AWS Auto Scaling

Planos de escalabilidade

AWS Auto Scaling: Planos de escalabilidade

Table of Contents

O que é um plano de escalabilidade?

Use um plano de escalabilidade para configurar a escalabilidade automática para recursos escaláveis relacionados ou associados em questão de minutos. Por exemplo, você pode usar etiquetas para agrupar recursos em categorias como produção, teste ou desenvolvimento. Em seguida, é possível pesquisar e configurar planos de escalabilidade para recursos escaláveis que pertencem a cada categoria. Ou, se sua infraestrutura de nuvem incluir o AWS CloudFormation, você pode definir modelos de pilha a serem usados para criar coleções de recursos. Então crie um plano de escalabilidade para os recursos escaláveis que pertencem a cada pilha.

Recursos compatíveis

O AWS Auto Scaling é compatível com o uso de planos de escalabilidade para os seguintes serviços e recursos:

- Amazon Aurora: aumente ou diminua o número de réplicas de leitura do Aurora provisionadas para um cluster de banco de dados do Aurora.
- Amazon EC2 Auto Scaling: inicie ou termine instâncias do EC2 aumentando ou diminuindo a capacidade desejada de um grupo do Auto Scaling.
- Amazon Elastic Container Service: aumente ou diminua a contagem de tarefas desejadas no Amazon ECS.
- Amazon DynamoDB: aumente ou diminua a capacidade provisionada de leitura e gravação do DynamoDB ou de um índice secundário global.
- Frota spot: inicie ou encerre instâncias do EC2 aumentando ou diminuindo a capacidade de destino de uma frota spot.

Recursos e benefícios do plano de escalabilidade

Os planos de escalabilidade fornecem estes recursos e benefícios:

- Detecção de recursos: o AWS Auto Scaling fornece detecção automática de recursos para ajudar a encontrar recursos de sua aplicação que podem ser escalados.
- Escalabilidade dinâmica: os planos de escalabilidade usam os serviços do Amazon EC2 Auto Scaling e do Application Auto Scaling para ajustar a capacidade de recursos escaláveis para lidar com alterações no tráfego ou na workload. As métricas de escalabilidade dinâmica podem ser métricas padrão de utilização ou de taxa de transferência ou métricas personalizadas.
- Recomendações de escalabilidade integradas: o AWS Auto Scaling fornece estratégias de escalabilidade com recomendações que você pode usar para otimizar a performance, os custos ou um equilíbrio entre os dois.
- Escalabilidade preditiva: os planos de escalabilidade também são compatíveis com a escalabilidade preditiva para grupos do Auto Scaling. Isso ajuda a escalar sua capacidade do Amazon EC2 mais rapidamente quando há picos de ocorrência regular.

> **Important**
>
> Se estiver usando planos de escalabilidade apenas para configurar a escalabilidade preditiva para seus grupos do Auto Scaling, recomendamos usar as políticas de escalabilidade preditiva dos grupos do Auto Scaling. Essa opção apresentada recentemente oferece recursos aprimorados, como o uso de agregações de métricas para criar novas métricas personalizadas ou reter dados

de métricas históricas em implantações azuis ou verdes. Para obter mais informações, consulte Escalabilidade predtiva o Amazon EC2 Auto Scaling no Manual do usuário do Amazon EC2 Auto Scaling.

Como começar a usar

Use os seguintes recursos para ajudar a criar e usar um plano de escalabilidade:

- Como funcionam os planos de escalabilidade (p. 3)
- Práticas recomendadas para planos de escalabilidade do (p. 5)
- Conceitos básicos dos planos de escalabilidade (p. 7)

Trabalhar com planos de escalabilidade

Você pode criar, acessar e gerenciar seus planos de escalabilidade usando qualquer uma das seguintes interfaces:

- AWS Management Console: fornece uma interface da Web que você pode usar para acessar os planos de escalabilidade. Se você se inscreveu para um Conta da AWS, poderá acessar seus planos de escalabilidade fazendo login no AWS Management Console, usando a caixa de pesquisa na barra de navegação para procurar AWS Auto Scaling e escolhendo AWS 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 a usar, consulte oAWS Command Line Interface User Guide (Guia do usuário da AWS Command Line Interface). Para obter mais informações, consulte planos de escalabilidade automática na Referência de comandos da AWS CLI.
- Ferramentas da AWS para Windows PowerShell – Fornece comandos para um conjunto amplo de produtos da AWS para os usuários que usam script no ambiente do PowerShell. Para começar a usar, consulte o Guia do usuário do AWS Tools for Windows PowerShell. Para obter mais informações, consulte Referência de Cmdlets do AWS Tools for PowerShell.
- AWS SDKs: fornecem operações de API específicas da linguagem e cuidam de muitos dos detalhes da conexão, como cálculo de assinaturas, tratamento de novas tentativas de solicitação e tratamento de erros. Para obter mais informações, consulte AWS SDKs.
- API de consulta: fornece ações de API de baixo nível que são chamadas usando solicitações HTTPS. Usar a API de consulta é a maneira mais direta de acessar a Serviços da AWS. No entanto, ela exige que a aplicação trate detalhes de baixo nível, como gerar o hash para assinar a solicitação e tratar erros. Para obter mais informações, consulte a Referência da API do AWS Auto Scaling.
- AWS CloudFormation— Oferece suporte à criação de planos de escalabilidade com o uso de modelos do CloudFormation. Para obter mais informações, consulte a referência para AWS::AutoScalingPlans::ScalingPlan no Guia do usuário do AWS CloudFormation.

Para se conectar a um AWS service (Serviço da AWS) de forma programática, use um endpoint. Para obter informações sobre endpoints para chamadas ao AWS Auto Scaling, consulteEndpoints e cotas do AWS Auto Scaling na Referência geral da AWS. Esta página também mostra a disponibilidade regional dos planos de escalabilidade.

Preços

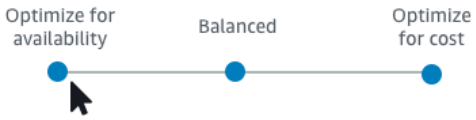
Todos os recursos do plano de escalabilidade estão habilitados para você usar. Os recursos são fornecidos sem custo adicional além das taxas de serviço do CloudWatch e dos outros recursos da Nuvem AWS que você usa.

Como funcionam os planos de escalabilidade

O AWS Auto Scaling permite que você utilize planos de escalabilidade para configurar um conjunto de instruções para escalar seus recursos. Se você usa o AWS CloudFormation ou adiciona etiquetas a recursos escaláveis, é possível configurar planos de escalabilidade para diferentes conjuntos de recursos por aplicação. O console do AWS Auto Scaling fornece recomendações para estratégias de escalabilidade personalizadas para cada recurso. Após criar o plano de escalabilidade, ele mescla escalabilidade dinâmica e métodos de escalabilidade preditiva para oferecer suporte à estratégia de escalabilidade.

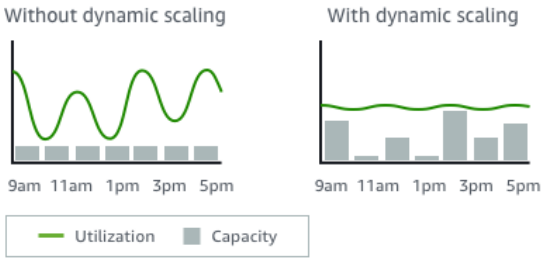
O que é uma estratégia de escalabilidade?

A estratégia de escalabilidade diz ao AWS Auto Scaling como otimizar a utilização dos recursos no plano de escalabilidade. Você pode otimizar para disponibilidade de custo ou um equilíbrio de ambos. Como alternativa, você também pode criar sua própria estratégia personalizada, de acordo com as métricas e os limites definidos por você. Você pode definir estratégias separadas para cada recurso ou tipo de recurso.



O que é a escalabilidade dinâmica?

A escalabilidade dinâmica cria políticas de escalabilidade de rastreamento de destino para os recursos em seu plano de escalabilidade. Essas políticas de escalabilidade ajustam a capacidade do recurso em resposta a alterações ativas na utilização de recursos. A intenção é fornecer capacidade suficiente para manter a utilização no valor de destino especificado pela estratégia de escalabilidade. Isso é semelhante à forma como o termostato mantém a temperatura da casa. Você escolhe a temperatura, e o termostato faz o resto.

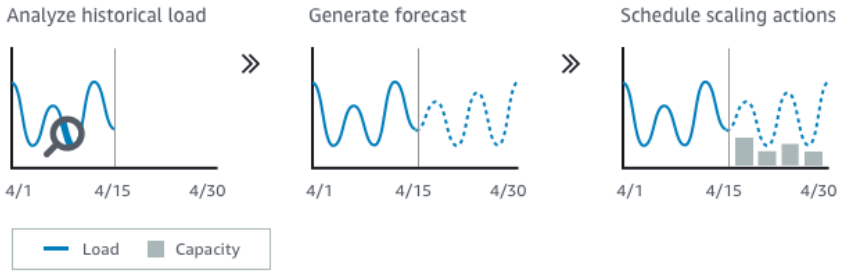


Por exemplo, você pode configurar seu plano de escalabilidade para manter o número de tarefas que o serviço do Amazon Elastic Container Service (Amazon ECS) executa em 75% da CPU. Quando a utilização da CPU do serviço ultrapassa 75% (o que significa que mais de 75% da CPU reservada para o serviço está sendo usada), o alarme de expansão aciona sua política de escalabilidade para adicionar outra tarefa ao serviço para ajudar com o aumento de carga.

O que é a escalabilidade preditiva?

A escalabilidade preditiva usa machine learning para analisar toda a workload histórica do recurso e faz previsões regulares sobre a carga futura. É um método semelhante ao das previsões meteorológicas. Usando a previsão, a escalabilidade preditiva gera ações de escalabilidade programadas para garantir

que a capacidade do recurso esteja disponível antes que o aplicativo precise dela. Assim como na escalabilidade dinâmica, a escalabilidade preditiva funciona para manter a utilização no valor de destino especificado pela estratégia de escalabilidade.



Por exemplo, você pode habilitar a escalabilidade preditiva e configurar a estratégia de escalabilidade para manter a utilização média da CPU do grupo do Auto Scaling em 50%. Sua previsão chama picos de tráfego para ocorrerem todos os dias às 8h. O plano de escalabilidade cria as ações de escalabilidade agendadas futuras para garantir que o grupo do Auto Scaling esteja pronto para lidar com o tráfego com antecedência. Isso ajuda a manter a performance do aplicativo constante, com o objetivo de sempre ter a utilização de recursos o mais próximo possível de 50% o tempo todo.

Veja a seguir os principais conceitos para entender escalabilidade preditiva:

- Carregar previsão: AWS Auto Scaling analisa até 14 dias de histórico de uma métrica de carga especificada e previsões futuras de demanda para os próximos dois dias. Esses dados estão disponíveis em intervalos de uma hora e são atualizados diariamente.
- Ações de escalabilidade programadas: o AWS Auto Scaling programa as ações de escalabilidade que aumentam e diminuem proativamente a capacidade para corresponder à previsão de carga. No horário programado, o AWS Auto Scaling atualiza a capacidade mínima com o valor especificado pela ação de escalabilidade programada. A intenção é manter a utilização de recursos no valor de destino especificado pela estratégia de escalabilidade. Se o seu aplicativo requer mais capacidade que previsão, escalabilidade dinâmica está disponível para adicionar capacidade adicional.
- Comportamento de capacidade máxima: limites de capacidade mínima e máxima para autoescalabilidade se aplicam a cada recurso. No entanto, é possível controlar se a aplicação pode aumentar a capacidade além de sua capacidade máxima quando a capacidade de previsão é maior que a capacidade máxima.

> **Note**
>
> Agora é possível usar as políticas de escalabilidade preditiva dos grupos do Auto Scaling. Para obter mais informações, consulte Escalabilidade predtiva o Amazon EC2 Auto Scaling no Manual do usuário do Amazon EC2 Auto Scaling.

