Este tópico descreve como conectar um balanceador de carga do Elastic Load Balancing a um grupo do Auto Scaling. Também descreve como ativar as verificações de saúde do Elastic Load Balancing para permitir que o Amazon EC2 Auto Scaling substitua instâncias que o Elastic Load Balancing relata como não íntegras.

Por padrão, o Amazon EC2 Auto Scaling substitui somente instâncias que não sejam íntegras ou inacessíveis com base nas verificações de integridade do Amazon EC2. Se você ativar as verificações de saúde do Elastic Load Balancing, o Amazon EC2 Auto Scaling poderá substituir uma instância em execução se algum dos load balancers do Elastic Load Balancing que você vincular ao grupo Auto Scaling relatar que ela não está íntegra.

Para ver um tutorial sobre como anexar um Application Load Balancer ao seu grupo de Auto Scaling, consulte. Tutorial: Configurar uma aplicação escalonada e com balanceamento de carga

Important

Antes de continuar, preencha todos os pré-requisitos na seção anterior.

Conteúdo

- Anexar um grupo-alvo ou Classic Load Balancer
- Separar um grupo-alvo ou Classic Load Balancer

Anexar um grupo-alvo ou Classic Load Balancer

Ao criar ou atualizar um grupo de Auto Scaling, você pode anexar um ou mais grupos-alvo ou Classic Load Balancers. Ao anexar um Application Load Balancer, Network Load Balancer ou Gateway Load Balancer, você anexa um grupo-alvo em vez do próprio balanceador de carga.

Siga as etapas nesta seção para usar o console para:

- Anexar um grupo-alvo ou Classic Load Balancer a um grupo de Auto Scaling
- Ative as verificações de saúde do Elastic Load Balancing

Para anexar um balanceador de carga existente enquanto cria um novo grupo do Auto Scaling

1. Abra o console do Amazon EC2 em https://console.aws.amazon.com/ec2/ e escolha Grupos do Auto Scaling no painel de navegação.
2. Na barra de navegação na parte superior da tela, escolha aquela em Região da AWS que você criou seu balanceador de carga.
3. Selecione Criar grupo do Auto Scaling.
4. Nas etapas 1 e 2, escolha as opções conforme desejado e prossiga para Etapa 3: Configurar opções avançadas.
5. Em Load balancing (Balanceamento de carga), escolha Attach to an existing load balancer (Anexar a um balanceador de carga existente).
6. Em Attach to an existing load balancer (Anexar a um balanceador de carga existente), siga um destes procedimentos:

   a. Para balanceadores de carga de aplicação, balanceadores de carga de rede e balanceadores de carga de gateway, especifique a propriedade:

      Escolha Choose from your load balancer target groups (Escolher entre os grupos de destino do balanceador de carga) e, em seguida, escolha um grupo de destino no campo Existing load balancer target groups (Grupos de destino do balanceador de carga existente).

   b. Para balanceadores de carga clássicos:

      Escolha Choose from Classic Load Balancers (Escolher entre balanceadores de carga clássicos) e, em seguida, escolha seu balanceador de carga no campo Classic Load Balancers (Balanceadores de carga clássicos).

7. (Opcional) Para verificações de integridade e tipos adicionais de verificação de integridade, selecione Ativar verificações de integridade do Elastic Load Balancing.
8. Opcional em Tempo de carência da verificação de integridade, insira a quantidade de tempo em segundos. Esse é o tempo que o Amazon EC2 Auto Scaling precisa aguardar antes de verificar o status de integridade de uma instância depois que ela entra no estado`InService`. Para ter mais informações, consulte Definir um período de carência da verificação de integridade para um grupo do Auto Scaling.

9. Prossiga para criar o grupo do Auto Scaling. Suas instâncias serão registradas automaticamente no balanceador de carga após a criação do grupo do Auto Scaling.

Para anexar um balanceador de carga existente ao seu grupo do Auto Scaling após ter sido criado

1. Abra o console do Amazon EC2 em https://console.aws.amazon.com/ec2/ e escolha 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 balanceadores de carga de aplicação, rede ou gateway), marque sua caixa de seleção e escolha um grupo de destino.

   b. Para Classic Load Balancers (Balanceadores de carga clássicos), marque sua caixa de seleção e escolha seu balanceador de carga.
5. Escolha Atualizar.

Ao terminar de conectar o balanceador de carga, você pode, opcionalmente, ativar as verificações de saúde que o usam.

Para ativar as verificações de saúde do Elastic Load Balancing

1. Na guia Detalhes, escolha Verificações de integridade, Editar.
2. Para Verificações de integridade, Tipos adicionais de verificação de integridade, selecione Ativar verificações de integridade do Elastic Load Balancing.
3. Em Período de carência da verificação de integridade, insira a quantidade de tempo em segundos. Esse é o tempo que o Amazon EC2 Auto Scaling precisa aguardar antes de verificar o status de integridade de uma instância depois que ela entra no estado`InService`. Para ter mais informações, consulte Definir um período de carência da verificação de integridade para um grupo do Auto Scaling.
4. Escolha 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 ter mais informações, consulte Verificar o status do anexo de seu balanceador de carga.

Separar um grupo-alvo ou Classic Load Balancer

Quando o balanceador de carga não for mais necessário, use o procedimento a seguir para desvinculá-lo do grupo do Auto Scaling.

Para desvincular um balanceador de carga de um grupo

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

Ao terminar de separar o grupo-alvo, você pode desativar as verificações de saúde do Elastic Load Balancing.

Para desativar as verificações de saúde do Elastic Load Balancing

1. Na guia Detalhes, escolha Verificações de integridade, Editar.

2. Para Verificações de saúde, Tipos adicionais de verificação de saúde, desmarque Ativar verificações de saúde do Elastic Load Balancing.
3. Escolha 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. (Opcional) Para verificações de integridade e tipos adicionais de verificação de integridade, selecione Ativar verificações de integridade do Elastic Load Balancing.

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

7. Prossiga para criar o grupo do Auto Scaling. Suas instâncias serão registradas automaticamente no balanceador de carga após a criação do grupo do Auto Scaling.

> **Note**
>
> Depois de criar o grupo do Auto Scaling, você poderá usar o console do Elastic Load Balancing para criar listeners adicionais. Isso será útil se você precisar criar um listener com um protocolo seguro, como HTTPS ou um ouvinte UDP. Você pode adicionar mais listeners aos balanceadores de carga existentes, desde que use portas distintas.

Verificar o status do anexo de seu balanceador de carga

Após você associar um balanceador de carga, ele entra no estado `Adding` ao registrar as instâncias no grupo. Quando todas as instâncias do grupo são registradas, ele entra no estado`Added`. Depois que pelo menos uma instância registrada passa nas verificações de integridade, ele entra no estado`InService`. Após o balanceador de carga entrar no estado `InService`, o Amazon EC2 Auto Scaling pode encerrar e substituir todas as instâncias relatadas como não íntegras. Se nenhuma instância registrada passar nas verificações de integridade (por exemplo, devido a um erro na configuração da verificação de integridade), o balanceador de carga não entrará no estado `InService`. O Amazon EC2 Auto Scaling não termina e substitui as instâncias.

Quando você desanexa um balanceador de carga, ele entra no estado `Removing` ao cancelar o registro das instâncias do grupo. As instâncias permanecem em execução depois que seus

registros são cancelados. Por padrão, a descarga da conexão (atraso de cancelamento de registro) é habilitada para Application Load Balancers, Network Load Balancers e Gateway Load Balancers. Se a descarga de conexão estiver habilitada, o Elastic Load Balancing aguardará que as solicitações em andamento sejam concluídas ou que o limite de tempo máximo expire (o que ocorrer primeiro) antes de cancelar o registro das instâncias.

Você pode verificar o status do anexo usando o AWS Command Line Interface (AWS CLI) ou AWS SDKs. Você não pode verificar o status do anexo no console.

Para usar o AWS CLI para verificar o status do anexo

O describe-traffic-sourcescomando a seguir retorna o status do anexo de todas as fontes de tráfego do grupo de Auto Scaling especificado.

```
aws autoscaling describe-traffic-sources --auto-scaling-group-name my-asg
```

O exemplo retorna o ARN do grupo-alvo do Elastic Load Balancing que está vinculado ao grupo do Auto Scaling, junto com o status do anexo do grupo-alvo no elementoState.

```
{
    "TrafficSources": [
        {
            "Identifier": "arn:aws:elasticloadbalancing:region:account-
id:targetgroup/my-targets/1234567890123456",
            "State": "InService",
            "Type": "elbv2"
        }
    ]
}
```

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).

Conteúdo

- Adicione uma Zona de disponibilidade
- Remover uma Zona de disponibilidade
- Recursos relacionados
- Limitações

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 Atualizar.
6. Para atualizar as zonas de disponibilidade do seu balanceador de carga para que ele compartilhe as mesmas zonas do seu grupo do Auto Scaling, execute as seguintes etapas:

   a. No painel de navegação, em Load Balancing (Balanceamento de carga), escolha Load balancers (Balanceadores de carga).

   b. Escolha seu balanceador de carga.

   c. Execute um destes procedimentos:

      - Para balanceadores de carga de aplicação e balanceadores de carga de rede:

        1. Na guia Description (Descrição), em Availability Zones (Zonas de disponibilidade), escolha Edit subnets (Editar sub-redes).
        2. Na página Edit subnets (Editar sub-redes), para Availability Zones (Zonas de disponibilidade), marque a caixa de seleção para a zona de disponibilidade a ser adicionada. Se houver somente uma sub-rede para essa zona de disponibilidade, ela estará selecionada. Se houver mais de uma sub-rede para essa zona de disponibilidade, selecione uma das opções disponíveis.

      - Para balanceadores de carga clássicos em uma VPC:

        1. Na guia Instâncias, selecione Editar zonas de disponibilidade.
        2. Na página Add and Remove Subnets (Adicionar e remover sub-redes), em Available subnets (Sub-redes disponíveis), selecione a sub-rede usando o ícone de adicionar (+). A sub-rede é movida sob Sub-redes selecionadas.

   d. Selecione 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 Atualizar.
6. Para atualizar as zonas de disponibilidade do seu balanceador de carga para que ele compartilhe as mesmas zonas do seu grupo do Auto Scaling, execute as seguintes etapas:

   a. No painel de navegação, em Load Balancing (Balanceamento de carga), escolha Load balancers (Balanceadores de carga).

   b. Escolha seu balanceador de carga.

   c. Execute um destes procedimentos:

      - Para balanceadores de carga de aplicação e balanceadores de carga de rede:

        1. Na guia Description (Descrição), em Availability Zones (Zonas de disponibilidade), escolha Edit subnets (Editar sub-redes).
        2. Na página Edit subnets (Editar sub-redes), para Availability Zones (Zonas de disponibilidade), desmarque a caixa de seleção para remover a sub-rede da zona de disponibilidade selecionada.

      - Para balanceadores de carga clássicos em uma VPC:

        1. Na guia Instâncias, selecione Editar zonas de disponibilidade.
        2. Na página Add and Remove Subnets (Adicionar e remover sub-redes), em Available subnets (Sub-redes disponíveis), remova a sub-rede usando o ícone de exclusão (-). A sub-rede é movida para Available subnets (Sub-redes disponíveis).

   d. Selecione Salvar.

Recursos relacionados

O Amazon EC2 Auto Scaling rebalanceia seu grupo quando você altera as zonas de disponibilidade. Isso significa substituir e redistribuir algumas instâncias. Para ter mais informações, consulte Exemplo: distribuir instâncias entre zonas de disponibilidade.

Se você registrou destinos em zonas de disponibilidade que não estão habilitados para o balanceador de carga, o balanceador de carga não roteará o tráfego para eles. Para obter mais informações, consulte Como o Elastic Load Balancing funciona no Manual do usuário do Elastic Load Balancing.

Limitações

Para atualizar quais zonas de disponibilidade estão habilitadas para seu balanceador de carga, é necessário estar ciente das seguintes limitações:

- Ao habilitar uma zona de disponibilidade para seu balanceador de carga, você especifica uma sub-rede nessa zona de disponibilidade. Observe que é possível habilitar no máximo uma sub-rede por zona de disponibilidade para seu balanceador de carga.
- Para balanceadores de carga voltados para a Internet, as sub-redes especificadas para o balanceador de carga devem ter pelo menos oito endereços IP disponíveis.
- Para balanceadores de carga de aplicação, é necessário habilitar pelo menos duas zonas de disponibilidade.
- Para balanceadores de carga de rede, você não pode desabilitar as zonas de disponibilidade habilitadas, mas pode habilitar zonas adicionais.
- Para balanceadores de carga de gateway, você não pode desativar as zonas de disponibilidade ativadas, mas pode ativar outras.

Exemplos de como trabalhar com o Elastic Load Balancing com o AWS Command Line Interface ()AWS CLI

Use o AWS CLI para anexar, separar e descrever balanceadores de carga e grupos-alvo, adicionar e remover verificações de saúde do Elastic Load Balancing e alterar quais zonas de disponibilidade estão habilitadas.

Este tópico mostra exemplos de AWS CLI comandos que executam tarefas comuns para o Amazon EC2 Auto Scaling.

> **Important**
>
> Para obter exemplos de comandos adicionais, consulte `aws elbv2` e `aws elb` na Referência de comandos AWS CLI .

Conteúdo

- Anexar seu grupo-alvo ou Classic Load Balancer
- Descreva seus grupos de destino ou Classic Load Balancers
- Adicionar verificações de integridade do Elastic Load Balancing
- Alterar suas zonas de disponibilidade
- Desvincular seu grupo-alvo ou Classic Load Balancer
- Remover as verificações de integridade do Elastic Load Balancing
- Comandos legados

Anexar seu grupo-alvo ou Classic Load Balancer

Use o create-auto-scaling-groupcomando a seguir para criar um grupo de Auto Scaling e anexar simultaneamente um grupo-alvo especificando seu Amazon Resource Name (ARN). O grupo de destino pode ser associado a um Application Load Balancer, um Network Load Balancer ou um balanceador de carga do Gateway.

Substitua os valores de amostra por,`--auto-scaling-group-name`,`--vpc-zone-identifier` `--min-size` e `--max-size`. Para a opção,`--launch-template` substitua *`my-launch-template`* e *`1`* pelo nome e versão de um modelo de execução para seu grupo do Auto Scaling. Para a opção,`--traffic-sources` substitua o ARN de amostra pelo ARN de um grupo de destino para um Application Load Balancer, Network Load Balancer ou balanceador de carga de gateway.

```
aws autoscaling create-auto-scaling-group --auto-scaling-group-name my-asg \
  --launch-template LaunchTemplateName=my-launch-template,Version='1' \
  --vpc-zone-identifier "subnet-5ea0c127,subnet-6194ea3b,subnet-c934b782" \
  --min-size 1 --max-size 5 \
  --traffic-sources "Identifier=arn:aws:elasticloadbalancing:region:account-
id:targetgroup/my-targets/12345678EXAMPLE1"
```

Use o attach-traffic-sourcescomando para anexar outros grupos-alvo ao grupo Auto Scaling depois que ele for criado.

O comando a seguir adiciona outro grupo-alvo ao mesmo grupo.

```
aws autoscaling attach-traffic-sources --auto-scaling-group-name my-asg \
  --traffic-sources "Identifier=arn:aws:elasticloadbalancing:region:account-
id:targetgroup/my-targets/12345678EXAMPLE2"
```

Como alternativa, para anexar um Classic Load Balancer ao seu grupo, especifique as opções `--traffic-sources` e `--type` ao usar create-auto-scaling-group ou,attach-traffic-sources como no exemplo a seguir. Substitua *my-classic-load-balancer* pelo nome de um Classic Load Balancer. Para a opção,`--type` especifique um valor de **elb**.

```
--traffic-sources "Identifier=my-classic-load-balancer" --type elb
```

Descreva seus grupos de destino ou Classic Load Balancers

Para descrever os balanceadores de carga ou os grupos-alvo vinculados ao seu grupo de Auto Scaling, use o comando a describe-traffic-sourcesseguir. Substitua *my-asg* pelo nome do seu grupo.

```
aws autoscaling describe-traffic-sources --auto-scaling-group-name my-asg
```

O exemplo retorna o ARN dos grupos de destino do Elastic Load Balancing que você anexou ao grupo do Auto Scaling.

```
{
    "TrafficSources": [
        {
            "Identifier": "arn:aws:elasticloadbalancing:region:account-
id:targetgroup/my-targets/12345678EXAMPLE1",
            "State": "InService",
            "Type": "elbv2"
        },
        {
            "Identifier": "arn:aws:elasticloadbalancing:region:account-
id:targetgroup/my-targets/12345678EXAMPLE2",
            "State": "InService",
            "Type": "elbv2"
        }
    ]
}
```

Para ver uma explicação do campo `State` na saída, consulte Verificar o status do anexo de seu balanceador de carga.

Adicionar verificações de integridade do Elastic Load Balancing

Para adicionar as verificações de saúde do Elastic Load Balancing às verificações de saúde que seu grupo de Auto Scaling realiza nas instâncias, use o comando a update-auto-scaling-groupseguir e

ELB especifique como o valor da opção. `--health-check-type` Substitua *my-asg* pelo nome do seu grupo.

```
aws autoscaling update-auto-scaling-group --auto-scaling-group-name my-asg \
  --health-check-type "ELB"
```

As novas instâncias geralmente precisam de tempo para um breve aquecimento antes de passarem por uma verificação de saúde. Se o período de carência não fornecer tempo de aquecimento suficiente, as instâncias podem não parecer prontas para atender ao tráfego. O Amazon EC2 Auto Scaling pode considerar essas instâncias não íntegras e substituí-las.

Para atualizar o período de carência da verificação de integridade, use a opção `--health-check-grace-period` ao usar,update-auto-scaling-group como no exemplo a seguir. Substitua *300* pelo número de segundos para manter as novas instâncias em serviço antes de encerrá-las, caso não estejam íntegras.

```
--health-check-grace-period 300
```

Para ter mais informações, consulte Verificações de integridade para instâncias em um grupo do Auto Scaling.

Alterar suas zonas de disponibilidade

Alterar suas zonas de disponibilidade apresenta algumas limitações das quais você deve estar ciente. Para ter mais informações, consulte Limitações.

Para alterar as zonas de disponibilidade de um Application Load Balancer ou Network Load Balancer

1. Antes de alterar as zonas de disponibilidade do balanceador de carga, é uma boa ideia primeiro atualizar as zonas de disponibilidade do grupo do Auto Scaling para verificar se há disponibilidade para seus tipos de instância nas zonas especificadas.

   Para atualizar as zonas de disponibilidade do seu grupo de Auto Scaling, use o comando a seguir update-auto-scaling-group. Substitua os IDs de sub-rede de amostra pelos IDs das sub-redes nas zonas de disponibilidade para habilitar. As sub-redes especificadas substituem as sub-redes habilitadas anteriormente. Substitua *my-asg* pelo nome do seu grupo.

   ```
   aws autoscaling update-auto-scaling-group --auto-scaling-group-name my-asg \
     --vpc-zone-identifier "subnet-41767929,subnet-cb663da2,subnet-8360a9e7"
   ```

2. Use o describe-auto-scaling-groupscomando a seguir para verificar se as instâncias nas novas sub-redes foram iniciadas. Se as instâncias tiverem sido iniciadas, você verá uma lista das instâncias e seus status. Substitua *my-asg* pelo nome do seu grupo.

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

3. Use o comando set-subnets a seguir para especificar as sub-redes do seu balanceador de carga. Substitua os IDs de sub-rede de amostra pelos IDs das sub-redes nas zonas de disponibilidade para habilitar. Você pode especificar somente uma sub-rede por Zona de disponibilidade. As sub-redes especificadas substituem as sub-redes habilitadas anteriormente. Substitua *my-lb-arn* pelo ARN do seu balanceador de carga.

```
aws elbv2 set-subnets --load-balancer-arn my-lb-arn \
  --subnets subnet-41767929 subnet-cb663da2 subnet-8360a9e7
```

Para alterar as zonas de disponibilidade de um Classic Load Balancer

1. Antes de alterar as zonas de disponibilidade do balanceador de carga, é uma boa ideia primeiro atualizar as zonas de disponibilidade do grupo do Auto Scaling para verificar se há disponibilidade para seus tipos de instância nas zonas especificadas.

   Para atualizar as zonas de disponibilidade do seu grupo de Auto Scaling, use o comando a seguir update-auto-scaling-group. Substitua os IDs de sub-rede de amostra pelos IDs das sub-redes nas zonas de disponibilidade para habilitar. As sub-redes especificadas substituem as sub-redes habilitadas anteriormente. Substitua *my-asg* pelo nome do seu grupo.

```
aws autoscaling update-auto-scaling-group --auto-scaling-group-name my-asg \
  --vpc-zone-identifier "subnet-41767929,subnet-cb663da2"
```

2. Use o describe-auto-scaling-groupscomando a seguir para verificar se as instâncias nas novas sub-redes foram iniciadas. Se as instâncias tiverem sido iniciadas, você verá uma lista das instâncias e seus status. Substitua *my-asg* pelo nome do seu grupo.

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

3. Use o comando attach-load-balancer-to-subnets a seguir para habilitar uma nova zona de disponibilidade para seu Classic Load Balancer. Substitua o ID de sub-rede de amostra pelo

ID da sub-rede para habilitar a zona de disponibilidade. Substitua *my-lb* pelo nome do seu balanceador de carga.

```
aws elb attach-load-balancer-to-subnets --load-balancer-name my-lb \
  --subnets subnet-cb663da2
```

Para desativar uma zona de disponibilidade, use o seguinte comando detach-load-balancer-from-subnets. Substitua o ID de sub-rede de amostra pelo ID da sub-rede para a zona de disponibilidade a ser desabilitada. Substitua *my-lb* pelo nome do seu balanceador de carga.

```
aws elb detach-load-balancer-from-subnets --load-balancer-name my-lb \
  --subnets subnet-8360a9e7
```

Desvincular seu grupo-alvo ou Classic Load Balancer

O detach-traffic-sourcescomando a seguir separa um grupo-alvo do seu grupo de Auto Scaling quando você não precisar mais dele.

Para a opção,`--auto-scaling-group-name` substitua *my-asg* pelo nome do seu grupo. Para a opção,`--traffic-sources` substitua o ARN de amostra pelo ARN de um grupo de destino para um Application Load Balancer, Network Load Balancer ou balanceador de carga de gateway.

```
aws autoscaling detach-traffic-sources --auto-scaling-group-name my-asg \
  --traffic-sources "Identifier=arn:aws:elasticloadbalancing:region:account-
id:targetgroup/my-targets/1234567890123456"
```

Para desvincular um Classic Load Balancer do seu grupo, especifique as opções `--traffic-sources` e,`--type` como no exemplo a seguir. Substitua *my-classic-load-balancer* pelo nome de um Classic Load Balancer. Para a opção,`--type` especifique um valor de **`elb`**.

```
--traffic-sources "Identifier=my-classic-load-balancer" --type elb
```

Remover as verificações de integridade do Elastic Load Balancing

Para remover as verificações de saúde do Elastic Load Balancing do seu grupo de Auto Scaling, use o comando a update-auto-scaling-groupseguir e **`EC2`** especifique como o valor da opção. `--health-check-type` Substitua *my-asg* pelo nome do seu grupo.

```
aws autoscaling update-auto-scaling-group --auto-scaling-group-name my-asg \
  --health-check-type "EC2"
```

Para ter mais informações, consulte Verificações de integridade para instâncias em um grupo do Auto Scaling.

Comandos legados

Os exemplos a seguir mostram como você pode usar comandos legados da CLI para anexar, desvincular e descrever balanceadores de carga e grupos de destino. Eles permanecem neste documento como referência para todos os clientes que desejam usá-los. Continuamos oferecendo suporte aos comandos antigos da CLI, mas recomendamos que você use os novos comandos "fontes de tráfego" da CLI, que podem anexar e desvincular vários tipos de fontes de tráfego. Você pode usar os comandos antigos da CLI e os comandos "fontes de tráfego" da CLI no mesmo grupo do Auto Scaling.

Anexar seu grupo de destino ou Classic Load Balancer (legado)

Para anexar um grupo de destino

O create-auto-scaling-groupcomando a seguir cria um grupo de Auto Scaling com um grupo-alvo anexado. Especifique o nome do recurso da Amazon (ARN) de um grupo de destino para um Application Load Balancer, Network Load Balancer ou balanceador de carga de gateway.

```
aws autoscaling create-auto-scaling-group --auto-scaling-group-name my-asg \
  --launch-template LaunchTemplateName=my-launch-template,Version='1' \
  --vpc-zone-identifier "subnet-5ea0c127,subnet-6194ea3b,subnet-c934b782" \
  --target-group-arns "arn:aws:elasticloadbalancing:region:account-id:targetgroup/my-
targets/1234567890123456" \
  --min-size 1 --max-size 5
```

O comando attach-load-balancer-target-groups a seguir anexa um grupo-alvo a um grupo existente do Auto Scaling.

```
aws autoscaling attach-load-balancer-target-groups --auto-scaling-group-name my-asg \
  --target-group-arns "arn:aws:elasticloadbalancing:region:account-id:targetgroup/my-
targets/1234567890123456"
```

Para anexar seu Classic Load Balancer

O create-auto-scaling-groupcomando a seguir cria um grupo de Auto Scaling com um Classic Load Balancer anexado.

```
aws autoscaling create-auto-scaling-group --auto-scaling-group-name my-asg \
  --launch-configuration-name my-launch-config \
  --vpc-zone-identifier "subnet-5ea0c127,subnet-6194ea3b,subnet-c934b782" \
  --load-balancer-names "my-load-balancer" \
  --min-size 1 --max-size 5
```

O attach-load-balancerscomando a seguir anexa o Classic Load Balancer especificado a um grupo de Auto Scaling existente.

```
aws autoscaling attach-load-balancers --auto-scaling-group-name my-asg \
  --load-balancer-names my-lb
```

Descrever seu grupo de destino ou Classic Load Balancer (legado)

Para descrever grupos de destino

Para descrever os grupos-alvo associados a um grupo do Auto Scaling, use o comando describe-load-balancer-target-groups. O exemplo a seguir lista os grupos de destino para *my-asg*.

