Guia do usuário de planos de escalabilidade

AWS Auto Scaling

AWS Auto Scaling: Guia do usuário de 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 AWS CloudFormation, você pode definir modelos de pilha para usar na criação de coleções de recursos. Então crie um plano de escalabilidade para os recursos escaláveis que pertencem a cada pilha.

Recursos compatíveis

AWS Auto Scaling suporta 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 encerre EC2 instâncias aumentando ou diminuindo a capacidade desejada de um grupo de 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 EC2 instâncias aumentando ou diminuindo a capacidade alvo de uma frota spot.

Recursos e benefícios do plano de escalabilidade

Os planos de escalabilidade fornecem estes recursos e benefícios:

- Descoberta de recursos — AWS Auto Scaling fornece descoberta automática de recursos para ajudar a encontrar recursos em seu aplicativo que possam ser escalados.
- Escalabilidade dinâmica — Os planos de escalabilidade usam os serviços Amazon Auto EC2 Scaling e Application Auto Scaling para ajustar a capacidade dos recursos escaláveis para lidar com mudanças no tráfego ou na carga de trabalho. As métricas de escalabilidade dinâmica podem ser métricas padrão de utilização ou de throughput 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 a EC2 capacidade da Amazon com mais rapidez quando ocorrem picos regulares.

> **Important**
>
> Se você usa planos de escalabilidade somente para escalabilidade preditiva, recomendamos que você defina políticas de escalabilidade preditiva diretamente em seus recursos de Auto Scaling. Essa opção oferece mais recursos, como o uso de agregações de métricas para criar novas métricas personalizadas ou reter dados métricos históricos em todas as blue/green implantações. Para obter mais informações sobre o Amazon EC2 Auto Scaling, consulte Escalabilidade preditiva para Amazon Auto EC2 Scaling no Guia do usuário do Amazon Auto EC2 Scaling. Para obter mais informações sobre o Application Auto Scaling, consulte Escalabilidade preditiva para Application Auto Scaling no Guia do Usuário do Application Auto Scaling.
>
> Para obter um guia sobre como migrar dos planos de escalabilidade para as políticas de escalabilidade preditiva EC2 do Amazon Auto Scaling, consulte. Migre seu plano de escalabilidade

Como começar

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

- Como funcionam os planos de escalabilidade
- Práticas recomendadas para planos de escalabilidade do
- Conceitos básicos dos planos de escalabilidade

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 em um Conta da AWS, você pode acessar seus planos de escalabilidade fazendo login no AWS Management Console, usando a caixa de pesquisa na barra de navegação para pesquisar e AWS Auto Scaling, em seguida, 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 o Guia do usuário do AWS Command Line Interface. Para obter mais informações, consulte planos de escalabilidade automática na Referência de comandos da AWS CLI .
- AWS Tools for Windows PowerShell— Fornece comandos para um amplo conjunto de AWS produtos para quem cria scripts no PowerShell ambiente. Para começar a usar, consulte o Guia do usuário do Ferramentas da AWS para PowerShell. Para obter mais informações, consulte Referência de Cmdlets do Ferramentas da AWS para PowerShell.
- AWS SDKs— fornece operações de API específicas do idioma e cuida de muitos detalhes da conexão, como calcular assinaturas, lidar com novas tentativas de solicitação e lidar com erros. Para obter mais informações, consulte AWS SDKs.
- API HTTPS: fornece ações de API de nível inferior que você chama usando solicitações HTTPS. Para obter mais informações, consulte a Referência da API do AWS Auto Scaling.
- AWS CloudFormation— Suporta a criação de planos de escalonamento usando CloudFormation modelos. Para obter mais informações, consulte a AWS::AutoScalingPlans::ScalingPlanreferência no Guia AWS CloudFormation do usuário.

Disponibilidade regional

A AWS Auto Scaling API está disponível em várias Regiões da AWS e fornece um endpoint para cada uma dessas regiões. Para obter uma lista de todas as regiões e endpoints em que a API está disponível atualmente, consulte AWS Auto Scaling endpoints e cotas nos endpoints e, para a Referência geral da AWSAmazon Web ARNs Services na .

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 CloudWatch e dos outros Nuvem AWS recursos que você usa.

> **Note**
>
> O recurso de escalabilidade preditiva depende da CloudWatch GetMetricDataoperação para coletar dados métricos históricos para previsão de capacidade, o que gera custos. No entanto, se você habilitar a escalabilidade preditiva com uma política de escalabilidade do Amazon EC2 Auto Scaling em vez de um plano de escalabilidade, não haverá cobranças pelas chamadas para. `GetMetricData`

Como funcionam os planos de escalabilidade

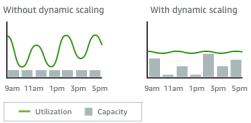
AWS Auto Scaling permite que você use planos de escalabilidade para configurar um conjunto de instruções para escalar seus recursos. Se você trabalha com AWS CloudFormation ou adiciona tags a recursos escaláveis, pode configurar planos de escalabilidade para diferentes conjuntos de recursos, por aplicativo. O AWS Auto Scaling console 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 explica AWS Auto Scaling como otimizar a utilização dos recursos em seu 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.

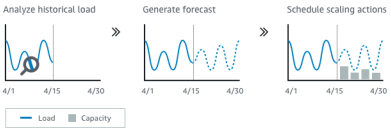


Por exemplo, você pode configurar seu plano de escalabilidade para manter o número de tarefas que o serviço do Amazon Elastic Container Service (Amazon ECS) executa em 75% da CPU. Quando a utilização da CPU do serviço ultrapassa 75% (o que significa que mais de 75% da CPU reservada para o serviço está 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.