Práticas recomendadas para planos de escalabilidade do

As práticas recomendadas a seguir podem ajudá-lo a obter o máximo dos planos de escalabilidade:

- Ao criar um modelo de execução ou uma configuração de execução, habilite o monitoramento detalhado para obter dados de métricas do CloudWatch para instâncias do EC2 em uma frequência de um minuto, pois isso garante uma resposta mais rápida às alterações de carga. Aumentar a escalabilidade das métricas com intervalos de cinco minutos pode resultar em tempo de resposta mais lento e aumentar a escalabilidade de dados obsoletos. Por padrão, as instâncias do EC2 são habilitadas para monitoramento básico, ou seja, os dados de métrica para instâncias estão disponíveis em intervalos de cinco minutos. Para uma cobrança adicional, habilite o monitoramento detalhado para obter dados de métrica para instâncias em intervalos de um minuto. Para obter mais informações, consulte Configurar o monitoramento de instâncias do Auto Scaling no Guia do usuário do Amazon EC2 Auto Scaling.
- Também recomendamos que você habilite as métricas do grupo do Auto Scaling. Caso contrário, a capacidade real dos dados não é mostrada nos gráficos de previsão de capacidade que são disponibilizados na conclusão assistente de criação do plano de dimensionamento. Para obter mais informações, consulte Monitorar métricas do CloudWatch para grupos de Auto Scaling e instâncias no Guia do usuário do Amazon EC2 Auto Scaling.
- Verifique qual tipo de instância o grupo do Auto Scaling usa e atente-se para o uso de um tipo de instância expansível. As instâncias expansíveis do Amazon EC2, como instâncias T3 e T2, foram criadas para oferecer um nível básico de performance de CPU com capacidade de expansão para um nível superior quando exigido pela workload. Dependendo da utilização de destino especificado pelo plano de escalabilidade, você pode executar o risco de exceder a linha de base e, em seguida, executar fora de créditos de CPU, que limita a performance. Para obter mais informações, consulte Créditos de CPU e performance básica para instâncias expansíveis. Para configurar as instâncias como `unlimited`, consulte Usar um grupo do Auto Scaling para executar uma instância expansível como ilimitada no Manual do usuário do Amazon EC2 para instâncias do Linux.

Outras considerações

> **Note**
>
> Agora é possível usar as políticas de escalabilidade preditiva dos grupos do Auto Scaling. Para obter mais informações, consulte Escalabilidade predtiva o Amazon EC2 Auto Scaling no Manual do usuário do Amazon EC2 Auto Scaling.

Tenha as seguintes considerações adicionais em mente:

- A escalabilidade preditiva usa previsões de carga para programar a capacidade no futuro. A qualidade das previsões varia com base na quantidade de ciclos da carga e na aplicabilidade do modelo de previsão treinado. O dimensionamento preditivo pode ser executado no modo somente previsão para avaliar a qualidade das previsões e das ações de dimensionamento criadas pela previsão. Você poderá definir o modo de dimensionamento preditivo para Forecast only (Somente previsão) ao criar o plano de dimensionamento e alterá-lo para Forecast and scale (Previsão e dimensionamento) quando a avaliação da qualidade da previsão for concluída. Para obter mais informações, consulte Configurações de dimensionamento preditivo (p. 14) e Monitorar e avaliar previsões (p. 16).
- Se você optar por especificar diferentes métricas para escalabilidade preditiva, é necessário garantir que a métrica de escalabilidade e a métrica de carga sejam altamente correlacionadas. O valor da métrica deve aumentar e diminuir em proporção ao número das instâncias no grupo do Auto Scaling.

Isso garante que os dados da métrica possam ser usados para expandir ou reduzir proporcionalmente o número de instâncias. Por exemplo, a métrica de carga é a contagem total da solicitação e a métrica de escalabilidade é a utilização média da CPU. Se a contagem total da solicitação aumenta em 50%, a média de utilização da CPU também deve aumentar em 50%, desde que a capacidade permaneça inalterada.

- Antes de criar o plano de dimensionamento, você deve excluir as ações de dimensionamento programadas anteriormente que não são mais necessárias acessando os consoles a partir dos quais elas foram criadas. O AWS Auto Scaling não cria uma ação de dimensionamento preditiva que substitui uma ação de dimensionamento programada existente.
- Suas configurações personalizadas para capacidade mínima e máxima, juntamente com outras configurações usadas para escalabilidade dinâmica, mostrados em outros consoles. No entanto, recomendamos que, após criar um plano de dimensionamento, você não modifique essas configurações a partir de outros consoles, pois o plano de dimensionamento não recebe as atualizações de outros consoles.
- Seu plano de dimensionamento pode conter recursos de vários serviços, mas cada recurso pode estar somente em um plano de dimensionamento por vez.

Evitar o erro ActiveWithProblems

Um erro “ActiveWithProblems” pode ocorrer quando um plano de escalabilidade é criado, ou quando recursos são adicionados a um plano de escalabilidade. O erro ocorre quando o plano de escalabilidade está ativo, mas não foi possível aplicar a configuração de escalabilidade a um ou mais recursos.

Geralmente, ele ocorre porque um recurso já tem uma política de escalabilidade ou um grupo do Auto Scaling não cumpre os requisitos mínimos para a escalabilidade preditiva.

Se algum dos recursos já tiver políticas de escalabilidade de vários consoles de serviços, o AWS Auto Scaling não substituirá as outras políticas de escalabilidade nem criará recursos por padrão. Você também pode excluir as políticas de escalabilidade existentes e substituí-las pelas políticas de escalabilidade de rastreamento de destino criadas no console do AWS Auto Scaling. Faça isso habilitando a configuração Replace external scaling policies (Substituir políticas externas de escalabilidade) de todos os recursos que tiverem políticas de escalabilidade a serem substituídas.

Com a escalabilidade preditiva, recomendamos aguardar 24 horas após a criação de um grupo do Auto Scaling para configurar a escalabilidade. Deve haver, no mínimo, 24 horas de dados históricos para gerar a previsão inicial. Se o grupo tiver menos de 24 horas de dados históricos e a escalabilidade preditiva estiver habilitada, o plano de escalabilidade não poderá gerar uma previsão até o próximo período de previsão após o grupo coletar a quantidade necessária de dados. No entanto, você também pode editar e salvar o plano de escalabilidade para reiniciar o processo de previsão assim que as 24 horas de dados estiverem disponíveis.

Conceitos básicos dos planos de escalabilidade

Antes de criar um plano de escalabilidade para usar com sua aplicação, analise-o detalhadamente ao executá-lo na Nuvem AWS. Anote o seguinte:

- Mesmo se você já tiver criado as políticas de escalabilidade de outros consoles. É possível substituir ou manter (sem permissão para fazer alterações nos valores) as políticas existentes de escalabilidade ao criar o plano de escalabilidade.
- A utilização de destino que faça sentido para cada recurso dimensionável em seu aplicativo com base no recurso como um todo. Por exemplo, a quantidade de CPU que as instâncias do EC2 em um grupo do Auto Scaling devem usar em comparação com a CPU disponível. Ou, no caso de um serviço como o DynamoDB, que usa um modelo de taxa de transferência provisionada, a quantidade de atividades de leitura e gravação que uma tabela ou índice deve usar em comparação com a taxa de transferência disponível. Em outras palavras, a proporção da capacidade consumida e da capacidade provisionada. É possível alterar a utilização de destino a qualquer momento depois de criar o plano de escalabilidade.
- Quanto tempo é necessário para iniciar e configurar um servidor. Essa informação ajuda a configurar um período para que cada instância do EC2 carregue após a inicialização e garantir que um novo servidor não seja iniciado enquanto o anterior ainda está em inicialização.
- Se o histórico de métricas é longo o suficiente para usar com a escalabilidade preditiva (se você estiver usando grupos do Auto Scaling recém-criados). Em geral, ter um ciclo completo de 14 dias de dados históricos se converte em previsões mais precisas. O mínimo é 24 horas.

Quanto melhor você entender seu aplicativo, mais eficaz você pode tornar seu plano de escalabilidade.

As tarefas a seguir ajudarão você a se familiarizar com os planos de escalabilidade. Você criará um plano de escalabilidade para um único grupo do Auto Scaling e habilitará as escalabilidades preditiva e dinâmica.

Tarefas

Etapa 1: Encontrar recursos escaláveis

Esta seção inclui uma introdução prática à criação de planos de escalabilidade no console do AWS Auto Scaling. Caso seja o seu primeiro plano de escalabilidade, é recomendável que você crie um plano de escalabilidade de exemplo usando um grupo do Amazon EC2 Auto Scaling.

Pré-requisitos

Para praticar o uso de um plano de escalabilidade, crie um grupo do Auto Scaling. Inicie pelo menos uma instância do Amazon EC2 no grupo do Auto Scaling. Para obter mais informações, consulte Conceitos básicos do Amazon EC2 Auto Scaling no Manual do usuário do Amazon EC2 Auto Scaling.

Use um grupo do Auto Scaling com as métricas do CloudWatch habilitadas para obter dados de capacidade nos grafos que estarão disponíveis quando você concluir o assistente Create Scaling Plan (Criar plano de escalabilidade). Para obter mais informações, consulte Habilitar métricas do grupo do Auto Scaling no Manual do usuário do Amazon EC2 Auto Scaling.

Gere alguma carga por alguns dias ou mais para ter dados de métricas do CloudWatch disponíveis para o recurso de escalabilidade preditiva, se possível.

Certifique-se de que você tenha as permissões necessárias para trabalhar com planos de escalabilidade. Para obter mais informações, consulte . Gerenciamento de Identidade e Acesso para planos de escalabilidade (p. 23).

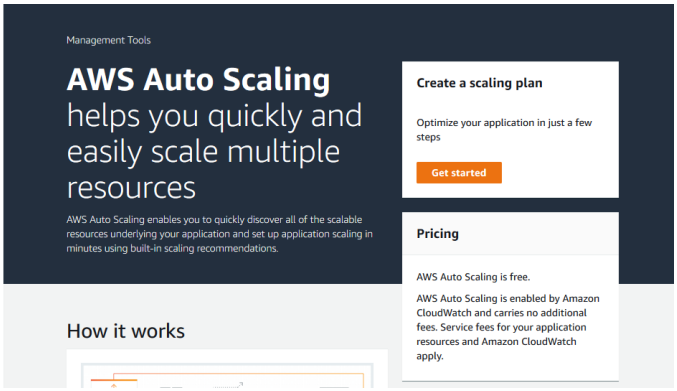
Adicionar o grupo do Auto Scaling ao novo plano de escalabilidade

Ao criar um plano de escalabilidade pelo console, isso ajuda você a encontrar os recursos escaláveis na primeira etapa. Antes de começar, confirme se os seguintes requisitos estão sendo atendidos:

- Você criou um grupo do Auto Scaling e iniciou pelo menos uma instância do EC2, conforme descrito na seção anterior.
- O grupo do Auto Scaling criado existe há pelo menos 24 horas.

Para começar a criar um plano de escalabilidade

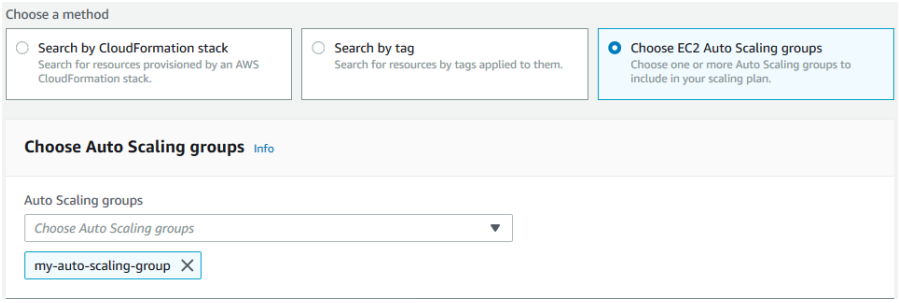
1. Abra a console do AWS Auto Scaling em https://console.aws.amazon.com/awsautoscaling/.
2. Na barra de navegação na parte superior da tela, escolha a mesma região usada ao criar o grupo do Auto Scaling.
3. Na página de boas-vindas, selecione Get started (Primeiros passos).