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

Descrever Classic Load Balancers

Para descrever os Classic Load Balancers associados a um grupo de Auto Scaling, use o describe-load-balancerscomando. O exemplo a seguir lista os balanceadores de carga clássicos para *my-asg*.

```
aws autoscaling describe-load-balancers --auto-scaling-group-name my-asg
```

Desvincular seu grupo de destino ou Classic Load Balancer (legado)

Para desanexar um grupo de destino

O comando detach-load-balancer-target-groups a seguir separa um grupo-alvo do seu grupo de Auto Scaling quando você não precisar mais dele.

```
aws autoscaling detach-load-balancer-target-groups --auto-scaling-group-name my-asg \
  --target-group-arns "arn:aws:elasticloadbalancing:region:account-id:targetgroup/my-targets/1234567890123456"
```

Desvincular um Classic Load Balancer

O detach-load-balancerscomando a seguir separa um Classic Load Balancer do seu grupo de Auto Scaling quando você não precisar mais dele.

```
aws autoscaling detach-load-balancers --auto-scaling-group-name my-asg \
  --load-balancer-names my-lb
```

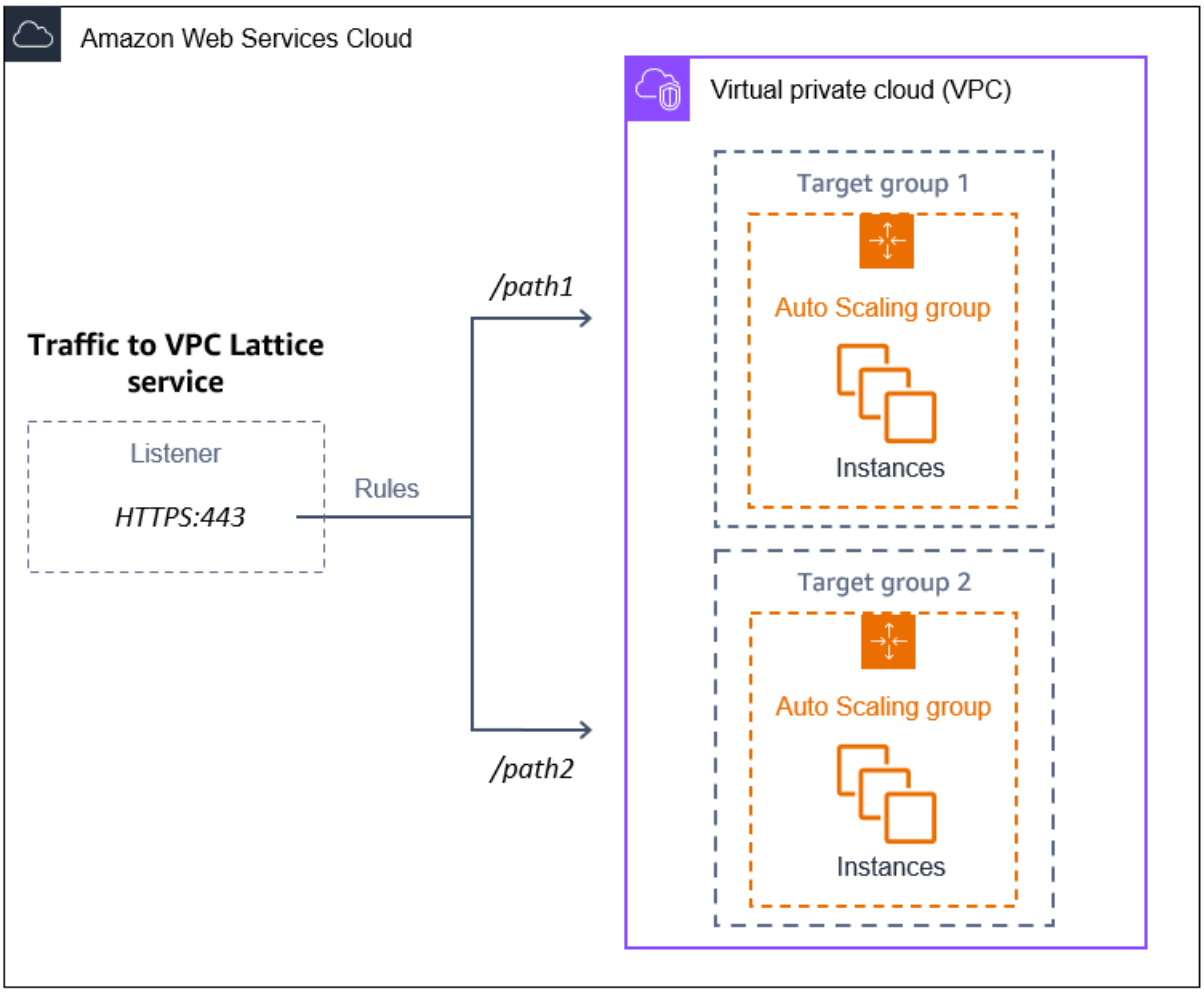
Rotear o tráfego para o grupo do Auto Scaling com um grupo de destino do VPC Lattice

Você pode usar o Amazon VPC Lattice para gerenciar o fluxo de tráfego e as chamadas de API entre seus aplicativos e serviços que são executados em recursos separados, como grupos do Auto Scaling ou funções do Lambda. VPC Lattice é um serviço de rede de aplicativos que permite conectar, proteger e monitorar todos os seus serviços em várias contas e nuvens privadas virtuais (VPCs). Para saber mais sobre o VPC Lattice, consulte o que é o VPC Lattice?

Para começar a usar o VPC Lattice, primeiro crie os recursos necessários do VPC Lattice que permitem que os recursos em uma VPC associada a uma rede de serviços se conectem entre si. Esses recursos incluem os serviços, os receptores, as regras de receptores e os grupos de destino.

Para associar um grupo do Auto Scaling a um serviço do VPC Lattice, crie um grupo de destino para o serviço que roteie as solicitações para instâncias registradas por ID de instância e adicione um receptor ao serviço que envie solicitações para o grupo de destino. Em seguida, anexe o grupo de destino ao grupo do Auto Scaling. O Amazon EC2 Auto Scaling registra automaticamente as instâncias do EC2 como destinos com o grupo de destino. Posteriormente, quando o Amazon EC2 Auto Scaling precisa encerrar uma instância, ele cancela automaticamente o registro da instância do grupo de destino antes do encerramento.

Depois de anexar o grupo-alvo, ele é o ponto de entrada para todas as solicitações recebidas no seu grupo do Auto Scaling. Como mostra o exemplo no diagrama a seguir, as solicitações recebidas podem então ser roteadas para o grupo-alvo apropriado usando regras do receptor especificadas para um serviço VPC Lattice.



Quando o tráfego é roteado pelo VPC Lattice para seu grupo do Auto Scaling, o VPC Lattice equilibra as solicitações entre as instâncias no grupo usando o balanceamento de carga ida e volta. O VPC Lattice também pode monitorar a integridade das suas instâncias registradas e rotear somente o tráfego para instâncias íntegras.

Para manter suas instâncias disponíveis para solicitações recebidas, você pode, opcionalmente, adicionar verificações de integridade do VPC Lattice ao seu grupo do Auto Scaling. Dessa forma, se uma das instâncias do EC2 falhar, o grupo do Auto Scaling iniciará automaticamente uma nova instância para substituí-la. O comportamento das verificações de integridade do VPC Lattice é semelhante ao comportamento das verificações de integridade do Elastic Load Balancing. As verificações de integridade padrão para um grupo do Auto Scaling são somente verificações de integridade do EC2.

Para saber mais sobre o VPC Lattice, consulte Simplifique a conectividade, a segurança e o monitoramento de serviço a serviço com o Amazon VPC Lattice — agora disponível ao público em geral no blog. AWS

Conteúdo

- Preparar para anexar um grupo de destino do VPC Lattice ao grupo do Auto Scaling
- Anexe um grupo de destino VPC Lattice ao seu grupo do Auto Scaling
- Verifique o status do anexo do grupo de destino do VPC Lattice

Preparar para anexar um grupo de destino do VPC Lattice ao grupo do Auto Scaling

Antes de anexar um grupo de destino do VPC Lattice ao grupo do Auto Scaling, será necessário concluir os pré-requisitos a seguir:

- Você já deve ter criado uma rede de serviços, um serviço, um receptor e um grupo-alvo do VPC Lattice. Para obter mais informações, consulte os tópicos a seguir no Guia do usuário do VPC Lattice.
  - Redes de serviços
  - Serviços
  - Listeners
  - Grupos de destino
- O grupo-alvo deve estar na mesma Conta da AWS VPC e região do seu grupo de Auto Scaling.
- O grupo de destino deve especificar um tipo de destino `instance`. Não é possível especificar um tipo de destino `ip` ao usar um grupo do Auto Scaling.
- É necessário ter permissões do IAM suficientes para anexar o grupo de destino ao grupo do Auto Scaling. O exemplo de política a seguir mostra as permissões mínimas necessárias para anexar e separar grupos de destino.

```
{
    "Version": "2012-10-17",
    "Statement": [
        {
            "Effect": "Allow",
            "Action": [
                "autoscaling:AttachTrafficSources",
```

```
                "autoscaling:DetachTrafficSources",
                "autoscaling:DescribeTrafficSources",
                "vpc-lattice:RegisterTargets",
                "vpc-lattice:DeregisterTargets"
            ],
            "Resource": "*"
        }
    ]
}
```

- Se o modelo de execução do seu grupo do Auto Scaling não contiver as configurações corretas para o VPC Lattice, como um grupo de segurança compatível, 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 as instâncias existentes, você pode iniciar uma atualização da instância para substituir as instâncias. Para ter mais informações, consulte Use uma atualização de instância para atualizar instâncias em um grupo de Auto Scaling.
- Antes de ativar as verificações de integridade do VPC Lattice no grupo do Auto Scaling, você pode configurar uma verificação de integridade baseada em aplicativos para verificar se o aplicativo está respondendo conforme o esperado. Para obter mais informações, consulte Verificações de integridade para seus grupos de destino no Guia do usuário do VPC Lattice.

Grupos de segurança: regras de entrada e saída

Os grupos de segurança atuam como firewall para instâncias EC2 associadas, controlando o tráfego de entrada e saída no nível da instância.

> **Note**
>
> A configuração da rede é suficientemente complexa e recomendamos fortemente que você crie um novo grupo de segurança para uso com o VPC Lattice. Também facilita a ajuda AWS Support se você precisar entrar em contato com eles. As seções a seguir se baseiam na suposição de que você segue essa recomendação.
>
> Para saber mais sobre a criação de grupos de segurança para o VPC Lattice que você pode usar com seu grupo do Auto Scaling, consulte Controle o tráfego usando grupos de segurança no Guia do usuário do VPC Lattice. Para solucionar problemas com o fluxo de tráfego, consulte o Guia do usuário do VPC Lattice para obter mais informações.

Para 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	Valor
Nome	Um nome fácil de lembrar.
Descrição	Uma descrição para ajudar você a identificar o grupo de segurança.
VPC	A mesma VPC do grupo do Auto Scaling.

Regras de entrada

Quando você cria um security group, ele não possui regras de entrada. Nenhum tráfego de entrada originário de clientes em uma rede de serviços do VPC Lattice para a instância será permitido até que você adicione regras de entrada ao grupo de segurança.

Para permitir que clientes em uma rede de serviços do VPC Lattice se conectem às instâncias no grupo do Auto Scaling, o grupo de segurança do grupo do Auto Scaling deve estar configurado corretamente. Nesse caso, defina uma regra de entrada para permitir o tráfego do nome da lista de prefixos AWS gerenciados do VPC Lattice, em vez de um endereço IP específico. A lista de prefixos do VPC Lattice é um intervalo de endereços IP usado pelo VPC Lattice na notação CIDR. Para obter mais informações, consulte Trabalhar com listas AWS de prefixos gerenciadas no Guia do usuário da Amazon VPC.

Para informações sobre como adicionar regras a um grupo de segurança, consulte Adicionar regras ao grupo de segurança no Guia do Usuário do Amazon VPC e use a tabela a seguir para determinar quais opções você deve selecionar.

Opção	Valor
Regra HTTP	Tipo: HTTP

Opção	Valor
	Fonte: com.amazonaws.*region*.vpc-lattice
Regra HTTPS	Tipo: HTTPS Fonte: com.amazonaws.*region*.vpc-lattice

O grupo de segurança é com estado: permite o tráfego de clientes dentro da rede de serviços do VPC Lattice para instâncias no grupo do Auto Scaling e envia a resposta de volta para o cliente que saiu anteriormente.

Regras de saída

Por padrão, um security group inclui uma regra de saída que permite todo o tráfego de saída. Opcionalmente, você pode remover essa regra padrão e adicionar uma regra de saída para acomodar necessidades específicas de segurança.

Limitações

- Não há suporte para grupos de instâncias mistas. Se você tentar vincular um grupo de destino do VPC Lattice a um grupo do Auto Scaling que tenha uma política de instâncias mistas, receberá a mensagem de erro Atualmente, os grupos do Auto Scaling com instâncias mistas não podem ser integrados a um serviço do VPC Lattice. Isso ocorre porque o algoritmo de balanceamento de carga distribui uniformemente a carga em todos os recursos disponíveis e pressupõe que as instâncias sejam semelhantes o suficiente para lidar com cargas iguais.

Anexe um grupo de destino VPC Lattice ao seu grupo do Auto Scaling

Este tópico descreve como anexar um grupo de destino do VPC Lattice a um grupo do Auto Scaling. Também descreve como ativar as verificações de integridade do VPC Lattice para permitir que o Amazon EC2 Auto Scaling substitua instâncias que o VPC Lattice relata como não íntegras.

Por padrão, o Amazon EC2 Auto Scaling substitui somente instâncias que não sejam íntegras ou inacessíveis com base nas verificações de integridade do Amazon EC2. Se você ativar as verificações de integridade do VPC Lattice, o Amazon EC2 Auto Scaling poderá substituir uma instância em execução se algum dos grupos de destino do VPC Lattice que você anexar ao grupo

do Auto Scaling relatar que ela não é íntegra. Para ter mais informações, consulte Verificações de integridade para instâncias em um grupo do Auto Scaling.

> **Important**
>
> Antes de continuar, preencha todos os pré-requisitos na seção anterior.

Anexar um grupo de destino do VPC Lattice

Você pode anexar um ou mais grupos-alvo a um grupo do Auto Scaling ao criar ou atualizar o grupo.

Console

Siga as etapas nesta seção para usar o console para:

- Anexar um grupo de destino do VPC Lattice ao grupo do Auto Scaling
- Ative as verificações de integridade do VPC Lattice

Para anexar um grupo de destino do VPC Lattice a um novo grupo do Auto Scaling

1. Abra o console do Amazon EC2 em https://console.aws.amazon.com/ec2/ e escolha Grupos do Auto Scaling no painel de navegação.
2. Na barra de navegação na parte superior da tela, escolha Região da AWS onde você criou seu grupo-alvo.
3. Selecione Criar grupo do Auto Scaling.
4. Nas etapas 1 e 2, escolha as opções conforme desejado e prossiga para a Etapa 3: Configurar opções avançadas.
5. Para opções de integração com o VPC Lattice, escolha Anexar ao serviço do VPC Lattice.
6. Em Escolher o grupo de destino do VPC Lattice, escolha seu grupo de destino.
7. (Opcional) Em Verificações de integridade, Tipos adicionais de verificação de integridade, selecione Ativar verificações de integridade do VPC Lattice.
8. (Opcional) Em Período de carência para verificação de integridade, insira o tempo, em segundos. Esse é o tempo que o Amazon EC2 Auto Scaling precisa aguardar antes de verificar o status de integridade de uma instância depois que ela entra no estado`InService`. Para ter mais informações, consulte Definir um período de carência da verificação de integridade para um grupo do Auto Scaling.

9. Prossiga para criar o grupo do Auto Scaling. Suas instâncias serão registradas automaticamente no grupo de destino VPC Lattice após a criação do grupo do Auto Scaling.

Para anexar um grupo de destino do VPC Lattice ao grupo do Auto Scaling existente

Use o procedimento a seguir para anexar um grupo de destino de um serviço a um grupo do Auto Scaling existente.

1. Abra o console do Amazon EC2 em https://console.aws.amazon.com/ec2/ e escolha 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. Sob as opções de integração com o VPC Lattice, escolha Anexar ao serviço do VPC Lattice.
5. Em Escolher o grupo de destino do VPC Lattice, escolha seu grupo de destino.
6. Escolha Atualizar.

Ao terminar de anexar o grupo de destino, você pode, opcionalmente, ativar as verificações de integridade que o usam.

Para ativar as verificações de integridade da VPC Lattice

1. Na guia Detalhes, escolha Verificações de integridade, Editar.
2. Em Verificações de integridade, Tipos adicionais de verificação de integridade, selecione Ativar verificações de integridade do VPC Lattice.
3. Em Período de carência da verificação de integridade, insira o tempo, em segundos. Esse é o tempo que o Amazon EC2 Auto Scaling precisa aguardar antes de verificar o status de integridade de uma instância depois que ela entra no estado`InService`. Para ter mais informações, consulte Definir um período de carência da verificação de integridade para um grupo do Auto Scaling.
4. Escolha Atualizar.

AWS CLI

Siga as etapas desta seção para usar o AWS CLI para:

- Anexar um grupo de destino do VPC Lattice ao grupo do Auto Scaling
- Ative as verificações de integridade do VPC Lattice

Anexar um grupo de destino do VPC Lattice ao grupo do Auto Scaling

Use o create-auto-scaling-groupcomando a seguir para criar um grupo de Auto Scaling e anexar simultaneamente um grupo-alvo do VPC Lattice especificando seu Amazon Resource Name (ARN).

Substitua os valores de amostra por,`--auto-scaling-group-name`,`--vpc-zone-identifier` `--min-size` e `--max-size`. Na opção,`--launch-template` substitua *`my-launch-template`* e *`1`* pelo nome e versão do modelo de execução que você criou para instâncias registradas em um grupo de destino do VPC Lattice. Na opção,`--traffic-sources` substitua o ARN de amostra pelo ARN do seu grupo de destino do VPC Lattice.

```
aws autoscaling create-auto-scaling-group --auto-scaling-group-name my-asg \
  --launch-template LaunchTemplateName=my-launch-template,Version='1' \
  --vpc-zone-identifier "subnet-5ea0c127,subnet-6194ea3b,subnet-c934b782" \
  --min-size 1 --max-size 5 \
  --traffic-sources "Identifier=arn:aws:vpc-lattice:region:account-id:targetgroup/
tg-0e2f2665eEXAMPLE"
```

Use o attach-traffic-sourcescomando a seguir para anexar um grupo-alvo do VPC Lattice a um grupo do Auto Scaling depois que ele já tiver sido criado.

```
aws autoscaling attach-traffic-sources --auto-scaling-group-name my-asg \
  --traffic-sources "Identifier=arn:aws:vpc-lattice:region:account-id:targetgroup/
tg-0e2f2665eEXAMPLE"
```

Para ativar as verificações de integridade do VPC Lattice

Se você configurou uma verificação de integridade baseada em aplicativos para seu grupo de destino do VPC Lattice, é possível ativar essas verificações de integridade. Use o update-auto-scaling-groupcomando create-auto-scaling-groupou com a `--health-check-type` opção e um valor de**`VPC_LATTICE`**. Para especificar o período de carência para as verificações de integridade realizadas pelo seu grupo do Auto Scaling, inclua a opção `--health-check-grace-period` e forneça seu valor em segundos.

```
--health-check-type "VPC_LATTICE" --health-check-grace-period 60
```

Desanexar um grupo de destino do VPC Lattice

Se o VPC Lattice não for mais necessário, use o procedimento a seguir para desanexar o grupo de destino do grupo do Auto Scaling.

Console

Siga as etapas nesta seção para usar o console para:

- Desanexar um grupo de destino do VPC Lattice de um grupo do Auto Scaling
- Desative as verificações de integridade do VPC Lattice

Para desanexar um grupo de destino do VPC Lattice de um grupo do Auto Scaling

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

Ao terminar de desanexar o grupo de destino, você pode desativar as verificações de integridade do VPC Lattice.

Para desativar as verificações de integridade do VPC Lattice

1. Na guia Detalhes, escolha Verificações de integridade, Editar.
2. Em Verificações de integridade, Tipos adicionais de verificação de integridade, desmarque Ativar verificações de integridade do VPC Lattice.
3. Escolha Atualizar.

AWS CLI

Siga as etapas desta seção para usar o AWS CLI para:

- Desanexar um grupo de destino do VPC Lattice de um grupo do Auto Scaling
- Desative as verificações de integridade do VPC Lattice

Use o detach-traffic-sourcescomando para separar um grupo-alvo do seu grupo de Auto Scaling quando você não precisar mais dele.

```
aws autoscaling detach-traffic-sources --auto-scaling-group-name my-asg \
  --traffic-sources "Identifier=arn:aws:vpc-lattice:region:account-id:targetgroup/
tg-0e2f2665eEXAMPLE"
```

Para atualizar as verificações de saúde em um grupo do Auto Scaling para que ele não use mais as verificações de saúde do VPC Lattice, use o comando. update-auto-scaling-group Inclua a opção `--health-check-type` e um valor de **`EC2`**.

```
aws autoscaling update-auto-scaling-group --auto-scaling-group-name my-asg \
  --health-check-type "EC2"
```

Verifique o status do anexo do grupo de destino do VPC Lattice

Após anexar um grupo de destino do VPC Lattice a um grupo do Auto Scaling, ele entra no estado `Adding` ao registrar as instâncias no grupo. Após todas as instâncias do grupo são registradas, ele entra no estado`Added`. Depois que pelo menos uma instância registrada passa nas verificações de integridade, ele entra no estado`InService`. Após o grupo de destino entrar no estado,`InService` o Amazon EC2 Auto Scaling pode encerrar e substituir todas as instâncias relatadas como não íntegras. Se nenhuma instância registrada passar nas verificações de integridade (por exemplo, devido a um erro na configuração da verificação de integridade), o grupo de destino não entrará no estado`InService`. O Amazon EC2 Auto Scaling não termina e substitui as instâncias.

Quando você desanexa um grupo de destino para um serviço, ele entra no estado `Removing` ao cancelar o registro das instâncias do grupo. As instâncias permanecem em execução após o cancelamento do registro. Por padrão, a drenagem da conexão (atraso de cancelamento de registro) é habilitada. Se a drenagem da conexão estiver habilitada, o VPC Lattice aguardará a conclusão das solicitações em andamento ou a expiração do tempo limite máximo (o que ocorrer primeiro) antes de cancelar o registro das instâncias.

Você pode verificar o status do anexo usando o AWS Command Line Interface (AWS CLI) ou AWS SDKs. Você não pode verificar o status do anexo no console.

Para usar o AWS CLI para verificar o status do anexo

O describe-traffic-sourcescomando a seguir retorna o status do anexo de todas as fontes de tráfego do grupo de Auto Scaling especificado.

```
aws autoscaling describe-traffic-sources --auto-scaling-group-name my-asg
```

O exemplo retorna o ARN do grupo-alvo do VPC Lattice que está anexado ao grupo do Auto Scaling, junto com o status do anexo do grupo de destino no elementoState.

```
{
    "TrafficSources": [
        {
            "Identifier": "arn:aws:vpc-lattice:region:account-
id:targetgroup/tg-0e2f2665eEXAMPLE",
            "State": "InService",
            "Type": "vpc-lattice"
        }
    ]
}
```

Use EventBridge para lidar com eventos do Auto Scaling

A Amazon EventBridge, anteriormente chamada de CloudWatch Eventos, ajuda você a configurar regras orientadas por eventos que monitoram recursos e iniciam ações direcionadas que usam outros serviços. AWS

Os eventos do Amazon EC2 Auto Scaling são entregues quase em EventBridge tempo real. Você pode estabelecer EventBridge regras que invocam ações programáticas e notificações em resposta a uma variedade desses eventos. Por exemplo, enquanto as instâncias estão em processo de inicialização ou encerramento, você pode invocar uma AWS Lambda função para realizar uma tarefa pré-configurada.

Os alvos das EventBridge regras podem incluir AWS Lambda funções, tópicos do Amazon SNS, destinos de API, barramentos de eventos e muito mais. Contas da AWS Para obter informações sobre as metas suportadas, consulte as EventBridge metas da Amazon no Guia EventBridge do usuário da Amazon.

Comece criando EventBridge regras com um exemplo usando um tópico e uma EventBridge regra do Amazon SNS. Em seguida, quando um usuário iniciar uma atualização de instância, o Amazon

SNS notificará você por e-mail sempre que um ponto de verificação for alcançado. Para ter mais informações, consulte Crie EventBridge regras, por exemplo, eventos de atualização.

Conteúdo

- Referência de eventos do Amazon EC2 Auto Scaling
- Exemplos de eventos e padrões de grupo de aquecimento
- Crie EventBridge regras

Referência de eventos do Amazon EC2 Auto Scaling

Usando a Amazon EventBridge, você pode criar regras que correspondam aos eventos recebidos e encaminhá-los aos alvos para processamento.

Conteúdo

- Eventos de ação do ciclo de vida
- Eventos de escalabilidade bem-sucedidos
- Eventos de escalabilidade sem êxito
- Eventos de atualização de instância

Eventos de ação do ciclo de vida

Quando você adiciona ganchos de ciclo de vida ao seu grupo de Auto Scaling, o Amazon EC2 Auto Scaling envia eventos EventBridge para quando uma instância passa para um estado de espera. Os eventos são emitidos com base no melhor esforço.

Tipos de eventos

- Expandir ação do ciclo de vida
- Reduzir ação do ciclo de vida

Expandir ação do ciclo de vida

O evento de exemplo a seguir mostra que o Amazon EC2 Auto Scaling moveu uma instância para um `Pending:Wait` estado devido a um hook do ciclo de vida de inicialização.

```
{
  "version": "0",
```

```
  "id": "12345678-1234-1234-1234-123456789012",
  "detail-type": "EC2 Instance-launch Lifecycle Action",
  "source": "aws.autoscaling",
  "account": "123456789012",
  "time": "yyyy-mm-ddThh:mm:ssZ",
  "region": "us-west-2",
  "resources": [
    "auto-scaling-group-arn"
  ],
  "detail": {
    "LifecycleActionToken": "87654321-4321-4321-4321-210987654321",
    "AutoScalingGroupName": "my-asg",
    "LifecycleHookName": "my-lifecycle-hook",
    "EC2InstanceId": "i-1234567890abcdef0",
    "LifecycleTransition": "autoscaling:EC2_INSTANCE_LAUNCHING",
    "NotificationMetadata": "additional-info",
    "Origin": "EC2",
    "Destination": "AutoScalingGroup"
  }
}
```

Reduzir ação do ciclo de vida

O evento de exemplo a seguir mostra que o Amazon EC2 Auto Scaling moveu uma instância para um `Terminating:Wait` estado devido a um hook de encerramento do ciclo de vida.

> **Important**
>
> Quando um grupo do Auto Scaling retorna instâncias para um grupo de aquecimento, o retorno de instâncias para o grupo de aquecimento também pode gerar eventos`EC2 Instance-terminate Lifecycle Action`. Eventos que são entregues quando uma instância passa para o estado de espera na redução têm `WarmPool` como valor para`Destination`. Para ter mais informações, consulte Instance reuse policy.

```
{
  "version": "0",
  "id": "12345678-1234-1234-1234-123456789012",
  "detail-type": "EC2 Instance-terminate Lifecycle Action",
  "source": "aws.autoscaling",
  "account": "123456789012",
```

```
  "time": "yyyy-mm-ddThh:mm:ssZ",
  "region": "us-west-2",
  "resources": [
    "auto-scaling-group-arn"
  ],
  "detail": {
    "LifecycleActionToken":"87654321-4321-4321-4321-210987654321",
    "AutoScalingGroupName":"my-asg",
    "LifecycleHookName":"my-lifecycle-hook",
    "EC2InstanceId":"i-1234567890abcdef0",
    "LifecycleTransition":"autoscaling:EC2_INSTANCE_TERMINATING",
    "NotificationMetadata":"additional-info",
    "Origin": "AutoScalingGroup",
    "Destination": "EC2"
  }
}
```

Eventos de escalabilidade bem-sucedidos

Os exemplos a seguir mostram os tipos de eventos para eventos de escalabilidade bem-sucedidos. Os eventos são emitidos com base no melhor esforço.

Tipos de eventos

- Evento de expansão bem-sucedido
- Evento de redução bem-sucedido

Evento de expansão bem-sucedido

O evento de exemplo a seguir mostra que o Amazon EC2 Auto Scaling executou uma instância com êxito.

```
{
  "version": "0",
  "id": "12345678-1234-1234-1234-123456789012",
  "detail-type": "EC2 Instance Launch Successful",
  "source": "aws.autoscaling",
  "account": "123456789012",
  "time": "yyyy-mm-ddThh:mm:ssZ",
  "region": "us-west-2",
  "resources": [
    "auto-scaling-group-arn",
```

```
    "instance-arn"
  ],
  "detail": {
    "StatusCode": "InProgress",
    "Description": "Launching a new EC2 instance: i-12345678",
    "AutoScalingGroupName": "my-asg",
    "ActivityId": "87654321-4321-4321-4321-210987654321",
    "Details": {
      "Availability Zone": "us-west-2b",
      "Subnet ID": "subnet-12345678"
    },
    "RequestId": "12345678-1234-1234-1234-123456789012",
    "StatusMessage": "",
    "EndTime": "yyyy-mm-ddThh:mm:ssZ",
    "EC2InstanceId": "i-1234567890abcdef0",
    "StartTime": "yyyy-mm-ddThh:mm:ssZ",
    "Cause": "description-text",
    "Origin": "EC2",
    "Destination": "AutoScalingGroup"
  }
}
```

Evento de redução bem-sucedido

O evento de exemplo a seguir mostra que o Amazon EC2 Auto Scaling encerrou uma instância com êxito.

> **Important**
>
> Quando um grupo do Auto Scaling retorna instâncias para um grupo de aquecimento, o retorno de instâncias para o grupo de aquecimento também pode gerar eventos`EC2 Instance Terminate Successful`. Os eventos que são entregues quando uma instância retorna com sucesso ao grupo de aquecimento têm `WarmPool` como valor para `Destination`. Para ter mais informações, consulte Instance reuse policy.

```
{
  "version": "0",
  "id": "12345678-1234-1234-1234-123456789012",
  "detail-type": "EC2 Instance Terminate Successful",
  "source": "aws.autoscaling",
  "account": "123456789012",
```

```
  "time": "yyyy-mm-ddThh:mm:ssZ",
  "region": "us-west-2",
  "resources": [
    "auto-scaling-group-arn",
    "instance-arn"
  ],
  "detail": {
    "StatusCode": "InProgress",
    "Description": "Terminating EC2 instance: i-12345678",
    "AutoScalingGroupName": "my-asg",
    "ActivityId": "87654321-4321-4321-4321-210987654321",
    "Details": {
      "Availability Zone": "us-west-2b",
      "Subnet ID": "subnet-12345678"
    },
    "RequestId": "12345678-1234-1234-1234-123456789012",
    "StatusMessage": "",
    "EndTime": "yyyy-mm-ddThh:mm:ssZ",
    "EC2InstanceId": "i-1234567890abcdef0",
    "StartTime": "yyyy-mm-ddThh:mm:ssZ",
    "Cause": "description-text",
    "Origin": "AutoScalingGroup",
    "Destination": "EC2"
  }
}
```

Eventos de escalabilidade sem êxito

Os exemplos a seguir mostram os tipos de eventos para eventos de escalabilidade malsucedidos. Os eventos são emitidos com base no melhor esforço.