> **Important**
>
> Se você usa planos de escalabilidade somente para escalabilidade preditiva, recomendamos que você defina políticas de escalabilidade preditiva diretamente em seus recursos de Auto Scaling. Essa opção oferece mais recursos, como o uso de agregações de métricas para criar novas métricas personalizadas ou reter dados métricos históricos em implantações azul/verdes. Para obter mais informações sobre o Amazon EC2 Auto Scaling, consulte Escalabilidade preditiva para Amazon Auto EC2 Scaling no Guia do usuário do Amazon Auto EC2 Scaling. Para obter mais informações sobre o Application Auto Scaling, consulte Escalabilidade preditiva para Application Auto Scaling no Guia do Usuário do Application Auto Scaling.
>
> Para obter um guia sobre como migrar dos planos de escalabilidade para as políticas de escalabilidade preditiva EC2 do Amazon Auto Scaling, consulte. Migre seu plano 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:

- Previsão de carga: AWS Auto Scaling analisa até 14 dias de histórico para uma métrica de carga especificada e prevê a demanda futura para os próximos dois dias. Esses dados estão disponíveis em intervalos de uma hora e são atualizados diariamente.
- Ações de escalonamento programadas: AWS Auto Scaling programa as ações de escalabilidade que aumentam e diminuem proativamente a capacidade de acordo com a previsão de carga. No horário programado, 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.

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 lançamento ou uma configuração de lançamento, ative o monitoramento detalhado para obter dados CloudWatch métricos das EC2 instâncias com 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 EC2 instâncias são habilitadas para monitoramento básico, o que significa que os dados métricos das 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 Auto EC2 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 CloudWatch Métricas de monitoramento para seus grupos e instâncias do Auto Scaling no Guia do usuário do Amazon Auto EC2 Scaling.
- Verifique qual tipo de instância o grupo do Auto Scaling usa e atente-se para o uso de um tipo de instância expansível. EC2 As instâncias da Amazon com desempenho intermitente, como instâncias T3 e T2, foram projetadas para fornecer um nível básico de desempenho de CPU com a capacidade de atingir um nível mais alto quando exigido pela sua carga de trabalho. 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 essas instâncias como`unlimited`, consulte Como usar um grupo de Auto Scaling para iniciar uma instância de desempenho intermitente como Ilimitada no Guia do usuário da Amazon EC2 .

Outras considerações

> **Important**
>
> Se você usa planos de escalabilidade somente para escalabilidade preditiva, recomendamos que você defina políticas de escalabilidade preditiva diretamente em seus recursos de Auto Scaling. Essa opção oferece mais recursos, como o uso de agregações de métricas para criar novas métricas personalizadas ou reter dados métricos históricos em implantações azul/verdes. Para obter mais informações sobre o Amazon EC2 Auto Scaling, consulte Escalabilidade preditiva para Amazon Auto EC2 Scaling no Guia do usuário do Amazon Auto EC2 Scaling. Para obter mais informações sobre o Application Auto Scaling, consulte Escalabilidade preditiva para Application Auto Scaling no Guia do Usuário do Application Auto Scaling.
>
> Para obter um guia sobre como migrar dos planos de escalabilidade para as políticas de escalabilidade preditiva EC2 do Amazon Auto Scaling, consulte. Migre seu plano de escalabilidade

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 ter mais informações, consulte Configurações de dimensionamento preditivo e Monitorar e avaliar previsões.
- 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 seu plano de escalabilidade, você deve excluir todas as ações de escalabilidade agendadas anteriormente que não sejam mais necessárias acessando os consoles a partir dos quais elas foram criadas. AWS Auto Scaling não cria uma ação de escalabilidade preditiva que se sobreponha a uma ação de escalabilidade 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.

Evitando o ActiveWithProblems erro

Um erro ActiveWithProblems "" pode ocorrer quando um plano de escalabilidade é criado ou 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. Opcionalmente, você pode excluir as políticas de escalabilidade existentes e substituí-las por políticas de escalabilidade de rastreamento de metas criadas no console. 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. Observe 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 se espera que as EC2 instâncias em um grupo de Auto Scaling usem em comparação com a CPU disponível. Ou, no caso de um serviço como o DynamoDB, que usa um modelo de throughput provisionado, a quantidade de atividades de leitura e gravação que uma tabela ou índice deve usar em comparação com o throughput 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. Saber disso ajuda você a configurar uma janela para que cada EC2 instância seja aquecida após a inicialização, a fim de garantir que um novo servidor não seja iniciado enquanto o anterior ainda estiver sendo executado.
- 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
- Etapa 2: Especificar a estratégia de escalabilidade
- Etapa 3: Definir configurações avançadas (opcional)

Etapa 1: Encontrar recursos escaláveis

Esta seção inclui uma introdução prática à criação de planos de escalabilidade no console do AWS Auto Scaling . Se este for seu primeiro plano de escalabilidade, recomendamos que você comece criando um exemplo de plano de escalabilidade usando um grupo do Amazon Auto EC2 Scaling.

Pré-requisitos

Para praticar o uso de um plano de escalabilidade, crie um grupo do Auto Scaling. Execute pelo menos uma EC2 instância da Amazon no grupo Auto Scaling. Para obter mais informações, consulte Comece a usar o Amazon EC2 Auto Scaling no Guia do usuário do Amazon Auto EC2 Scaling.