4. Na página Find scalable resources (Encontrar recursos escaláveis), siga um destes procedimentos:
   - Escolha Search by CloudFormation stack (Pesquisar por pilha do CloudFormation) e selecione a pilha do AWS CloudFormation a ser usada.
   - Selecione Search by tag (Pesquisar por etiqueta). Para cada etiqueta, selecione uma chave de etiqueta em Key (Chave) e os valores de etiqueta em Value (Valor). Para adicionar tags, escolha Add another row (Adicionar outra linha). Para remover tags, escolha Remove (Remover).
   - Selecione Choose EC2 Auto Scaling groups (Escolher grupos de Auto Scaling do EC2) e selecione um ou mais grupos do Auto Scaling.

Note

Para obter um tutorial introdutório, selecione Choose EC2 Auto Scaling groups (Escolher grupos do EC2 Auto Scaling) e escolha o grupo do Auto Scaling que você criou.



5. Selecione Next (Próximo) para continuar com o processo de criação do plano de escalabilidade.

Saiba mais sobre como identificar os recursos escaláveis

Se você já criou um plano de escalabilidade de exemplo e deseja criar outros, consulte os casos a seguir para usar uma pilha do CloudFormation ou um conjunto de etiquetas com mais detalhes. Use esta seção para decidir se quer escolher a opção Search by CloudFormation stack (Pesquisar por pilha do CloudFormation) ou Search by tag (Pesquisar por etiqueta) para identificar os recursos escaláveis ao usar a console para criar o plano de escalabilidade.

Ao escolher a opção Search by CloudFormation stack (Pesquisar por pilha do CloudFormation) ou Search by tag (Pesquisar por etiqueta) na etapa 1 do assistente Create Scaling Plan (Criar plano de escalabilidade), os recursos escaláveis associados à pilha ou ao conjunto de etiquetas serão disponibilizados para o plano de escalabilidade. À medida que você define seu plano de dimensionamento, é possível escolher quais desses recursos incluir ou excluir.

Identificar os recursos escaláveis usando uma pilha do CloudFormation

Ao usar o CloudFormation, você trabalha com pilhas para provisionar recursos. Todos os recursos em uma pilha são definidos pelo modelo da pilha. O seu plano de escalabilidade adiciona uma camada de orquestração no início da pilha que facilita a configuração da escalabilidade para múltiplos recursos. Sem um plano de escalabilidade você precisaria definir a escalabilidade de cada recurso dimensionável individualmente. Isso significa descobrir a ordem do provisionamento de recursos e políticas de escalabilidade e entender as sutilezas de como essas dependências funcionam.

Na console do AWS Auto Scaling, é possível selecionar uma pilha existente para verificar se ela tem recursos que podem ser configurados para a escalabilidade automática. O AWS Auto Scaling identifica somente recursos que estão definidos na pilha selecionada. Ele não passa por pilhas aninhadas.

Para que os serviços do ECS sejam identificados em uma pilha do CloudFormation, a console do AWS Auto Scaling precisa saber qual cluster do ECS está executando o serviço. Isso exige que os serviços do ECS estejam na mesma pilha do CloudFormation que o cluster do ECS que está executando o serviço. Do contrário, eles devem fazer parte do cluster padrão. Para ser identificado corretamente, o nome do serviço do ECS também deve ser exclusivo em cada um desses clusters do ECS.

Para obter mais informações sobre o CloudFormation, consulte O que é o AWS CloudFormation? no Manual do usuário do AWS CloudFormation.

Identificar recursos escaláveis usando etiquetas

As etiquetas apresentam metadados que podem ser usados para identificar recursos escaláveis relacionados na console do AWS Auto Scaling, usando filtros de etiquetas.

Use etiquetas para identificar os seguintes recursos:

- Clusters de banco de dados do Aurora
- Grupos do Auto Scaling
- Tabelas e índices secundários globais do DynamoDB

Ao pesquisar por mais de uma tag, cada recurso deverá ter todas as tags listadas para ser descoberto.

Para obter mais informações sobre marcação, consulte a documentação a seguir.

- Aprenda a etiquetar clusters do Aurora no Guia do usuário do Amazon Aurora.
- Aprenda a etiquetar grupos do Auto Scaling no Guia do usuário do Amazon EC2 Auto Scaling.
- Aprenda a etiquetar recursos do DynamoDB no Guia do desenvolvedor do Amazon DynamoDB.
- Saiba mais sobre as práticas recomendadas de marcação de recursos da AWS na Referência geral da AWS.

Etapa 2: Especificar a estratégia de escalabilidade

Use o procedimento a seguir para especificar as estratégias de dimensionamento para os recursos que foram encontrados na etapa anterior.

Para cada tipo de recurso, o AWS Auto Scaling escolhe a métrica que é mais usada para determinar a quantidade do recurso que está sendo usada em determinado momento. Você escolhe a estratégia de dimensionamento mais apropriada para otimizar a performance do aplicativo com base nessa métrica. Quando você habilita o recurso de dimensionamento dinâmico e o recurso de dimensionamento preditivo, a estratégia de dimensionamento é compartilhada entre eles. Para obter mais informações, consulte . Como funcionam os planos de escalabilidade (p. 3).

As seguintes estratégias de dimensionamento estão disponíveis:

- Optimize for availability (Otimizar para disponibilidade): o AWS Auto Scaling aumenta e reduz automaticamente a escala do recurso na horizontal para manter a utilização em 40%. Essa opção é útil quando o aplicativo tem necessidades de dimensionamento urgentes e, às vezes, imprevisíveis.
- Balance availability and cost (Equilibrar disponibilidade e custo): o AWS Auto Scaling aumenta e reduz automaticamente a escala do recurso na horizontal para manter a utilização em 50%. Essa opção ajuda a manter a alta disponibilidade ao mesmo tempo que reduz os custos.
- Optimize for cost (Otimizar para custo): o AWS Auto Scaling aumenta e reduz automaticamente a escala do recurso na horizontal para manter a utilização em 70%. Essa opção é útil para reduzir custos, caso o aplicativo possa lidar com a necessidade de ter a capacidade de buffer reduzida quando houver alterações inesperadas na demanda.

Por exemplo, o plano de escalabilidade configura o grupo do Auto Scaling para adicionar ou remover instâncias do Amazon EC2 com base na quantidade de CPU usada em média para todas as instâncias do grupo. Você escolhe se deseja otimizar a utilização para disponibilidade, custo ou uma combinação de ambos alterando a estratégia de dimensionamento.

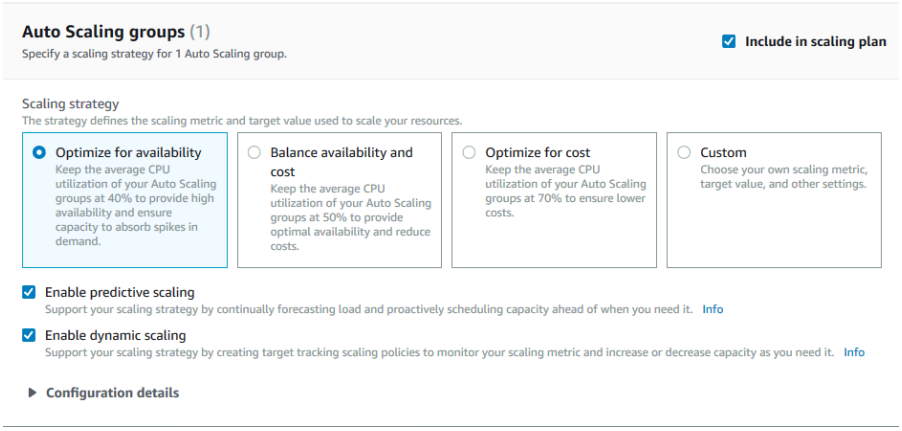
Se preferir, você poderá configurar uma estratégia personalizada, caso uma estratégia existente não atenda às suas necessidades. Com uma estratégia personalizada, é possível alterar o valor da utilização pretendida, escolher outra métrica ou ambos.

Important

Para o tutorial introdutório, conclua somente a primeira etapa do procedimento a seguir e selecione Next (Próximo) para continuar.

Para especificar uma estratégia de escalabilidade

1. Na página Specify scaling strategy (Especificar estratégia de escalabilidade), para Scaling plan details (Detalhes do plano de escalabilidade), Name (Nome), insira um nome para o plano de escalabilidade. O nome do plano de escalabilidade deve ser exclusivo em seu conjunto de planos de escalabilidade da região. Pode ter no máximo 128 caracteres e não deve conter barras verticais “|”, barras “/” ou dois pontos “:”.
2. Todos os recursos incluídos são listados por tipo de recurso. Em Auto Scaling groups (Grupos do Auto Scaling), faça o seguinte:



   a. Ignore esta etapa para usar a estratégia de escalabilidade e métricas padrão. Para usar uma estratégia de escalabilidade ou métricas diferentes, realize as seguintes etapas:

      i. Em Scaling strategy (Estratégia de escalabilidade), escolha a estratégia de escalabilidade desejada.

         No tutorial introdutório, escolha Optimize for availability (Otimizar para disponibilidade). Essa opção especifica que a utilização média da CPU de seu grupo do Auto Scaling seja mantida em 40%.

      ii. Se você escolher Custom (Personalizado), expanda Configuration details (Detalhes da configuração) para escolher as métricas e o valor de destino desejados.

         - Para Scaling metric (Escalar métrica), escolha a métrica de escalabilidade desejada.
         - Em Target value (Valor de destino), escolha o valor de destino desejado, como a utilização de destino ou a taxa de transferência de destino durante qualquer intervalo de um minuto.
         - Em Load metric (Métrica de carga) [apenas para grupos do Auto Scaling], escolha a métrica de carga desejada para usar a escalabilidade preditiva.
         - Selecione Replace external scaling policies (Substituir as políticas de escalabilidade externas) para especificar que o AWS Auto Scaling pode excluir políticas de escalabilidade criadas fora do plano de escalabilidade (como de outros consoles) e substituí-las por novas políticas de dimensionamento com monitoramento do objetivo criadas pelo plano de escalabilidade.

   b. (Opcional) Por padrão, a escalabilidade preditiva está habilitada para os grupos do Auto Scaling. Para desativar a escalabilidade preditiva dos grupos do Auto Scaling, desmarque Enable predictive scaling (Habilitar escalabilidade preditiva).

c. (Opcional) Por padrão, a escalabilidade dinâmica é habilitada para cada tipo de recurso. Para desativar a escalabilidade dinâmica de um tipo de recurso, desmarque a opção Enable dynamic scaling (Habilitar escalabilidade dinâmica).

d. (Opcional) Por padrão, quando você especifica a origem de um aplicativo a partir da qual vários recursos dimensionáveis são descobertos, todos os tipos de recursos são automaticamente incluídos no seu plano de escalabilidade. Para omitir um tipo de recurso do seu plano de dimensionamento, desmarque a opção Include in scaling plan (Incluir no plano de dimensionamento).

3. (Opcional) Para especificar uma estratégia de escalabilidade para outro tipo de recurso, repita as etapas anteriores.
4. Quando concluir, selecione Next (Próximo) para continuar com o processo de criação do plano de escalabilidade.

Etapa 3: Definir configurações avançadas (opcional)

Agora que especificou a estratégia de dimensionamento a ser usada para cada tipo de recurso, você pode optar por personalizar qualquer uma das configurações padrão para cada recurso usando a etapa Configure advanced settings (Definir configurações avançadas). Para cada tipo de recurso, há vários grupos de configurações que você pode personalizar. Na maioria dos casos, no entanto, as configurações padrão devem ser mais eficientes, com a possível exceção dos valores para a capacidade mínima e a capacidade máxima, que devem ser ajustados com cuidado.

Ignore esse procedimento se quiser manter as configurações padrão. Você pode alterar essas configurações a qualquer momento, editando o plano de escalabilidade.

> Important
>
> No tutorial introdutório, vamos fazer algumas alterações para atualizar a capacidade máxima do grupo do Auto Scaling e habilitar a escalabilidade preditiva no modo somente previsão. Embora não seja necessário personalizar todas as configurações para o tutorial, vamos também examinar brevemente as configurações de cada seção.

Configurações gerais

Use este procedimento para visualizar e personalizar as configurações que você especificou na etapa anterior para cada recurso. Você também pode personalizar a capacidade mínima e capacidade máxima para cada recurso.