Tipos de eventos

- Evento de expansão sem êxito
- Evento de redução sem êxito

Evento de expansão sem êxito

O evento de exemplo a seguir mostra que o Amazon EC2 Auto Scaling falhou ao executar uma instância.

```
{
```

```
  "version": "0",
  "id": "12345678-1234-1234-1234-123456789012",
  "detail-type": "EC2 Instance Launch Unsuccessful",
  "source": "aws.autoscaling",
  "account": "123456789012",
  "time": "yyyy-mm-ddThh:mm:ssZ",
  "region": "us-west-2",
  "resources": [
    "auto-scaling-group-arn",
    "instance-arn"
  ],
  "detail": {
    "StatusCode": "Failed",
    "AutoScalingGroupName": "my-asg",
    "ActivityId": "87654321-4321-4321-4321-210987654321",
    "Details": {
      "Availability Zone": "us-west-2b",
      "Subnet ID": "subnet-12345678"
    },
    "RequestId": "12345678-1234-1234-1234-123456789012",
    "StatusMessage": "message-text",
    "EndTime": "yyyy-mm-ddThh:mm:ssZ",
    "EC2InstanceId": "i-1234567890abcdef0",
    "StartTime": "yyyy-mm-ddThh:mm:ssZ",
    "Cause": "description-text",
    "Origin": "EC2",
    "Destination": "AutoScalingGroup"
  }
}
```

Evento de redução sem êxito

O evento de exemplo a seguir mostra que o Amazon EC2 Auto Scaling falhou ao encerrar uma instância.

> **Important**
>
> Quando um grupo do Auto Scaling retorna instâncias para um grupo de aquecimento em redução, deixar de retornar as instâncias ao grupo de aquecimento também pode gerar eventos. `EC2 Instance Terminate Unsuccessful` Os eventos que são entregues quando uma instância falha ao retornar ao pool quente têm `WarmPool` como valor`Destination`. Para obter mais informações, consulteInstance reuse policy.

```
{
  "version": "0",
  "id": "12345678-1234-1234-1234-123456789012",
  "detail-type": "EC2 Instance Terminate Unsuccessful",
  "source": "aws.autoscaling",
  "account": "123456789012",
  "time": "yyyy-mm-ddThh:mm:ssZ",
  "region": "us-west-2",
  "resources": [
    "auto-scaling-group-arn",
    "instance-arn"
  ],
  "detail": {
    "StatusCode": "Failed",
    "AutoScalingGroupName": "my-asg",
    "ActivityId": "87654321-4321-4321-4321-210987654321",
    "Details": {
      "Availability Zone": "us-west-2b",
      "Subnet ID": "subnet-12345678"
    },
    "RequestId": "12345678-1234-1234-1234-123456789012",
    "StatusMessage": "message-text",
    "EndTime": "yyyy-mm-ddThh:mm:ssZ",
    "EC2InstanceId": "i-1234567890abcdef0",
    "StartTime": "yyyy-mm-ddThh:mm:ssZ",
    "Cause": "description-text",
    "Origin": "AutoScalingGroup",
    "Destination": "EC2"
  }
}
```

Eventos de atualização de instância

Os seguintes exemplos mostram eventos do recurso de atualização de instância. Os eventos são emitidos com base no melhor esforço.

Tipos de eventos

- Ponto de verificação alcançado
- Atualização de instância iniciada
- Atualização de instância bem-sucedida
- Falha na atualização de instância

- Atualização de instância cancelada

Ponto de verificação alcançado

Quando o número de instâncias substituídas atinge o limite percentual definido para o ponto de verificação, o Amazon EC2 Auto Scaling envia o evento a seguir.

```
{
  "version": "0",
  "id": "12345678-1234-1234-1234-123456789012",
  "detail-type": "EC2 Auto Scaling Instance Refresh Checkpoint Reached",
  "source": "aws.autoscaling",
  "account": "123456789012",
  "time": "yyyy-mm-ddThh:mm:ssZ",
  "region": "us-west-2",
  "resources": [
    "auto-scaling-group-arn"
  ],
  "detail": {
    "InstanceRefreshId": "ab00cf8f-9126-4f3c-8010-dbb8cad6fb86",
    "AutoScalingGroupName": "my-asg",
    "CheckpointPercentage": "50",
    "CheckpointDelay": "300"
  }
}
```

Atualização de instância iniciada

Quando o status de uma atualização de instância muda para, o Amazon EC2 Auto Scaling envia o seguinte eventoInProgress.

```
{
  "version": "0",
  "id": "12345678-1234-1234-1234-123456789012",
  "detail-type": "EC2 Auto Scaling Instance Refresh Started",
  "source": "aws.autoscaling",
  "account": "123456789012",
  "time": "yyyy-mm-ddThh:mm:ssZ",
  "region": "us-west-2",
  "resources": [
    "auto-scaling-group-arn"
  ],
```

```
  "detail": {
    "InstanceRefreshId": "c613620e-07e2-4ed2-a9e2-ef8258911ade",
    "AutoScalingGroupName": "my-asg"
  }
}
```

Atualização de instância bem-sucedida

Quando o status de uma atualização de instância muda para, o Amazon EC2 Auto Scaling envia o seguinte eventoSucceeded.

```
{
  "version": "0",
  "id": "12345678-1234-1234-1234-123456789012",
  "detail-type": "EC2 Auto Scaling Instance Refresh Succeeded",
  "source": "aws.autoscaling",
  "account": "123456789012",
  "time": "yyyy-mm-ddThh:mm:ssZ",
  "region": "us-west-2",
  "resources": [
    "auto-scaling-group-arn"
  ],
  "detail": {
    "InstanceRefreshId": "c613620e-07e2-4ed2-a9e2-ef8258911ade",
    "AutoScalingGroupName": "my-asg"
  }
}
```

Falha na atualização de instância

Quando o status de uma atualização de instância muda para, o Amazon EC2 Auto Scaling envia o seguinte eventoFailed.

```
{
  "version": "0",
  "id": "12345678-1234-1234-1234-123456789012",
  "detail-type": "EC2 Auto Scaling Instance Refresh Failed",
  "source": "aws.autoscaling",
  "account": "123456789012",
  "time": "yyyy-mm-ddThh:mm:ssZ",
  "region": "us-west-2",
  "resources": [
```

```
    "auto-scaling-group-arn"
  ],
  "detail": {
    "InstanceRefreshId": "c613620e-07e2-4ed2-a9e2-ef8258911ade",
    "AutoScalingGroupName": "my-asg"
  }
}
```

Atualização de instância cancelada

Quando o status de uma atualização de instância muda para, o Amazon EC2 Auto Scaling envia o seguinte eventoCancelled.

```
{
  "version": "0",
  "id": "12345678-1234-1234-1234-123456789012",
  "detail-type": "EC2 Auto Scaling Instance Refresh Cancelled",
  "source": "aws.autoscaling",
  "account": "123456789012",
  "time": "yyyy-mm-ddThh:mm:ssZ",
  "region": "us-west-2",
  "resources": [
    "auto-scaling-group-arn"
  ],
  "detail": {
    "InstanceRefreshId": "c613620e-07e2-4ed2-a9e2-ef8258911ade",
    "AutoScalingGroupName": "my-asg"
  }
}
```

Exemplos de eventos e padrões de grupo de aquecimento

O Amazon EC2 Auto Scaling suporta vários padrões predefinidos na Amazon. EventBridge Isso simplifica a forma como um padrão de evento é criado. Você seleciona valores de campo em um formulário e EventBridge gera o padrão para você. No momento, o Amazon EC2 Auto Scaling não oferece suporte a padrões predefinidos para eventos emitidos por um grupo do Auto Scaling com um grupo de alta atividade. Você deve inserir o padrão como um objeto JSON. Esta seção e o tópico Crie EventBridge regras para eventos de piscina aquecida mostram como usar um padrão de evento para selecionar eventos e enviá-los para destinos.

Para criar EventBridge regras que filtrem os eventos relacionados ao pool aquecido para os quais o Amazon EC2 Auto Scaling EventBridge envia, `Origin` inclua os campos e `Destination` da seção `detail` do evento.

Os valores de `Origin` e `Destination` podem ser:

`EC2` | `AutoScalingGroup` | `WarmPool`

Conteúdo

- Eventos de exemplo
- Exemplo de padrões de eventos

Eventos de exemplo

Quando você adiciona ganchos de ciclo de vida ao seu grupo de Auto Scaling, o Amazon EC2 Auto Scaling envia eventos EventBridge para quando uma instância passa para um estado de espera. Para ter mais informações, consulte Usar ganchos do ciclo de vida com um grupo de alta atividade.

Esta seção inclui exemplos desses eventos quando seu grupo do Auto Scaling tem um grupo de aquecimento. Os eventos são emitidos com base no melhor esforço.

> **Note**
>
> Para eventos para os quais o Amazon EC2 Auto Scaling envia quando EventBridge a escalabilidade é bem-sucedida, consulte. Eventos de escalabilidade bem-sucedidos Para eventos em que a escalabilidade não é bem-sucedida, consulteEventos de escalabilidade sem êxito.

Exemplos de evento

- Expandir ação do ciclo de vida
- Reduzir ação do ciclo de vida

Expandir ação do ciclo de vida

Os eventos que são entregues quando uma instância faz a transição para um estado de espera por eventos de expansão têm `EC2 Instance-launch Lifecycle Action` como valor`detail-`

`type`. No objeto`detail`, os valores dos atributos `Origin` e `Destination` mostram de onde a instância vem e para onde está indo.

Neste exemplo de evento de expansão, uma nova instância é iniciada e seu estado muda para `Warmed:Pending:Wait` porque ela foi adicionada ao grupo de aquecimento. Para obter mais informações, consulte Transições de estado do ciclo de vida para instâncias em um grupo de alta atividade.

```
{
  "version": "0",
  "id": "12345678-1234-1234-1234-123456789012",
  "detail-type": "EC2 Instance-launch Lifecycle Action",
  "source": "aws.autoscaling",
  "account": "123456789012",
  "time": "2021-01-13T00:12:37.214Z",
  "region": "us-west-2",
  "resources": [
    "auto-scaling-group-arn"
  ],
  "detail": {
    "LifecycleActionToken": "71514b9d-6a40-4b26-8523-05e7eEXAMPLE",
    "AutoScalingGroupName": "my-asg",
    "LifecycleHookName": "my-launch-lifecycle-hook",
    "EC2InstanceId": "i-1234567890abcdef0",
    "LifecycleTransition": "autoscaling:EC2_INSTANCE_LAUNCHING",
    "NotificationMetadata": "additional-info",
    "Origin": "EC2",
    "Destination": "WarmPool"
  }
}
```

Neste exemplo de evento de expansão, o estado da instância muda para `Pending:Wait` porque ela foi adicionada ao grupo do Auto Scaling a partir do pool quente. Para ter mais informações, consulte Transições de estado do ciclo de vida para instâncias em um grupo de alta atividade.

```
{
  "version": "0",
  "id": "12345678-1234-1234-1234-123456789012",
  "detail-type": "EC2 Instance-launch Lifecycle Action",
  "source": "aws.autoscaling",
  "account": "123456789012",
  "time": "2021-01-19T00:35:52.359Z",
```

```
  "region": "us-west-2",
  "resources": [
    "auto-scaling-group-arn"
  ],
  "detail": {
    "LifecycleActionToken": "19cc4d4a-e450-4d1c-b448-0de67EXAMPLE",
    "AutoScalingGroupName": "my-asg",
    "LifecycleHookName": "my-launch-lifecycle-hook",
    "EC2InstanceId": "i-1234567890abcdef0",
    "LifecycleTransition": "autoscaling:EC2_INSTANCE_LAUNCHING",
    "NotificationMetadata": "additional-info",
    "Origin": "WarmPool",
    "Destination": "AutoScalingGroup"
  }
}
```

Reduzir ação do ciclo de vida

Os eventos que são entregues quando uma instância faz a transição para um estado de espera em eventos de redução têm `EC2 Instance-terminate Lifecycle Action` como valor para `detail-type`. No objeto`detail`, os valores dos atributos `Origin` e `Destination` mostram de onde a instância vem e para onde está indo.

Neste evento de exemplo, o estado de uma instância muda para `Warmed:Pending:Wait` quando ela é retornada ao grupo de alta atividade. Para ter mais informações, consulte Transições de estado do ciclo de vida para instâncias em um grupo de alta atividade.

```
{
  "version": "0",
  "id": "12345678-1234-1234-1234-123456789012",
  "detail-type": "EC2 Instance-terminate Lifecycle Action",
  "source": "aws.autoscaling",
  "account": "123456789012",
  "time": "2022-03-28T00:12:37.214Z",
  "region": "us-west-2",
  "resources": [
    "auto-scaling-group-arn"
  ],
  "detail": {
    "LifecycleActionToken": "42694b3d-4b70-6a62-8523-09a1eEXAMPLE",
    "AutoScalingGroupName": "my-asg",
    "LifecycleHookName": "my-termination-lifecycle-hook",
    "EC2InstanceId": "i-1234567890abcdef0",
```

```
    "LifecycleTransition": "autoscaling:EC2_INSTANCE_TERMINATING",
    "NotificationMetadata": "additional-info",
    "Origin": "AutoScalingGroup",
    "Destination": "WarmPool"
  }
}
```

Exemplo de padrões de eventos

A seção anterior fornece exemplos de eventos emitidos pelo Amazon EC2 Auto Scaling.

EventBridge os padrões de eventos têm a mesma estrutura dos eventos aos quais eles correspondem. O padrão menciona os campos com os quais você deseja fazer a correspondência e fornece os valores que você está procurando.

Os seguintes campos no evento formam o padrão de evento definido na regra para invocar uma ação:

`"source": "aws.autoscaling"`

Identifica que o evento é do Amazon EC2 Auto Scaling.

`"detail-type": "EC2 Instance-launch Lifecycle Action"`

Identifica o tipo de evento.

`"Origin": "EC2"`

Identifica a origem da instância.

`"Destination": "WarmPool"`

Identifica o destino da instância.

Use o padrão de evento de exemplo a seguir para capturar todos os `EC2 Instance-launch Lifecycle Action` eventos associados a instâncias que entram no 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 expansão.

```
{
  "source": [ "aws.autoscaling" ],
  "detail-type": [ "EC2 Instance-launch Lifecycle Action" ],
  "detail": {
      "Origin": [ "WarmPool" ],
      "Destination": [ "AutoScalingGroup" ]
   }
}
```

Use o padrão de evento de exemplo a seguir para capturar todos os `EC2 Instance-launch Lifecycle Action` eventos associados a instâncias que são iniciadas diretamente no grupo do Auto Scaling.

```
{
  "source": [ "aws.autoscaling" ],
  "detail-type": [ "EC2 Instance-launch Lifecycle Action" ],
  "detail": {
      "Origin": [ "EC2" ],
      "Destination": [ "AutoScalingGroup" ]
   }
}
```

Use o padrão de evento de exemplo a seguir para capturar todos os `EC2 Instance-terminate Lifecycle Action` eventos associados a instâncias que retornam ao grupo de alta atividade ao reduzir a escala vertical.

```
{
  "source": [ "aws.autoscaling" ],
  "detail-type": [ "EC2 Instance-terminate Lifecycle Action" ],
  "detail": {
      "Origin": [ "AutoScalingGroup" ],
      "Destination": [ "WarmPool" ]
   }
}
```

Use o exemplo de padrão de evento a seguir para capturar todos os eventos associados a `EC2 Instance-launch Lifecycle Action`, independentemente da origem ou do destino.

```
{
  "source": [ "aws.autoscaling" ],
  "detail-type": [ "EC2 Instance-launch Lifecycle Action" ]
}
```

Crie EventBridge regras

Quando um evento é emitido pelo Amazon EC2 Auto Scaling, uma notificação de evento é enviada para a EventBridge Amazon como um arquivo JSON. Você pode escrever uma EventBridge regra para automatizar as ações a serem tomadas quando um padrão de evento corresponder à regra. Se EventBridge detectar um padrão de evento que corresponda a um padrão definido em uma regra, EventBridge invoca o alvo (ou alvos) especificado na regra.

Você pode usar os procedimentos de exemplo nesta seção como ponto de partida.

Talvez a documentação a seguir também seja útil.

- Para executar ações personalizadas em instâncias conforme elas estão sendo iniciadas ou antes que sejam encerradas usando uma função do Lambda, consulte Tutorial: Configurar um gancho do ciclo de vida que invoca uma função do Lambda.
- Para invocar uma função Lambda em chamadas de API registradas CloudTrail, consulte Tutorial: AWS Registrar chamadas EventBridge de API usando o Amazon EventBridge User Guide.
- Para obter mais informações sobre como criar regras de eventos, consulte Criação de EventBridge regras da Amazon que reagem a eventos no Guia EventBridge do usuário da Amazon.

Tópicos

- Crie EventBridge regras, por exemplo, eventos de atualização
- Crie EventBridge regras para eventos de piscina aquecida

Crie EventBridge regras, por exemplo, eventos de atualização

O exemplo a seguir cria uma EventBridge regra para enviar uma notificação por e-mail. Ele faz isso sempre que seu grupo do Auto Scaling emite um evento quando um ponto de verificação é atingido durante uma atualização de instância. O procedimento para configurar notificações por e-

mail usando o Amazon SNS está incluído. Para usar o Amazon SNS para enviar notificações por e-mail, você deve primeiro criar um tópico e, em seguida, assinar seus endereços de e-mail para o tópico.

Para obter mais informações sobre o recurso de atualização de instância, consulte Use uma atualização de instância para atualizar instâncias em um grupo de Auto Scaling.

Crie um tópico do Amazon SNS

Um tópico do SNS é um ponto de acesso lógico, um canal de comunicação que seu grupo do Auto Scaling usa para enviar notificações. Você cria um tópico especificando um nome para o tópico.

Os nomes de tópico devem atender aos seguintes requisitos:

- Ter de 1 a 256 caracteres
- Conter letras maiúsculas e minúsculas ASCIIs, números, sublinhados ou hífens

Para obter mais informações, consulte Criação de um tópico do Amazon SNS no Guia do desenvolvedor do Amazon Simple Notification Service.

Assinar o tópico do Amazon SNS

Para receber as notificações que seu grupo do Auto Scaling envia ao tópico, você deve assinar um endpoint para o tópico. Neste procedimento, em Endpoint, especifique o endereço de e-mail no qual você deseja receber as notificações do Amazon EC2 Auto Scaling.

Para obter instruções, consulte Assinatura de um tópico do Amazon SNS no Guia do desenvolvedor do Amazon Simple Notification Service.

Confirmar sua assinatura do Amazon SNS

O Amazon SNS envia um e-mail de confirmação para o endereço de e-mail especificado na etapa anterior.

Certifique-se de abrir o e-mail em AWS Notificações e escolher o link para confirmar a assinatura antes de continuar com a próxima etapa.

Você receberá uma mensagem de confirmação de. AWS O Amazon SNS agora está configurado para receber notificações e enviar a notificação como um e-mail para o endereço de e-mail que você especificou.

Encaminhar eventos para seu tópico do Amazon SNS

Crie uma regra que corresponda aos eventos selecionados e os encaminhe para o tópico do Amazon SNS para notificar os endereços de e-mail inscritos.

Para criar uma regra que envie notificações para o tópico do Amazon SNS

1. Abra o EventBridge console da Amazon em https://console.aws.amazon.com/events/.
2. No painel de navegação, escolha Regras.
3. Selecione 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 serviço na sua conta gera um evento, ele sempre vai para o ônibus de eventos padrão da sua conta.
   c. Em Rule type (Tipo de regra), selecione Rule with an event pattern (Regra com um padrão de evento).
   d. Selecione Next (Próximo).
5. Em Build event pattern (Criar padrão de evento), faça o seguinte:
   a. Em Origem do evento, escolha AWS eventos ou eventos de EventBridge parceiros.
   b. Em Event pattern (Padrão de evento), faça o seguinte:
      i. Em Event source, escolha Serviços da AWS.
      ii. Em AWS service (Serviço da AWS), escolha Auto Scaling.
      iii. Em Event type (Tipo de evento), escolha Instance Refresh (Atualização de instância).
      iv. Por padrão, a regra corresponde a qualquer instância na região. Para criar uma regra que notifique você quando um ponto de verificação for atingido durante uma atualização de instância, escolha Specific instance event(s) (Eventos específicos de instância) e selecione EC2 Auto Scaling Instance Refresh Checkpoint Reached (Ponto de verificação de atualização de instância do EC2 Auto Scaling atingido).
      v. Por padrão, a regra corresponde a qualquer grupo do Auto Scaling na região. Para fazer com que a regra corresponda a um grupo do Auto Scaling específico, escolha

Specific group name(s) (Nomes de grupos específicos) e selecione um ou mais grupos do Auto Scaling.

vi. Escolha Próximo.

6. Em Select target(s) (Selecionar destino(s)), faça o seguinte:

   a. Em Target types (Tipos de destino), escolha AWS service (Serviço da AWS).

   b. Em Select a target (Selecionar um destino), escolha SNS topic (Tópico do SNS).

   c. Em Topic (Tópico), escolha o tópico do Amazon SNS.

   d. (Opcional) Em Additional settings (Configurações adicionais), é possível, opcionalmente, definir configurações adicionais. Para obter mais informações, consulte Criação de EventBridge regras da Amazon que reagem a eventos no Guia EventBridge do usuário da Amazon.

   e. Escolha Próximo.

7. (Opcional) Em Tags (Etiquetas), é possível atribuir, opcionalmente, uma ou mais etiquetas à sua regra e, em seguida, escolher Next (Próximo).

8. Em Review and create (Revisar e criar), revise os detalhes da regra e modifique-os conforme necessário. Em seguida, escolha Create rule (Criar regra).

Crie EventBridge regras para eventos de piscina aquecida

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 AWS Lambda função que você deseja que a regra use como destino. Você deve especificar essa função como o destino da regra. O procedimento a seguir fornece somente as etapas para criar a EventBridge regra que atua quando novas instâncias entram no pool aquecido. Para obter um tutorial introdutório que mostra como criar uma função simples do Lambda a ser invocada quando um evento recebido corresponder a uma regra, consulte Tutorial: Configurar um gancho do ciclo de vida que invoca uma função do Lambda.

Para obter mais informações sobre como criar e trabalhar com grupos de alta atividade, consulte Grupos de alta atividade do Amazon EC2 Auto Scaling.

Para criar uma regra de evento para invocar uma função do Lambda

1. Abra o EventBridge console da Amazon em https://console.aws.amazon.com/events/.

2. No painel de navegação, escolha Regras.
3. Selecione 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 gera um evento, ele sempre vai para o ônibus de eventos padrão da sua conta.
   c. Em Rule type (Tipo de regra), selecione Rule with an event pattern (Regra com um padrão de evento).
   d. Selecione Next (Próximo).
5. Em Build event pattern (Criar padrão de evento), faça o seguinte:
   a. Em Origem do evento, escolha AWS eventos ou eventos de EventBridge parceiros.
   b. Em Event pattern (Padrão de evento), 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 ter mais informações, consulte Exemplo de padrões de eventos.
   c. Escolha Próximo.
6. Em Select target(s) (Selecionar destino(s)), faça o seguinte:
   a. Em Target types (Tipos de destino), escolha AWS service (Serviço da AWS).

b. Em Select a target (Selecionar um destino), escolha Lambda function (Função do Lambda).

c. Em Function (Função), escolha a função para a qual deseja enviar os eventos.

d. (Opcional) Em Configure version/alias (Configurar versão/alias), insira as configurações de versão e alias para a função do Lambda de destino.

e. (Opcional) Em Additional settings (Configurações adicionais), insira qualquer configuração adicional conforme adequado para seu aplicativo. Para obter mais informações, consulte Criação de EventBridge regras da Amazon que reagem a eventos no Guia EventBridge do usuário da Amazon.

f. Escolha Próximo.

7. (Opcional) Em Tags (Etiquetas), é possível atribuir, opcionalmente, uma ou mais etiquetas à sua regra e, em seguida, escolher Next (Próximo).

8. Em Review and create (Revisar e criar), revise os detalhes da regra e modifique-os conforme necessário. Em seguida, escolha Create rule (Criar regra).

Fornecer conectividade de rede para suas instâncias do Auto Scaling usando a Amazon VPC

A Amazon Virtual Private Cloud (Amazon VPC) é um serviço que permite lançar AWS recursos como grupos de Auto Scaling em uma rede virtual logicamente isolada que você define.

Uma sub-rede na Amazon VPC é uma subdivisão em uma zona de disponibilidade definida por um segmento de intervalo de endereços IP da VPC. Usando sub-redes, você pode agrupar suas instâncias com base em suas necessidades operacionais e de segurança. Uma sub-rede reside totalmente dentro da Zona de disponibilidade em que foi criada. Você ativa instâncias do Auto Scaling dentro das sub-redes.

Para habilitar a comunicação entre a internet e as instâncias em suas sub-redes, 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 Guia do usuário da Amazon VPC.

Conteúdo

- VPC padrão
- VPC não padrão
- Considerações sobre a escolha de sub-redes da VPC
- Endereçamento IP em uma VPC
- Interfaces de rede em uma VPC
- Locação de localização de instância
- AWS Outposts
- Mais recursos para saber mais sobre VPCs

VPC padrão

Se você criou o seu Conta da AWS depois de 4 de dezembro de 2013 ou está criando seu grupo de Auto Scaling em um novo Região da AWS, criamos uma VPC padrão para você. Sua VPC padrão é fornecida com uma sub-rede padrão em cada Zona de disponibilidade. Por padrão, se você tiver uma VPC padrão, seu grupo do Auto Scaling será criado na VPC padrão.

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 VPCs padrão no Guia do usuário da 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 balanceador de carga Elastic Load Balancing ao seu grupo do Auto Scaling, as instâncias poderão ser iniciadas em sub-redes públicas ou privadas. No entanto, o balanceador de carga deve ser criado somente nas sub-redes para suportar a resolução de DNS.
- Se você estiver acessando suas instâncias do Auto Scaling diretamente por meio do SSH, as instâncias poderão ser iniciadas somente em sub-redes públicas.
- Se você estiver acessando instâncias do Auto Scaling sem entrada AWS Systems Manager usando o Session Manager, as instâncias podem ser executadas em sub-redes públicas ou privadas.
- Se você estiver usando sub-redes privadas, poderá permitir que as instâncias do Auto Scaling acessem a Internet usando um gateway NAT público.
- Por padrão, as sub-redes padrão em uma VPC padrão são sub-redes públicas.

Endereçamento IP em uma VPC

Quando você ativa suas instâncias do Auto Scaling em uma VPC, um endereço IP privado no intervalo de CIDR da sub-rede na qual a instância foi executada é automaticamente atribuído às suas instâncias. Isso permite que suas instâncias se comuniquem com outras instâncias na VPC.

Você pode configurar um modelo de execução ou uma configuração de execução para atribuir 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 AWS serviços.

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? no Centro de AWS Conhecimento.

Locação de localização de instância

Por padrão, todas as instâncias na VPC são executadas como instâncias de locação compartilhada. O Amazon EC2 Auto Scaling também oferece suporte a instâncias dedicadas e hosts dedicados. Para ter mais informações, consulte Criar um modelo de execução usando configurações avançadas.

AWS Outposts

AWS Outposts estende uma Amazon VPC de uma AWS região para um posto avançado com os componentes da VPC que são acessíveis na região, incluindo gateways de internet, gateways privados virtuais, Amazon VPC Transit Gateways e VPC endpoints. Um Outpost fica hospedado em uma zona de disponibilidade na região e é uma extensão dessa zona de disponibilidade que você pode usar para resiliência.

Para 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
  - Exemplo: VPC com servidores em sub-redes privadas e NAT
  - Gateways NAT
- Sub-redes públicas em uma VPC
  - Exemplo: VPC para um ambiente de teste
  - Exemplo: VPC para servidores web e de banco de dados
- Sub-redes para seu Application Load Balancer
  - Sub-redes para seu balanceador de carga
- Informações gerais da VPC
  - Guia do usuário da Amazon VPC
  - Conectar VPCs usando emparelhamento da VPC
  - Interfaces de rede elástica
  - Usar endpoints da VPC para conectividade privada

Segurança no Amazon EC2 Auto Scaling

A segurança na nuvem AWS é a maior prioridade. Como AWS cliente, você se beneficia de uma arquitetura de data center e rede criada para atender aos requisitos das organizações mais sensíveis à segurança.