Use um grupo de Auto Scaling com CloudWatch métricas habilitadas para ter dados de capacidade nos gráficos que estão disponíveis quando você conclui o assistente Create Scaling Plan. Para obter mais informações, consulte Monitore CloudWatch métricas para seus grupos e instâncias do Auto Scaling no Guia do usuário do Amazon Auto EC2 Scaling.

Gere alguma carga por alguns dias ou mais para ter dados CloudWatch métricos 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.

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:

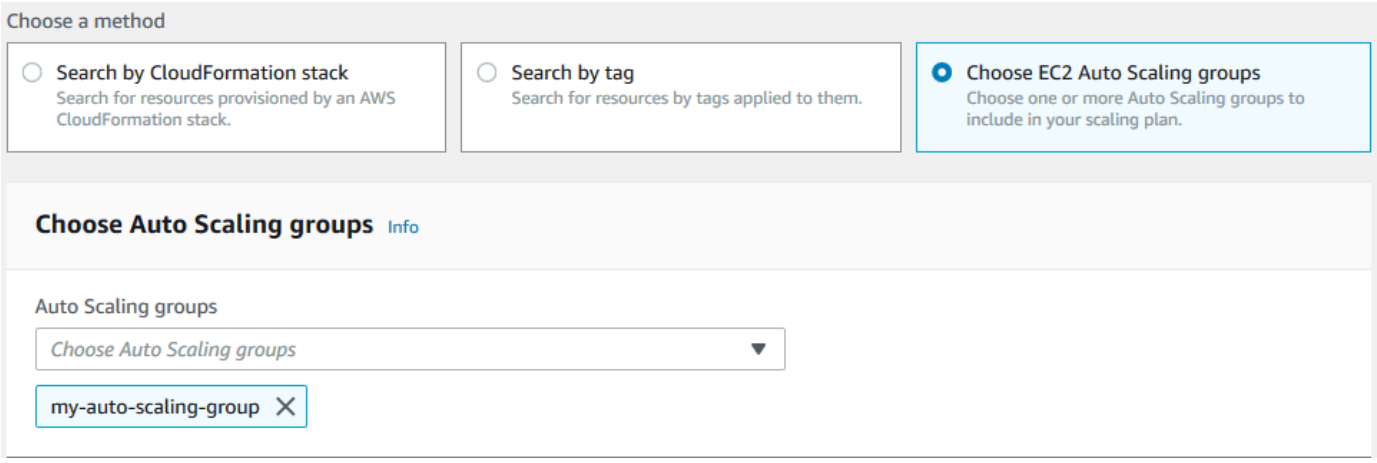
- Você criou um grupo de Auto Scaling e executou pelo menos uma EC2 instância, conforme descrito na seção anterior.
- O grupo do Auto Scaling criado existe há pelo menos 24 horas.

Para começar a criar um plano de escalabilidade

1. Abra o AWS Auto Scaling console 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 Pesquisar por CloudFormation pilha e, em seguida, escolha a AWS CloudFormation pilha 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 Remover.
   - Escolha Escolher grupos de EC2 Auto Scaling e, em seguida, escolha um ou mais grupos de Auto Scaling.

   > **Note**
   >
   > Para um tutorial introdutório, escolha Escolher grupos de EC2 Auto Scaling e, em seguida, escolha o grupo de 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 exemplo de plano de escalabilidade e gostaria de criar mais, consulte os cenários a seguir para usar uma CloudFormation pilha ou um conjunto de tags com mais detalhes. Você pode usar esta seção para decidir se deseja escolher a opção Pesquisar por CloudFormation pilha ou Pesquisar por tag para descobrir seus recursos escaláveis ao usar o console para criar seu plano de escalabilidade.

Quando você escolhe a opção Pesquisar por CloudFormation pilha ou Pesquisar por tag na etapa 1 do assistente Create Scaling Plan, isso torna os recursos escaláveis associados à pilha ou ao conjunto de tags disponíveis para o plano de escalabilidade. À medida que você define seu plano de dimensionamento, é possível escolher quais desses recursos incluir ou excluir.

Descobrindo recursos escaláveis usando uma pilha CloudFormation

Ao usar 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.

No AWS Auto Scaling console, você pode selecionar uma pilha existente para examiná-la em busca de recursos que possam ser configurados para escalabilidade automática. AWS Auto Scaling encontra somente os recursos definidos na pilha selecionada. Ele não passa por pilhas aninhadas.

Para que seus serviços do ECS possam ser descobertos em uma CloudFormation pilha, o AWS Auto Scaling console deve saber qual cluster do ECS está executando o serviço. Isso exige que seus serviços do ECS estejam na mesma CloudFormation pilha do 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 CloudFormation, consulte O que é AWS CloudFormation? no Guia do AWS CloudFormation usuário.

Identificar recursos escaláveis usando etiquetas

As tags fornecem metadados que podem ser usados para descobrir recursos escaláveis relacionados no AWS Auto Scaling console, usando filtros de tags.

Use etiquetas para identificar os seguintes recursos:

- clusters de bancos de dados 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.
- Saiba como marcar grupos do Auto Scaling no Guia do usuário do Amazon Auto EC2 Scaling.
- Aprenda a etiquetar recursos do DynamoDB no Guia do desenvolvedor do Amazon DynamoDB.

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, AWS Auto Scaling escolhe a métrica mais comumente usada para determinar quanto do recurso está em uso em um 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.