Para visualizar e personalizar as configurações gerais

1. Na página Configure advanced settings (Definir configurações avançadas), selecione a seta à esquerda de qualquer um dos cabeçalhos de seção para expandir a seção. Para o tutorial, expanda a seção Auto Scaling groups (Grupos do Auto Scaling).
2. Na tabela exibida, escolha o grupo do Auto Scaling que você está usando neste tutorial.
3. Deixe a opção Include in scaling plan (Incluir no plano de dimensionamento) selecionada. Se essa opção não estiver selecionada, o recurso será omitido do plano de dimensionamento. Se você não incluir pelo menos um recurso, o plano de dimensionamento não poderá ser criado.
4. Para expandir a visualização e ver os detalhes da seção General Settings (Configurações gerais), selecione a seta à esquerda do cabeçalho da seção.
5. Você pode optar por qualquer um dos itens a seguir. Para este tutorial, localize a configuração Maximum capacity (Capacidade máxima) e insira o valor `3` no lugar do valor atual.

- Scaling strategy (Estratégia de escalabilidade): permite que você otimize para disponibilidade, custo ou um equilíbrio de ambos ou que especifique uma estratégia personalizada.
- Enable dynamic scaling (Habilitar escalabilidade dinâmica): se essa configuração estiver desmarcada, o recurso selecionado não poderá ser escalado usando uma configuração de escalabilidade com monitoramento do objetivo.
- Enable predictive scaling (Habilitar escalabilidade preditiva): [apenas para grupos do Auto Scaling] se essa configuração estiver desmarcada, o grupo selecionado não poderá ser escalado usando a escalabilidade preditiva.
- Scaling metric (Métrica de escalabilidade): especifica a métrica de escalabilidade a ser usada. Se você selecionar Custom (Personalizada), poderá especificar uma métrica personalizada a ser usada em vez das métricas predefinidas disponíveis na console. Para obter mais informações, consulte o próximo tópico desta seção.
- Target value (Valor de destino): especifica o valor de utilização de destino a ser usado.
- Load metric (Métrica de carga): [apenas para grupos do Auto Scaling] especifica a métrica de carga a ser usada. Se você selecionar Custom (Personalizada), poderá especificar uma métrica personalizada a ser usada em vez das métricas predefinidas disponíveis na console. Para obter mais informações, consulte o próximo tópico desta seção.
- Minimum capacity (Capacidade mínima): especifica a capacidade mínima para o recurso. O AWS Auto Scaling garante que o recurso nunca fique abaixo desse tamanho.
- Maximum capacity (Capacidade máxima): especifica a capacidade máxima para o recurso. O AWS Auto Scaling garante que o recurso nunca fique acima desse tamanho.

  Note

  Ao usar o dimensionamento preditivo, se preferir, você poderá escolher outro comportamento de capacidade máxima a ser usado com base na capacidade da previsão. Essa configuração está na seção Predictive scaling settings (Configurações de dimensionamento preditivo).

Métricas personalizadas

AWS Auto Scaling fornece as métricas mais comumente usadas para escalabilidade automática. No entanto, dependendo das suas necessidades, você pode preferir obter dados de métricas diferentes em vez das métricas na console. O Amazon CloudWatch apresenta várias métricas diferentes que você pode escolher. O CloudWatch também permite publicar suas próprias métricas.

Você pode usar o JSON para especificar uma métrica personalizada do CloudWatch. Antes de seguir as instruções, recomendamos que você se familiarize com o Manual do usuário Amazon CloudWatch.

Para especificar uma métrica personalizada, cie uma carga útil em formato JSON usando um conjunto de parâmetros exigidos de um modelo. Adicione os valores para cada parâmetro do CloudWatch. Nós fornecemos o modelo como parte das opções personalizadas para Scaling metric (Métrica de dimensionamento) e Load metric (Métrica de carga) nas configurações avançadas do seu plano de dimensionamento.

JSON representa dados de duas formas:

- Um objeto, que é uma coleção não ordenada de pares de nome/valor. Um objeto é definido nas chaves esquerda e direita ({) (}). Cada par de nome e valor começa com o nome seguido por uma vírgula seguido pelo valor. O pares de nome-valor são separados por vírgulas.
- Uma matriz, que é uma coleção ordenada de valores. Uma matriz é definida nas chaves esquerda ([) e direita (]). Os itens na matriz são separados por vírgulas.

Este é um exemplo do modelo JSON com valores de amostra para cada parâmetro:

```
{
  "MetricName": "MyBackendCPU",
  "Namespace": "MyNamespace",
  "Dimensions": [
    {
      "Name": "MyOptionalMetricDimensionName",
      "Value": "MyOptionalMetricDimensionValue"
    }
  ],
  "Statistic": "Sum"
}
```

Para obter mais informações, consulte Especificação da métrica personalizada de escalabilidade e Especificação da métrica personalizada de carga na Referência da API do AWS Auto Scaling.

Configurações de dimensionamento dinâmico

Use este procedimento para visualizar e personalizar as configurações para a política de dimensionamento de rastreamento de destino que o AWS Auto Scaling cria.

Para visualizar e personalizar as configurações do dimensionamento dinâmico

1. Para expandir a visualização e ver os detalhes da seção Dynamic scaling settings (Configurações do dimensionamento dinâmico), selecione a seta à esquerda do cabeçalho da seção.
2. Você pode optar pelos itens a seguir. No entanto, as configurações padrão são adequadas para este tutorial.

   - Replace external scaling policies (Substituir as políticas externas de escalabilidade): se essa configuração estiver desmarcada, as políticas existentes de escalabilidade criadas ficarão de fora do plano de escalabilidade, e não serão criadas outras.
   - Disable scale-in (Desabilitar redução da escala na horizontal): se essa configuração estiver desmarcada, a redução automática da escala na horizontal para diminuir a capacidade atual do recurso será permitida quando a métrica especificada estiver abaixo do valor pretendido.
   - Cooldown (Desaquecimento): cria períodos de desaquecimento para o aumento e a redução da escala na horizontal. O período de desaquecimento é o tempo de espera que a política de escalabilidade aguarda para que uma ação de escalabilidade anterior entre em vigor. Para obter mais informações, consulte Período de desaquecimento no Manual do usuário do Application Auto Scaling. (Essa configuração não será exibida se o recurso for um grupo do Auto Scaling.)
   - Instance warmup (Carregamento da instância): [apenas para grupos do Auto Scaling] controla o período decorrido antes que uma instância recém-executada comece a contribuir para as métricas do CloudWatch. Para obter mais informações, consulte Carregamento da instância no Manual do usuário do Amazon EC2 Auto Scaling.

Configurações de dimensionamento preditivo

Se o recurso for um grupo do Auto Scaling, siga este procedimento para ver e personalizar as configurações que o AWS Auto Scaling deve usar para a escalabilidade preditiva.

Para visualizar e personalizar as configurações do dimensionamento preditivo

1. Para expandir a visualização e ver os detalhes da seção Predictive scaling settings (Configurações do dimensionamento preditivo), selecione a seta à esquerda do cabeçalho da seção.
2. Você pode optar pelos itens a seguir. Para este tutorial, altere o Predictive scaling mode (Modo de dimensionamento preditivo) para Forecast only (Somente previsão).

- Predictive scaling mode (Modo de escalabilidade preditiva): especifica o modo de escalabilidade. O padrão é Forecast and scale (Previsão e escala). Se você alterá-lo para Forecast only (Somente previsão), o plano de dimensionamento vai prever a capacidade futura, mas não vai aplicar as ações de dimensionamento.
- Pre-launch instances (Pré-executar instâncias): ajusta as ações de escalabilidade para serem executadas mais cedo com a redução da escala. Por exemplo, a previsão diz para adicionar capacidade às 10h e o tempo de buffer é de 5 minutos (300 segundos). A hora da execução da ação de escalabilidade correspondente será às 9h55. Essa opção é útil para grupos do Auto Scaling, em que uma instância pode levar alguns minutos para entrar em serviço depois de ser iniciada. O tempo real pode variar porque depende de vários fatores, como o tamanho da instância e se há scripts de startup a serem concluídos. O padrão é 300 segundos.
- Max capacity behavior (Comportamento de capacidade máxima): controla se a escala do recurso selecionado poderá ser aumentada na vertical acima da capacidade máxima quando a capacidade da previsão estiver próxima ou exceder a capacidade máxima especificada no momento. O padrão é Enforce the maximum capacity setting (Aplicar a configuração de capacidade máxima).
  - Enforce the maximum capacity setting (Aplicar a configuração de capacidade máxima): o AWS Auto Scaling não pode escalar a capacidade do recurso acima da capacidade máxima. A capacidade máxima é imposta como um limite fixo.
  - Set the maximum capacity to equal forecast capacity (Definir a capacidade máxima para igualar a capacidade da previsão): o AWS Auto Scaling pode escalar a capacidade do recurso acima da capacidade máxima para igualar, mas não exceder, a capacidade da previsão.
  - Increase maximum capacity above forecast capacity (Aumentar a capacidade máxima acima da capacidade da previsão): o AWS Auto Scaling pode escalar a capacidade do recurso acima da capacidade máxima por um valor de buffer especificado. A intenção é dar à política de escalabilidade de rastreamento de destino capacidade extra se ocorrer tráfego inesperado.
- Max capacity behavior buffer (Buffer de comportamento da capacidade máxima): se você escolheu Increase maximum capacity above forecast capacity (Aumentar a capacidade máxima acima da capacidade da previsão), escolha o tamanho do buffer da capacidade a ser usado quando a capacidade da previsão estiver próxima ou exceder a capacidade máxima. O valor é especificado como uma porcentagem em relação à capacidade de previsão. Por exemplo, com um buffer de 10%, se a capacidade da previsão for 50, e a capacidade máxima for 40, a capacidade máxima efetiva será 55.

3. Ao concluir as configurações personalizadas, selecione Next (Próximo).

   Note

   Para reverter qualquer alteração, selecione os recursos e, em seguida, selecione Revert to original (Reverter para original). Isso redefine os recursos selecionados para o estado conhecido mais recentemente dentro do plano de escalabilidade.

Etapa 4: Criar o plano de escalabilidade

Na página Review and create (Revisar e criar), revise os detalhes do seu plano de escalabilidade e selecione Create scaling plan (Criar plano de escalabilidade). Você é direcionado para uma página que mostra o status do plano de dimensionamento. O plano de dimensionamento pode levar um tempo para terminar de ser criado enquanto os recursos são atualizados.

Com o dimensionamento preditivo, o AWS Auto Scaling analisa o histórico da métrica de carga especificada pelos últimos 14 dias (é necessário o mínimo de 24 horas de dados) para gerar uma previsão para os dois próximos dias. Então, ele programa ações de dimensionamento para ajustar a capacidade do recurso a fim de corresponder à previsão para cada hora do período da previsão.

Depois que a criação do plano de dimensionamento for concluída, visualize os detalhes desse plano selecionando o nome dele na tela Scaling plans (Planos de dimensionamento).

(Opcional) Ver as informações de escalabilidade de um recurso

Use este procedimento para visualizar as informações de dimensionamento criadas para um recurso.

Os dados são apresentados das seguintes maneiras:

- Gráficos que mostram dados recentes do histórico de métricas do CloudWatch.
- Gráficos de dimensionamento preditivo mostrando previsões de carga e previsões de capacidade com base nos dados do AWS Auto Scaling.
- Uma tabela que lista todas as ações de dimensionamento preditivo programadas para o recurso.