A segurança é uma responsabilidade compartilhada entre você AWS e você. O modelo de responsabilidade compartilhada descreve isso como a segurança da nuvem e segurança na nuvem:

- Segurança da nuvem — AWS é responsável por proteger a infraestrutura que executa AWS os serviços na AWS nuvem. AWS também fornece serviços que você pode usar com segurança. Auditores terceirizados testam e verificam regularmente a eficácia de nossa segurança como parte dos AWS programas de de . Para saber mais sobre os programas de conformidade que se aplicam ao Amazon EC2 Auto Scaling, AWS consulte serviços em escopo por programa de conformidade serviços em escopo por AWS de conformidade.
- Segurança na nuvem — Sua responsabilidade é determinada pelo AWS serviço que você usa. Você também é responsável por outros fatores, incluindo a confidencialidade de seus dados, os requisitos da sua empresa e as leis e normas aplicáveis.

Esta documentação ajuda a entender como aplicar o modelo de responsabilidade compartilhada ao usar o Amazon EC2 Auto Scaling. Os tópicos a seguir mostram como configurar o Amazon EC2 Auto Scaling para atender aos seus objetivos de segurança e de compatibilidade. Você também aprende a usar outros AWS serviços que ajudam você a monitorar e proteger seus recursos do Amazon EC2 Auto Scaling.

Tópicos

- Segurança da infraestrutura no Amazon EC2 Auto Scaling
- Resiliência no Amazon EC2 Auto Scaling
- Proteção de dados no Amazon EC2 Auto Scaling
- Gerenciamento de identidade e acesso para o Amazon EC2 Auto Scaling
- Validação de compatibilidade do Amazon EC2 Auto Scaling
- Amazon EC2 Auto Scaling e endpoints da VPC da interface

Segurança da infraestrutura no Amazon EC2 Auto Scaling

Como um serviço gerenciado, o Amazon EC2 Auto Scaling é protegido AWS pela segurança de rede global. Para obter informações sobre serviços AWS de segurança e como AWS proteger a infraestrutura, consulte AWS Cloud Security. Para projetar seu AWS ambiente usando as melhores práticas de segurança de infraestrutura, consulte Proteção de infraestrutura no Security Pillar AWS Well-Architected Framework.

Você usa chamadas de API AWS publicadas para acessar o Amazon EC2 Auto Scaling pela rede. Os clientes precisam oferecer suporte para:

- Transport Layer Security (TLS). Exigimos TLS 1.2 e recomendamos TLS 1.3.
- Conjuntos de criptografia com sigilo de encaminhamento perfeito (perfect forward secrecy, ou PFS) como DHE (Ephemeral Diffie-Hellman, ou Efêmero Diffie-Hellman) ou ECDHE (Ephemeral Elliptic Curve Diffie-Hellman, ou Curva elíptica efêmera Diffie-Hellman). A maioria dos sistemas modernos, como Java 7 e versões posteriores, comporta esses modos.

Além disso, as solicitações devem ser assinadas utilizando um ID da chave de acesso e uma chave de acesso secreta associada a uma entidade principal do IAM. Ou é possível 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 obter mais informações, consulte Amazon EC2 Auto Scaling e endpoints da VPC da interface.

Recursos relacionados

Para obter informações sobre os recursos para isolar o tráfego de serviços fornecidos pelo Amazon EC2, consulte Segurança da infraestrutura no Amazon EC2 no Guia do usuário do Amazon EC2 para instâncias Linux.

Resiliência no Amazon EC2 Auto Scaling

A infraestrutura AWS global é construída em torno Regiões da AWS de zonas de disponibilidade. Regiões da AWS fornecem várias zonas de disponibilidade fisicamente separadas e isoladas,

conectadas a redes de baixa latência, alta taxa de transferência e alta redundância. Com as zonas de disponibilidade, é possível projetar e operar aplicações e bancos de dados que automaticamente executam o failover entre as zonas sem interrupção. As zonas de disponibilidade são mais altamente disponíveis, tolerantes a falhas e escaláveis que uma ou várias infraestruturas de datacenter tradicionais.

Para obter mais informações sobre zonas de disponibilidade Regiões da AWS e zonas de disponibilidade, consulte Infraestrutura AWS global.

Para se beneficiar da redundância geográfica do design da zona de disponibilidade, faça o seguinte:

- Estenda seu grupo de Auto Scaling em várias zonas de disponibilidade.
- Mantenha pelo menos uma instância em cada zona de disponibilidade.
- Anexe um balanceador de carga para distribuir o tráfego de entrada nas mesmas zonas de disponibilidade. Se você usar um Application Load Balancer, certifique-se de que cada instância do EC2 obtenha uma quantidade semelhante de tráfego, mantendo o balanceamento de carga entre zonas ativado. Isso ajuda a limitar o impacto do aumento da carga nas instâncias existentes durante um evento de failover e resulta em maior resiliência do que sem o balanceamento de carga entre zonas.
- Certifique-se de que as verificações de integridade do Elastic Load Balancing estejam configuradas corretamente e também de que estejam habilitadas no grupo do Auto Scaling. Então, se uma instância falhar em sua verificação de integridade, o Elastic Load Balancing para de enviar tráfego para ela e redireciona o tráfego para instâncias íntegras, enquanto o Amazon EC2 Auto Scaling substitui a instância não íntegra.

O Amazon EC2 Auto Scaling ajuda a manter disponibilidade de 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 suportam políticas de escalabilidade e interrompe as atividades de escalabilidade quando métricas confiáveis não estão disponíveis, como quando faltam pontos de dados.
- Tenta manter números equivalentes de instâncias em cada zona de disponibilidade habilitada à medida que seu grupo aumenta.

- Usa zonas de disponibilidade para manter alta disponibilidade. Quando uma zona de disponibilidade se torna não íntegra, o Amazon EC2 Auto Scaling fará o seguinte:
  - Inicia novas instâncias em 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 balanceadores de carga associados ao seu grupo do Auto Scaling. Dessa forma, você não precisa registrar e cancelar o registro de instâncias separadamente.

Recursos relacionados

Para obter informações sobre recursos para ajudar a suportar suas necessidades de resiliência de dados fornecidos pelo Amazon EBS, consulte Resiliência no Amazon Elastic Block Store no Guia do usuário do Amazon EBS.

Proteção de dados no Amazon EC2 Auto Scaling

O modelo de responsabilidade AWS compartilhada O modelo se aplica à proteção de dados no Amazon EC2 Auto Scaling. Conforme descrito neste modelo, AWS é responsável por proteger a infraestrutura global que executa todos os Nuvem AWS. Você é responsável por manter o controle sobre seu conteúdo hospedado nessa infraestrutura. Você também é responsável pelas tarefas de configuração e gerenciamento de segurança dos Serviços da AWS que usa. Para obter mais informações sobre a privacidade de dados, consulte as Perguntas frequentes sobre privacidade de dados. Para mais informações sobre a proteção de dados na Europa, consulte o artigo AWS Shared Responsibility Model and GDPR no Blog de segurança da AWS .

Para fins de proteção de dados, recomendamos que você proteja Conta da AWS as credenciais e configure usuários individuais com AWS IAM Identity Center ou AWS Identity and Access Management (IAM). Dessa maneira, cada usuário receberá apenas as permissões necessárias para cumprir suas obrigações de trabalho. Recomendamos também que você proteja seus dados das seguintes formas:

- Use uma autenticação multifator (MFA) com cada conta.

- Use SSL/TLS para se comunicar com os recursos. AWS Exigimos TLS 1.2 e recomendamos TLS 1.3.
- Configure a API e o registro de atividades do usuário com AWS CloudTrail.
- Use soluções de AWS criptografia, juntamente com todos os controles de segurança padrão Serviços da AWS.
- Use serviços gerenciados de segurança avançada, como o Amazon Macie, que ajuda a localizar e proteger dados sigilosos armazenados no Amazon S3.
- Se você precisar de módulos criptográficos validados pelo FIPS 140-2 ao acessar AWS por meio de uma interface de linha de comando ou de uma API, use um endpoint 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 e-mail dos seus clientes, em marcações ou campos de formato livre, como um campo Nome. Isso inclui quando você trabalha com o Amazon EC2 Auto Scaling ou Serviços da AWS outro usando o console, a API AWS CLI ou os SDKs. 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 iniciar uma instância do Amazon EC2, você tem a opção de passar dados do usuário para a instância para fazer configurações adicionais quando a instância for inicializada. Também recomendamos que você nunca forneça informações confidenciais ou sigilosas nos dados do usuário que serão passados para uma instância.

Use AWS KMS keys para criptografar volumes do Amazon EBS

Você pode configurar seu grupo do Auto Scaling para criptografar dados de volume do Amazon EBS armazenados na nuvem com o AWS KMS keys. O Amazon EC2 Auto Scaling AWS suporta chaves gerenciadas e gerenciadas pelo cliente para criptografar seus dados. Observe que a opção `KmsKeyId` para especificar uma chave gerenciada pelo cliente não está disponível quando você usa uma configuração de execução. Para especificar sua chave gerenciada pelo cliente, use um modelo de execução. Para ter mais informações, consulte Criar um modelo de execução para um grupo do Auto Scaling. Para obter informações sobre como criar, armazenar e gerenciar suas chaves de AWS KMS criptografia, consulte o Guia do AWS Key Management Service desenvolvedor.

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 mais informações, consulte Usar criptografia com AMIs baseadas em EBS no Guia do usuário do Amazon EC2 para instâncias Linux e criptografia por padrão no Guia do usuário do Amazon EBS.

> **Note**
>
> Para obter informações sobre como configurar a política de chave necessária para iniciar instâncias do Auto Scaling ao usar uma chave gerenciada pelo cliente para criptografia, consulte Política de AWS KMS chaves necessária para uso com volumes criptografados.

Recursos relacionados

Para as diretrizes de proteção de dados fornecidas pelo Amazon EBS, consulte Proteção de dados no Amazon Elastic Block Store no Guia do usuário do Amazon EBS.

Política de AWS KMS chaves necessária para uso com volumes criptografados

O Amazon EC2 Auto Scaling usa funções vinculadas a serviços para delegar permissões a outras pessoas. Serviços da AWS As funções vinculadas ao serviço Amazon EC2 Auto Scaling são predefinidas e incluem permissões que o Amazon EC2 Auto Scaling exige para ligar para outras pessoas em seu nome. Serviços da AWS As permissões predefinidas também incluem acesso ao seu Chaves gerenciadas pela AWS. No entanto, elas não incluem acesso às chaves gerenciadas pelo cliente, permitindo que você mantenha o controle total sobre essas chaves.

Este tópico descreve como configurar a política de chaves de que você precisa para iniciar instâncias do Auto Scaling ao especificar uma chave gerenciada pelo cliente para a criptografia do Amazon EBS.

> **Note**
>
> O Amazon EC2 Auto Scaling não precisa de autorização adicional para usar a Chave gerenciada pela AWS padrão para proteger os volumes criptografados em sua conta.

Conteúdo

- Visão geral
- Configurar políticas de chave
- Exemplo 1: seções da política de chaves que permitem acesso à chave gerenciada pelo cliente
- Exemplo 2: seções da política de chaves que permitem acesso entre contas à chave gerenciada pelo cliente
- Editar políticas de chaves no console do AWS KMS

Visão geral

O seguinte AWS KMS keys pode ser usado para a criptografia do Amazon EBS quando o Amazon EC2 Auto Scaling inicia instâncias:

- Chave gerenciada pela AWS— Uma chave de criptografia em sua conta que o Amazon EBS cria, possui e gerencia. Essa é a chave de criptografia padrão para uma nova conta. O Chave gerenciada pela AWS é usado para criptografia, a menos que você especifique uma chave gerenciada pelo cliente.
- Chave gerenciada pelo cliente — Uma chave de criptografia personalizada que você cria, possui e gerencia. Para obter mais informações, consulte Criação de chaves Guia do desenvolvedor do AWS Key Management Service .

  Observação: a chave deve ser simétrica. O Amazon EBS não oferece suporte a chaves gerenciadas pelo cliente assimétricas.

Você configura chaves gerenciadas pelo cliente ao criar snapshots criptografados ou um modelo de execução que especifica volumes criptografados ou ao habilitar a criptografia por padrão.

Configurar políticas de chave

Suas chaves do KMS devem ter uma política de chaves que permita que o Amazon EC2 Auto Scaling execute instâncias com volumes do Amazon EBS criptografados com uma chave gerenciada pelo cliente.

Use os exemplos nesta página para configurar uma política de chaves para conceder ao Amazon EC2 Auto Scaling acesso à sua chave gerenciada pelo cliente. Você pode modificar a política de chaves da chave gerenciada pelo cliente no momento em que a chave é criada ou posteriormente.

É necessário adicionar, no mínimo, duas declarações de política à política de chaves para que ela funcione com o Amazon EC2 Auto Scaling.

- A primeira declaração permite que a identidade do IAM especificada no elemento `Principal` use a chave gerenciada pelo cliente diretamente. Inclui permissões para realizar as `DescribeKey` operações AWS KMS `Encrypt` `DecryptReEncrypt*`,`GenerateDataKey*`,, e na chave.
- A segunda declaração permite que a identidade do IAM especificada no `Principal` elemento use a `CreateGrant` operação para gerar concessões que delegam um subconjunto de suas próprias permissões para aqueles Serviços da AWS que estão integrados com AWS KMS ou outro principal. Isso permite que eles usem a chave para criar recursos criptografados em seu nome.

Ao adicionar as novas declarações de política à sua política de chave, não altere as declarações existentes na política.

Para cada um dos exemplos a seguir, os argumentos que devem ser substituídos, como uma ID de chave ou o nome de uma função vinculada ao serviço, são mostrados como texto de *`espaço reservado para o usuário`*. 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 obter mais informações, consulte os seguintes recursos do :

- Para criar uma chave com o AWS CLI, consulte create-key.
- Para atualizar uma política de chaves com o AWS CLI, consulte put-key-policy.
- Para encontrar um nome do recurso da Amazon (ARN) e um ID de chave, consulte Como encontrar o ID de chave e o ARN no Guia do desenvolvedor do AWS Key Management Service .
- Para obter informações sobre as funções vinculadas ao serviço do Amazon EC2 Auto Scaling, consulte Funções vinculadas ao serviço do Amazon EC2 Auto Scaling.
- Para obter informações sobre a criptografia do Amazon EBS e o KMS em geral, a criptografia do Amazon EBS no Guia do usuário do Amazon EBS e no Guia do desenvolvedor.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, certificando-se de substituir *111122223333* pelo ID real da conta na Conta da AWS qual você deseja criar o grupo 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 chaves é usar a visualização padrão do AWS KMS console para políticas de chaves e tornar uma identidade (usuário ou função) do IAM um dos principais usuários da política de chaves apropriada. Para obter mais informações, consulte Usando a visualização AWS Management Console padrão no Guia do AWS Key Management Service desenvolvedor.

> **Important**
>
> Tenha cuidado. As declarações de política de visualização padrão do console incluem permissões para realizar AWS KMS `Revoke` operações na chave gerenciada pelo cliente. Se você conceder Conta da AWS acesso a uma chave gerenciada pelo cliente em sua conta e acidentalmente revogar a concessão que lhes deu essa permissão, os usuários externos não poderão mais acessar seus dados criptografados ou a chave que foi usada para criptografar seus dados.

Gerenciamento de identidade e acesso para o Amazon EC2 Auto Scaling

AWS Identity and Access Management (IAM) é uma ferramenta AWS service (Serviço da AWS) que ajuda o administrador a controlar com segurança o acesso aos AWS recursos. Os administradores do IAM controlam quem pode ser autenticado (conectado) e autorizado (ter permissões) para usar os recursos do Amazon EC2 Auto Scaling. O IAM é um AWS service (Serviço da AWS) que você pode usar sem custo adicional.

Para usar o Amazon EC2 Auto Scaling, você precisa de Conta da AWS um e de suas credenciais de segurança para entrar na sua conta. Para obter mais informações, consulte as credenciais de AWS segurança no Guia do usuário do IAM.

Para concluir a documentação do IAM, consulte o Guia do usuário do IAM.

Controle de acesso

É possível ter credenciais válidas para autenticar suas solicitações. No entanto, a menos que você tenha permissões, não poderá criar nem acessar os recursos do Amazon EC2 Auto Scaling. Por exemplo, é necessário ter permissões para criar grupos do Auto Scaling, executar instâncias com modelos de execução e assim por diante.

As seções a seguir apresentam detalhes sobre como um administrador do IAM pode usá-lo para ajudar a proteger seus recursos do Amazon EC2 Auto Scaling controlando quem pode executar ações do Amazon EC2 Auto Scaling.

Recomendamos ler os tópicos do Amazon EC2 primeiro. Consulte Gerenciamento de identidade e acesso para o Amazon EC2 no Manual do usuário do Amazon EC2 para instâncias do Linux. Depois de ler os tópicos desta seção, você terá uma boa noção do que as permissões de controle de acesso do Amazon EC2 oferecem e como elas podem se adequar às permissões de recursos do Amazon EC2 Auto Scaling.

Tópicos

- Como o Amazon EC2 Auto Scaling funciona com o IAM
- Permissões de API para o Amazon EC2 Auto Scaling
- AWS políticas gerenciadas para o Amazon EC2 Auto Scaling

- Funções vinculadas ao serviço do Amazon EC2 Auto Scaling
- Exemplos de políticas baseadas em identidade do Amazon EC2 Auto Scaling
- Prevenção contra o ataque “Confused deputy” entre serviços
- Suporte a modelo de execução
- Funções do IAM para aplicações que são executadas em instâncias do Amazon EC2

Como o Amazon EC2 Auto Scaling funciona com o IAM

Antes de usar o IAM para gerenciar o acesso ao Amazon EC2 Auto Scaling, saiba quais recursos do IAM estão disponíveis para uso com o Amazon EC2 Auto Scaling.

Recursos do IAM que você pode usar com o Amazon EC2 Auto Scaling

Recurso do IAM	Suporte a Amazon EC2 Auto Scaling
Políticas baseadas em identidade	Sim
Políticas baseadas em recursos	Não
Ações de políticas	Sim
Recursos de políticas	Sim
Chaves de condição de política (específicas do serviço)	Sim
ACLs	Não
ABAC (tags em políticas)	Parcial
Credenciais temporárias	Sim
Perfis de serviço	Sim
Perfis vinculados ao serviço	Sim

Para obter uma visão de alto nível de como o Amazon EC2 Auto Scaling e Serviços da AWS outros funcionam com a maioria dos recursos do IAM, Serviços da AWS consulte esse trabalho com o IAM no Guia do usuário do IAM.

Políticas baseadas em identidade para o Amazon EC2 Auto Scaling

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 perfil do IAM. Essas políticas controlam quais ações os usuários e perfis podem realizar, em quais recursos e em que condições. Para saber como criar uma política baseada em identidade, consulte 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 anexada. Para saber mais sobre todos os elementos que podem ser usados em uma política JSON, consulte Referência de elementos da política JSON do IAM no Guia do Usuário do IAM.

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 perfil do IAM e as políticas de bucket do Amazon S3. Em serviços compatíveis com políticas baseadas em recursos, os administradores de serviço podem usá-las para controlar o acesso a um recurso específico. Para o recurso ao qual a política está anexada, a política define quais ações uma entidade principal especificada pode executar nesse recurso e em que condições. Você deve especificar uma entidade principal em uma política baseada em recursos. Os diretores podem incluir contas, usuários, funções, usuários federados ou. Serviços da AWS

Para permitir o acesso entre contas, 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 o principal e o recurso são diferentes Contas da AWS, um

administrador do IAM na conta confiável também deve conceder permissão à entidade principal (usuário ou função) 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 os perfis 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 políticas AWS JSON para especificar quem tem acesso ao quê. Ou seja, qual entidade principal pode executar ações em quais recursos e em que condições.

O elemento `Action` de uma política JSON descreve as ações que você pode usar para permitir ou negar acesso em uma política. As ações de política geralmente têm o mesmo nome da operação de AWS API 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.

Incluem ações em uma política para conceder permissões para executar a operação associada.

Para ver uma lista das ações do Amazon EC2 Auto Scaling, consulte Ações definidas pelo Amazon EC2 Auto Scaling na Referência de autorização do serviço.

As ações de política no Amazon EC2 Auto Scaling usam o seguinte prefixo antes da ação:

```
autoscaling
```

Para especificar várias ações em uma única instrução, separe-as com vírgulas.

```
"Action": [
  "autoscaling:action1",
  "autoscaling:action2"
]
```

Você pode especificar várias ações usando caracteres curinga (*). Por exemplo, para especificar todas as ações que começam com a palavra `Describe`, inclua a seguinte ação:

```
"Action": "autoscaling:Describe*"
```

Recursos de política para o Amazon EC2 Auto Scaling

Oferece suporte a recursos de políticas	Sim

Os administradores podem usar políticas AWS JSON para especificar quem tem acesso ao quê. Ou seja, qual entidade principal pode executar ações em quais recursos e em que condições.

O elemento `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 caractere 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.

> **Note**
>
> Para ver um exemplo de uma política do IAM que usa ARNs para controlar o acesso aos grupos do Auto Scaling, consulte Controlar quais grupos do Auto Scaling podem ser excluídos.

Nem todas as ações do Amazon EC2 Auto Scaling oferecem suporte a permissões em nível de recurso. Para ações que não são compatíveis com permissões em nível de recurso, você precisa usar um curinga (*) como o recurso.

As ações do Amazon EC2 Auto Scaling a seguir oferecem suporte a permissões em nível de recurso.

- `DescribeAccountLimits`
- `DescribeAdjustmentTypes`
- `DescribeAutoScalingGroups`
- `DescribeAutoScalingInstances`
- `DescribeAutoScalingNotificationTypes`
- `DescribeInstanceRefreshes`
- `DescribeLaunchConfigurations`
- `DescribeLifecycleHooks`
- `DescribeLifecycleHookTypes`
- `DescribeLoadBalancers`

- `DescribeLoadBalancerTargetGroups`
- `DescribeMetricCollectionTypes`
- `DescribeNotificationConfigurations`
- `DescribePolicies`
- `DescribeScalingActivities`
- `DescribeScalingProcessTypes`
- `DescribeScheduledActions`
- `DescribeTags`
- `DescribeTerminationPolicyTypes`
- `DescribeWarmPool`

Chaves de condição de política do Amazon EC2 Auto Scaling

Compatível com chaves de condição de política específicas do serviço	Sim

Os administradores podem usar políticas AWS JSON para especificar quem tem acesso ao quê. Ou seja, qual entidade principal pode executar ações em quais recursos e em que condições.

O elemento `Condition` (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 atendentes 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, AWS avalia a condição usando uma `OR` operação lógica. 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.

AWS suporta chaves de condição globais e chaves de condição específicas do serviço. Para ver todas as chaves de condição AWS globais, consulte as chaves de contexto de condição AWS global no Guia do usuário do IAM.

O Amazon EC2 Auto Scaling oferece suporte às seguintes chaves de condição que podem ser usadas para controlar o acesso a ações compatíveis e impor a configuração de grupos do Auto Scaling:

- `autoscaling:InstanceTypes`
- `autoscaling:LaunchConfigurationName`
- `autoscaling:LaunchTemplateVersionSpecified`
- `autoscaling:LoadBalancerNames`
- `autoscaling:MaxSize`
- `autoscaling:MinSize`
- `autoscaling:ResourceTag/key-name: tag-value`
- `autoscaling:TargetGroupARNs`
- `autoscaling:VPCZoneIdentifiers`

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 ter mais informações, consulte Etiquetar grupos e instâncias do Auto Scaling.

- `aws:RequestTag/key-name: tag-value`
- `aws:ResourceTag/key-name: tag-value`

- `aws:TagKeys: [tag-key, ...]`

Para saber com que ações da API Amazon EC2 Auto Scaling você pode usar uma chave de condição, consulte Ações definidas pelo Amazon EC2 Auto Scaling na Referência de autorização do serviço. Para obter mais informações sobre chaves de condição do Amazon EC2 Auto Scaling, consulte Chaves de condição para Amazon EC2 Auto Scaling.

> **Note**
>
> Para exemplos de políticas do IAM que usam chaves de condição para controlar o acesso a ações compatíveis e impor a configuração de grupos do Auto Scaling, consulte os seguintes recursos:
>
> - Exigir um modelo de execução e um número de versão— Este exemplo exige que um modelo de execução e o número da versão do modelo de execução sejam especificados ao criar ou atualizar grupos de Auto Scaling.
> - Controlar o tamanho de grupos do Auto Scaling que podem ser criados— Este exemplo impõe restrições aos valores possíveis das `MaxSize` propriedades `MinSize` e ao criar ou atualizar grupos de Auto Scaling com uma tag específica.
> - Controlar quais políticas de escalabilidade podem ser excluídas— Este exemplo afirma que a exclusão de políticas de escalabilidade é permitida somente para grupos de Auto Scaling sem uma tag específica.

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 perfis 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. Em AWS, esses atributos são chamados de tags. Você pode anexar tags a entidades do IAM (usuários ou funções) e a vários AWS recursos. 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 Sim para o serviço. Se um serviço oferecer suporte às três chaves de condição somente para alguns tipos de recursos, o valor será 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) 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 ter mais informações, consulte Etiquetar grupos e instâncias do Auto Scaling.

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 “Trabalhe com o IAM” no Guia do usuário do IAM.

Você está usando credenciais temporárias se fizer login AWS Management Console usando qualquer método, exceto um nome de usuário e senha. Por exemplo, quando você acessa AWS

usando o link de login único (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 perfis. Para obter mais informações sobre como alternar perfis, consulte Alternar para uma função (console) no Guia do usuário do IAM.

Você pode criar manualmente credenciais temporárias usando a AWS API AWS CLI ou. Em seguida, você pode usar essas credenciais temporárias para acessar AWS. 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 perfis de serviço	Sim

O perfil de serviço é um perfil 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 uma função 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 ter mais informações, consulte Receba notificações usando o Amazon SNS e Receba notificações usando o Amazon SQS.

Ao criar um grupo de Auto Scaling, você pode, opcionalmente, passar uma função de serviço para permitir que instâncias do Amazon EC2 acessem outras em seu nome. Serviços da AWS O perfil de serviço para instâncias do Amazon EC2 (também chamado de perfil de instância do Amazon EC2 para um modelo de execução ou configuração de execução) é um tipo especial de perfil de serviço que é atribuído a cada instância do EC2 em um grupo do Auto Scaling quando a instância é executada. Você pode usar o console do IAM e AWS CLI criar ou editar essa função de serviço. Para ter mais informações, consulte Funções do IAM para aplicações que são executadas em instâncias do Amazon EC2.

> **Warning**
>
> A alteração das permissões de um perfil de serviço pode interromper a funcionalidade do Amazon EC2 Auto Scaling. Edite perfis de serviço somente quando o Amazon EC2 Auto Scaling fornecer orientação para isso.

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. As funções vinculadas ao serviço aparecem em você Conta da AWS e são de propriedade do serviço. Um administrador do IAM pode visualizar, mas não pode editar as permissões para perfis vinculados 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.

Permissões de API para o Amazon EC2 Auto Scaling

Você deve conceder permissões aos usuários para chamar as ações da API do Amazon EC2 Auto Scaling de que eles precisam, como indicado em Ações de políticas para o Amazon EC2 Auto Scaling. Além disso, para algumas ações do Amazon EC2 Auto Scaling, você deve conceder aos usuários permissão para chamar ações específicas de outras APIs. AWS

Permissões necessárias de outras APIs AWS

Além das permissões da API do Amazon EC2 Auto Scaling, os usuários devem ter as seguintes permissões de AWS outras APIs para realizar com sucesso a ação associada.

Criar um grupo do Auto Scaling (`autoscaling:CreateAutoScalingGroup`)

- `iam:CreateServiceLinkedRole`— Para criar a função vinculada ao serviço padrão, caso essa função ainda não exista.

- `iam:PassRole`— Para passar uma função do IAM para o serviço ou para instâncias do EC2 no lançamento. Necessário quando um perfil vinculado ao serviço não padrão, um perfil do IAM para um hook do ciclo de vida ou um modelo de execução que especifica um perfil de instância (um contêiner para um perfil do IAM) é fornecido.
- `ec2:RunInstance`— Para iniciar instâncias quando um modelo de lançamento é fornecido.
- `ec2:CreateTags`— Marcar instâncias e volumes no lançamento quando um modelo de lançamento 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 um perfil do IAM é fornecido.

Anexar um grupo de destino do VPC Lattice (`autoscaling:AttachTrafficSources`)

- `vpc-lattice:RegisterTargets`— Para registrar automaticamente as instâncias com o grupo-alvo.

Desanexar um grupo de destino 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`— Para 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 um perfil do IAM).

AWS políticas gerenciadas para o Amazon EC2 Auto Scaling

Uma política AWS gerenciada é uma política autônoma criada e administrada por AWS. AWS as políticas gerenciadas são projetadas para fornecer permissões para muitos casos de uso comuns, para que você possa começar a atribuir permissões a usuários, grupos e funções.