As seguintes estratégias de dimensionamento estão disponíveis:

- Otimize a disponibilidade —AWS Auto Scaling expande o recurso para dentro e para fora automaticamente para manter a utilização dos recursos em 40 por cento. Essa opção é útil quando o aplicativo tem necessidades de dimensionamento urgentes e, às vezes, imprevisíveis.
- Equilibre disponibilidade e custo —AWS Auto Scaling expande o recurso para dentro e para fora automaticamente para manter a utilização dos recursos em 50%. Essa opção ajuda a manter a alta disponibilidade ao mesmo tempo que reduz os custos.
- Otimize o custo —AWS Auto Scaling expande o recurso para dentro e para fora automaticamente para manter a utilização dos recursos em 70 por cento. 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.

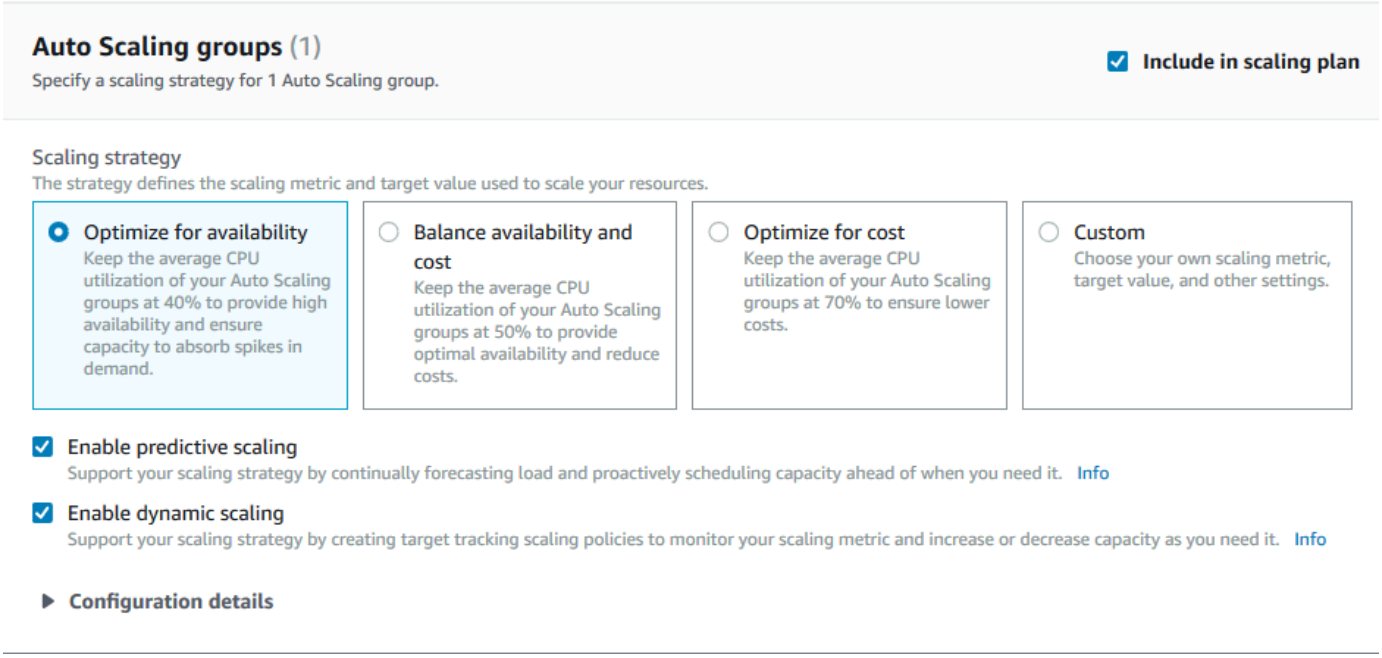
Por exemplo, o plano de escalabilidade configura seu grupo de Auto Scaling para adicionar ou remover instâncias da EC2 Amazon com base na quantidade média da CPU usada para todas as instâncias do grupo. Você escolhe se deseja otimizar a utilização para disponibilidade, custo ou uma combinação de ambos 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 o throughput 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 Substituir políticas de escalabilidade externas para especificar que AWS Auto Scaling podem excluir políticas de escalabilidade criadas anteriormente fora do plano de escalabilidade (como de outros consoles) e substituí-las por novas políticas de escalabilidade de rastreamento de metas 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.
   - Capacidade mínima — especifica a capacidade mínima do recurso. AWS Auto Scaling garante que seu recurso nunca fique abaixo desse tamanho.
   - Capacidade máxima — especifica a capacidade máxima do recurso. AWS Auto Scaling garante que seu recurso nunca ultrapasse esse 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 usadas para escalonamento automático. No entanto, dependendo das suas necessidades, você pode preferir obter dados de métricas diferentes em vez das métricas na console. A Amazon CloudWatch tem muitas métricas diferentes para escolher. CloudWatch também permite que você publique suas próprias métricas.

Você usa o JSON para especificar uma métrica CloudWatch personalizada. Antes de seguir essas instruções, recomendamos que você se familiarize com o Guia do CloudWatch usuário da Amazon.

Para especificar uma métrica personalizada, cie uma carga útil em formato JSON usando um conjunto de parâmetros exigidos de um modelo. Você adiciona os valores para cada parâmetro de 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 esse procedimento para visualizar e personalizar as configurações da política de escalabilidade de rastreamento de destino AWS Auto Scaling criada.