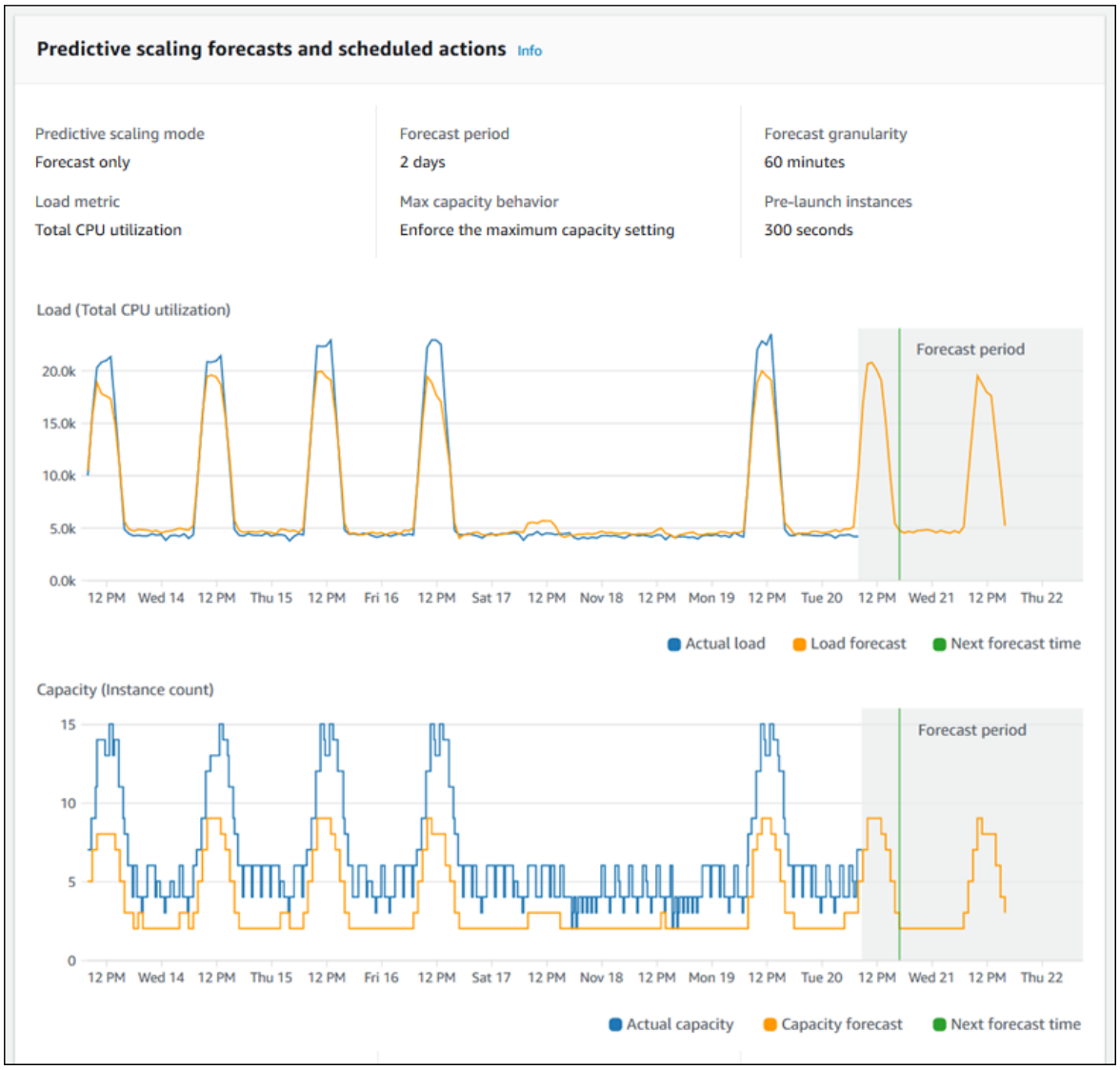
Para visualizar informações de dimensionamento de um recurso

1. Abra a console do AWS Auto Scaling em https://console.aws.amazon.com/awsautoscaling/.
2. Na página Scaling plans (Planos de dimensionamento), escolha o plano de escalabilidade.
3. Na página Scaling plan details (Detalhes de plano de escalabilidade), escolha o recurso para exibir.

Monitorar e avaliar previsões

Quando seu plano de escalabilidade estiver em funcionamento, você poderá monitorar a previsão de carga, a previsão de capacidade e as ações de escalabilidade para examinar a performance da escalabilidade preditiva. Todos esses dados são disponibilizados na console do AWS Auto Scaling para todos os grupos do Auto Scaling que estão habilitados para escalabilidade preditiva. Lembre-se de que o plano de dimensionamento exige pelo menos 24 horas de dados de carga históricos para fazer a previsão inicial.

No exemplo a seguir, o lado esquerdo de cada gráfico mostra um padrão histórico. O lado direito mostra a previsão que foi gerada pelo plano de dimensionamento para o período de previsão. Tanto os valores reais e previstos (em azul e laranja) são representados.



AWS Auto ScalingO aprende com seus dados automaticamente. Primeiro, ele faz uma previsão de carga. Em seguida, um cálculo da previsão de capacidade determina o número mínimo de instâncias que são necessárias para oferecer suporte ao aplicativo. Com base na previsão de capacidade, o AWS Auto Scaling agenda ações de escalabilidade que escalam o grupo do Auto Scaling antes das alterações de carga previstas. Se a escalabilidade dinâmica estiver habilitada (recomendado), o grupo do Auto Scaling poderá aumentar a escala da capacidade adicional na horizontal (ou remover a capacidade) com base na utilização atual do grupo de instâncias.

Ao avaliar o grau de sucesso da escalabilidade preditiva, monitore a correspondência da previsão e os valores reais ao longo do tempo. Quando você cria um plano de escalabilidade, o AWS Auto Scaling fornece gráficos com base nos dados reais mais recentes. Ele também fornece uma previsão inicial para as próximas 48 horas. No entanto, quando o plano de escalabilidade é criado, há muito poucos dados previstos para comparar aos dados reais. Aguarde até que o plano de escalabilidade obtenha valores de previsão por alguns períodos antes de comparar os valores de previsão históricos com os valores reais. Após alguns dias de previsões diárias, você terá uma amostra maior de valores de previsão para comparar com os valores reais.

Para padrões que ocorrem diariamente, o intervalo de tempo entre a criação do seu plano de escalabilidade e a avaliação da eficiência da previsão pode ser de apenas alguns dias. No entanto, esse

período não é suficiente para avaliar a previsão com base em uma alteração de padrão recente. Por exemplo, digamos que você esteja visualizando a previsão para um grupo do Auto Scaling que iniciou uma nova campanha de marketing na semana passada. A campanha aumenta significativamente o tráfego da web nos mesmos dois dias a cada semana. Em situações como essa, recomendamos aguardar que o grupo colete uma semana ou duas de novos dados antes de avaliar a eficácia da previsão. A mesma recomendação se aplica a um novo grupo do Auto Scaling que tenha apenas começado a coletar dados de métrica.

Se os valores previstos e reais não corresponderem após seu monitoramento ao longo de um período adequado, você também deverá considerar sua opção de métrica de carga. Para garantir a eficácia, a métrica de carga precisa representar uma medida confiável e precisa da carga total em todas as instâncias no grupo do Auto Scaling. A métrica de carga é essencial do dimensionamento preditivo. Se você escolher uma métrica de carga que não seja ideal, ela poderá impedir que a escalabilidade preditiva faça previsões precisas de carga e de capacidade e agende os ajustes de capacidade corretos para o grupo do Auto Scaling.

Etapa 5: Limpar

Depois de concluir o tutorial de conceitos básicos, você poderá optar por manter o plano de escalabilidade. Contudo, se não estiver usando ativamente seu plano de escalabilidade, você deve considerar a remoção deles para que sua conta não incorra em cobranças desnecessárias.

A exclusão de um plano de escalabilidade exclui as políticas de escalabilidade com monitoramento do objetivo, os alarmes do CloudWatch associados e as ações de escalabilidade preditiva que o AWS Auto Scaling criou em seu nome.

A exclusão de um plano de escalabilidade não exclui a pilha do AWS CloudFormation, o grupo do Auto Scaling nem outros recursos escaláveis.

Para excluir um plano de dimensionamento

1. Abra a console do AWS Auto Scaling em https://console.aws.amazon.com/awsautoscaling/.
2. Na página Scaling plans (Planos de dimensionamento), selecione o plano de dimensionamento que você criou para este tutorial e selecione Delete (Excluir).
3. Quando a confirmação for solicitada, escolha Excluir.

Depois de excluir seu plano de escalabilidade, os recursos não serão revertidos para a capacidade original. Por exemplo, se o grupo do Auto Scaling for escalado para 10 instâncias quando você excluir o plano de escalabilidade, o grupo ainda estará escalado para 10 instâncias após a exclusão do plano de escalabilidade. Você pode atualizar a capacidade de recursos específicos acessando o console para cada serviço individual.

Excluir o grupo do Auto Scaling

Para evitar que sua conta acumule cobranças do Amazon EC2, exclua também o grupo do Auto Scaling criado para este tutorial.

Para obter instruções detalhadas, consulte Excluir o grupo do Auto Scaling no Manual do usuário do Amazon EC2 Auto Scaling.

Etapa 6: próximas etapas

Agora que você se familiarizou com os planos de escalabilidade e alguns de seus recursos, recomendamos que crie seu próprio template de plano de escalabilidade usando o AWS CloudFormation.

Um modelo do AWS CloudFormation é um arquivo de texto em formato YAML ou JSON que descreve a infraestrutura da Amazon Web Services necessária para executar uma aplicação ou um serviço com as interconexões entre os componentes da infraestrutura. Com o AWS CloudFormation, você implanta e gerencia uma coleção associada de recursos como uma pilha. O AWS CloudFormation está disponível sem custo adicional e você paga apenas pelos recursos da AWS necessários para executar seus aplicativos. Os recursos podem consistir em qualquer recurso da AWS definido no modelo. Para obter mais informações, consulte Conceitos do AWS CloudFormation no Manual do usuário do AWS CloudFormation.

No Manual do usuário do AWS CloudFormation, apresentamos um modelo simples para você começar. O modelo de exemplo está disponível como exemplo na seção AWS::AutoScalingPlans::ScalingPlan da documentação de referência de modelos do AWS CloudFormation. O modelo de exemplo cria um plano de escalabilidade para um único grupo do Auto Scaling e habilita as escalabilidades preditiva e dinâmica.

Para obter mais informações, consulte Conceitos básicos do AWS CloudFormation no Manual do usuário do AWS CloudFormation.

Segurança do plano de escalabilidade

A segurança da nuvem na AWS é a nossa maior prioridade. Como cliente da AWS, você se contará com um datacenter e uma arquitetura de rede criados para atender aos requisitos das organizações com as maiores exigências de segurança.

A segurança é uma responsabilidade compartilhada entre a AWS e você. O modelo de responsabilidade compartilhada descreve isso como segurança da nuvem e segurança na nuvem:

- Segurança da nuvem: a AWS é responsável pela proteção da infraestrutura que executa produtos da AWS na Nuvem AWS. A AWS também fornece serviços que podem ser usados com segurança. Auditores externos testam e verificam regularmente a eficácia da nossa segurança como parte dos Programas de conformidade da AWS. Para saber mais sobre os programas de compatibilidade que se aplicam ao AWS Auto Scaling, consulte Serviços da AWS em escopo por programa de compatibilidade.
- Segurança da nuvem: sua responsabilidade é determinada pelo serviço da AWS que você usa. Você também é responsável por outros fatores, incluindo a confidencialidade de seus dados, os requisitos da sua empresa e as leis e regulamentos aplicáveis.

Esta documentação ajuda você a entender como aplicar o modelo de responsabilidade compartilhada ao usar planos de escalabilidade e como gerenciar o acesso aos planos de escalabilidade.

Tópicos

Planos de escalabilidade e endpoints de interface da VPC (AWS PrivateLink)

É possível estabelecer uma conexão privada entre planos de escalabilidade e outros Serviços da AWS e serviços de endpoint criando um endpoint de interface da VPC. É possível usar essa conexão para chamar a API do AWS Auto Scaling na VPC sem enviar tráfego pela Internet. O endpoint fornece conectividade confiável e dimensionável para a API do AWS Auto Scaling. Ele faz isso sem exigir um gateway da internet, uma instância NAT ou uma conexão VPN.

Os endpoints da VPC de interface são fornecidos pelo AWS PrivateLink, um recurso que permite a comunicação privada entre os Serviços da AWS usando endereços IP privados. Para mais informações, consulte AWS PrivateLink.

Criar um endpoint de interface da VPC para planos de escalabilidade

É possível criar um endpoint da VPC para o AWS Auto Scaling (planos de escalabilidade) usando o console da Amazon VPC ou a AWS Command Line Interface (AWS CLI). Crie um endpoint para o AWS Auto Scaling usando o seguinte nome de serviço:

- com.amazonaws.*region*.autoscaling-plans: cria um endpoint para as operações de API do AWS Auto Scaling.