Lembre-se de que as políticas AWS gerenciadas podem não conceder permissões de privilégio mínimo para seus casos de uso específicos porque estão disponíveis para uso de todos os AWS clientes. 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 nas políticas AWS gerenciadas. Se AWS atualizar as permissões definidas em uma política AWS gerenciada, a atualização afetará todas as identidades principais (usuários, grupos e funções) às quais a política está anexada. AWS é mais provável que atualize uma política AWS gerenciada quando uma nova AWS service (Serviço da AWS) é lançada ou novas operações de API são disponibilizadas para serviços existentes.

Para mais informações, consulte Políticas gerenciadas pela AWS no Manual do usuário do IAM.

Políticas gerenciadas do Amazon EC2 Auto Scaling

Você pode anexar as seguintes políticas gerenciadas às suas identidades AWS Identity and Access Management (IAM) (usuários ou funções). Cada política fornece acesso a todas ou a algumas das ações de API para o Amazon EC2 Auto Scaling.

- AutoScalingFullAccess— Concede acesso total ao Amazon EC2 Auto Scaling para identidades do IAM que precisam de acesso total ao Amazon EC2 Auto Scaling a partir dos SDKs ou, mas não de acesso AWS CLI . AWS Management Console
- AutoScalingReadOnlyAccess— Concede acesso somente de leitura ao Amazon EC2 Auto Scaling para identidades do IAM que estão fazendo chamadas somente para os SDKs ou. AWS CLI
- 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. 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.

Política gerenciada pelo AutoScalingServiceRolePolicy AWS

Você não pode se associar AutoScalingServiceRolePolicyàs 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 ter mais informações, consulte Funções vinculadas ao serviço do Amazon EC2 Auto Scaling.

Atualizações do Amazon EC2 Auto Scaling para políticas gerenciadas AWS

Veja detalhes sobre as atualizações das políticas AWS gerenciadas do Amazon EC2 Auto Scaling desde que esse serviço começou a rastrear essas alterações. Para receber alertas automáticos sobre mudanças nesta página, assine o RSS feed na página de histórico de documentos do Amazon EC2 Auto Scaling.

Alteração	Descrição	Data
O Amazon EC2 Auto Scaling adiciona permissões à respectiva função vinculada ao serviço	A `AutoScalingServiceRolePolicy` política agora concede permissões para chamar a ação da GetSecurityGroupsForVpcAPI do Amazon EC2 para obter todos os grupos de segurança de uma VPC para melhorar a validação, e a ação da GetInstanceTypesFromInstanceRequirementsAPI do Amazon EC2 para obter informações sobre quais tipos de instância atendem a um determinado conjunto de requisitos de instância. Para ter mais informações, consulte Funções vinculadas ao serviço do Amazon EC2 Auto Scaling.	29 de fevereiro de 2024

Alteração	Descrição	Data
O Amazon EC2 Auto Scaling adiciona permissões à respectiva função vinculada ao serviço	A política `AutoScalingServiceRolePolicy` agora concede permissões ao serviço para acessar as ações de API necessárias para uma integração com o VPC Lattice. • Ações `GetTargetGroup` e `ListTargetGroup` . Obrigatório para recuperar informações sobre grupos de destino VPC Lattice. • Ações `RegisterTargets` e `DeregisterTargets` . Obrigatório para registrar e cancelar o registro de instâncias de grupos de destino VPC Lattice. • `ListTargets` . Permite que o Amazon EC2 Auto Scaling recupere informações de integridade para instâncias registradas em grupos de destino VPC Lattice. Para ter mais informações, consulte Funções vinculadas ao serviço do Amazon EC2 Auto Scaling.	6 de dezembro de 2022

Alteração	Descrição	Data
O Amazon EC2 Auto Scaling adiciona permissões à respectiva função vinculada ao serviço	Para oferecer suporte ao uso de um AWS Systems Manager parâmetro como alias para uma ID de AMI ao criar um modelo de lançamento, a `AutoScalingServiceRolePolicy` política agora concede permissão para chamar a ação da AWS Systems Manager GetParametersAPI. Para ter mais informações, consulte Funções vinculadas ao serviço do Amazon EC2 Auto Scaling.	28 de março de 2022
O Amazon EC2 Auto Scaling adiciona permissões à respectiva função vinculada ao serviço	Para oferecer suporte à escalabilidade preditiva, a `AutoScalingServiceRolePolicy` política agora inclui permissão para chamar a ação da CloudWatch GetMetricDataAPI. Para ter mais informações, consulte Funções vinculadas ao serviço do Amazon EC2 Auto Scaling.	19 de maio de 2021
O Amazon EC2 Auto Scaling começou a monitorar alterações	O Amazon EC2 Auto Scaling começou a monitorar as alterações em suas políticas gerenciadas 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. Uma função vinculada ao serviço é um tipo exclusivo de função do IAM vinculada diretamente a uma. AWS service (Serviço da AWS)

Os perfis vinculados a serviços oferecem uma maneira segura de delegar permissões a outros serviços da Serviços da AWS , pois somente o serviço vinculado pode assumir uma função vinculada ao serviço. Para obter mais informações, consulte Usar perfis vinculados ao serviço no Guia do usuário do IAM. As funções vinculadas ao serviço também permitem que todas as chamadas de API sejam visíveis por meio de. AWS CloudTrail Isso ajuda com os requisitos de monitoramento e auditoria porque você pode rastrear todas as ações que o Amazon EC2 Auto Scaling executa em seu nome. Para ter mais informações, consulte Registre chamadas da API do Amazon EC2 Auto Scaling com AWS CloudTrail.

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 obter mais informações, consulte Usar perfis vinculados ao serviço no Guia do usuário do IAM.

Conteúdo

- Visão geral
- Permissões concedidas pela função vinculada ao serviço
- Criar uma função vinculada ao serviço (automática)
- Criar uma função vinculada ao serviço (manual)
- Editar a função vinculada ao serviço
- Excluir a função vinculada ao serviço
- Regiões compatíveis com funções vinculadas ao serviço do Amazon EC2 Auto Scaling

Visão geral

Há dois tipos de funções vinculadas ao serviço do Amazon EC2 Auto Scaling:

- A função padrão vinculada ao serviço da sua conta, chamada. AWSServiceRoleForAutoScaling Essa função é automaticamente atribuída aos seus grupos do Auto Scaling, a menos que você especifique outra função vinculada ao serviço.

- ***Uma função vinculada ao serviço com um sufixo personalizado que você especifica ao criar a função, por exemplo, AWSServiceRoleForAutoScaling _ mysuffix.***

As permissões de 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 qualquer uma das funções ao editar suas políticas de AWS Key Management Service chaves para permitir que as instâncias lançadas pelo Amazon EC2 Auto Scaling sejam criptografadas com sua chave gerenciada pelo cliente. No entanto, se você planeja conceder acesso granular a uma determinada CMK gerenciada pelo cliente, 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 seus registros CloudTrail

Se você criar chaves gerenciadas pelo cliente às quais nem todos os usuários devem ter acesso, siga estas etapas para permitir o uso de uma função vinculada ao serviço com sufixo personalizado:

1. Crie uma função vinculada ao serviço com um sufixo personalizado. Para ter mais informações, consulte Criar uma função vinculada ao serviço (manual).
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 AWS KMS chaves necessária para uso com volumes criptografados.
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 (usando) PassRole. 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 AWSServiceRoleForAutoScalingchamada ou seu sufixo personalizado função vinculada ao serviço.

A função vinculada ao serviço confia no seguinte serviço para assumir a função:

- `autoscaling.amazonaws.com`

A função usa a política `AutoScalingServiceRolePolicy`, 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:GetInstanceTypesFromInstanceRequirements",
        "ec2:GetSecurityGroupsForVpc",
        "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/pare e encerre instâncias do EC2.
- `iam`— Passe funções do IAM para instâncias do EC2 para que os aplicativos em execução nas instâncias possam acessar credenciais temporárias para a função.
- `iam`— Crie a função AWSServiceRoleForEC2Spotvinculada 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`— crie, descreva, modifique e exclua CloudWatch alarmes para políticas de escalabilidade e recupere métricas usadas para escalabilidade preditiva.
- `sns`— Publique notificações no Amazon SNS quando as instâncias são iniciadas ou encerradas.
- `events`— Crie, descreva, atualize e exclua EventBridge regras em seu nome.
- `ssm`— Leia os parâmetros do Parameter Store ao usar um parâmetro do Systems Manager como 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 AWSServiceRoleForAutoScalinga função vinculada ao serviço para você na primeira vez que você cria um grupo de Auto Scaling, a menos que você crie manualmente uma função vinculada ao serviço com sufixo personalizado e a especifique 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 um perfil vinculado ao serviço 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 de Auto Scaling antes disso, o Amazon EC2 Auto Scaling AWSServiceRoleForAutoScalingcriou a função 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 Revisar, deixe o nome da função em branco para criar uma função vinculada ao serviço com o nome AWSServiceRoleForAutoScalingou digite um sufixo para criar uma função vinculada ao serviço com o sufixo nome _. AWSServiceRoleForAutoScaling***
7. (Opcional) Em Role description (Descrição da função), edite a descrição para a função vinculada ao serviço.
8. Selecione Criar função.

Para criar uma função vinculada a serviço (AWS CLI)

Use o seguinte comando da create-service-linked-roleCLI para criar uma função vinculada ao serviço para o Amazon EC2 Auto Scaling com o sufixo nome _. AWSServiceRoleForAutoScaling

```
aws iam create-service-linked-role --aws-service-name autoscaling.amazonaws.com --
custom-suffix suffix
```

A saída desse comando inclui o ARN da função vinculada ao serviço, o qual você pode usar para conceder acesso à chave gerenciada pelo cliente para a função vinculada ao serviço.

```
{
    "Role": {
        "RoleId": "ABCDEF0123456789ABCDEF",
        "CreateDate": "2018-08-30T21:59:18Z",
```

```
        "RoleName": "AWSServiceRoleForAutoScaling_suffix",
        "Arn": "arn:aws:iam::123456789012:role/aws-service-role/
autoscaling.amazonaws.com/AWSServiceRoleForAutoScaling_suffix",
        "Path": "/aws-service-role/autoscaling.amazonaws.com/",
        "AssumeRolePolicyDocument": {
            "Version": "2012-10-17",
            "Statement": [
                {
                    "Action": [
                        "sts:AssumeRole"
                    ],
                    "Principal": {
                        "Service": [
                            "autoscaling.amazonaws.com"
                        ]
                    },
                    "Effect": "Allow"
                }
            ]
        }
    }
}
```

Para 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.

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 ter mais informações, consulte Excluir infraestrutura do Auto Scaling.

É possível usar o IAM para excluir uma função vinculada ao serviço. Para obter mais informações, consulte Excluir um perfil vinculado ao serviço no Guia do usuário do IAM.

Se você excluir a função AWSServiceRoleForAutoScalingvinculada ao serviço, o Amazon EC2 Auto Scaling criará a função novamente quando você criar um grupo de Auto Scaling e não especificar uma função vinculada ao serviço diferente.

Regiões compatíveis com funções vinculadas ao serviço do Amazon EC2 Auto Scaling

O Amazon EC2 Auto Scaling oferece suporte ao uso de funções vinculadas a serviços em todos os lugares em que Regiões da AWS o serviço está disponível.

Exemplos de políticas baseadas em identidade do Amazon EC2 Auto Scaling

Por padrão, um novo usuário não Conta da AWS tem permissão para fazer nada. Um administrador do IAM deve criar e atribuir políticas do IAM que concedam a uma identidade do IAM (como um usuário ou perfil) permissão para executar ações de API do Amazon EC2 Auto Scaling.

Para saber como criar uma política do IAM usando esses exemplos de documentos de política JSON, consulte Criar políticas na aba JSON no Manual do usuário do IAM.

A seguir, um exemplo de uma política de permissões.

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

```
    },
    {
       "Effect": "Allow",
       "Action": "autoscaling:Describe*",
       "Resource": "*"
    }]
}
```

Este exemplo de política concede permissões para criar, atualizar e excluir grupos do Auto Scaling, mas somente se o grupo usar a etiqueta **`purpose=testing`**. Como as ações `Describe` não oferecem suporte a permissões em nível de recurso, é necessário especificá-las em uma declaração separada sem condições. Para iniciar instâncias com um modelo de execução, o usuário também precisa ter a permissão/`ec2:RunInstances`. Para ter mais informações, consulte Suporte a modelo de execução.

> **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

- Controlar o tamanho de grupos do Auto Scaling que podem ser criados
- Controlar quais chaves de tag e valores de tag podem ser usados
- Controlar quais grupos do Auto Scaling podem ser excluídos
- Controlar quais políticas de escalabilidade podem ser excluídas
- Controlar o acesso às ações de atualização da instância
- Criar um perfil vinculado ao serviço
- Controle qual função vinculada ao serviço pode ser passada (usando) PassRole

Controlar o tamanho de grupos do Auto Scaling que podem ser criados

A política a seguir concede permissões para criar e atualizar todos os grupos do Auto Scaling com a tag,**`environment=development`** desde que o solicitante não especifique um tamanho mínimo menor que **`1`** ou um tamanho máximo maior que **`10`**. Sempre que possível, use tags para ajudar você a controlar o acesso aos grupos do Auto Scaling na sua conta.

```
{
   "Version": "2012-10-17",
   "Statement": [{
      "Effect": "Allow",
      "Action": [
          "autoscaling:CreateAutoScalingGroup",
          "autoscaling:UpdateAutoScalingGroup"
      ],
      "Resource": "*",
      "Condition": {
          "StringEquals": { "autoscaling:ResourceTag/environment": "development" },
          "NumericGreaterThanEqualsIfExists": { "autoscaling:MinSize": 1 },
          "NumericLessThanEqualsIfExists": { "autoscaling:MaxSize": 10 }
      }
   }]
}
```

Como alternativa, se você não estiver usando tags para controlar o acesso a grupos do Auto Scaling, poderá usar ARNs para identificar os grupos do Auto Scaling aos quais a política do IAM se aplica.

Um grupo do Auto Scaling tem o ARN a seguir.

```
"Resource": "arn:aws:autoscaling:region:account-
id:autoScalingGroup:*:autoScalingGroupName/my-asg"
```

Também possível especificar vários ARNs incluindo-os em uma lista. Para obter mais informações sobre como especificar os ARNs dos recursos do Amazon EC2 Auto Scaling no elemento,`Resource` consulte Recursos de política para o Amazon EC2 Auto Scaling.

Controlar quais chaves de tag e valores de tag podem ser usados

Também é possível usar condições adicionais em suas políticas do IAM para controlar as chaves de tag e os valores que podem ser aplicados aos grupos do Auto Scaling. Para conceder permissões para criar ou etiquetar um grupo do Auto Scaling somente se o solicitante especificar determinadas

etiquetas, use a chave de condição `aws:RequestTag`. Para permitir somente chaves de tags específicas, use a chave de condição `aws:TagKeys` com o modificador `ForAllValues`.

A política a seguir requer que o solicitante especifique uma etiqueta com a chave **`environment`** na solicitação. O valor **`"?*"`** impõe que haja um valor para a chave de tag. Para usar um caractere curinga, é necessário usar o operador de condição `StringLike`.

```
{
   "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.

Controlar quais grupos do Auto Scaling podem ser excluídos

A política a seguir permite a exclusão de um grupo do Auto Scaling somente se o grupo tiver a tag **`environment=development`**.

```
{
   "Version": "2012-10-17",
   "Statement": [{
      "Effect": "Allow",
      "Action": "autoscaling:DeleteAutoScalingGroup",
      "Resource": "*",
      "Condition": {
```

```
            "StringEquals": { "aws:ResourceTag/environment": "development" }
        }
    }]
}
```

Como alternativa, se você não estiver usando chaves de condição para controlar o acesso aos grupos do Auto Scaling poderá, em vez disso, especificar os ARNs dos recursos no elemento `Resource` para controlar o acesso.

A política a seguir dá aos usuários permissões para usar a ação `DeleteAutoScalingGroup` da API, mas somente para grupos do Auto Scaling cujo nome comece 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"
]
```

Controlar quais políticas de escalabilidade podem ser excluídas

A política a seguir concede permissões para usar a ação `DeletePolicy` para excluir uma política de escalabilidade. No entanto, ela também negará a ação se o grupo do Auto Scaling que está recebendo a ação tiver a tag **`environment=production`**. Sempre que possível, use tags para ajudar você a controlar o acesso aos grupos do Auto Scaling na sua conta.

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

Controlar o acesso às ações de atualização da instância

A política a seguir concede permissão para iniciar, reverter e cancelar uma atualização de instância somente se o grupo do Auto Scaling que está recebendo a ação tiver a tag **`environment=testing`**. Como as ações `Describe` não oferecem suporte a permissões em nível de recurso, é necessário especificá-las em uma declaração separada sem condições.

```
{
   "Version": "2012-10-17",
   "Statement": [{
      "Effect": "Allow",
      "Action": [
          "autoscaling:StartInstanceRefresh",
          "autoscaling:CancelInstanceRefresh",
          "autoscaling:RollbackInstanceRefresh"
      ],
      "Resource": "*",
      "Condition": {
          "StringEquals": { "autoscaling:ResourceTag/environment": "testing" }
      }
   },
   {
      "Effect": "Allow",
      "Action": "autoscaling:DescribeInstanceRefreshes",
      "Resource": "*"
```

```
    }]
}
```

Para especificar uma configuração desejada na chamada,`StartInstanceRefresh` é possível que os usuários precisem de algumas permissões relacionadas, como:

- ec2: RunInstances — Para iniciar instâncias do EC2 usando um modelo de execução, o usuário deve ter a `ec2:RunInstances` permissão em uma política do IAM. Para ter mais informações, consulte Suporte a modelo de execução.
- ec2: CreateTags — Para iniciar instâncias do EC2 a partir de um modelo de execução que adiciona tags às instâncias e volumes na criação, o usuário deve ter a `ec2:CreateTags` permissão em uma política do IAM. Para ter mais informações, consulte Permissões necessárias para marcar instâncias e volumes.
- iam: PassRole — Para iniciar instâncias do EC2 a partir de um modelo de execução que contém um perfil de instância (um contêiner para uma função do IAM), o usuário também deve ter a `iam:PassRole` permissão em uma política do IAM. Para obter mais informações e um exemplo de política do IAM, consulte Funções do IAM para aplicações que são executadas em instâncias do Amazon EC2.
- ssm: GetParameters — Para iniciar instâncias do EC2 a partir de um modelo de execução que usa um AWS Systems Manager parâmetro, o usuário também deve ter a `ssm:GetParameters` permissão em uma política do IAM. Para ter mais informações, consulte Use AWS Systems Manager parâmetros em vez de IDs de AMI nos modelos de lançamento.

Criar um perfil vinculado ao serviço

O Amazon EC2 Auto Scaling exige permissões para criar uma função vinculada ao serviço na primeira vez que qualquer usuário em você chama as ações da API do Amazon Conta da AWS EC2 Auto Scaling. Se a função vinculada ao serviço ainda não existir, o Amazon EC2 Auto Scaling a criará em sua conta. A função vinculada ao serviço concede permissões ao Amazon EC2 Auto Scaling para que ele possa Serviços da AWS ligar para outras pessoas 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 (usando) PassRole

Os usuários que criam ou atualizam grupos do Auto Scaling e especificam um perfil vinculado ao serviço de sufixo personalizado na solicitação necessitam da permissão `iam:PassRole`.

Você pode usar a `iam:PassRole` permissão para proteger a segurança de suas chaves gerenciadas pelo AWS KMS cliente se conceder a diferentes funções vinculadas ao serviço acesso a chaves diferentes. Dependendo das necessidades de sua organização, talvez você tenha uma chave para a equipe de desenvolvimento, outra para a equipe de QA e outra para a equipe financeira. Primeiro, crie uma função vinculada ao serviço que tenha acesso à chave necessária, por exemplo, uma função vinculada ao serviço chamada. AWSServiceRoleForAutoScaling_devteamkeyaccess Em seguida, anexe a política a uma identidade do IAM, como um usuário ou um perfil.

A política a seguir concede permissões para passar o perfil **`AWSServiceRoleForAutoScaling_devteamkeyaccess`** para qualquer grupo do Auto Scaling cujo nome comece com **`devteam-`**. Se a identidade do IAM que cria o grupo do Auto Scaling tentar especificar um perfil vinculado ao serviço diferente, ela receberá um erro. Se eles optarem por não especificar uma função vinculada ao serviço, a AWSServiceRoleForAutoScalingfunção padrão será usada em seu lugar.

```
{
   "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.

Prevenção contra o ataque “Confused deputy” entre serviços

O problema “confused deputy” é um problema de segurança em que uma entidade que não tem permissão para executar uma ação pode coagir uma entidade mais privilegiada a executá-la.

Em AWS, a falsificação de identidade entre serviços pode resultar em um problema confuso de delegado. A imitação entre serviços pode ocorrer quando um serviço (o serviço de chamada) chama outro serviço (o serviço chamado). O serviço de chamada pode ser manipulado para utilizar as suas permissões para atuar nos recursos de outro cliente em que, de outra forma, ele não teria permissão para acessar.

Para evitar isso, AWS fornece ferramentas que ajudam você a proteger seus dados para todos os serviços com diretores de serviços que receberam acesso aos recursos em sua conta. Recomendamos o uso das chaves de contexto de condição global `aws:SourceArn` e `aws:SourceAccount` nas políticas de confiança para funções do serviço do Amazon EC2 Auto Scaling. Essas chaves limitam as permissões que o Amazon EC2 Auto Scaling concede a outro serviço ao recurso.

Os valores dos `SourceAccount` campos `SourceArn` e são definidos quando o Amazon EC2 Auto Scaling AWS Security Token Service usa 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 que o Amazon EC2 Auto Scaling armazena. Sempre que possível, use `aws:SourceArn`, que é mais específico. Defina o valor como o ARN ou um padrão de ARN com curinga (`*`) para as partes desconhecidas do ARN. Se você não conhece o ARN do recurso, use `aws:SourceAccount` em vez disso.

O exemplo a seguir mostra como é possível usar as chaves de contexto de condição global `aws:SourceArn` e `aws:SourceAccount` no Amazon EC2 Auto Scaling para evitar o problema do substituto confuso.

Exemplo: uso das chaves de condição **`aws:SourceArn`** e **`aws:SourceAccount`**

A função de serviço é uma função que um serviço assume para realizar ações em seu nome. Nos casos em que você quiser criar ganchos de ciclo de vida que enviem notificações para qualquer lugar que não seja a Amazon EventBridge, você deve criar uma função de serviço para permitir que o Amazon EC2 Auto Scaling envie notificações para um tópico do Amazon SNS ou fila do Amazon SQS em seu nome. Se quiser que apenas um grupo do Auto Scaling seja associado ao acesso entre serviços, você pode especificar a política de confiança da do perfil de serviço da seguinte forma.

Este exemplo de política de confiança usa declarações de condição para limitar a capacidade de `AssumeRole` na função de serviço somente para as ações que afetam o grupo do Auto Scaling especificado na conta especificada. As condições `aws:SourceArn` e `aws:SourceAccount` são avaliadas de forma independente. Qualquer solicitação para usar o perfil de serviço deve atender às duas condições.

Antes de usar essa política, substitua os valores de Região, ID da conta, UUID e nome de grupo por valores válidos da sua conta.

```
{
  "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 elemento `Principal` especifica a entidade principal de serviço do serviço (`autoscaling.amazonaws.com`).
- O elemento `Action` especifica a ação`sts:AssumeRole`.
- O elemento `Condition` especifica as chaves de condição globais `aws:SourceArn` e `aws:SourceAccount`. O ARN da fonte inclui o ID da conta, portanto, não é necessário usar `aws:SourceAccount` com `aws:SourceArn`.

Mais informações

Para obter mais informações, consulte Chaves de contexto de condição globais da AWS, O problema 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, atualizar e excluir todos os grupos do Auto Scaling, mas somente se o grupo usa a tag **`purpose=testing`**. Em seguida, concede permissão para todas as ações `Describe`. Como as ações `Describe` não oferecem suporte a permissões em nível de recurso, é necessário especificá-las em uma declaração separada sem condições.

Identidades do IAM (usuários ou perfis) com esta política têm permissão para criar ou atualizar um grupo do Auto Scaling usando um modelo de execução porque eles também têm permissão para usar a ação `ec2:RunInstances`.

```
{
    "Version": "2012-10-17",
    "Statement": [
          {
            "Effect": "Allow",
            "Action": [
                "autoscaling:CreateAutoScalingGroup",
                "autoscaling:UpdateAutoScalingGroup",
                "autoscaling:DeleteAutoScalingGroup"
            ],
            "Resource": "*",
            "Condition": {
                "StringEquals": { "autoscaling:ResourceTag/purpose": "testing" }
            }
         },
         {
            "Effect": "Allow",
            "Action": [
                "autoscaling:Describe*",
                "ec2:RunInstances"
            ],
            "Resource": "*"
        }
    ]
}
```

Os usuários que criam ou atualizam grupos do Auto Scaling podem precisar de algumas permissões relacionadas, como:

- ec2: CreateTags — Para adicionar tags às instâncias e volumes na criação, o usuário deve ter a `ec2:CreateTags` permissão em uma política do IAM. Para ter mais informações, consulte Permissões necessárias para marcar instâncias e volumes.
- iam: PassRole — Para iniciar instâncias do EC2 a partir de um modelo de execução que contém um perfil de instância (um contêiner para uma função do IAM), o usuário também deve ter a `iam:PassRole` permissão em uma política do IAM. Para obter mais informações e um exemplo de política do IAM, consulte Funções do IAM para aplicações que são executadas em instâncias do Amazon EC2.

- ssm: GetParameters — Para iniciar instâncias do EC2 a partir de um modelo de execução que usa um AWS Systems Manager parâmetro, o usuário também deve ter a `ssm:GetParameters` permissão em uma política do IAM. Para ter mais informações, consulte Use AWS Systems Manager parâmetros em vez de IDs de AMI nos modelos de lançamento.

Essas permissões para ações a serem concluídas ao iniciar instâncias são verificadas quando o usuário interage com um grupo do Auto Scaling. Para ter mais informações, consulte Validação de permissões para `ec2:RunInstances` e `iam:PassRole`.

Os exemplos a seguir mostram declarações de políticas que você pode usar para controlar os acessos que os usuários do IAM têm para usar modelos de execução.

Tópicos

- Exigir modelos de execução que têm uma tag específica
- Exigir um modelo de execução e um número de versão
- Exigir o uso do Instance Metadata Service Version 2 (IMDSv2)
- Restringir o acesso aos recursos do Amazon EC2
- Permissões necessárias para marcar instâncias e volumes
- Permissões adicionais do modelo de execução
- Validação de permissões para ec2:RunInstances e iam:PassRole
- Recursos relacionados

Exigir modelos de execução que têm uma tag específica

Ao conceder permissões,`ec2:RunInstances` é possível especificar que os usuários só poderão usar modelos de execução com tags ou IDs específicos para limitar permissões ao executar instâncias com um modelo de execução. Você também pode controlar a AMI e outros recursos aos quais qualquer pessoa que use modelos de execução possa fazer referência e usar ao iniciar instâncias especificando permissões adicionais em nível de recurso para a chamada`RunInstances`.

O exemplo a seguir restringe permissões à ação `ec2:RunInstances` para executar modelos que estão localizados na região especificada e que têm a tag **`purpose=testing`**. Ele também dá aos usuários acesso aos recursos especificados em um modelo de execução: AMIs, tipos de instância, volumes, pares de chaves, interfaces de rede e grupos de segurança.

```
{
```

```
    "Version": "2012-10-17",
    "Statement": [
        {
            "Effect": "Allow",
            "Action": "ec2:RunInstances",
            "Resource": "arn:aws:ec2:region:account-id:launch-template/*",
            "Condition": {
                "StringEquals": { "aws:ResourceTag/purpose": "testing" }
            }
        },
        {
            "Effect": "Allow",
            "Action": "ec2:RunInstances",
            "Resource": [
                "arn:aws:ec2:region::image/ami-*",
                "arn:aws:ec2:region:account-id:instance/*",
                "arn:aws:ec2:region:account-id:subnet/*",
                "arn:aws:ec2:region:account-id:volume/*",
                "arn:aws:ec2:region:account-id:key-pair/*",
                "arn:aws:ec2:region:account-id:network-interface/*",
                "arn:aws:ec2:region:account-id:security-group/*"
            ]
        }
    ]
}
```

Para obter mais informações sobre o uso de políticas baseadas em tags com modelos de execução, consulte Controlar o acesso aos modelos de execução com permissões do IAM no Guia do usuário do Amazon EC2 para instâncias do Linux.