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.)
   - Aquecimento da instância — [Somente grupos do Auto Scaling] Controla o tempo decorrido até que uma instância recém-lançada comece a contribuir com as métricas. CloudWatch Para obter mais informações, consulte Aquecimento de instâncias no Guia do usuário do Amazon Auto EC2 Scaling.

Configurações de dimensionamento preditivo

Se seu recurso for um grupo de Auto Scaling, use esse procedimento para visualizar e personalizar as configurações AWS Auto Scaling usadas para 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 trezentos 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).
     - Imponha a configuração de capacidade máxima —AWS Auto Scaling não é possível escalar a capacidade dos recursos acima da capacidade máxima. A capacidade máxima é imposta como um limite fixo.
     - Defina a capacidade máxima como igual à capacidade prevista —AWS Auto Scaling pode escalar a capacidade dos recursos acima da capacidade máxima para igualar, mas não exceder, a capacidade prevista.
     - Aumentar a capacidade máxima acima da capacidade prevista —AWS Auto Scaling pode escalar a capacidade dos recursos acima da capacidade máxima de acordo com 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 a escala preditiva, AWS Auto Scaling analisa o histórico da métrica de carga especificada nos últimos 14 dias (é necessário um mínimo de 24 horas de dados) para gerar uma previsão para dois dias à frente. 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 mostrando dados recentes do histórico métrico de CloudWatch.
- Gráficos de escala preditiva mostrando previsões de carga e previsões de capacidade com base em dados de. AWS Auto Scaling

- Uma tabela que lista todas as ações de dimensionamento preditivo programadas para o recurso.

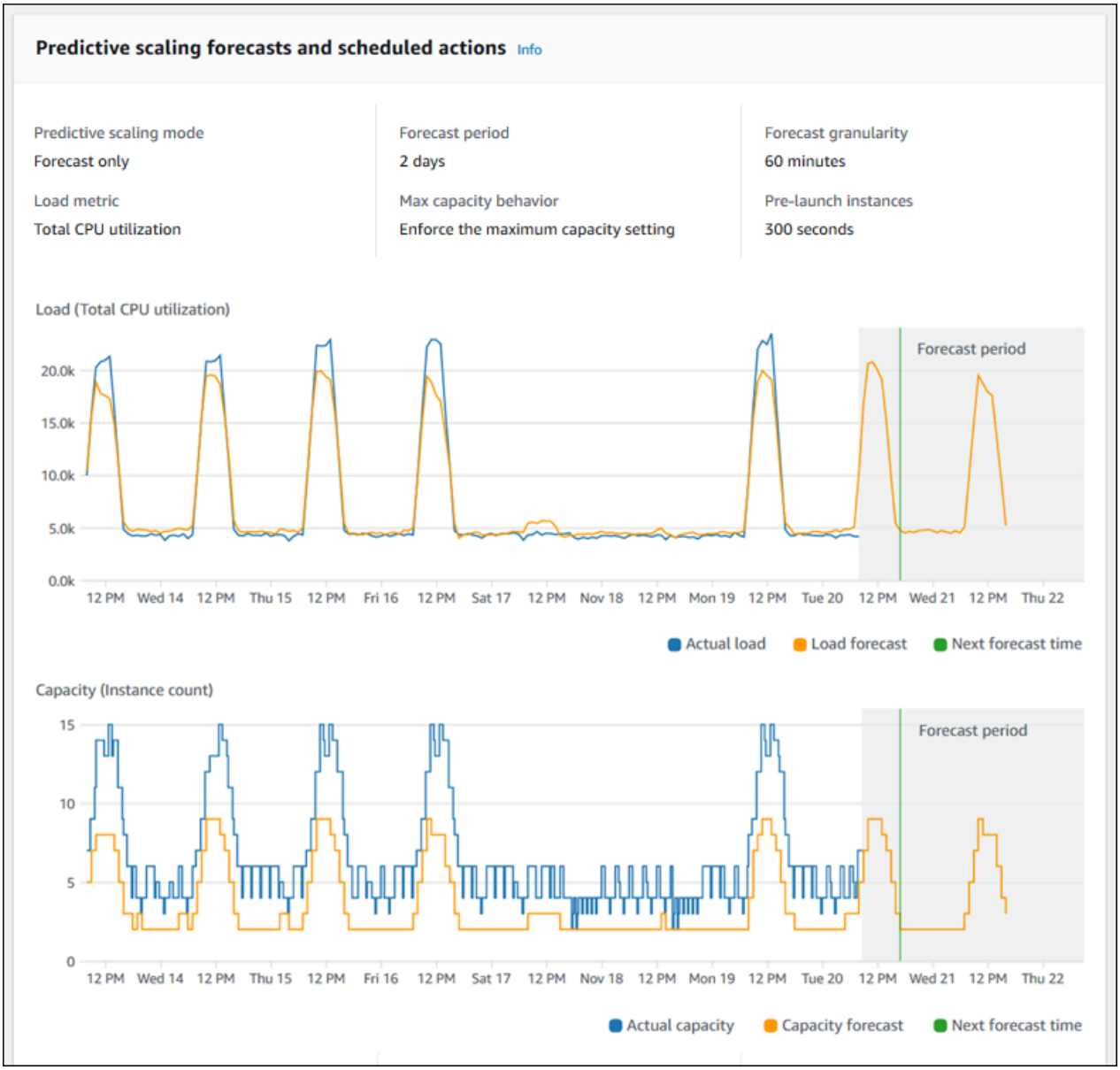
Para visualizar informações de dimensionamento de um recurso

1. Abra o AWS Auto Scaling console 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 estão disponíveis no AWS Auto Scaling console para todos os grupos de 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 Scaling 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, 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 de rastreamento de metas, seus CloudWatch alarmes associados e as ações de escalabilidade preditiva criadas em seu nome. AWS Auto Scaling

A exclusão de um plano de escalabilidade não exclui sua AWS CloudFormation pilha, grupo de Auto Scaling ou outros recursos escaláveis.