Para obter mais informações, consulte Criar um endpoint de interface no Guia do usuário da Amazon VPC.

Habilite o DNS privado para que o endpoint faça solicitações de API para o serviço compatível usando o nome do host DNS padrão (por exemplo, `autoscaling-plans.us-east-1.amazonaws.com`). Na criação de um endpoint para Serviços da AWS, essa configuração é habilitada por padrão. Para obter mais informações, consulte Acessar um serviço por um endpoint de interface no Manual do usuário da Amazon VPC.

Não é necessário alterar as configurações do AWS Auto Scaling. AWS Auto Scaling A API chama outros Serviços da AWS usando endpoints de serviço públicos ou endpoints de interface da VPC privados, o que estiver em uso.

Criar uma política de endpoint da VPC para planos de escalabilidade

Você pode anexar uma política ao VPC endpoint para controlar o aceso à API do AWS 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 do VPC endpoint que nega a todos permissão para excluir um plano 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-plans:DeleteScalingPlan",
            "Effect": "Deny",
            "Resource": "*",
            "Principal": "*"
        }
    ]
}
```

Para obter mais informações, consulte Usar políticas de endpoint da VPC no Manual do usuário do Amazon VPC.

Migração de endpoints

Em 22 de novembro de 2019, apresentamos `autoscaling-plans.region.amazonaws.com` como o nome de host DNS padrão e endpoint para chamadas à API do AWS Auto Scaling. O novo endpoint é compatível com a última versão da AWS CLI e dos SDKs. Caso você ainda não tenha feito isso, instale a AWS CLI e os SDKs mais recentes para usar o novo endpoint. Para atualizar a AWS CLI, consulte

Instalar a AWS CLI usando pip no Manual do usuário da AWS Command Line Interface. Para obter mais informações sobre os SDKs da AWS, consulte Ferramentas para a Amazon Web Services.

> **Important**
>
> Para a compatibilidade com versões anteriores, o endpoint `autoscaling.region.amazonaws.com` existente continuará tendo suporte para chamadas à API do AWS Auto Scaling. Para configurar o endpoint `autoscaling.region.amazonaws.com` como um endpoint privado da VPC de interface, consulte Amazon EC2 Auto Scaling e endpoints da VPC de interface no Manual do usuário do Amazon EC2 Auto Scaling.

Endpoint a ser chamado ao usar a CLI ou a API do AWS Auto Scaling

Para a versão atual do AWS Auto Scaling, as chamadas à API do AWS Auto Scaling vão automaticamente para o endpoint `autoscaling-plans.region.amazonaws.com` em vez de `autoscaling.region.amazonaws.com`.

Você pode chamar o novo endpoint na CLI usando o parâmetro a seguir com cada comando para especificar o endpoint: `--endpoint-url https://autoscaling-plans.region.amazonaws.com`.

Embora não seja recomendado, também é possível chamar o endpoint antigo na CLI usando o seguinte parâmetro com cada comando para especificar o endpoint: `--endpoint-url https://autoscaling.region.amazonaws.com`.

Para os vários SDKs usados para chamar as APIs, consulte a documentação do SDK de interesse para saber mais sobre como direcionar as solicitações para um endpoint específico. Para mais informações, consulte Ferramentas para a Amazon Web Services.

Proteção de dados

O modelo de responsabilidade compartilhada da AWS se aplica à proteção de dados no AWS Auto Scaling. Conforme descrito nesse modelo, a AWS é responsável por proteger a infraestrutura global que executa toda a Nuvem AWS. Você é responsável por manter o controle sobre seu conteúdo hospedado nessa infraestrutura. Esse conteúdo inclui as tarefas de configuração e gerenciamento de segurança dos Serviços da AWS que você usa. Para obter mais informações sobre a privacidade de dados, consulte as Perguntas frequentes sobre privacidade de dados. Para obter mais informações sobre a proteção de dados na Europa, consulte a postagem do blog AWS Shared Responsibility Model and GDPR no Blog de segurança da AWS.

Para fins de proteção de dados, recomendamos que você proteja as credenciais da conta da Conta da AWS e configure as contas de usuário individuais com o AWS Identity and Access Management (IAM). Dessa maneira, cada usuário receberá apenas as permissões necessárias para cumprir suas obrigações de trabalho. Recomendamos também que você proteja seus dados das seguintes formas:

- Use uma autenticação multifator (MFA) com cada conta.
- Use SSL/TLS para se comunicar com os recursos da AWS. Recomendamos TLS 1.2 ou posterior.
- Configure o registro em log das atividades da API e do usuário com o AWS CloudTrail.
- Use as soluções de criptografia da AWS, juntamente com todos os controles de segurança padrão nos serviços da AWS.
- Use serviços gerenciados de segurança avançada, como o Amazon Macie, que ajuda a localizar e proteger dados pessoais armazenados no Amazon S3.
- Se você precisar de módulos criptográficos validados pelo FIPS 140-2 ao acessar a AWS por meio de uma interface de linha de comando ou uma API, use um endpoint do FIPS. Para obter mais informações sobre endpoints do FIPS, consulte Federal Information Processing Standard (FIPS) 140-2.

É altamente recomendável que você nunca coloque informações de identificação confidenciais, como endereços de e-mail dos seus clientes, em marcações ou campos de formato livre, como um campo Name (Nome). Isso também vale para o uso do AWS Auto Scaling ou de outros serviços da AWS com o console, a API, a AWS CLI ou os AWS SDKs. Quaisquer dados inseridos em marcações ou campos de formato livre usados para nomes podem ser usados para logs de cobrança ou diagnóstico. Se você fornecer um URL para um servidor externo, recomendemos fortemente que não sejam incluídas informações de credenciais no URL para validar a solicitação a esse servidor.

Gerenciamento de Identidade e Acesso para planos de escalabilidade

O AWS Identity and Access Management (IAM) é um serviço da AWS service (Serviço da AWS) que ajuda a controlar o acesso aos recursos da AWS de forma segura. Os administradores do IAM controlam quem pode ser autenticado (conectado) e autorizado (ter permissões) a usar os recursos do AWS Auto Scaling. O IAM é um AWS service (Serviço da AWS) que pode ser usado sem custo adicional.

Para usar planos de escalabilidade, você precisa de uma Conta da AWS e de credenciais. Para aumentar a segurança da sua conta, recomendamos usar um usuário do IAM para fornecer credenciais de acesso, em vez de usar as credenciais de sua conta da Conta da AWS. Para obter mais informações, consulte Credenciais do usuário raiz da conta da Amazon Web Services e do usuário do IAM na Referência geral da AWS e as Práticas recomendadas do IAM no Manual do usuário do IAM.

Para ter uma visão geral dos usuários do IAM e saber por que eles são importantes para a segurança da sua conta, consulte Credenciais de segurança da AWS na Referência geral da AWS.

Para obter detalhes sobre como usar o IAM, consulte o Manual do usuário do IAM.

Controle de acesso

É possível ter credenciais válidas para autenticar suas solicitações. No entanto, a menos que tenha permissões, não é possível criar nem acessar os planos de escalabilidade. Por exemplo, é necessário ter permissões para criar planos de escalabilidade, configurar escalabilidade preditiva etc.

As seções a seguir apresentam detalhes sobre como um administrador do IAM pode usar o IAM para ajudar a proteger seus planos de escalabilidade, controlando quem pode trabalhar com planos de escalabilidade.

Tópicos

Como os planos de escalabilidade funcionam com o IAM

Antes de usar o IAM para gerenciar quem pode criar, acessar e gerenciar planos de escalabilidade do AWS Auto Scaling, você precisa saber quais recursos do IAM estão disponíveis para uso com planos de escalabilidade.

Tópicos

Políticas baseadas em identidade

Com as políticas baseadas em identidade do IAM, é possível especificar ações ou recursos permitidos ou negados, além das condições sob as quais as ações são permitidas ou negadas. Os planos de escalabilidade são compatíveis com ações, recursos e chaves de condição específicas. Para saber mais sobre todos os elementos usados em uma política JSON, consulte Referência de elementos de política JSON do IAM no Manual do usuário do IAM.

Ações

Os administradores podem usar AWS as políticas JSON da para especificar quem tem acesso a quê. Ou seja, qual principal pode executar ações em quais recursos, e em que condições.

O elemento `Action` de uma política JSON descreve as ações que você pode usar para permitir ou negar acesso em uma política. As ações de política geralmente têm o mesmo nome que a operação de API da AWS associada. Existem algumas exceções, como ações somente de permissão, que não têm uma operação de API correspondente. Há também algumas operações que exigem várias ações em uma política. Essas ações adicionais são chamadas de ações dependentes.

Inclua ações em uma política para conceder permissões para executar a operação associada.

As ações de plano de escalabilidade em instruções de políticas do IAM usam este prefixo antes da ação: `autoscaling-plans:`. As instruções de política devem incluir um elemento `Action` ou `NotAction`. Os planos de escalabilidade têm seus próprios conjuntos de ações que descrevem as tarefas que podem ser executadas com esse serviço.

Para especificar várias ações em uma única declaração, separe-as com vírgulas, conforme exibido no exemplo a seguir.

```
"Action": [
      "autoscaling-plans:DescribeScalingPlans",
      "autoscaling-plans:DescribeScalingPlanResources"
```

Você também pode especificar várias ações usando caracteres curinga (*). Por exemplo, para especificar todas as ações que começam com a palavra `Describe`, inclua a ação a seguir.