Exigir um modelo de execução e um número de versão

Você também pode usar as permissões do IAM para obrigar que um modelo de execução e o número da versão do modelo de execução sejam especificados ao criar ou atualizar grupos do Auto Scaling.

O exemplo a seguir permite que os usuários criem e atualizem grupos do Auto Scaling somente se um modelo de execução e o número da versão do modelo de execução forem especificados. Se usuários com essa política omitirem o número da versão para especificar a versão `$Latest` ou `$Default` do modelo de execução ou tentarem usar uma configuração de execução, a ação falhará.

```
{
```

```
    "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 ajuste de escala automático que usam um novo modelo de execução ou uma nova versão de um modelo de execução com as opções de metadados de instância configuradas, você pode iniciar uma atualização de instâncias. Para ter mais informações, consulte Atualizar instâncias do Auto Scaling.

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. Para especificar permissões em nível de recurso para recursos especificados em um modelo de execução, você deve incluir os recursos na declaração de ação`RunInstances`.

```
{
    "Version": "2012-10-17",
    "Statement": [
        {
            "Effect": "Allow",
            "Action": "ec2:RunInstances",
            "Resource": [
                "arn:aws:ec2:region:account-id:launch-template/*",
                "arn:aws:ec2:region::image/ami-04d5cc9b88example",
                "arn:aws:ec2:region:account-id:subnet/subnet-1a2b3c4d",
                "arn:aws:ec2:region:account-id:volume/*",
                "arn:aws:ec2:region:account-id:key-pair/*",
                "arn:aws:ec2:region:account-id:network-interface/*",
                "arn:aws:ec2:region:account-id:security-group/sg-903004f88example"
            ]
        },
```

```
        {
            "Effect": "Allow",
            "Action": "ec2:RunInstances",
            "Resource": "arn:aws:ec2:region:account-id:instance/*",
            "Condition": {
                "StringEquals": { "ec2:InstanceType": ["t2.micro", "t2.small"] }
            }
        }
    ]
}
```

Neste exemplo, há duas declarações:

- A primeira declaração requer que os usuários executem instâncias em uma sub-rede específica (**`subnet-1a2b3c4d`**), usando um grupo de segurança (**`sg-903004f88example`**) específico e usando uma AMI (**`ami-04d5cc9b88example`**) específica. Ele também dá aos usuários acesso aos recursos especificados em um modelo de execução: interfaces de rede, pares de chaves e volumes.
- A segunda instrução permite que os usuários executem instâncias usando somente os tipos de instância **`t2.micro`** e,**`t2.small`** o que é possível fazer para controlar os custos.

  No entanto, observe que atualmente não há uma maneira eficaz de impedir completamente que os usuários que têm permissão para iniciar instâncias com um modelo de execução executem outros tipos de instância. Isso ocorre porque um tipo de instância especificado em um modelo de execução pode ser substituído para usar tipos de instância definidos usando a seleção de tipo de instância baseada em atributos.

Para obter uma lista completa das permissões em nível de recurso que você pode usar para controlar a configuração das instâncias que um usuário pode executar, consulte Ações, recursos e chaves de condição do Amazon EC2 na Referência de autorização do serviço.

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 do modelo de execução

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. É possível especificar essas ações adicionais no elemento `Action` de uma instrução de política do IAM.

Validação de permissões para **`ec2:RunInstances`** e **`iam:PassRole`**

Os usuários podem especificar qual versão de um modelo de execução seu grupo do Auto Scaling usa. Dependendo de suas permissões, essa pode ser uma versão numerada específica ou a versão `$Latest` ou `$Default` do modelo de execução. Se for o último, tome cuidado especial. Isso pode substituir as permissões `ec2:RunInstances` e `iam:PassRole` que você pretendia restringir.

Esta seção explica o cenário de uso da versão mais recente ou padrão do modelo de execução com um grupo do Auto Scaling.

Quando um usuário chama as APIs,`CreateAutoScalingGroup` `UpdateAutoScalingGroup` ou,`StartInstanceRefresh` o Amazon EC2 Auto Scaling verifica suas permissões em relação à versão do modelo de execução que é a versão mais recente ou padrão no momento antes de prosseguir com a solicitação. Isso valida as permissões para ações a serem concluídas ao iniciar instâncias, como as ações `ec2:RunInstances` e `iam:PassRole`. Para fazer isso, emitimos uma chamada de RunInstancesdry run do Amazon EC2 para validar se o usuário tem as permissões necessárias para a ação, sem realmente fazer a solicitação. Quando uma resposta é retornada, ela é lida pelo Amazon EC2 Auto Scaling. Se as permissões do usuário não permitirem uma determinada

ação, haverá falha na solicitação do Amazon EC2 Auto Scaling, que retornará um erro ao usuário contendo informações sobre a permissão ausente.

Depois que a verificação inicial e a solicitação forem concluídas, sempre que as instâncias forem executadas, o Amazon EC2 Auto Scaling as executará com a versão mais recente ou padrão, mesmo que ela tenha sido alterada, usando as permissões de seu perfil vinculado ao serviço. Isso significa que um usuário que esteja usando o modelo de execução pode atualizá-lo para transferir um perfil do IAM para uma instância, mesmo que não tenha a permissão`iam:PassRole`.

Use a chave de condição `autoscaling:LaunchTemplateVersionSpecified` se quiser limitar quem tem acesso à configuração de grupos para usar a versão `$Latest` ou `$Default`. Isso garante que o grupo do Auto Scaling só aceite uma versão numerada específica quando um usuário chama as APIs `CreateAutoScalingGroup` e `UpdateAutoScalingGroup`. Para ver um exemplo que mostra como adicionar essa chave de condição a uma política do IAM, consulte Exigir um modelo de execução e um número de versão.

Para grupos do Auto Scaling configurados para usar a versão `$Latest` ou `$Default` do modelo de execução, considere limitar quem pode criar e gerenciar versões do modelo de execução, incluindo a ação `ec2:ModifyLaunchTemplate` que permite ao usuário especificar a versão padrão do modelo de execução. Para obter mais informações, consulte Controlar permissões de versionamento no Guia do usuário do Amazon EC2 para instâncias do Linux.

Recursos relacionados

Para saber mais sobre permissões para visualizar, criar e excluir modelos de execução e versões de modelos de execução, consulte Controlar o acesso aos modelos de execução com permissões do IAM no Guia do usuário do Amazon EC2 para instâncias do Linux.

Para obter mais informações sobre as permissões em nível de recurso que você pode usar para controlar o acesso à chamada,`RunInstances` consulte Ações, recursos e chaves de condição do Amazon EC2 na Referência de autorização do serviço.

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 AWS recursos. Depois, crie o perfil da instância e atribua a função a ele.

> **Note**
>
> Como prática recomendada, é altamente recomendável que você crie a função para que ela tenha as permissões mínimas para outras Serviços da AWS que seu aplicativo exige.

Conteúdos

- Pré-requisitos
- Criar um modelo de inicialização
- Consulte também

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.

Se você usa o console do IAM em vez do AWS CLI ou um dos AWS SDKs, o console cria um perfil de instância automaticamente e dá a ele o mesmo nome da função à qual ele corresponde.

Para criar uma função do IAM (console)

1. Abra o console IAM em https://console.aws.amazon.com/iam/.
2. No painel de navegação à esquerda, escolha Roles (Funções).
3. Selecione 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 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 AWS recursos podem fazer referência à função, você não pode editar o nome da função após sua criação.
10. Reveja a função e escolha Criar função.

Permissões do IAM

Use uma política baseada em identidade do IAM para controlar o acesso ao novo perfil do IAM. A permissão `iam:PassRole` é necessária na identidade do IAM (usuário ou perfil) que cria ou atualiza um grupo do Auto Scaling usando um modelo de execução que especifica um perfil de instância.

O exemplo de política a seguir concede permissões para passar somente perfis do IAM cujo nome comece com **`qateam-`**.

```
{
    "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 permissões para a ação `iam:PassRole` de um grupo do Auto Scaling que usa um modelo de execução, consulte Validação de permissões para `ec2:RunInstances` e `iam:PassRole`.

Criar um modelo de inicialização

Ao criar o modelo de execução usando o AWS Management Console, na seção Detalhes avançados, selecione a função no perfil da instância do IAM. Para ter mais informações, consulte Criar um modelo de execução usando configurações avançadas.

Ao criar o modelo de execução usando o create-launch-templatecomando do AWS CLI, especifique o nome do perfil da instância da sua função do IAM, conforme mostrado no exemplo a seguir.

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

Consulte também

Para obter mais informações para começar a aprender e usar 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

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, consulteAWS Programas de conformidade.

É possível fazer download de relatórios de auditoria de terceiros usando o AWS Artifact. Para obter mais informações, consulte Downloading Reports inAWS 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 atributos para ajudar com a conformidade:

- Guias de referência rápida de conformidade e segurança - 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.
- 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 obter mais informações, consulte a Referência dos serviços qualificados pela HIPAA.

- atributos de conformidade da AWS: essa coleção de manuais e guias pode ser aplicada a seu setor e local.
- Guias de conformidade do cliente da AWS: entenda o modelo de responsabilidade compartilhada sob a ótica da conformidade. Os guias resumem as práticas recomendadas para proteção de Serviços da AWS e mapeiam as diretrizes para controles de segurança em várias estruturas (incluindo o Instituto Nacional de Padrões e Tecnologia (NIST), o Conselho de Padrões de Segurança do Setor de Cartões de Pagamento (PCI) e a Organização Internacional de Padronização (ISO)).
- Avaliar atributos com regras no AWS Config Guia do desenvolvedor: o serviço AWS Config avalia como as configurações de atributos 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 atributos 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 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

Amazon EC2 Auto Scaling e endpoints da VPC da interface

É possível melhorar o procedimento de segurança da sua VPC configurando o Amazon EC2 Auto Scaling para usar um endpoint da VPC de interface. Os endpoints de interface são alimentados por AWS PrivateLink uma tecnologia que permite que você acesse de forma privada as APIs do Amazon EC2 Auto Scaling restringindo todo o tráfego de rede entre sua VPC e o Amazon EC2 Auto Scaling à rede. AWS Com endpoints de interface, também não são necessários um gateway da Internet, um dispositivo NAT nem um gateway privado virtual.

Não é necessário configurar AWS PrivateLink, mas é recomendado. Para obter mais informações sobre AWS PrivateLink endpoints de VPC, consulte O que é? AWS PrivateLink no AWS PrivateLink Guia.

Tópicos

- Criar um VPC endpoint de interface
- Criar uma política de endpoint da VPC

Criar um VPC endpoint de interface

Crie um endpoint para o Amazon EC2 Auto Scaling usando o seguinte nome de serviço:

```
com.amazonaws.region.autoscaling
```

Para obter mais informações, consulte Acessar um AWS serviço usando uma interface VPC endpoint no Guia.AWS PrivateLink

Você não precisa alterar nenhuma configuração do Amazon EC2 Auto Scaling. O Amazon EC2 Auto Scaling chama AWS outros serviços usando endpoints de serviço ou endpoints VPC de interface privada, os que estiverem em uso.

Criar uma política de endpoint da VPC

É possível associar uma política ao seu endpoint da VPC para controlar o acesso à API do Amazon EC2 Auto Scaling. A política especifica:

- O principal que pode executar ações.
- As ações que podem ser executadas.
- O recurso no qual as ações podem ser executadas.

O exemplo a seguir mostra uma política de VPC endpoint que nega a todos permissão para excluir uma política de escalabilidade por meio do endpoint. O exemplo de política também concede a todos permissão para executar todas as outras ações.

```
{
   "Statement": [
        {
            "Action": "*",
            "Effect": "Allow",
            "Resource": "*",
            "Principal": "*"
        },
        {
            "Action": "autoscaling:DeleteScalingPolicy",
            "Effect": "Deny",
            "Resource": "*",
            "Principal": "*"
```

```
        }
    ]
}
```

Para obter mais informações, consulte Controlar o acesso aos endpoints da VPC usando políticas de endpoint no Guia AWS PrivateLink .

Solucionar problemas do Amazon EC2 Auto Scaling

O Amazon EC2 Auto Scaling fornece erros específicos e descritivos para ajudar a solucionar problemas. Você pode encontrar as mensagens de erro na descrição das ações de escalabilidade.

Tópicos

- Recuperar uma mensagem de erro de ações de escalabilidade
- Desative as atividades de escalabilidade
- Recursos adicionais para solução de problemas
- Solucionar problemas do Amazon EC2 Auto Scaling: falhas ao iniciar instâncias do EC2
- Solucionar problemas do Amazon EC2 Auto Scaling: problemas de AMI
- Solucionar problemas do Amazon EC2 Auto Scaling: problemas do balanceador de carga
- Solucionar problemas do Amazon EC2 Auto Scaling: modelos de execução
- Solucionar problemas com as verificações de integridade do Amazon EC2 Auto Scaling

Recuperar uma mensagem de erro de ações de escalabilidade

Para recuperar uma mensagem de erro da descrição das atividades de escalabilidade, use o describe-scaling-activitiescomando. Você tem um registro de atividades de escalabilidade que data de 6 semanas atrás. As ações de escalabilidade são ordenadas por hora de início, com as ações de escalabilidade mais recentes listadas primeiro.

> **Note**
>
> As ações de escalabilidade também são exibidas no histórico de atividades no console do Amazon EC2 Auto Scaling, na guia Activity (Atividades) do grupo do Auto Scaling.

Para ver as ações de escalabilidade de um grupo específico do Auto Scaling, use o comando a seguir.

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

O exemplo a seguir é de uma resposta, em que `StatusCode` contém o status atual da atividade e `StatusMessage` contém a mensagem de erro.

```
{
    "Activities": [
        {
            "ActivityId": "3b05dbf6-037c-b92f-133f-38275269dc0f",
            "AutoScalingGroupName": "my-asg",
            "Description": "Launching a new EC2 instance: i-003a5b3ffe1e9358e.  Status
 Reason: Instance failed to complete user's Lifecycle Action: Lifecycle Action with
 token e85eb647-4fe0-4909-b341-a6c42d8aba1f was abandoned: Lifecycle Action Completed
 with ABANDON Result",
            "Cause": "At 2021-01-11T00:35:52Z a user request created an
 AutoScalingGroup changing the desired capacity from 0 to 1.  At 2021-01-11T00:35:53Z
 an instance was started in response to a difference between desired and actual
 capacity, increasing the capacity from 0 to 1.",
            "StartTime": "2021-01-11T00:35:55.542Z",
            "EndTime": "2021-01-11T01:06:31Z",
            "StatusCode": "Cancelled",
            "StatusMessage": "Instance failed to complete user's Lifecycle Action:
 Lifecycle Action with token e85eb647-4fe0-4909-b341-a6c42d8aba1f was abandoned:
 Lifecycle Action Completed with ABANDON Result",
            "Progress": 100,
            "Details": "{\"Subnet ID\":\"subnet-5ea0c127\",\"Availability Zone\":\"us-
west-2b\"...}",
            "AutoScalingGroupARN": "arn:aws:autoscaling:us-
west-2:123456789012:autoScalingGroup:283179a2-
f3ce-423d-93f6-66bb518232f7:autoScalingGroupName/my-asg"
        },
     ...
    ]
}
```

Para obter uma descrição dos campos na saída, consulte Atividade na Referência da API do Amazon EC2 Auto Scaling.

Para visualizar as ações de dimensionamento para um grupo excluído

Para visualizar as atividades de escalabilidade após a exclusão do grupo Auto Scaling, adicione `--include-deleted-groups` a opção ao comando describe-scaling-activitiesda seguinte maneira.

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

O exemplo a seguir é uma resposta com uma ação de escalabilidade para um grupo excluído.

```
{
    "Activities": [
        {
            "ActivityId": "e1f5de0e-f93e-1417-34ac-092a76fba220",
            "AutoScalingGroupName": "my-asg",
            "Description": "Launching a new EC2 instance.  Status Reason: Your Spot
 request price of 0.001 is lower than the minimum required Spot request fulfillment
 price of 0.0031. Launching EC2 instance failed.",
            "Cause": "At 2021-01-13T20:47:24Z a user request update of AutoScalingGroup
 constraints to min: 1, max: 5, desired: 3 changing the desired capacity from 0 to 3.
  At 2021-01-13T20:47:27Z an instance was started in response to a difference between
 desired and actual capacity, increasing the capacity from 0 to 3.",
            "StartTime": "2021-01-13T20:47:30.094Z",
            "EndTime": "2021-01-13T20:47:30Z",
            "StatusCode": "Failed",
            "StatusMessage": "Your Spot request price of 0.001 is lower than the
 minimum required Spot request fulfillment price of 0.0031. Launching EC2 instance
 failed.",
            "Progress": 100,
            "Details": "{\"Subnet ID\":\"subnet-5ea0c127\",\"Availability Zone\":\"us-
west-2b\"...}",
            "AutoScalingGroupState": "Deleted",
            "AutoScalingGroupARN": "arn:aws:autoscaling:us-
west-2:123456789012:autoScalingGroup:283179a2-
f3ce-423d-93f6-66bb518232f7:autoScalingGroupName/my-asg"
        },
     ...
    ]
}
```

Desative as atividades de escalabilidade

Você tem as seguintes opções se precisar investigar um problema sem interferência de políticas de escalabilidade ou ações programadas:

- Evite que todas as políticas de escalabilidade e ações programadas alterem a capacidade desejada do grupo suspendendo os processos `AlarmNotification` e. `ScheduledActions` Para ter mais informações, consulte Suspender e retomar os processos do Amazon EC2 Auto Scaling.

- Desative as políticas de escalabilidade individuais para que elas não alterem a capacidade desejada do grupo em resposta às mudanças na carga. Para ter mais informações, consulte Desabilitar uma política de escalabilidade para um grupo do Auto Scaling.
- Atualize as políticas individuais de escalabilidade de rastreamento de metas para escalar apenas para fora (adicionar capacidade) desativando a parte de expansão da política. Esse método evita que a capacidade desejada do grupo diminua, mas permite que ela seja aumentada quando a carga aumenta. Para ter mais informações, consulte Políticas de escalabilidade com monitoramento do objetivo para o Amazon EC2 Auto Scaling.

Recursos adicionais para solução de problemas

As páginas a seguir apresentam mais informações para solucionar problemas com o Amazon EC2 Auto Scaling.

- Verificar uma ação de escalabilidade para um grupo do Auto Scaling
- Visualizar grafos de monitoramento no console do Amazon EC2 Auto Scaling
- Verificações de integridade para instâncias em um grupo do Auto Scaling
- Considerações e limitações dos ganchos do ciclo de vida
- Concluir uma ação do ciclo de vida
- Fornecer conectividade de rede para suas instâncias do Auto Scaling usando a Amazon VPC
- Remover temporiamente instâncias do grupo do Auto Scaling
- Desabilitar uma política de escalabilidade para um grupo do Auto Scaling
- Suspender e retomar os processos do Amazon EC2 Auto Scaling
- Controlar quais instâncias do Auto Scaling serão terminadas durante uma redução de escala na horizontal
- Excluir infraestrutura do Auto Scaling
- Cotas do Amazon EC2 Auto Scaling

Os seguintes AWS recursos também podem ajudar:

- Tópicos do Amazon EC2 Auto Scaling no Centro de conhecimento AWS
- Perguntas sobre o Amazon EC2 Auto Scaling no re:POST AWS
- Publicações do Amazon EC2 Auto Scaling no blog de computação AWS

- Solução de problemas CloudFormation no Guia AWS CloudFormation do usuário

Geralmente, a solução de problemas requer consulta e descoberta iterativas por um especialista ou de uma comunidade de ajudantes. Se você continuar enfrentando problemas depois de tentar as sugestões desta seção, entre em contato AWS Support (no, clique em Support AWS Management Console, Support Center) ou faça uma pergunta no AWS re:POST usando a tag Amazon EC2 Auto Scaling.

Solucionar problemas do Amazon EC2 Auto Scaling: falhas ao iniciar instâncias do EC2

Esta página fornece informações sobre suas instâncias do EC2 que falham ao ativar, as possíveis causas e as etapas que você pode realizar para resolver o problema.

Para recuperar uma mensagem de erro, consulte Recuperar uma mensagem de erro de ações de escalabilidade.

Quando suas instâncias EC2 falham ao ativar, você pode obter uma ou mais das seguintes mensagens de erro:

Problemas ao iniciar

- A configuração solicitada não é suportada atualmente.
- O grupo de segurança <nome do grupo de segurança> não existe. Falha ao ativar a instância EC2.
- O par de chaves <par de chaves associado à sua instância do EC2> não existe. Falha ao ativar a instância EC2.
- O tipo de instância solicitado (<tipo de instância>) não tem suporte na Zona de disponibilidade solicitada (<Zona de disponibilidade da instância>)...
- Seu preço de solicitação spot de 0,015 é inferior ao preço mínimo de atendimento de solicitação spot exigido de 0,0735...
- Nome de dispositivo inválido <nome do dispositivo> / Carregamento do nome de dispositivo inválido. Falha ao ativar a instância EC2.
- O valor (<nome associado ao dispositivo de armazenamento de instâncias>) do parâmetro virtualName é inválido... Falha ao ativar a instância EC2.
- Mapeamentos de dispositivos de blocos do EBS não suportados para AMIs de armazenamento de instância.

- Os grupos de posicionamento não podem ser usados com instâncias do tipo '<instance type>'. Falha ao ativar a instância EC2.
- Cliente. InternalError: Erro do cliente na inicialização.
- No momento, não temos capacidade de <tipo de instância> suficiente para tipo de instância na zona de disponibilidade solicitada. Falha ao ativar a instância EC2.
- A reserva solicitada não tem capacidade compatível e disponível suficiente para essa solicitação. Falha ao ativar a instância EC2.
- Sua reserva do bloco de capacidade <reservation id> ainda não está ativa. Falha ao ativar a instância EC2.
- Não há capacidade spot disponível que corresponda à sua solicitação. Falha ao ativar a instância EC2.
- <número de instâncias> instância(s) já estão em execução. Falha ao ativar a instância EC2.

A configuração solicitada não é suportada atualmente.

Causa: algumas opções em seu modelo de execução ou configuração de execução podem não ser compatíveis com o tipo de instância, ou a configuração da instância pode não ser suportada na AWS região ou nas zonas de disponibilidade solicitadas.

Solução: tente uma configuração de instância diferente. Para pesquisar um tipo de instância que atenda aos seus requisitos, consulte Como encontrar tipos de instâncias do Amazon EC2 no Manual do usuário do Amazon EC2 para 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 AWS Graviton baseado em ARM em vez de um processador Intel Xeon, você precisa de uma AMI compatível com ARM. Para obter mais informações sobre como escolher um tipo de instância compatível, consulte Compatibilidade para alterar o tipo de instância no Guia do usuário do Amazon EC2 para instâncias do Linux.
- 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 os tipos de instância oferecidos por localização (região ou zona de disponibilidade), use o describe-

instance-type-offeringscomando. 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 describe-security-groupscomando para obter a lista dos grupos de segurança associados à sua conta.
2. Na lista, selecione os security groups a serem usados. Em vez disso, para criar um grupo de segurança, use o create-security-groupcomando.
3. Crie um novo modelo de execução ou uma nova configuração de execução.
4. Atualize seu grupo de Auto Scaling com o novo modelo de lançamento ou configuração de lançamento usando o update-auto-scaling-groupcomando.

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 describe-key-pairscomando para obter a lista dos pares de chaves disponíveis para você.
2. Na lista, selecione o par de chaves a ser usado. Em vez disso, para criar um key pair, use o create-key-paircomando.
3. Crie um novo modelo de execução ou uma nova configuração de execução.
4. Atualize seu grupo de Auto Scaling com o novo modelo de lançamento ou configuração de lançamento usando o update-auto-scaling-groupcomando.

O tipo de instância solicitado (<tipo de instância>) não tem suporte na Zona de disponibilidade solicitada (<Zona de disponibilidade da instância>)...

Mensagem de erro: O tipo de instância solicitado (<tipo de instância>) não tem suporte na zona de disponibilidade solicitada (<zona de disponibilidade da instância>)…Falha na execução da instância EC2.

Causa: as zonas de disponibilidade especificadas em seu grupo do Auto Scaling não são compatíveis com o tipo de instância escolhido.

Solução:

1. Verifique quais zonas de disponibilidade oferecem suporte ao tipo de instância escolhido usando o describe-instance-type-offeringscomando ou no console do Amazon EC2 verificando o valor das zonas de disponibilidade no painel de rede da página Tipos de instância.
2. Atualize ou remova a sub-rede de qualquer zona não suportada nas configurações do seu grupo de Auto Scaling usando o comando. update-auto-scaling-group Para ter mais informações, consulte Adicionar e remover zonas de disponibilidade.

Seu preço de solicitação spot de 0,015 é inferior ao preço mínimo de atendimento de solicitação spot exigido de 0,0735...

Causa: o preço máximo spot na solicitação é inferior ao preço spot do tipo de instância que você selecionou.

Solução: envie uma nova solicitação com um preço máximo spot mais alto (possivelmente o preço sob demanda). Anteriormente, o preço spot pago era baseado em lances. Hoje, você paga o preço spot atual. Ao definir seu preço máximo mais alto, a chance do serviço spot do Amazon EC2 iniciar e manter a quantidade necessária de capacidade é maior.

Nome de dispositivo inválido <nome do dispositivo> / Carregamento do nome de dispositivo inválido. Falha ao ativar a instância 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 de Auto Scaling com o novo modelo de lançamento ou configuração de lançamento usando o update-auto-scaling-groupcomando.

O valor (<nome associado ao dispositivo de armazenamento de instâncias>) do parâmetro virtualName é inválido... 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 de Auto Scaling com o novo modelo de lançamento ou configuração de lançamento usando o update-auto-scaling-groupcomando.

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 de Auto Scaling com o novo modelo de lançamento ou configuração de lançamento usando o update-auto-scaling-groupcomando.

Os grupos de posicionamento não podem ser usados com instâncias do tipo '<instance type>'. 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 de Auto Scaling com um novo grupo de posicionamento, modelo de lançamento ou configuração de lançamento usando o update-auto-scaling-groupcomando.

Cliente. InternalError: Erro do cliente na inicialização.

Problema: O Amazon EC2 Auto Scaling tenta iniciar uma instância que tem um volume do EBS criptografado, mas a função vinculada ao serviço não tem acesso à chave gerenciada AWS KMS

pelo cliente usada para criptografá-la. Para ter mais informações, consulte Política de AWS KMS chaves necessária para uso com volumes criptografados.

Causa 1: você precisa de uma política de chaves que conceda permissão para usar a chave gerenciada pelo cliente para a função vinculada ao serviço adequada.

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.

Causa 2: Se a chave gerenciada pelo cliente e o grupo Auto Scaling estiverem em AWS contas diferentes, você precisará configurar o acesso entre contas à chave gerenciada pelo cliente para dar permissão para usar a chave gerenciada pelo cliente para a função vinculada ao serviço adequada.

Solução 2: permita que 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 acesse 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 ter mais informações, consulte Exemplo 2: seções da política de chaves que permitem acesso entre contas à chave gerenciada pelo cliente.

Solução 3: Use uma chave gerenciada pelo cliente na mesma conta da AWS que o grupo do Auto Scaling.

1. Copie e criptografe novamente o snapshot com outra chave gerenciada pelo cliente pertencente à mesma conta que o grupo do Auto Scaling.
2. Permita que 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, a combinação do tipo de instância solicitada e da zona de disponibilidade não é compatível.

Solução: Para resolver o problema, tente o seguinte:

- Aguarde alguns minutos para que o Amazon EC2 Auto Scaling tenha capacidade para esse tipo de instância em outras zonas de disponibilidade habilitadas.
- Expanda seu grupo do Auto Scaling para zonas de disponibilidade adicionais. Para ter mais informações, consulte Adicionar e remover zonas de disponibilidade.
- Siga as práticas recomendadas de uso de um conjunto diversificado de tipos de instância para não depender de um tipo de instância específico. Para ter mais informações, consulte Grupos do Auto Scaling com vários tipos de instâncias e opções de compra.

A reserva solicitada não tem capacidade compatível e disponível suficiente para essa solicitação. Falha ao ativar a instância EC2.

Causa 1: você atingiu o limite do número total de instâncias que pode executar com uma reserva de capacidade `targeted` sob demanda.

Solução 1: aumente o número de instâncias que você pode executar com a reserva de capacidade `targeted` sob demanda ou use um grupo de reservas de capacidade para que qualquer coisa além da capacidade reservada seja executada como capacidade sob demanda regular. Para ter mais informações, consulte Use reservas de capacidade sob demanda para reservar capacidade em zonas de disponibilidade específicas.

Causa 2: você atingiu o limite do número total de instâncias que pode executar com um bloco de capacidade.