Para excluir um plano de escalabilidade

1. Abra o AWS Auto Scaling console 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 EC2 cobranças da Amazon, você também deve excluir o grupo Auto Scaling que você criou para este tutorial.

Para step-by-step obter instruções, consulte Excluir seu grupo de Auto Scaling no Guia do usuário do Amazon Auto EC2 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 AWS CloudFormation modelo é um arquivo de texto em formato JSON ou YAML que descreve a infraestrutura da Amazon Web Services necessária para executar um aplicativo ou serviço junto com quaisquer interconexões entre os componentes da infraestrutura. Com AWS CloudFormation, você implanta e gerencia uma coleção associada de recursos como uma pilha. AWS CloudFormation está disponível sem custo adicional e você paga somente pelos AWS recursos necessários para executar seus aplicativos. Os recursos podem consistir em qualquer AWS recurso definido no modelo. Para obter mais informações, consulte Como AWS CloudFormation funciona no Guia AWS CloudFormation do usuário.

No Manual do usuário do AWS CloudFormation , apresentamos um modelo simples para você começar. O modelo de amostra está disponível como exemplo na

AWS::AutoScalingPlans::ScalingPlanseção da documentação de referência do AWS CloudFormation modelo. 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 .

Migre seu plano de escalabilidade

Você pode migrar de um plano de escalabilidade para as políticas de escalabilidade do Amazon Auto EC2 Scaling e do Application Auto Scaling.

Processo de migração

- Etapa 1: revisar sua configuração existente
- Etapa 2: criar políticas de escalabilidade preditiva
- Etapa 3: Analise as previsões geradas pelas políticas de escalabilidade preditiva
- Etapa 4: Prepare-se para excluir o plano de escalabilidade
- Etapa 5: excluir o plano de escalabilidade
- Etapa 6: reativar o escalonamento dinâmico
- Etapa 7: reativar a escala preditiva
- Referência do Amazon EC2 Auto Scaling para migrar políticas de escalabilidade de rastreamento de metas
- Referência do Application Auto Scaling para migrar políticas de escalabilidade de rastreamento de metas
- Mais informações

> **Important**
>
> Para migrar um plano de escalabilidade, você deve concluir várias etapas na ordem exata. Ao migrar seu plano de escalabilidade, não o atualize, pois isso interrompe a ordem das operações e pode causar comportamentos indesejáveis.

Etapa 1: revisar sua configuração existente

Para determinar quais configurações de escala você deve mover, use o describe-scaling-planscomando.

```
aws autoscaling-plans describe-scaling-plans \
  --scaling-plan-names my-scaling-plan
```

Anote os itens que você deseja preservar do plano de escalabilidade existente, que pode incluir o seguinte:

- **`MinCapacity`**— A capacidade mínima do recurso escalável.
- **`MaxCapacity`**— A capacidade máxima do recurso escalável.
- **`PredefinedLoadMetricType`**— Uma métrica de carga para escalonamento preditivo.
- **`PredefinedScalingMetricType`**— Uma métrica de escalabilidade para rastreamento de metas, escalabilidade (dinâmica) e escala preditiva.
- **`TargetValue`**— O valor alvo para a métrica de escala.

Diferenças entre planos de escalabilidade e políticas de escalabilidade

Há algumas diferenças importantes entre planos de escalabilidade e políticas de escalabilidade:

- Uma política de escalabilidade pode permitir somente um tipo de escalabilidade: escalabilidade de rastreamento de metas ou escalabilidade preditiva. Para usar os dois métodos de escalabilidade, você deve criar políticas separadas.
- Da mesma forma, você deve definir a métrica de escalabilidade para escalabilidade preditiva e a métrica de escala para escalabilidade de rastreamento de metas separadamente dentro de suas respectivas políticas.

Etapa 2: criar políticas de escalabilidade preditiva

Se você não usa escalabilidade preditiva, vá em frente para. Etapa 4: Prepare-se para excluir o plano de escalabilidade

Para fornecer tempo para avaliar a previsão, recomendamos que você crie políticas de escalabilidade preditiva antes de outras políticas de escalabilidade.

Para qualquer grupo de Auto Scaling com uma especificação de métrica de carga existente, faça o seguinte para transformá-la em uma política de escalabilidade preditiva baseada no Amazon EC2 Auto Scaling.

Para criar políticas de escalabilidade preditiva

1. Em um arquivo JSON, defina uma `MetricSpecifications` estrutura conforme mostrado no exemplo a seguir:

```
{
  "MetricSpecifications":[
    {
      ...
    }
  ]
}
```

2. Na `MetricSpecifications` estrutura, para cada métrica de carga em seu plano de escalabilidade, crie uma `PredefinedLoadMetricSpecification` ou `CustomizedLoadMetricSpecification` use as configurações equivalentes do plano de escalabilidade.

   A seguir estão exemplos da estrutura da seção métrica de carga.