```
"Action": "autoscaling-plans:Describe*"
```

Para ver uma lista completa de ações do plano de escalabilidade que podem ser usadas em declarações de políticas, consulte Ações, recursos e chaves de condição para o AWS Auto Scaling na Referência de autorização de serviço.

Recursos

O elemento `Resource` especifica o objeto ou os objetos aos quais a ação se aplica.

Os planos de escalabilidade não têm recursos definidos pelo serviço que podem ser usados como o elemento `Resource` de uma declaração de política do IAM. Portanto, não há nomes do recurso da Amazon (ARNs) para uso em uma política do IAM. Para controlar o acesso a ações do plano de escalabilidade, use sempre um * (asterisco) como recurso ao escrever uma política do IAM.

Chaves de condição

O elemento `Condition` (ou bloco de `Condition`) permite que você especifique condições nas quais uma instrução está em vigor. Por exemplo, é recomendável aplicar uma política somente após uma data específica. Para expressar condições, use chaves de condição predefinidas.

Os planos de escalabilidade não fornecem nenhuma chave de condição específica ao serviço, mas são compatíveis com o uso de algumas chaves de condição globais. Para ver todas AWS as chaves de condição globais, consulte AWS Chaves de contexto de condição globais no Manual do usuário do IAM.

O elemento `Condition` é opcional.

Exemplos

Para visualizar exemplos de políticas baseadas em identidade para planos e escalabilidade, consulte Exemplos de políticas baseadas em identidade para planos de escalabilidade (p. 27).

Políticas baseadas em recursos

Outros serviços da Amazon Web Services, como o Amazon Simple Storage Service, oferecem suporte a políticas de permissões baseadas em recursos. Por exemplo: você pode anexar uma política de permissões a um bucket do S3 para gerenciar permissões de acesso a esse bucket.

Os planos de escalabilidade não são compatíveis com as políticas baseadas em recurso.

Listas de controle de acesso (ACLs)

Os planos de escalabilidade não são compatíveis com listas de controle de acesso (ACLs).

Autorização baseada em tags

Não é possível etiquetar os planos de escalabilidade. Também não contam com recursos definidos pelo serviço que possam ser marcados. Portanto, não oferecem suporte ao controle de acesso com base em etiquetas de um recurso.

Os planos de escalabilidade podem conter recursos etiquetáveis, como grupos do Auto Scaling, que oferecem suporte a controle de acesso com base em etiquetas. Para obter mais informações, consulte a documentação do IAM para esse produto da AWS.

Funções do IAM

Uma função do IAM é uma entidade dentro da sua Conta da AWS que tem permissões específicas.

Usar credenciais temporárias

É possível usar credenciais temporárias para fazer login com federação, assumir uma função do IAM ou assumir uma função entre contas. As credenciais de segurança temporárias são obtidas chamando AWS STS operações da API como AssumeRole ou GetFederationToken.

Os planos de escalabilidade são compatíveis com o uso de credenciais temporárias.

Funções vinculadas a serviço para planos de escalabilidade

O AWS Auto Scaling usa funções vinculadas a serviço para as permissões de que ela precisa para chamar outros serviços da AWS em seu nome. As funções vinculadas a serviço facilitam a configuração dos planos de escalabilidade, já que não é preciso adicionar as permissões necessárias manualmente. Para obter mais informações, consulte Usar funções vinculadas a serviço no Manual do usuário do IAM.

O AWS Auto Scaling usa alguns tipos de funções vinculadas a serviço para chamar outros produtos da AWS em seu nome quando você trabalha com um plano de escalabilidade:

- Função vinculada ao serviço de escalabilidade preditiva: permite que o AWS Auto Scaling acesse dados de métricas históricas do CloudWatch. Também permite a criação de ações agendadas para grupos do Auto Scaling com base em uma previsão de carga e previsão de capacidade. Para mais informações, consulte Função vinculada ao serviço de escalabilidade preditiva (p. 26).
- Função vinculada ao serviço do Amazon EC2 Auto Scaling: permite que o AWS Auto Scaling acesse e gerencie políticas de dimensionamento com monitoramento do objetivo para grupos do Auto Scaling. Para obter mais informações, consulte Funções vinculadas a serviços do Amazon EC2 Auto Scaling no Guia do usuário do Amazon EC2 Auto Scaling.
- Função vinculada ao serviço do Application Auto Scaling: permite que o AWS Auto Scaling acesse e gerencie políticas de dimensionamento com monitoramento do objetivo para outros recursos escaláveis. Há uma função vinculada ao serviço para cada serviço. Para obter mais informações, consulte Funções vinculadas a serviço do Application Auto Scaling, no Guia do usuário do Application Auto Scaling.

É possível usar o procedimento a seguir para determinar se sua conta já tem uma função vinculada ao serviço.

Como determinar se uma função vinculada ao serviço já existe

1. Abra o console do IAM em https://console.aws.amazon.com/iam/.
2. No painel de navegação, selecione Roles (Funções).
3. Procure na lista `AWSServiceRole` para localizar as funções vinculadas a serviços existentes em sua conta. Procure o nome da função vinculada ao serviço que você deseja verificar.

Funções de serviço

O AWS Auto Scaling não conta com funções de serviço para planos de escalabilidade.

Função vinculada ao serviço de escalabilidade preditiva

O AWS Auto Scaling usa funções vinculadas a serviço para as permissões necessárias para chamar outros produtos da AWS em seu nome quando você trabalha com um plano de escalabilidade. Para mais informações, consulte Funções vinculadas a serviço para planos de escalabilidade (p. 25).

As seções a seguir descrevem como criar e gerenciar a função vinculada a serviço para escalabilidade preditiva. Primeiro, configure permissões para que uma entidade do IAM (por exemplo, um usuário, um grupo ou uma função) crie, edite ou exclua uma função vinculada ao serviço.

Permissões concedidas pela função vinculada ao serviço

O AWS Auto Scaling usa a função vinculada a serviço chamada AWSServiceRoleForAutoScalingPlans_EC2AutoScaling para chamar as seguintes ações para você, quando habilitar a escalabilidade preditiva:

- `cloudwatch:GetMetricData`
- `autoscaling:DescribeAutoScalingGroups`
- `autoscaling:DescribeScheduledActions`
- `autoscaling:BatchPutScheduledUpdateGroupAction`
- `autoscaling:BatchDeleteScheduledAction`

AWSServiceRoleForAutoScalingPlans_EC2AutoScaling confia no serviço `autoscaling-plans.amazonaws.com` para assumir a função.

Criar a função vinculada ao serviço (automática)

Não é preciso criar manualmente a função AWSServiceRoleForAutoScalingPlans_EC2AutoScaling. O AWS Auto Scaling cria essa função quando você cria um plano de escalabilidade em sua conta e habilita a escalabilidade preditiva.

Para o AWS Auto Scaling criar uma função vinculada ao serviço em seu nome, você deve ter as permissões necessárias. Para obter mais informações, consulte Service-linked role permissions (Permissões de função vinculada a serviços) no Guia do usuário do IAM.

Criar a função vinculada ao serviço (manual)

Para criar a função vinculada a serviço manualmente, é possível usar o console do IAM, a CLI do IAM ou a API do IAM. Para obter mais informações, consulte Criar uma função vinculada ao serviço no Manual do usuário do IAM.

Para criar uma função vinculada a serviço (AWS CLI)

Use o comando create-service-linked-role da CLI para criar a função vinculada a serviço.

```
aws iam create-service-linked-role --aws-service-name autoscaling-plans.amazonaws.com
```

Editar a função vinculada ao serviço

É possível editar a descrição de AWSServiceRoleForAutoScalingPlans_EC2AutoScaling usando o IAM. 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 precisar mais usar os planos de escalabilidade, é recomendável excluir a função AWSServiceRoleForAutoScalingPlans_EC2AutoScaling.

Somente é possível excluir uma função vinculada a serviço depois de excluir todos os planos de escalabilidade da Conta da AWS que tenham escalabilidade preditiva habilitada. Isso evita que a permissão para acessar os planos de escalabilidade seja removida por engano.

Você pode usar o console, a CLI do IAM ou a API do IAM para excluir a função vinculada a serviço. Para obter mais informações, consulte Excluir uma função vinculada ao serviço no Guia do usuário do IAM.

Depois de excluir a função vinculada a serviço AWSServiceRoleForAutoScalingPlans_EC2AutoScaling, o AWS Auto Scaling cria a função novamente quando você cria um plano de escalabilidade com a escalabilidade preditiva habilitada.

Regiões compatíveis

O AWS Auto Scaling oferece suporte a funções vinculadas a serviço em todas as Regiões da AWS em que os planos de escalabilidade estão disponíveis. Para obter informações sobre a disponibilidade regional de planos de escalabilidade, consulte Endpoints e cotas do AWS Auto Scaling na Referência geral da AWS.

Exemplos de políticas baseadas em identidade para planos de escalabilidade

Por padrão, um novo usuário do IAM não tem permissões para fazer nada. Um administrador do IAM deve criar políticas do IAM que concedam aos usuários e às funções permissão para trabalhar com planos de

escalabilidade. O administrador deve anexar as políticas aos usuários ou às funções do IAM que exijam essas permissões.

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.

Caso você não tenha familiaridade com a criação de políticas, recomendamos primeiro criar um usuário do IAM na sua conta e anexar políticas a ele. Você pode usar o console para verificar os efeitos de cada política à medida que anexa a política ao usuário.

Tópicos

Práticas recomendadas de políticas

As políticas baseadas em identidade são muito eficientes. Elas determinam se alguém pode criar, acessar ou excluir recursos do AWS Auto Scaling em sua conta. Essas ações podem incorrer em custos para a Conta da AWS. Ao criar ou editar políticas baseadas em identidade, siga estas diretrizes e recomendações:

- Primeiro, use políticas gerenciadas pela AWS: para começar a usar o AWS Auto Scaling rapidamente, use as políticas gerenciadas pela AWS para conceder a seus funcionários as permissões de que precisam. Essas políticas já estão disponíveis em sua conta e são mantidas e atualizadas pela AWS. Para obter mais informações, consulte Começar a usar permissões com políticas gerenciadas da AWS no Manual do usuário do IAM.
- Conceder privilégio mínimo: ao criar políticas personalizadas, conceda apenas as permissões necessárias para executar uma tarefa. Comece com um conjunto mínimo de permissões e conceda permissões adicionais conforme necessário. Fazer isso é mais seguro do que começar com permissões que são muito lenientes e tentar restringi-las superiormente. Para obter mais informações, consulte Conceder privilégio mínimo no Manual do usuário do IAM.
- Habilitar MFA para operações confidenciais: para aumentar a segurança, exija que os usuários do IAM usem Multi-Factor Authentication (MFA) para acessar recursos ou operações de API confidenciais. Para obter mais informações, consulte Usar autenticação multifator (MFA) AWS no Guia do usuário do IAM.
- Usar condições de política para segurança adicional: na medida do possível, defina as condições sob as quais suas políticas baseadas em identidade permitem o acesso a um recurso. Por exemplo, você pode gravar condições para especificar um intervalo de endereços IP permitidos do qual a solicitação deve partir. Você também pode escrever condições para permitir somente solicitações em uma data especificada ou período ou para exigir o uso de SSL ou MFA. Para obter mais informações, consulte Elementos de política JSON do IAM: condição no Manual do usuário do IAM.

Permitir que os usuários criem planos de escalabilidade

Veja a seguir um exemplo de política de permissões que permite que os usuários criem planos de escalabilidade.

```
{
    "Version": "2012-10-17",
    "Statement": [
        {
```

```
            "Effect": "Allow",
            "Action": [
              "autoscaling-plans:*",
              "cloudwatch:PutMetricAlarm",
              "cloudwatch:DeleteAlarms",
              "cloudwatch:DescribeAlarms",
              "cloudformation:ListStackResources"
            ],
            "Resource": "*"
        }
    ]
}
```

Para trabalhar com um plano de escalabilidade, o usuário precisa ter permissões adicionais que permitam trabalhar com recursos específicos na conta dele. Essas permissões estão listadas em Permissões adicionais necessárias (p. 29).

Cada usuário do console também precisa de permissões para identificar os recursos escaláveis na conta deles e visualizar grafos de dados de métricas do CloudWatch no console do AWS Auto Scaling. O conjunto completo de permissões necessárias para trabalhar com o console do AWS Auto Scaling está listado abaixo:

- `cloudformation:ListStacks`: para listar pilhas.
- `tag:GetTagKeys`: para encontrar recursos escaláveis que contêm determinadas chaves de tag.
- `tag:GetTagValues`: para encontrar recursos que contêm determinados valores de tag.
- `autoscaling:DescribeTags`: para encontrar grupos do Auto Scaling que contêm determinadas etiquetas.
- `cloudwatch:GetMetricData`: para ver dados em gráficos de métricas.

Permitir que os usuários habilitem a escalabilidade preditiva

Veja a seguir um exemplo de política de permissões que permite que os usuários habilitem a escalabilidade preditiva. Essas permissões estendem os recursos dos planos de escalabilidade configurados para escalar grupos do Auto Scaling.