Com os Blocos de Capacidade, você fica limitado pela quantidade de capacidade adquirida originalmente. Se você observar um número de inicializações superior ao previsto e usar toda a capacidade disponível, isso causará falha nas inicializações. As instâncias encerradas passam por um longo processo de limpeza antes de serem totalmente encerradas. Durante esse período, elas não podem ser reutilizadas. Isso também pode causar falha nas inicializações. Para ter mais informações, consulte Use blocos de capacidade para cargas de trabalho de aprendizado de máquina.

Solução 2: Para resolver o problema, tente o seguinte:

- Mantenha a solicitação como está. Se uma instância do Capacity Block estiver sendo encerrada, você deverá esperar alguns minutos para que a instância termine e a capacidade fique disponível novamente. O Amazon EC2 Auto Scaling continuará a fazer a solicitação de execução automaticamente até que a capacidade seja disponibilizada.
- Certifique-se de adquirir capacidade suficiente para acomodar sua workload de pico, para que você não encontre esse erro com frequência.

Sua reserva do bloco de capacidade <reservation id> ainda não está ativa. Falha ao ativar a instância EC2.

Causa: O bloco de capacidade especificado ainda não está ativo.

Solução: siga a abordagem recomendada para blocos de capacidade e use a escalabilidade programada. Isso ajuda a garantir que você aumente a capacidade desejada do grupo do Auto Scaling somente quando a reserva estiver ativa e a diminua antes que a reserva termine.

Não há capacidade spot disponível que corresponda à sua solicitação. Falha ao ativar a instância EC2.

Causa: no momento, não há capacidade de reserva suficiente para atender à sua solicitação de instâncias spot.

Solução: Para resolver o problema, tente o seguinte:

- Aguarde alguns minutos; a capacidade pode mudar com frequência. O Amazon EC2 Auto Scaling continuará a fazer a solicitação de execução automaticamente até que a capacidade seja disponibilizada.
- Expanda seu grupo do Auto Scaling para zonas de disponibilidade adicionais. Para ter mais informações, consulte Adicionar e remover zonas de disponibilidade.
- Siga as práticas recomendadas de uso de um conjunto diversificado de tipos de instância para não depender de um tipo de instância específico. Para ter mais informações, consulte Grupos do Auto Scaling com vários tipos de instâncias e opções de compra.

<número de instâncias> instância(s) já estão em execução. Falha ao ativar a instância EC2.

Causa: você atingiu o limite do número total de instâncias que pode iniciar em uma região. Quando você cria sua AWS conta, definimos limites padrão para o número de instâncias que você pode executar por região.

Solução: Para resolver o problema, tente 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.

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.
- A AMI <ID da AMI> está pendente e não pode ser executada. Falha ao ativar a instância EC2.
- Nome do dispositivo inválido <device name>. Falha ao ativar a instância EC2.
- A arquitetura 'arm64' do tipo de instância especificado não corresponde à arquitetura 'x86_64' da AMI especificada... Falha na execução da instância EC2.
- A AMI '<AMI ID>' está desabilitada e não pode ser executada. Falha ao ativar a instância EC2.

> **Important**
>
> AWS suporta o compartilhamento privado de uma AMI com outra AWS conta modificando as permissões da AMI. Se uma AMI se tornar privada sem ser compartilhada, isso pode resultar em um erro de autorização ao iniciar novas instâncias. Para obter mais informações sobre o compartilhamento de AMIs privadas, consulte Compartilhar uma AMI com AWS contas específicas no Guia do usuário do Amazon EC2 para instâncias Linux.

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 de Auto Scaling com o novo modelo de lançamento ou configuração de lançamento usando o update-auto-scaling-groupcomando.

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.

Nome do dispositivo inválido <device name>. Falha ao ativar a instância EC2.

Causa: Ao conectar um volume do EBS a uma instância do EC2, você deve fornecer um nome de dispositivo válido para o volume. A AMI selecionada deve ser compatível com esse nome de dispositivo.

Solução:

1. Crie um novo modelo de inicialização ou configuração de inicialização e especifique o nome de dispositivo correto para sua AMI. A convenção de nomenclatura recomendada varia de acordo com o tipo de virtualização da AMI. Para obter mais informações, consulte Nomes de dispositivos no Guia do usuário do Amazon EC2 para instâncias do Linux.
2. Atualize seu grupo de Auto Scaling com o novo modelo de lançamento ou configuração de lançamento usando o update-auto-scaling-groupcomando.

A arquitetura 'arm64' do tipo de instância especificado não corresponde à arquitetura 'x86_64' da AMI especificada... Falha na execução da instância EC2.

Causa 1: Se a arquitetura da AMI e o tipo de instância usado em seu modelo de execução ou configuração de execução não forem os mesmos, você receberá um erro quando o Amazon EC2 Auto Scaling tentar iniciar uma instância usando a configuração de instância incompatível.

Solução 1:

1. Verifique a arquitetura da sua AMI usando o comando describe-images ou no console do Amazon EC2 verificando o valor da arquitetura no painel de detalhes da página Amazon Machine Images (AMIs).
2. Encontre um tipo de instância que tenha a mesma arquitetura da sua AMI usando o describe-instance-typescomando ou no console do Amazon EC2 verificando a coluna Arquitetura na tela Tipos de instância. Para obter mais informações sobre como escolher um tipo de instância compatível, consulte Compatibilidade para alterar o tipo de instância no Guia do usuário do Amazon EC2 para instâncias do Linux.
3. Crie um novo modelo ou uma nova configuração de execução usando um tipo de instância que tenha a mesma arquitetura da sua AMI.
4. Atualize seu grupo de Auto Scaling com o novo modelo de lançamento ou configuração de lançamento usando o update-auto-scaling-groupcomando.

Causa 2: O Amazon EC2 Auto Scaling tenta iniciar um tipo de instância especificado na política de instâncias mistas do seu grupo do Auto Scaling, mas o tipo de instância não tem a mesma arquitetura da AMI especificada em seu modelo de execução.

Solução 1: não inclua tipos de instância que tenham arquiteturas diferentes em sua política de instâncias mistas.

1. Verifique a arquitetura da sua AMI usando o comando describe-images ou no console do Amazon EC2 verificando o valor da arquitetura no painel de detalhes da página Amazon Machine Images (AMIs).
2. Verifique a arquitetura de cada tipo de instância que você pretende incluir em sua política de instâncias mistas usando o describe-instance-typescomando ou a partir do console do Amazon EC2, verificando a coluna Arquitetura na tela Tipos de instância. Para obter informações sobre como escolher um tipo de instância compatível, consulte Compatibilidade para alterar o tipo de instância no Guia do usuário do Amazon EC2 para instâncias do Linux.
3. Atualize ou remova os tipos de instância incompatíveis do seu grupo de Auto Scaling usando o update-auto-scaling-groupcomando.

Solução 2: para iniciar instâncias Arm (Graviton2) e x86_64 (Intel) no mesmo grupo do Auto Scaling, você deve usar modelos de execução que ofereçam suporte a uma AMI compatível com ARM e uma AMI compatível com Intel x86, respectivamente, para corresponder aos tipos de instância em sua política de instâncias mistas.

1. Verifique a arquitetura da AMI no seu modelo de execução existente usando o comando describe-images ou do console do Amazon EC2 verificando o valor da arquitetura no painel de detalhes da página Amazon Machine Images (AMIs).
2. Crie um novo modelo de execução usando uma AMI que corresponda à outra arquitetura que você pretende usar.
3. Atualize seu grupo de Auto Scaling para substituir o modelo de execução existente e especificar o novo modelo de execução para cada tipo de instância compatível usando o comando. update-auto-scaling-group Para ter mais informações, consulte Usar um modelo de execução diferente para um tipo de instância.

A AMI '<AMI ID>' está desabilitada e não pode ser executada. Falha ao ativar a instância EC2.

Causa: você está tentando executar instâncias de uma AMI que foi desabilitada. Para obter mais informações, consulte Desabilitar uma AMI no Guia do usuário do Amazon EC2 para instâncias do Linux.

Solução:

1. Crie um novo modelo ou uma nova configuração de execução e especifique uma AMI que não esteja desabilitada.
2. Atualize seu grupo de Auto Scaling com o novo modelo de lançamento ou configuração de lançamento usando o update-auto-scaling-groupcomando.

Solucionar problemas do Amazon EC2 Auto Scaling: problemas do balanceador de carga

Esta página fornece informações sobre os problemas causados pelo balanceador de carga associados a seu grupo do Auto Scaling, as possíveis causas e as etapas que você pode realizar para resolver os problemas.

Para recuperar uma mensagem de erro, consulte Recuperar uma mensagem de erro de ações de escalabilidade.

Quando houver falha ao iniciar suas instâncias EC2 devido a problemas com o balanceador de carga associados a seu grupo do Auto Scaling, você poderá receber uma ou mais das seguintes mensagens de erro.

Problemas do balanceador de carga

- Um ou mais grupos de destino não encontrados. Falha na validação da configuração do balanceador de carga.
- Não é possível encontrar o Load Balancer <seu load balancer>. Falha na validação da configuração do balanceador de carga.
- Não há nenhum balanceador de carga ATIVO chamado <nome do balanceador de carga>. Falha ao atualizar a configuração do balanceador de carga.
- A instância do EC2 <ID da instância> não está na VPC. Falha ao atualizar a configuração do balanceador de carga.

> **Note**
>
> Você pode usar o Reachability Analyzer para solucionar problemas de conectividade verificando se as instâncias do seu grupo do Auto Scaling podem ser acessadas por meio do balanceador de carga. Para saber mais sobre os diferentes problemas de configuração incorreta de rede que são automaticamente detectados pelo Reachability Analyzer, consulte Reachability Analyzer no Guia do usuário do Reachability Analyzer.

Um ou mais grupos de destino não encontrados. Falha na validação da configuração do balanceador de carga.

Problema: quando seu grupo do Auto Scaling inicia instâncias, o Amazon EC2 Auto Scaling tenta validar a existência dos recursos do Elastic Load Balancing associados ao grupo do Auto Scaling. Quando um grupo-alvo não pode ser encontrado, a atividade de escalabilidade falha e você obtém o erro `One or more target groups not found. Validating load balancer configuration failed.`.

Causa 1: um grupo-alvo vinculado ao seu grupo do Auto Scaling foi excluído.

Solução 1: Você pode criar um novo grupo de Auto Scaling sem o grupo-alvo ou remover o grupo-alvo não utilizado do grupo Auto Scaling usando o console do Amazon EC2 Auto Scaling ou o comando -groups. detach-load-balancer-target

Causa 2: O grupo-alvo existe, mas houve um problema ao tentar especificar o ARN do grupo-alvo ao criar o grupo doAuto Scaling. Os recursos não estão criados na ordem correta.

Solução 2: crie um novo grupo do Auto Scaling e especifique o nome do balanceador de carga no final.

Não é possível encontrar o Load Balancer <seu load balancer>. Falha na validação da configuração do balanceador de carga.

Problema: quando seu grupo do Auto Scaling inicia instâncias, o Amazon EC2 Auto Scaling tenta validar a existência dos recursos do Elastic Load Balancing associados ao grupo do Auto Scaling. Quando não é possível encontrar um Classic Load Balancer, a atividade de escalabilidade falha e você recebe o erro `Cannot find Load Balancer <your load balancer>. Validating load balancer configuration failed.`.

Causa 1: o Classic Load Balancer foi excluído.

Solução 1: Você pode criar um novo grupo de Auto Scaling sem o balanceador de carga ou remover o balanceador de carga não utilizado do grupo Auto Scaling usando o console do Amazon EC2 Auto Scaling ou o comando. detach-load-balancers

Causa 2: o Classic Load Balancer existe, mas houve um problema ao tentar especificar o nome do balanceador de carga durante a criação do grupo do Auto Scaling. Os recursos não estão criados na ordem correta.

Solução 2: crie um novo grupo do Auto Scaling e especifique o nome do balanceador de carga no final.

Não há nenhum balanceador de carga ATIVO chamado <nome do balanceador de carga>. Falha ao atualizar a configuração do balanceador de carga.

Causa: O balanceador de carga especificado pode ter sido excluído.

Solução: você pode criar um novo balanceador de carga e, em seguida, criar um novo grupo do Auto Scaling ou criar um novo grupo do Auto Scaling sem o balanceador de carga.

A instância do EC2 <ID da instância> não está na VPC. Falha ao atualizar a configuração do balanceador de carga.

Causa: A instância especificada não existe na VPC.

Solution: você pode excluir o balanceador de carga associado à instância ou criar um novo grupo do Auto Scaling.

Solucionar problemas do Amazon EC2 Auto Scaling: modelos de execução

Use as informações a seguir para ajudar a diagnosticar e corrigir problemas comuns que podem ser encontrados ao tentar especificar um modelo de inicialização para o grupo do Auto Scaling.

Não é possível iniciar instâncias

Se você não conseguir iniciar instâncias com um modelo de inicialização já especificado, verifique o seguinte para a solução de problemas em geral: Solucionar problemas do Amazon EC2 Auto Scaling: falhas ao iniciar instâncias do EC2.

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 ter mais informações, consulte Criar um modelo de execução para um grupo do Auto Scaling.

Você não está autorizado a usar o modelo de execução (permissões insuficientes)

Problema: quando você tenta especificar um modelo de inicialização para um grupo do Auto Scaling, recebe o erro `You are not authorized to use launch template` (Você não está autorizado a usar o modelo de inicialização).

Causa 1: se você estiver tentando usar um modelo de execução e as credenciais do IAM não tiverem permissões suficientes, você receberá um erro informando que não está autorizado a usar o modelo de execução.

Solução 1: Para resolver o problema, tente o seguinte:

- Verifique se as credenciais do IAM que você está usando para fazer a solicitação têm permissões para chamar as ações de API do EC2 necessárias, incluindo a ação`ec2:RunInstances`. Se você especificou qualquer tag no seu modelo de execução, também deverá ter permissão para usar a ação`ec2:CreateTags`.

- Como alternativa, verifique se as credenciais do IAM que você está usando para fazer a solicitação estão atribuídas à política `AmazonEC2FullAccess`. Essa política AWS gerenciada concede acesso total a todos os recursos do Amazon EC2 e serviços relacionados, incluindo Amazon EC2 Auto Scaling e Elastic CloudWatch Load Balancing.

Para obter mais informações sobre as permissões necessárias para usar modelos de execução, incluindo exemplos de políticas do IAM, consulte Controlar o acesso aos modelos de execução com permissões do IAM no Guia do usuário do Amazon EC2 para instâncias do Linux. Para obter exemplos de políticas do IAM, consulte Suporte a modelo de execução.

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 2: verifique se as credenciais do IAM que você está usando para fazer a solicitação têm a `iam:PassRole` permissão correta para transmitir 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. Para obter mais tópicos de solução de problemas relacionados a perfis de instância, consulte Solução de problemas do Amazon EC2 e IAM no Manual do usuário do IAM.

Causa 3: Se você estiver tentando usar um modelo de execução que especifica uma AMI em outra Conta da AWS, e a AMI é privada e não é compartilhada com a pessoa que Conta da AWS você está usando, você recebe um erro informando que não está autorizado a usar o modelo de execução.

Solução 3: verifique se as permissões na AMI incluem a conta que você está usando. Para obter mais informações, consulte Share an AMI with specific Contas da AWS (Compartilhar uma AMI com específico) no Guia do usuário do Amazon EC2 para instâncias do Linux.

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.

Problemas de verificação de integridade

- Uma instância foi retirada de serviço em resposta a uma falha de verificação de status de instância do EC2
- Uma instância foi retirada de serviço em resposta a uma reinicialização programada do EC2
- Uma instância foi retirada de serviço em resposta a uma verificação de integridade do EC2 que indicou que ela tinha sido terminada ou interrompida
- Uma instância foi retirada de serviço em resposta a uma falha na verificação de integridade do sistema ELB

> **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 ter mais informações, consulte Opções de notificação do Amazon SNS para o Amazon EC2 Auto Scaling.
>
> 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 lançadas em um grupo de Auto Scaling exigem tempo de aquecimento suficiente (período de carência) para evitar a rescisão antecipada devido à substituição do exame de saúde. Para ter mais informações, consulte Definir um período de carência da verificação de integridade para um grupo do Auto Scaling.

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

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 ter mais informações, consulte Suspender e retomar os processos do Amazon EC2 Auto Scaling.

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 reinicializar as instâncias no seu grupo do Auto Scaling, recomendamos colocar as instâncias em espera primeiro. Para ter mais informações, consulte Remover temporiamente instâncias do grupo do Auto Scaling.

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 ter mais informações, consulte Grupos do Auto Scaling com vários tipos de instâncias e opções de compra.
- Se você usar vários tipos de instâncias, considere ativar o recurso de rebalanceamento de capacidade. Ele será útil se você quiser que o serviço spot do Amazon EC2 tente iniciar uma nova

instância spot antes que uma instância em execução seja encerrada. Para ter mais informações, consulte Usar o rebalanceamento de capacidade para lidar com interrupções de spot do Amazon EC2.

Causa 3: Com os blocos de capacidade, o Amazon EC2 encerra todas as instâncias que ainda estão em execução 30 minutos antes do horário final do bloco de capacidade. Esse encerramento abrupto faz com que seu grupo de Auto Scaling tente iniciar novas instâncias para manter a capacidade desejada, mesmo quando o bloco de capacidade estiver terminando.

Solução 3: Para resolver esse problema, tente o seguinte:

- Diminua a capacidade desejada do grupo Auto Scaling para evitar que ele tente iniciar novas instâncias. Para ter mais informações, consulte Escalabilidade manual para o Amazon EC2 Auto Scaling.
- Certifique-se de escalar seu grupo de Auto Scaling 30 minutos antes do horário de término do bloco de capacidade para que você não encontre esse erro com frequência. Certifique-se de que todos os ganchos do ciclo de vida tenham sido concluídos 30 minutos antes do término do bloco de capacidade. Para ter mais informações, consulte Use blocos de capacidade para cargas de trabalho de aprendizado de máquina.

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 ter mais informações, consulte Verificações de integridade para instâncias em um grupo do Auto Scaling.

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 gancho 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 balanceadores 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 do Amazon EC2 Auto Scaling	A documentação de cada operação de API mostra os parâmetros de solicitação e a resposta XML e fornece links para tópicos de referência do SDK específicos da linguagem.
escalonamento automático na AWS CLI referência de comando	Descrições dos AWS CLI comandos que você pode usar para trabalhar com grupos do Auto Scaling.
Referência do cmdlet do AWS Tools for PowerShell	As AWS Ferramentas PowerShell permitem que você crie scripts de operações em seus AWS recursos a partir da linha de PowerShell comando.
Criar um grupo do Auto Scaling com AWS CloudFormation	O recurso AWS::AutoScaling::AutoScalingAWS::AutoScaling::AutoScalingGroup permite criar, modelar e gerenciar seus grupos de Auto Scaling sem ações manuais.
Endpoints e cotas do Amazon EC2 Auto Scaling no Referência geral da AWS	Informações sobre regiões e endpoints do Amazon EC2 Auto Scaling.
Páginas do produtos	A principal página da Web para obter informaçõ es sobre o Amazon EC2 Auto Scaling.
AWS re:Post	Serviço gerenciado de perguntas e respostas da AWS que oferece respostas obtidas coletivamente e revisadas por especialistas para suas perguntas técnicas.

Recurso	Descrição
Crie uma AMI no Guia do usuário do Amazon EC2 para instâncias do Linux	Aprenda como criar uma imagem de máquina da Amazon (AMI) a partir de uma instância personalizada.
Como se conectar à sua instância Linux no 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 no Guia do usuário do Amazon EC2 para instâncias do Windows	Saiba como se conectar às instâncias do Windows que você inicia.
Criação de um alarme de cobrança para monitorar suas AWS cobranças estimadas no Guia do CloudWatch usuário da Amazon	Saiba como monitorar suas cobranças estimadas usando CloudWatch.
Guia do usuário do Application Auto Scaling	Saiba como configurar o auto scaling para recursos escaláveis para a Amazon Web Services além do Amazon EC2.

Os seguintes recursos gerais estão disponíveis para ajudar você a saber mais sobre a 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 — Siga os tutoriais para iniciar seu step-by-step primeiro aplicativo no. AWS

- 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 principal página da web com informações sobre AWS Support um one-on-one canal de suporte de resposta rápida para ajudá-lo 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
Atualização de segurança do IAM	A política AutoScalingServiceRolePolicygerenciada agora concede permissões adicionais ao Amazon EC2 (`ec2:GetSecurityGroupsForVpc` `eec2:GetInstanceTypesFromInstanceRequirements`).	29 de fevereiro de 2024
Hibernação de piscina quente suportada em adicionais Regiões da AWS	Agora você pode hibernar instâncias em um pool aquecido em duas regiões adicionais: AWS GovCloud (Leste dos EUA) e AWS GovCloud (Oeste dos EUA). Para obter mais informações sobre grupos de alta atividade , consulte Grupos de alta atividade para o Amazon EC2 Auto Scaling no Guia do usuário do Amazon EC2 Auto Scaling.	26 de fevereiro de 2024
Hibernação de piscina quente suportada em adicionais Regiões da AWS	Agora você pode hibernar instâncias em uma piscina aquecida em duas regiões adicionais: Europa (Zurique) e Oriente Médio (Emirados Árabes Unidos). Para obter	21 de fevereiro de 2024

	mais informações sobre grupos de alta atividade , consulte Grupos de alta atividade para o Amazon EC2 Auto Scaling no Guia do usuário do Amazon EC2 Auto Scaling.	
Support para uso de parâmetros entre contas	Agora você pode usar um AWS Systems Manager parâmetro compartilhado de outro Conta da AWS com o Amazon EC2 Auto Scaling. Para obter mais informações, consulte Usar AWS Systems Manager parâmetros em vez de IDs de AMI em modelos de execução no Guia do usuário do Amazon EC2 Auto Scaling.	21 de fevereiro de 2024
Nova opção de proteção de preço spot	Agora você pode definir seu limite de proteção de preço para instâncias spot como uma porcentagem do preço sob demanda ao usar a seleção de tipo de instância baseada em atributos. Para obter mais informaçõ es, consulte Proteção de preços no Guia do usuário do Amazon EC2 Auto Scaling.	29 de janeiro de 2024

Políticas de manutenção de instância	Agora você pode usar uma política de manutenção de instâncias para definir se as instâncias são executadas antes ou depois do encerramento das instâncias existentes durante eventos que fazem com que suas instâncias sejam substituídas, incluindo uma atualização da instância. Para obter mais informações, consulte Políticas de manutenção de instância no Guia do usuário do Amazon EC2 Auto Scaling.	15 de novembro de 2023
Blocos de capacidade para ML	Agora você pode executar instâncias em um bloco de capacidade especificando o ID de reserva do bloco de capacidade ao criar um modelo de execução. Com os blocos de capacidade, você pode reservar instâncias de GPU em uma data futura para apoiar suas cargas de trabalho de machine learning de curta duração (ML). Para obter mais informações, consulte Use blocos de capacidade para cargas de trabalho de aprendizado de máquina no Guia do usuário do Amazon EC2 Auto Scaling.	31 de outubro de 2023

Novos recursos de atualização de instância	Agora você pode configurar uma atualização de instância para definir seu status como falha e, opcionalmente, reverter quando detectar que um CloudWatch alarme especificado entrou no estado. `ALARM` Para obter mais informações, consulte Desfazer alterações com uma reversão no Guia do usuário do Amazon EC2 Auto Scaling.	31 de julho de 2023
Alterações do guia	Um novo tópico sobre a execução de instâncias sob demanda nas reservas de capacidade foi adicionado ao guia. Para obter mais informações, consulte Use reservas de capacidade sob demanda para reservar capacidade em zonas de disponibilidade específicas no guia do usuário de Amazon EC2 Auto Scaling.	28 de julho de 2023

Alterações do guia	Um novo tópico sobre como migrar suas AWS CloudFormation pilhas das configurações de lançamento para os modelos de lançamento foi adicionado ao guia. Para obter mais informações, consulte Migre AWS CloudFormation Stacks das configurações de lançamento para os modelos de lançamento no Guia do usuário do Amazon EC2 Auto Scaling.	18 de abril de 2023

Suporte para novas operações de API	Esta versão adiciona três novas operações de API,,`AttachTrafficSources` `DetachTrafficSources` e `DescribeTrafficSources` . Além disso, um novo campo `TrafficSources` foi adicionado aos resultados das `DescribeAutoScalingGroups` operações. Um novo status de atividade ,,`WaitingForConnectionDraining` foi adicionado aos resultados das operações `DescribeScalingActivities` . O Amazon EC2 Auto Scaling também oferece suporte a um novo valor,,`VPC_LATTICE` para o campo `HealthCheckType` nas operações,`CreateAutoScalingGroup` `UpdateAutoScalingGroup` e `DescribeAutoScalingGroups` . Para obter mais informações, consulte a Referência da API do Amazon EC2 Auto Scaling.	31 de março de 2023

Suporte para o Amazon VPC Lattice	É a versão de disponibilidade geral do VPC Lattice para o Amazon EC2 Auto Scaling. Para obter mais informaçõ es, consulte o tráfego de rota para o seu grupo de escala automático com um grupo de destino de treliça VPC no guia do usuário do Amazon EC2 Auto Scaling.	31 de março de 2023
Alterações do guia	A seção com AWS CLI exemplos para trabalhar com o Elastic Load Balancing agora inclui exemplos novos e atualizados. Para obter mais informações, consulte Exemplos de como trabalhar com o Elastic Load Balancing com o AWS Command Line Interface (AWS CLI) no Guia do usuário do Amazon EC2 Auto Scaling.	31 de março de 2023
Support para escalabilidade preditiva, além de Regiões da AWS	Agora você pode criar políticas de escalabilidade preditiva nas regiões do Oriente Médio (EAU) e AWS GovCloud (Leste dos EUA). Para obter mais informações, consulte Escalabilidade predtiva o Amazon EC2 Auto Scaling no Manual do usuário do Amazon EC2 Auto Scaling.	16 de março de 2023

Novos recursos de atualização de instância	Agora você pode optar por terminar ou ignorar instâncias em espera e substituir ou ignorar instâncias protegidas de redução da escala horizontalmente, em vez de esperar que elas se tornem substituíveis. Também é possível reverter as alterações de uma atualização de instância com falha. Como parte da atualização, a documentação foi expandida para incluir tópicos sobre reverter uma atualização de instância, cancelar uma atualização de instância e entender os valores padrão dos parâmetros configuráveis de uma atualização de instância. Para obter mais informações, consulte Substituir instâncias do Auto Scaling com base na atualização de uma instância no Manual do usuário do Amazon EC2 Auto Scaling.	10 de fevereiro de 2023

Support para usar um AWS Systems Manager parâmetro para uma ID de AMI	Agora você pode usar um parâmetro do Systems Manager em vez de um ID de AMI no modelo de execução. Para obter mais informaçõ es, consulte Usar parâmetro s do AWS Systems Manager em vez de IDs de AMI em modelos de execução no Guia do usuário do Amazon EC2 Auto Scaling.	19 de janeiro de 2023
Recomendações de escalabil idade preditiva	Já é possível obter recomenda ções para avaliar e escolher a política de escalabilidade preditiva correta no console do Amazon EC2 Auto Scaling. Para obter mais informações, consulte Avaliar políticas de escalabilidade preditiva no Guia do usuário do Amazon EC2 Auto Scaling.	18 de janeiro de 2023
Previsões de escalabilidade preditiva	As previsões geradas pelo dimensionamento preditivo agora são atualizadas a cada seis horas, em vez de diariamente. Para obter mais informações, consulte Escalabilidade predtiva o Amazon EC2 Auto Scaling no Manual do usuário do Amazon EC2 Auto Scaling.	6 de janeiro de 2023

Support para matemática CloudWatch métrica	Agora você pode usar a matemática métrica ao criar políticas de dimensionamento de rastreamento de destino. Com a matemática métrica, você pode consultar várias CloudWatch métricas e usar expressões matemáticas para criar novas séries temporais com base nessas métricas. Para obter mais informações, consulte Escalabilidade para o Amazon EC2 Auto Scaling usando a escalabilidade para o Amazon EC2 Auto Scaling usando matemática de escalabilidade para o Amazon EC2 Auto Scaling no Guia do usuário do Amazon EC2 Auto Scaling.	8 de dezembro de 2022
Atualizar permissões de função vinculada a serviços do IAM	A política `AutoScalingServiceRolePolicy` concede permissões adicionais ao Amazon EC2 Auto Scaling. Para obter mais informações, consulte Políticas gerenciadas pela AWS para o Amazon EC2 Auto Scaling no Guia do usuário do Amazon EC2 Auto Scaling.	6 de dezembro de 2022