   With predefined metrics

   ```
   {
     "MetricSpecifications":[
       {
         "PredefinedLoadMetricSpecification":{
           "PredefinedMetricType":"ASGTotalCPUUtilization"
         },
         ...
       }
     ]
   }
   ```

   Para obter mais informações, consulte PredictiveScalingPredefinedLoadMetrica Amazon EC2 Auto Scaling API Reference.

   With custom metrics

   ```
   {
     "MetricSpecifications":[
       {
         "CustomizedLoadMetricSpecification":{
           "MetricDataQueries":[
             {
               "Id":"load_metric",
               "MetricStat":{
   ```

```
            "Metric":{
              "MetricName":"MyLoadMetric",
              "Namespace":"MyNameSpace",
              "Dimensions":[
                {
                  "Name":"MyOptionalMetricDimensionName",
                  "Value":"MyOptionalMetricDimensionValue"
                }
              ]
            },
            "Stat":"Sum"
          }
        }
      ]
    },
    ...
  }
 ]
}
```

Para obter mais informações, consulte PredictiveScalingCustomizedLoadMetrica Amazon EC2 Auto Scaling API Reference.

3. Adicione a especificação métrica de escala ao `MetricSpecifications` e defina um valor alvo.

   A seguir estão exemplos da estrutura das seções de métrica de escala e valor alvo.

   With predefined metrics

```
{
  "MetricSpecifications":[
    {
      "PredefinedLoadMetricSpecification":{
        "PredefinedMetricType":"ASGTotalCPUUtilization"
      },
      "PredefinedScalingMetricSpecification":{
        "PredefinedMetricType":"ASGCPUUtilization"
      },
      "TargetValue":50
    }
  ],
  ...
```

```
}
```

Para obter mais informações, consulte PredictiveScalingPredefinedScalingMetrica Amazon EC2 Auto Scaling API Reference.

With custom metrics

```
{
  "MetricSpecifications":[
    {
      "CustomizedLoadMetricSpecification":{
        "MetricDataQueries":[
          {
            "Id":"load_metric",
            "MetricStat":{
              "Metric":{
                "MetricName":"MyLoadMetric",
                "Namespace":"MyNameSpace",
                "Dimensions":[
                  {
                    "Name":"MyOptionalMetricDimensionName",
                    "Value":"MyOptionalMetricDimensionValue"
                  }
                ]
              },
              "Stat":"Sum"
            }
          }
        ]
      },
      "CustomizedScalingMetricSpecification":{
        "MetricDataQueries":[
          {
            "Id":"scaling_metric",
            "MetricStat":{
              "Metric":{
                "MetricName":"MyUtilizationMetric",
                "Namespace":"MyNameSpace",
                "Dimensions":[
                  {
                    "Name":"MyOptionalMetricDimensionName",
                    "Value":"MyOptionalMetricDimensionValue"
                  }
```

```
                ]
              },
              "Stat":"Average"
            }
          }
        ]
      },
      "TargetValue":50
    }
  ],
  ...
}
```

Para obter mais informações, consulte PredictiveScalingCustomizedScalingMetrica Amazon EC2 Auto Scaling API Reference.

4. Para fazer apenas uma previsão, adicione a propriedade `Mode` com um valor de`ForecastOnly`. Depois de concluir a migração da escala preditiva e garantir que a previsão seja precisa e confiável, você pode alterar o modo para permitir a escalabilidade. Para obter mais informações, consulte Etapa 7: reativar a escala preditiva.

```
{
  "MetricSpecifications":[
  ...
  ],
  "Mode":"ForecastOnly",
  ...
}
```

Para obter mais informações, consulte PredictiveScalingConfigurationa Amazon EC2 Auto Scaling API Reference.

5. Se a **`ScheduledActionBufferTime`** propriedade estiver presente em seu plano de escalabilidade, copie seu valor para a `SchedulingBufferTime` propriedade em sua política de escalabilidade preditiva.

```
{
  "MetricSpecifications":[
  ...
  ],
  "Mode":"ForecastOnly",
  "SchedulingBufferTime":300,
```

```
  ...
}
```

Para obter mais informações, consulte PredictiveScalingConfigurationa Amazon EC2 Auto Scaling API Reference.

6. Se as **`PredictiveScalingMaxCapacityBuffer`** propriedades **`PredictiveScalingMaxCapacityBehavior`** e estiverem presentes em seu plano de escalabilidade, você poderá configurar as `MaxCapacityBuffer` propriedades `MaxCapacityBreachBehavior` e em sua política de escalabilidade preditiva. Essas propriedades definem o que deve acontecer se a capacidade prevista se aproximar ou exceder a capacidade máxima especificada para o grupo Auto Scaling.

> **Warning**
>
> Se você definir a `MaxCapacityBreachBehavior` propriedade como`IncreaseMaxCapacity`, mais instâncias poderão ser iniciadas do que o pretendido, a menos que você monitore e gerencie o aumento da capacidade máxima. A capacidade máxima aumentada se torna a nova capacidade máxima normal para o grupo Auto Scaling até que você a atualize manualmente. A capacidade máxima não diminui automaticamente de volta ao máximo original.

```
{
  "MetricSpecifications":[
  ...
  ],
  "Mode":"ForecastOnly",
  "SchedulingBufferTime":300,
  "MaxCapacityBreachBehavior": "IncreaseMaxCapacity",
  "MaxCapacityBuffer": 10
}
```

Para obter mais informações, consulte PredictiveScalingConfigurationa Amazon EC2 Auto Scaling API Reference.