```
{
    "Version": "2012-10-17",
    "Statement": [
        {
            "Effect": "Allow",
            "Action": [
              "cloudwatch:GetMetricData",
              "autoscaling:DescribeAutoScalingGroups",
              "autoscaling:DescribeScheduledActions",
              "autoscaling:BatchPutScheduledUpdateGroupAction",
              "autoscaling:BatchDeleteScheduledAction"
            ],
            "Resource": "*"
        }
    ]
}
```

Permissões adicionais necessárias

Ao conceder permissões para planos de escalabilidade, é necessário decidir para quais recursos os usuários obterão permissões. Dependendo dos cenários aos quais você quer oferecer suporte, é possível especificar as ações a seguir no elemento `Action` de uma declaração de política do IAM.

Grupos do Auto Scaling

Para adicionar grupos do Auto Scaling a um plano de escalabilidade, os usuários precisam ter as seguintes permissões do Amazon EC2 Auto Scaling:

- `autoscaling:UpdateAutoScalingGroup`
- `autoscaling:DescribeAutoScalingGroups`
- `autoscaling:PutScalingPolicy`
- `autoscaling:DescribePolicies`
- `autoscaling:DeletePolicy`

serviços da ECS

Para adicionar serviços do ECS a um plano de escalabilidade, os usuários precisam ter as seguintes permissões do Amazon ECS e do Application Auto Scaling:

- `ecs:DescribeServices`
- `ecs:UpdateService`
- `application-autoscaling:RegisterScalableTarget`
- `application-autoscaling:DescribeScalableTargets`
- `application-autoscaling:DeregisterScalableTarget`
- `application-autoscaling:PutScalingPolicy`
- `application-autoscaling:DescribeScalingPolicies`
- `application-autoscaling:DeleteScalingPolicy`

Frota spot

Para adicionar frotas spot a um plano de escalabilidade, os usuários precisam ter as seguintes permissões do Amazon EC2 e do Application Auto Scaling:

- `ec2:DescribeSpotFleetRequests`
- `ec2:ModifySpotFleetRequest`
- `application-autoscaling:RegisterScalableTarget`
- `application-autoscaling:DescribeScalableTargets`
- `application-autoscaling:DeregisterScalableTarget`
- `application-autoscaling:PutScalingPolicy`
- `application-autoscaling:DescribeScalingPolicies`
- `application-autoscaling:DeleteScalingPolicy`

Tabelas ou índices globais do DynamoDB

Para adicionar tabelas ou índices globais do DynamoDB a um plano de escalabilidade, os usuários precisam ter as seguintes permissões do DynamoDB e do Application Auto Scaling:

- `dynamodb:DescribeTable`
- `dynamodb:UpdateTable`
- `application-autoscaling:RegisterScalableTarget`
- `application-autoscaling:DescribeScalableTargets`
- `application-autoscaling:DeregisterScalableTarget`

- `application-autoscaling:PutScalingPolicy`
- `application-autoscaling:DescribeScalingPolicies`
- `application-autoscaling:DeleteScalingPolicy`

clusters de bancos de dados Aurora

Para adicionar clusters de banco de dados do Aurora a um plano de escalabilidade, os usuários precisam ter as seguintes permissões do Amazon Aurora e do Application Auto Scaling:

- `rds:AddTagsToResource`
- `rds:CreateDBInstance`
- `rds:DeleteDBInstance`
- `rds:DescribeDBClusters`
- `rds:DescribeDBInstances`
- `application-autoscaling:RegisterScalableTarget`
- `application-autoscaling:DescribeScalableTargets`
- `application-autoscaling:DeregisterScalableTarget`
- `application-autoscaling:PutScalingPolicy`
- `application-autoscaling:DescribeScalingPolicies`
- `application-autoscaling:DeleteScalingPolicy`

Permissões necessárias para criar uma função vinculada ao serviço

O AWS Auto Scaling requer permissões para criar uma função vinculada a serviço na primeira vez que qualquer usuário em sua Conta da AWS cria um plano de escalabilidade com escalabilidade preditiva habilitada. Se a função vinculada ao serviço ainda não existir, o AWS Auto Scaling a criará em sua conta. A função vinculada ao serviço concede permissões ao AWS Auto Scaling, para que ele possa chamar todos os outros serviços 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"
```

Veja a seguir um exemplo de política de permissões para que um usuário crie uma função vinculada a serviço.

```
{
    "Version": "2012-10-17",
    "Statement": [
        {
            "Effect": "Allow",
            "Action": "iam:CreateServiceLinkedRole",
            "Resource": "arn:aws:iam::*:role/aws-service-role/autoscaling-
plans.amazonaws.com/AWSServiceRoleForAutoScalingPlans_EC2AutoScaling",
            "Condition": {
                "StringLike": {
                    "iam:AWSServiceName":"autoscaling-plans.amazonaws.com"
                }
            }
        }
    ]
```

```
}
```

Para mais informações, consulte Função vinculada ao serviço de escalabilidade preditiva (p. 26).

Validação de conformidade

Auditores externos avaliam a segurança e a conformidade dos serviços da Serviços da AWS como parte de vários programas de conformidade da AWS, como SOC, PCI, FedRAMP e HIPAA.

Para saber se o AWS Auto Scaling ou outros Serviços da AWS estão no escopo de programas específicos de conformidade, consulte Serviços da AWS no escopo por programa de conformidade. Para obter informações gerais, consulte Programas de conformidade da AWS.

Você pode fazer download de relatórios de auditoria de terceiros usando o AWS Artifact. Para obter mais informações, consulte Downloading Reports in AWS Artifact.

Sua responsabilidade de conformidade ao usar o Serviços da AWS é determinada pela confidencialidade dos seus dados, pelos objetivos de conformidade da sua empresa e pelos regulamentos e leis aplicáveis. A AWS fornece os seguintes recursos para ajudar com a conformidade:

- Guias de referência rápida de conformidade e segurança: esses guias de implantação abordam as considerações de arquitetura e fornecem etapas para implantação de ambientes de lista de referência na AWS concentrados em conformidade e segurança.
- Whitepaper Elaboração de arquitetura para segurança e conformidade com HIPAA: esse whitepaper descreve como as empresas podem usar a AWS para criar aplicações adequadas aos padrões HIPAA.

  Note

  Nem todos os Serviços da AWS estão qualificados pela HIPAA. Para mais informações, consulte a Referência dos serviços qualificados pela HIPAA.
- Recursos de conformidade da AWS: essa coleção de manuais e guias pode ser aplicada a seu setor e local.
- Avaliar recursos com regras no AWS Config Developer Guide (Guia do desenvolvedor do CCI): o serviço AWS Config avalia como as configurações de recursos estão em conformidade com práticas internas, diretrizes do setor e regulamentos.
- AWS Security Hub: esse AWS service (Serviço da AWS) fornece uma visão abrangente do estado de sua segurança na AWS que ajuda você a conferir sua conformidade com padrões e práticas recomendadas de segurança do setor.
- AWS Audit Manager: esse AWS service (Serviço da AWS) ajuda a auditar continuamente seu uso da AWS para simplificar a forma como você gerencia os riscos e a conformidade com regulamentos e padrões do setor.

Segurança da infraestrutura

Como serviço gerenciado, o AWS Auto Scaling é protegido pelos procedimentos de segurança da rede global da AWS que estão descritos no whitepaper Amazon Web Services: Overview of security processes.

Você usa chamadas de API publicadas pela AWS para acessar o AWS Auto Scaling por meio da rede. Os clientes devem oferecer suporte a Transport Layer Security (TLS) 1.0 ou posterior. Recomendamos TLS 1.2 ou posterior. Os clientes também devem ter suporte a conjuntos de criptografia com perfect forward secrecy (PFS) como Ephemeral Diffie-Hellman (DHE) ou Ephemeral Elliptic Curve Diffie-Hellman (ECDHE). A maioria dos sistemas modernos como Java 7 e versões posteriores oferece suporte a esses modos.

Além disso, as solicitações devem ser assinadas usando um ID da chave de acesso e uma chave de acesso secreta associada a uma entidade principal do IAM. Ou você pode usar o AWS Security Token Service (AWS STS) para gerar credenciais de segurança temporárias para assinar solicitações.

Cotas para planos de escalabilidade

Sua Conta da AWS tem as seguintes cotas de serviço (anteriormente chamadas de limites) relacionadas aos planos de escalabilidade.

Para solicitar um aumento, use o formulário Limites do Auto Scaling. Lembre-se de especificar o tipo de recurso na solicitação de aumento, por exemplo, Amazon EC2 Auto Scaling, Amazon ECS ou DynamoDB.

Cotas-padrão por região por conta

Item	Padrão
Número máximo de recursos escaláveis por tipo de recurso	As cotas variam dependendo do tipo de recurso. Amazon DynamoDB: 3000 Grupos do Amazon EC2 Auto Scaling: 200 Todos os outros tipos de recursos: 500
Número máximo de planos de escalabilidade	100
Número máximo de instruções de escalabilidade por plano de escalabilidade	500
Número máximo de configuração de rastreamento de destino por escalabilidade	10

Tenha em mente as cotas de serviço ao aumentar suas cargas de trabalho. Por exemplo, quando você atingir o número máximo de unidades de capacidade permitidas por um serviço, a expansão será interrompida. Se a demanda cair e a capacidade atual diminuir, o AWS Auto Scaling poderá expandir novamente. Para evitar atingir esse limite de cota de serviço novamente, é possível solicitar um aumento. Cada serviço tem suas próprias cotas padrão para a capacidade máxima do recurso. Para obter informações sobre as cotas-padrão para outros serviços da Amazon Web Services, consulte Endpoints e cotas de serviço na Referência geral da Amazon Web Services.

Recursos de planos de escalabilidade

Os recursos adicionais a seguir podem ajudar você a trabalhar com planos de escalabilidade.

- Saiba mais sobre o uso do monitoramento do objetivo para autoescalabilidade dos grupos do Amazon EC2 Auto Scaling.
- Saiba mais sobre o uso do monitoramento do objetivo para autoescalabilidade de seus recursos escaláveis além do Amazon EC2, como índices e tabelas do DynamoDB e serviços do Amazon ECS.
- Saiba como usar o AWS CloudTrail para chamadas de log feitas para a API do AWS Auto Scaling e armazená-las no Amazon S3. Você pode usar esses logs do CloudTrail para determinar quais chamadas foram feitas, o endereço IP de origem da chamada, quem fez a chamada, quando ela foi feita, etc.

Os recursos adicionais a seguir estão disponíveis para ajudar você a saber mais sobre a Amazon Web Services.

- Aulas e workshops – links para cursos de especialidades e baseados em função, bem como laboratórios autoguiados para ajudar a aperfeiçoar suas habilidades na AWS e a obter experiência prática.
- Ferramentas do desenvolvedor da AWS: links para ferramentas de desenvolvedor, SDKs, toolkits de IDE e ferramentas da linha de comando para desenvolver e gerenciar aplicações da 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: o centro para criar e gerenciar os seus casos da AWS Support. Também inclui links para outros recursos úteis, como fóruns, perguntas frequentes técnicas, status de integridade do serviço e AWS Trusted Advisor.
- AWS Support: a página Web principal para obter informações sobre o AWS Support, um canal de suporte de resposta rápida e com atendimento individual para ajudar a construir e a executar aplicações na nuvem.
- Entre em contato conosco: um ponto central de contato para consultas relativas a faturamento, conta, eventos, abuso e outros problemas da AWS.
- AWSTermos do site da : informações detalhadas sobre os nossos direitos autorais e marca registrada. Sua conta, licença e acesso ao site, entre outros tópicos.

Histórico do documento dos planos de escalabilidade

A tabela a seguir descreve adições importantes na documentação do AWS Auto Scaling. Para receber notificações sobre atualizações dessa documentação, você pode se inscrever em o feed RSS.

update-history-change	update-history-description	update-history-date
Novo capítulo "Segurança (p. 36)	Um novo capítulo Segurança no Manual do usuário do AWS Auto Scaling ajuda a entender como aplicar o modelo de responsabilidade compartilhada ao usar o AWS Auto Scaling. Como parte da atualização, o capítulo “Autenticação e controle de acesso” do manual do usuário foi substituído por uma seção nova e mais simples, Gerenciamento de identidade e acesso para o AWS Auto Scaling.	12 de março de 2020
Suporte para endpoints da Amazon VPC (p. 36)	Agora, você pode estabelecer uma conexão privada entre sua VPC e o AWS Auto Scaling. Para ver as considerações e instruções de migração, consulte AWS Auto Scaling e endpoints da VPC de interface.	22 de novembro de 2019
Suporte para aumentar a capacidade máxima acima da capacidade de previsão, além de alterações no guia (p. 36)	Adiciona suporte ao console para permitir que o plano de dimensionamento aumente a capacidade máxima acima da capacidade da previsão por um valor de buffer especificado. Para obter mais informações, consulte Configurações da escalabilidade preditiva no Manual do usuário do AWS Auto Scaling. Essa versão também inclui várias seções reescritas no tutorial Conceitos básicos do AWS Auto Scaling.	9 de março de 2019
Escalabilidade preditiva e melhorias (p. 36)	Agora, você pode usar a escalabilidade preditiva para escalar grupos do Amazon EC2 Auto Scaling. Esta versão também inclui suporte para a substituição de políticas de escalabilidade criadas fora do	20 de novembro de 2018

	plano de escalabilidade (como a partir de outros consoles) e controla se você ativa o recurso de escalabilidade dinâmico do plano. Para obter mais informações, consulte Conceitos básicos do AWS Auto Scaling.	
Suporte para a configuração de recursos personalizados (p. 36)	Suporte adicionado para a personalização de várias configurações para cada recurso individual ou vários recursos ao mesmo tempo. Para obter mais informações, consulte Conceitos básicos do AWS Auto Scaling.	9 de outubro de 2018
Tags como origem do aplicativo (p. 36)	Esta versão adiciona suporte para especificar um conjunto de tags como origem do aplicativo.	23 de abril de 2018
Novo serviço (p. 36)	Versão inicial do AWS Auto Scaling.	16 de janeiro de 2018