Nova estratégia de alocação spot	Agora você pode usar a estratégia de alocação otimizada de preço e capacidade para solicitar instâncias spot dos pools spot com menor probabilidade de interrupção e com o preço mais baixo possível. Para obter mais informações, consulte Allocation strategies (Estratégias de alocação) no Guia do usuário do Amazon EC2 Auto Scaling.	10 de novembro de 2022
Compatibilidade com escalação preditiva na região Ásia-Pacífico (Jacarta)	Agora você pode criar políticas de escalação preditiva na Região Ásia-Pacífico (Jacarta). Para obter mais informações, consulte Escalabilidade predtiva o Amazon EC2 Auto Scaling no Manual do usuário do Amazon EC2 Auto Scaling.	13 de outubro de 2022
Compatibilidade com métricas personalizadas para escalação preditiva no console	Agora você pode usar métricas personalizadas ao criar políticas de escalação preditiva no console do Amazon EC2 Auto Scaling. Para obter mais informações, consulte Escalabilidade predtiva o Amazon EC2 Auto Scaling no Manual do usuário do Amazon EC2 Auto Scaling.	13 de outubro de 2022

CloudWatch monitoramento de métricas de escalabilidade preditiva	Agora você pode acessar dados de monitoramento para escalabilidade preditiva usando. CloudWatch Isso permite que você use a matemática métrica para criar novas séries temporais que exibem a precisão dos dados de previsão. Para obter mais informações, consulte Monitore métricas de escalabilidade preditiva CloudWatch no Guia do usuário do Amazon EC2 Auto Scaling.	7 de julho de 2022
Compatibilidade com escalação preditiva na região Ásia-Pacífico (Osaka)	Agora você pode criar políticas de escalação preditiva na Região Ásia-Pacífico (Osaka). Para obter mais informaçõ es, consulte Escalabilidade predtiva o Amazon EC2 Auto Scaling no Manual do usuário do Amazon EC2 Auto Scaling.	6 de julho de 2022

Compatibilidade com hibernação de grupo de alta atividade passa a ser oferecida em regiões adicionais	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 Grupos de alta atividade para o Amazon EC2 Auto Scaling no Guia do usuário do Amazon EC2 Auto Scaling.	5 de julho de 2022
Atualização das verificações de integridade	Ao executar verificações de integridade, agora o Amazon EC2 Auto Scaling ajuda a minimizar qualquer tempo de inatividade que possa ocorrer devido a problemas temporários ou verificações de integridade mal configuradas. Para obter mais informações, consulte Como o Amazon EC2 Auto Scaling minimiza o tempo de inatividade no Guia do usuário do Amazon EC2 Auto Scaling.	21 de maio de 2022

Aquecimento de instância padrão	Agora você pode unificar todas as configurações de aquecimento e resfriamento de um grupo de Auto Scaling e otimizar o desempenho das políticas de escalabilidade que escalam continuamente ativando o aquecimento padrão da instância. Para obter mais informações, consulte Set the default instance warmup for an Auto Scaling group (Definir o aquecimento de instância padrão para um grupo do Auto Scaling) no Guia do usuário do Amazon EC2 Auto Scaling.	19 de abril de 2022
Alterações do guia	Um novo capítulo sobre integração com outros AWS serviços foi adicionado ao guia. Para obter mais informações, consulte AWS serviços integrados ao Amazon EC2 Auto Scaling no Guia do usuário do Amazon EC2 Auto Scaling.	29 de março de 2022

Atualizar permissões de função vinculada a serviços do IAM	A política `AutoScalingServiceRolePolicy` concede permissões de leitura adicionais ao Amazon EC2 Auto Scaling. Para obter mais informações, consulte Políticas gerenciadas pela AWS para o Amazon EC2 Auto Scaling no Guia do usuário do Amazon EC2 Auto Scaling.	28 de março de 2022
Os metadados da instância fornecem o estado de destino do ciclo de vida	É possível recuperar o estado de destino do ciclo de vida de uma instância do Auto Scaling nos metadados de instância . Para mais informações, consulte Retrieve the target lifecycle state through instance metadata (Recuperar o estado de destino do ciclo de vida por meio de metadados da instância) no Guia do usuário do Amazon EC2 Auto Scaling.	24 de março de 2022

Suporte à nova funcionalidade de grupo de alta atividade	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 Grupos de alta atividade para o Amazon EC2 Auto Scaling no Manual do usuário do Amazon EC2 Auto Scaling.	24 de fevereiro de 2022
Alterações do guia	O console do Amazon EC2 Auto Scaling foi atualizado com outras opções para ajudar você a iniciar uma atualização de instância com ignorar correspondência habilitado e uma configuração desejada especificada. Para obter mais informações, consulte Iniciar ou cancelar uma atualização de instância (console) no Guia do usuário do Amazon EC2 Auto Scaling.	3 de fevereiro de 2022

Métricas personalizadas para políticas de escalabilidade preditiva	Agora, você pode escolher se deseja usar métricas personalizadas ao criar políticas de escalabilidade preditiva. Também é possível usar a métrica matemática para personalizar ainda mais as métricas incluídas na política. Para mais informações, consulte Advanced predictive scaling policy configurations using custom metrics (Configurações avançadas de política de escalabilidade preditiva usando métricas personalizadas).	24 de novembro de 2021
Nova estratégia de alocação sob demanda	Agora, é possível escolher se deseja 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 Allocation strategies (Estratégias de alocação) no Guia do usuário do Amazon EC2 Auto Scaling.	27 de outubro de 2021

Seleção de tipo de instância baseada em atributos	O Amazon EC2 Auto Scaling adiciona suporte à seleção de tipo de instância baseada em atributos. Em vez de escolher manualmente os tipos de instância, você pode expressar seus requisitos de instância como um conjunto de atributos, como vCPU, memória e armazenamento. Para mais informações, consulte Creating an Auto Scaling group using attribute-based instance type selection (Criar um grupo do Auto Scaling usando seleção de tipo de instância baseada em atributo) no Guia do usuário do Amazon EC2 Auto Scaling.	27 de outubro de 2021
Suporte para filtragem de grupos por etiquetas	Agora você pode filtrar seus grupos do Auto Scaling usando filtros de etiquetas ao recuperar informaçõ es sobre grupos do Auto Scaling usando o comando `describe-auto-scal ing-groups` . Para mais informações, consulte Use tags to filter Auto Scaling groups (Usar etiquetas para filtrar grupos do Auto Scaling) no Guia do usuário do Amazon EC2 Auto Scaling.	14 de outubro de 2021

Alterações do guia	O console do Amazon EC2 Auto Scaling foi atualizado para ajudar você a criar políticas de rescisão personalizadas com. AWS Lambda A documentação do console foi adequadamente revisada. Para mais informações, consulte Using different termination policies (console) (Usar diferentes políticas de encerramento [console]).	14 de outubro de 2021
Suporte para cópia de configurações de execução para modelos de execução	Agora você pode copiar todas as configurações de lançamento em uma AWS região para novos modelos de lançamento do console do Amazon EC2 Auto Scaling. 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

Expande a funcionalidade de atualização de instância	Agora você pode incluir atualizações, como uma nova versão de um modelo de execução, ao substitui r instâncias, adicionando a configuração desejada ao comando `start-ins tance-refresh` . Você também pode ignorar a substituição de instâncias que já têm a configuração desejada ativando a correspon dência de ignorar. Para obter mais informações, consulte Substituir instâncias do Auto Scaling com base na atualizaç ão de uma instância no Manual do usuário do Amazon EC2 Auto Scaling.	5 de agosto de 2021
Suporte para políticas de término personalizadas	Agora você pode criar políticas de rescisão personalizadas com AWS Lambda. Para obter mais informações, consulte Criação de uma política de término personalizada com o Lambda. A documentação para especificar políticas de término foi atualizada de maneira adequada.	29 de julho de 2021

Alterações do guia	O console do Amazon EC2 Auto Scaling foi atualizado e aprimorado com outros recursos para ajudá-lo a criar ações programadas com um fuso horário especificado. A documentação para Escalabilidade programada foi revisada em conformidade.	3 de junho de 2021
volumes gp3 em configurações de execução	Agora você pode especificar volumes gp3 nos mapeamentos de dispositivos de bloco para configurações de execução.	2 de junho de 2021
Suporte para escalabilidade preditiva	Agora, você pode usar a escalabilidade preditiva para escalar proativamente grupos do Amazon EC2 Auto Scaling usando uma política de escalabilidade. Para obter mais informações, consulte Escalabilidade predtiva o Amazon EC2 Auto Scaling no Manual do usuário do Amazon EC2 Auto Scaling. Com essa atualização, a política AutoScalingServiceRolePolicygerenciada agora inclui permissão para chamar a ação `cloudwatch:GetMetricData` da API.	19 de maio de 2021

Alterações do guia	Agora você pode acessar modelos de exemplo para ganchos de ciclo de vida em. GitHub Para obter mais informações, consulte Ganchos do ciclo de vida do Amazon EC2 Auto Scaling no Manual do usuário do Amazon EC2 Auto Scaling.	9 de abril de 2021
Suporte a grupos de alta atividade	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 Grupos de alta atividade para o Amazon EC2 Auto Scaling no Manual do usuário do Amazon EC2 Auto Scaling.	8 de abril de 2021

Suporte para pontos de verificação	Agora você pode adicionar pontos de verificação a uma atualização de instância para substituir instâncias em fases e executar verificações em suas instâncias em pontos específicos. Para obter mais informações, consulte Adição de pontos de verificação a uma atualização de instância no Manual do usuário do Amazon EC2 Auto Scaling.	18 de março de 2021
Alterações do guia	Documentação aprimorada para uso EventBridge com eventos e ganchos de ciclo de vida do Amazon EC2 Auto Scaling. Para obter mais informações, consulte Como usar o Amazon EC2 Auto Scaling EventBridge com um tutorial: Configurar um gancho de ciclo de vida que invoque uma função Lambda no Guia do usuário do Amazon EC2 Auto Scaling.	18 de março de 2021

Suporte para fusos horários locais	Agora você pode criar ações programadas recorrentes no fuso horário local adicionando a opção `--time-zone` ao comando put-scheduled-update-group-action. Se o seu fuso horário seguir o horário de verão, a ação recorrente ajustará automaticamente o horário de verão (DST). Para obter mais informações, consulte Escalabilidade programada no Manual do usuário do Amazon EC2 Auto Scaling.	9 de março de 2021
Expande a funcionalidade para políticas de instâncias mistas	Agora, você pode priorizar tipos de instância para sua capacidade spot quando usar uma política de instâncias mistas. O Amazon EC2 Auto Scaling tenta atender as prioridades com base no melhor esforço, mas primeiro otimiza a capacidade. Para obter mais informações, consulte Grupos de Auto Scaling com vários tipos de instância e opções de compra no Manual do usuário do Amazon EC2 Auto Scaling.	8 de março de 2021

Escalabilidade de atividades para grupos excluídos	Agora você pode visualizar atividades de escalabilidade para grupos do Auto Scaling excluídos adicionando a opção `--include-deleted-groups` ao comando describe-scaling-activities. Para obter mais informações, consulte Solução de problemas do Amazon EC2 Auto Scaling no Manual do usuário do Amazon EC2 Auto Scaling.	23 de fevereiro de 2021
Melhorias no console	Agora você pode criar e anexar um Application Load Balancer ou Network Load Balancer ao console do Amazon EC2 Auto Scaling. Para obter mais informações, consulte Criar e anexar um novo Application Load Balancer ou Network Load Balancer (console) no Guia do usuário do Amazon EC2 Auto Scaling.	24 de novembro de 2020
Várias interfaces de rede	Agora você pode configurar um modelo de execução para um grupo do Auto Scaling que especifique várias interfaces de rede. Para obter mais informações, consulte Interfaces de rede em uma VPC.	23 de novembro de 2020

Vários modelos de execução	Vários modelos de execução podem agora ser usados com grupos do Auto Scaling. Para obter mais informaçõ es, consulte Especificar um modelo de execução diferente para um tipo de instância no Manual do usuário do Amazon EC2 Auto Scaling.	19 de novembro de 2020
Balanceadores de carga de gateway	Guia atualizado para mostrar como anexar um balancead or de carga de gateway a um grupo do Auto Scaling para garantir que as instância s de dispositivo executada s pelo Amazon EC2 Auto Scaling sejam registradas automaticamente e cancelada s no balanceador de carga. Para obter mais informações, consulte Tipos de Elastic Load Balancing e Anexação de um balanceador de carga ao seu grupo do Auto Scaling no Manual do usuário do Amazon EC2 Auto Scaling.	10 de novembro de 2020

Vida útil máxima da instância	Agora você pode reduzir a duração máxima da instância para um dia (86.400 segundos). Para obter mais informações, consulte Substituir instâncias do Auto Scaling com base na vida útil máxima da instância no Manual do usuário do Amazon EC2 Auto Scaling.	9 de novembro de 2020
Rebalanceamento de capacidade	Você pode configurar seu grupo do Auto Scaling para iniciar uma instância spot de substituição quando o Amazon EC2 emitir uma recomenda ção de rebalanceamento. Para obter mais informações, consulte Rebalanceamento de capacidade do Amazon EC2 Auto Scaling no Manual do usuário do Amazon EC2 Auto Scaling.	4 de novembro de 2020
Instance Metadata Service Version 2	É possível usar o Instance Metadata Service Version 2, que é um método orientado a sessão para solicitação de metadados da instância ao usar as configurações de execução. Para obter mais informações, consulte Configurar as opções de metadados de instância no Manual do usuário do Amazon EC2 Auto Scaling.	28 de julho de 2020

Alterações do guia	Várias melhorias e novos procedimentos de console nas seções Controle de quais instâncias do Auto Scaling são terminadas durante a redução de escala na horizontal, Monitoramento das instâncias e grupos do Auto Scaling, Modelos de execução e Configurações de execução do Manual do usuário do Amazon EC2 Auto Scaling.	28 de julho de 2020
Atualização de instância	Inicie uma atualização de instância para atualizar todas as instâncias no seu grupo do Auto Scaling quando você fizer uma alteração de configuração. Para obter mais informações, consulte Substituir instâncias do Auto Scaling com base na atualização de uma instância no Manual do usuário do Amazon EC2 Auto Scaling.	16 de junho de 2020

Alterações do guia	Várias melhorias nas seções Substituição de instâncias do Auto Scaling com base no tempo de vida máximo da instância, Grupos do Auto Scaling com vários tipos de instância e opções de compra, Escalabilidade baseada no Amazon SQS e Marcação de instância e grupos do Auto Scaling do Manual do usuário do Amazon EC2 Auto Scaling.	6 de maio de 2020
Alterações do guia	Várias melhorias na documentação do IAM. Para obter mais informações, consulte Suporte a modelos de execução e Exemplos de políticas baseadas em identidade do Amazon EC2 Auto Scaling no Manual do usuário do Amazon EC2 Auto Scaling.	4 de março de 2020

Desabilitar políticas de escalabilidade	Agora você pode desabilitar e reabilitar as políticas de escalabilidade. Esse recurso permite desabilitar temporariamente uma política de escalabilidade enquanto preserva os detalhes de configuração para que você possa habilitar a política novamente mais tarde. Para obter mais informações, consulte Desativação de uma política de escalabilidade para um grupo do Auto Scaling no Manual do usuário do Amazon EC2 Auto Scaling.	18 de fevereiro de 2020
Adicionar funcionalidade de notificação	O Amazon EC2 Auto Scaling agora envia eventos para AWS Health Dashboard você quando seus grupos do Auto Scaling não conseguem escalar devido à falta de um grupo de segurança ou modelo de lançamento. Para obter mais informaçõ es, consulte Notificações do AWS Health Dashboard para o Amazon EC2 Auto Scaling no Manual do usuário do Amazon EC2 Auto Scaling.	12 de fevereiro de 2020

Alterações do guia	Várias melhorias e correções nas seções Como o Amazon EC2 Auto Scaling funciona com o IAM, Exemplo de políticas baseadas em identidade do Amazon EC2 Auto Scaling, Política de chave da CMK necessári a para uso com volumes criptografados e Monitoram ento das suas instâncias e grupos do Auto Scaling no Manual do usuário do Amazon EC2 Auto Scaling.	10 de fevereiro de 2020
Alterações do guia	Melhoria na documentação para grupos do Auto Scaling que usam peso de instância s. Saiba como usar políticas de escalabilidade ao usar “unidades de capacidade” para medir a capacidade desejada. Para obter mais informações, consulte Como funcionam as políticas de escalabilidade e Tipos de ajuste da escalabilidade no Manual do usuário do Amazon EC2 Auto Scaling.	6 de fevereiro de 2020

Novo capítulo “Segurança”	Um novo capítulo sobre Segurança no Guia do usuário do Amazon EC2 Auto Scaling ajuda você a entender como aplicar o modelo de responsabilidade compartilhada ao usar o Amazon EC2 Auto Scaling. Como parte dessa atualização, o capítulo do Manual do usuário "Controle de acesso aos recursos do Amazon EC2 Auto Scaling" foi substituído por uma seção nova e mais útil, Gerenciamento de identidades e acesso para o Amazon EC2 Auto Scaling.	4 de fevereiro de 2020
Recomendações para tipos de instância	AWS Compute Optimizer fornece recomendações de instâncias do Amazon EC2 para ajudá-lo a melhorar o desempenho, economizar dinheiro ou ambos. Para obter mais informações, consulte Obtenção de recomendações de um tipo de instância no Manual do usuário do Amazon EC2 Auto Scaling.	3 de dezembro de 2019

Hosts dedicados e grupos de recursos de host	Guia atualizado para mostrar como criar um modelo de execução que especifica um grupo de recursos de host. Isso permite criar um grupo do Auto Scaling com um modelo de execução que especific a uma AMI de BYOL a ser usada em hosts dedicados . Para obter mais informaçõ es, consulte Criação de um modelo de execução para um grupo do Auto Scaling no Manual do usuário do Amazon EC2 Auto Scaling.	3 de dezembro de 2019
Suporte a endpoints da Amazon VPC	Agora você pode estabelec er uma conexão privada entre sua VPC e o Amazon EC2 Auto Scaling. Para obter mais informações, consulte Amazon EC2 Auto Scaling e endpoints da VPC da interface no Manual do usuário do Amazon EC2 Auto Scaling.	22 de novembro de 2019

Vida útil máxima da instância	Agora você pode substituir instâncias de forma automática especificando o período máximo que uma instância pode estar em serviço. Se alguma instância estiver se aproximando desse limite, o Amazon EC2 Auto Scaling gradualmente as substituirá. Para obter mais informações, consulte Substituir instâncias do Auto Scaling com base na vida útil máxima da instância no Manual do usuário do Amazon EC2 Auto Scaling.	19 de novembro de 2019
Ponderação de instâncias	Para grupos do Auto Scaling com vários tipos de instância, agora você pode especificar opcionalmente o número de unidades de capacidade com que cada tipo de instância contribui para a capacidade do grupo. Para obter mais informações, consulte Ponderação de instâncias do Auto Scaling do Amazon EC2 no Manual do usuário do Amazon EC2 Auto Scaling.	19 de novembro de 2019

Número mínimo de tipos de instância	Você não precisa mais especificar tipos de instância adicionais para grupos de instâncias spot, sob demanda e reservadas. Para todos os grupos de Auto Scaling, o mínimo agora é um tipo de instância. Para obter mais informações, consulte Grupos de Auto Scaling com vários tipos de instância e opções de compra no Manual do usuário do Amazon EC2 Auto Scaling.	16 de setembro de 2019
Suporte para a nova estratégia de alocação spot	O Amazon EC2 Auto Scaling agora oferece suporte a uma nova estratégia de alocação spot "otimizada para capacidade" que atende à sua solicitação usando grupos de instâncias spot escolhidos de forma ideal com base na capacidade spot disponível. Para obter mais informações, consulte Grupos de Auto Scaling com vários tipos de instância e opções de compra no Manual do usuário do Amazon EC2 Auto Scaling.	12 de agosto de 2019

Alterações do guia	Documentação do Amazon EC2 Auto Scaling melhorada nos tópicos 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	Agora, o Amazon EC2 Auto Scaling adiciona tags às instâncias do Amazon EC2 como parte da mesma chamada de API que inicia as instâncias. Para obter mais informações, consulte Marcação de grupos e instâncias do Auto Scaling.	26 de julho de 2019
Alterações do guia	Melhoria na documentação do Amazon EC2 Auto Scaling no tópico Suspensão e retomada de processos de escalabilidade. Atualização nos Exemplos de políticas gerenciadas pelo cliente para incluir um exemplo de política que permite que os usuários transmitam apenas 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	Adicionado suporte para novos recursos do Amazon EBS no tópico de modelo de execução. Altere o estado de criptografia de um volume do EBS ao restaurar um snapshot. Para obter mais informações, consulte Criação de um modelo de execução para um grupo do Auto Scaling no Manual do usuário do Amazon EC2 Auto Scaling.	13 de maio de 2019
Alterações do guia	Melhoria na documenta ção do Amazon EC2 Auto Scaling nas seguintes seções: Controle de quais instâncias do Auto Scaling são terminada s durante uma redução de escala na horizontal, Grupos do Auto Scaling, Grupos do Auto Scaling com vários tipos de instâncias e opções de compra e Escalabilidade dinâmica para o Amazon EC2 Auto Scaling.	12 de março de 2019

Suporte para a combinação de tipos de instâncias e opções de compra	Provisione e escale instâncias automaticamente nas opções de compra (spot, sob demanda e instâncias reservadas) e tipos de instância em um único grupo do Auto Scaling. Para obter mais informações, consulte Grupos de Auto Scaling com vários tipos de instância e opções de compra no Manual do usuário do Amazon EC2 Auto Scaling.	13 de novembro de 2018
Tópico atualizado para escalabilidade baseada no Amazon SQS	Guia atualizado para explicar como você pode usar métricas personalizadas para escalar um grupo do Auto Scaling em resposta à mudança na demanda de uma fila do Amazon SQS. Para obter mais informações, consulte Escalabilidade baseada no Amazon SQS no Manual do usuário do Amazon EC2 Auto Scaling.	26 de julho de 2018

A tabela a seguir descreve alterações importantes feitas na documentação do Amazon EC2 Auto Scaling antes de julho de 2018.

Atributo	Descrição	Data de lançamento
Suporte para as políticas de escalabilidade	Configure a escalabilidade dinâmica da sua aplicação em apenas algumas etapas. Para obter mais informações,	12 de julho de 2017

Atributo	Descrição	Data de lançamento
de rastreamento de destino	consulte Políticas de escalabilidade com monitoramento do objetivo do Amazon EC2 Auto Scaling.	
Suporte para permissões em nível de recurso	Criar políticas do IAM para controlar acesso em nível de recurso. Para obter mais informações, consulte Controle do acesso aos seus recursos do Amazon EC2 Auto Scaling.	15 de maio de 2017
Monitoramento de melhorias	As métricas de grupo do Auto Scaling não precisam mais que você habilite monitoramento detalhado. Agora você pode habilitar a coleta de métricas do grupo e visualizar gráficos de métricas na guia Monitoramento no console. Para obter mais informações, consulte Monitoramento de seus grupos e instâncias do Auto Scaling usando a Amazon. CloudWatch	18 de agosto de 2016
Suporte para Application Load Balancers	Anexar um ou mais grupos de destino a um grupo novo ou existente do Auto Scaling. Para obter mais informaçõ es, consulte Anexação de um balanceador de carga ao seu grupo do Auto Scaling.	11 de agosto de 2016
Eventos para ganchos do ciclo de vida	O Amazon EC2 Auto Scaling envia eventos EventBrid ge para quando chama ganchos de ciclo de vida. Para obter mais informações, consulte Como obter EventBrid ge quando seu grupo de Auto Scaling é dimensionado.	24 de fevereiro de 2016
Proteção de instância	Impedir que o Amazon EC2 Auto Scaling selecione instâncias específicas para término ao reduzir a escala. Para obter mais informações, consulte Proteção de instância.	07 de dezembro de 2015
Políticas de escalabilidade em etapas	Criar uma política de escalabilidade que permita escalonar com base no tamanho da violação do alarme. Para obter mais informações, consulte Tipos de política de escalabilidade.	06 de julho de 2015

Atributo	Descrição	Data de lançamento
Atualizar balanceador de carga	Anexar ou desvincular um balanceador de carga de um grupo do Auto Scaling existente. Para obter mais informações, consulte Anexação de um balanceador de carga ao seu grupo do Auto Scaling.	11 de junho de 2015
Support for ClassicLink	Vincular instâncias EC2-Classic em seu grupo do Auto Scaling a uma VPC permitindo comunicação entre essas instâncias EC2-Classic vinculadas e instâncias na VPC usando endereços IP privados. Para obter mais informações, consulte Vinculação de instâncias do EC2-Classic a uma VPC.	19 de janeiro de 2015
Ganchos do ciclo de vida	Manter suas instâncias recém-ativadas ou encerradas em um estado pendente enquanto você realiza ações nelas. Para obter mais informações, consulte Ganchos do ciclo de vida do Amazon EC2 Auto Scaling.	30 de julho de 2014
Desvincular instâncias	Desvincular instâncias de um grupo do Auto Scaling. Para obter mais informações, consulte Desvincular instâncias do EC2 do seu grupo do Auto Scaling.	30 de julho de 2014
Colocar instâncias em um estado de Standby	Colocar instâncias que estão em um estado de `InService` em um estado de `Standby`. Para obter mais informações, consulte Remoção temporária de instâncias do seu grupo do Auto Scaling.	30 de julho de 2014
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

Atributo	Descrição	Data de lançamento
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, consulteCriação de um grupo do Auto Scaling usando uma instância do EC2.	02 de janeiro de 2014
Anexar instâncias	Habilite a escalabilidade automática para uma instância do EC2 anexando a instância a um grupo do Auto Scaling existente. Para obter mais informações, consulte Anexar instâncias do EC2 ao seu grupo do Auto Scaling.	02 de janeiro de 2014
Visualizar limites da conta	Visualize os limites dos recursos do Auto Scaling para a sua conta. Para obter mais informações, consulte 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

Atributo	Descrição	Data de lançamento
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 aplicaçõe s 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 Requestin g Spot Instances for fault-tolerant and flexible applicati ons (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

Atributo	Descrição	Data de lançamento
Suporte para o Amazon SNS	Use o Amazon SNS para receber notificações sempre que o Amazon EC2 Auto Scaling iniciar ou terminar instâncias do EC2. Para obter mais informações, consulte Obtenção de notificações do SNS quando o grupo do Auto Scaling é escalado. O Amazon EC2 Auto Scaling também adicionou os seguintes novos recursos: • A capacidade de configurar ações de escalabilidade recorrentes usando a sintaxe cron. Para obter mais informações, consulte a operação da API `PutScheduledUpdateGroupAction`. • Uma nova configuração que permite escalar sem adicionar a instância executada ao balanceador de carga (LoadBalancer). Para obter mais informações, consulte o tipo de dados da API `ProcessType`. • O sinalizador `ForceDelete` na operação `DeleteAutoScalingGroup` que informa ao Amazon EC2 Auto Scaling para excluir o grupo do Auto Scaling com as instâncias associadas a ele sem esperar que as instâncias sejam terminadas primeiro. Para obter mais informações, consulte a operação da API `DeleteAutoScalingGroup`.	20 de julho de 2011
Ações de escalabilidade programadas	Suporte adicional para ações de escalabilidade programadas. Para obter mais informações, consulte Escalabilidade programada do Amazon EC2 Auto Scaling.	2 de dezembro de 2010
Suporte para a Amazon VPC	Adicionado suporte para a Amazon VPC. Para obter mais informações, consulte Execução de instâncias do Auto Scaling em uma VPC.	2 de dezembro de 2010

Atributo	Descrição	Data de lançamento
O suporte a clusters HPC	Suporte adicionado para clusters de computação de alta performance (HPC).	2 de dezembro de 2010
Suporte a verificações de integridade	Suporte adicionado para o uso de verificações de integridade do Elastic Load Balancing com instâncias do EC2 gerenciadas pelo Amazon EC2 Auto Scaling. Para obter mais informações, consulte Verificações de saúde para instâncias em um grupo de Auto Scaling.	2 de dezembro de 2010
Support para CloudWatch alarmes	O mecanismo de acionamento antigo foi removido e o Amazon EC2 Auto Scaling foi redesenhado para usar o recurso de alarme. CloudWatch Para obter mais informações, consulte Escalabilidade dinâmica do Amazon EC2 Auto Scaling.	2 de dezembro de 2010
Suspender e retomar a escalabilidade	Suporte adicional para suspender e retomar processos de escalabilidade.	2 de dezembro de 2010
Suporte ao IAM	Adicionado suporte ao IAM. Para obter mais informaçõ es, consulte Controle do acesso aos seus recursos do Amazon EC2 Auto Scaling.	2 de dezembro de 2010

As traduções são geradas por tradução automática. Em caso de conflito entre o conteúdo da tradução e da versão original em inglês, a versão em inglês prevalecerá.