7. Salve o arquivo JSON com um nome exclusivo. Anote o nome do arquivo. Você precisará dela na próxima etapa e novamente no final do procedimento de migração ao reativar suas políticas

de escalabilidade preditiva. Para obter mais informações, consulte Etapa 7: reativar a escala preditiva.

8. Depois de salvar o arquivo JSON, execute o put-scaling-policycomando. No exemplo a seguir, substitua cada *user input placeholder* por suas próprias informações.

```
aws autoscaling put-scaling-policy --policy-name my-predictive-scaling-policy \
  --auto-scaling-group-name my-asg --policy-type PredictiveScaling \
  --predictive-scaling-configuration file://my-predictive-scaling-config.json
```

Se bem-sucedido, esse comando gerará o nome do recurso da Amazon (ARN) da política.

```
{
  "PolicyARN": "arn:aws:autoscaling:region:account-id:scalingPolicy:2f4f5048-
d8a8-4d14-b13a-d1905620f345:autoScalingGroupName/my-asg:policyName/my-predictive-
scaling-policy",
  "Alarms": []
}
```

9. Repita essas etapas para cada especificação de métrica de carga que você está migrando para uma política de escalabilidade preditiva baseada no Amazon EC2 Auto Scaling.

Etapa 3: Analise as previsões geradas pelas políticas de escalabilidade preditiva

Se você não usa a escala preditiva, pule o procedimento a seguir.

Uma previsão está disponível logo após você criar uma política de escalabilidade preditiva. Depois que o Amazon EC2 Auto Scaling gerar a previsão, você poderá revisar a previsão da política por meio do console do Amazon Auto EC2 Scaling e ajustá-la conforme necessário.

Para revisar a previsão de uma política de escalabilidade preditiva

1. Abra o EC2 console da Amazon em https://console.aws.amazon.com/ec2/.
2. No painel de navegação, escolha Grupos de Auto Scaling e, em seguida, escolha o nome do seu grupo de Auto Scaling na lista.
3. Na guia Escalabilidade automática, em Políticas de escalabilidade preditiva, escolha sua política.
4. Na seção Monitorar, você pode visualizar as previsões passadas e futuras de carga e de capacidade da política em relação aos valores reais.

Para obter mais informações, consulte Analise os gráficos de monitoramento de escalabilidade preditiva no Guia do usuário do Amazon EC2 Auto Scaling.

5. Repita essas etapas para cada política de escalabilidade preditiva que você criou.

Etapa 4: Prepare-se para excluir o plano de escalabilidade

Para quaisquer recursos com uma configuração de escalabilidade de rastreamento de metas existente, faça o seguinte para coletar todas as informações adicionais necessárias do plano de escalabilidade antes de excluí-lo.

Para descrever as informações da política de escalabilidade do plano de escalabilidade, use o describe-scaling-plan-resourcescomando. No exemplo de comando a seguir, *my-scaling-plan* substitua por suas próprias informações.

```
aws autoscaling-plans describe-scaling-plan-resources \
  --scaling-plan-name my-scaling-plan \
  --scaling-plan-version 1
```

Analise a saída e confirme que você deseja migrar as políticas de escalabilidade descritas. Use essas informações para criar novas políticas de escalabilidade de rastreamento de metas baseadas no Amazon EC2 Auto Scaling e no Application Auto Scaling em. Etapa 6: reativar o escalonamento dinâmico

Etapa 5: excluir o plano de escalabilidade

Antes de criar novas políticas de escalabilidade de rastreamento de metas, você deve excluir o plano de escalabilidade para excluir as políticas de escalabilidade que ele criou.

Para excluir seu plano de escalabilidade, use o delete-scaling-plancomando. No exemplo de comando a seguir, *my-scaling-plan* substitua por suas próprias informações.

```
aws autoscaling-plans delete-scaling-plan \
  --scaling-plan-name my-scaling-plan \
  --scaling-plan-version 1
```

Depois de excluir o plano de escalabilidade, a escalabilidade dinâmica é desativada. Portanto, se houver picos repentinos no tráfego ou na carga de trabalho, a capacidade disponível para

cada recurso escalável não aumentará sozinha. Como precaução, talvez você queira aumentar manualmente a capacidade de seus recursos escaláveis no curto prazo.

Para aumentar a capacidade de um grupo de Auto Scaling

1. Abra o EC2 console da Amazon em https://console.aws.amazon.com/ec2/.
2. No painel de navegação, escolha Grupos de Auto Scaling e, em seguida, escolha o nome do seu grupo de Auto Scaling na lista.
3. Na guia Detalhes, escolha Detalhes do grupo, Editar.
4. Para Capacidade desejada, aumente a capacidade desejada.
5. Ao terminar, escolha Atualizar.

Como adicionar uma réplica do Aurora a um cluster de banco de dados

1. Abra o console do Amazon RDS em https://console.aws.amazon.com/rds/.
2. No painel de navegação, escolha Bancos de dados e selecione seu cluster de banco de dados.
3. Verifique se o cluster e a instância primária estão no estado Disponível.
4. Escolha Ações, Adicionar leitor.
5. Na página Adicionar leitor, especifique as opções para sua nova réplica do Aurora.
6. Escolha Adicionar leitor.

Para aumentar a capacidade de leitura e gravação provisionada de uma tabela do DynamoDB ou índice secundário global

1. Abra o console do DynamoDB em. https://console.aws.amazon.com/dynamodb/
2. No painel de navegação, escolha Tabelas e, em seguida, escolha o nome da tabela na lista.
3. Na guia Configurações adicionais, escolha Capacidade de leitura/gravação, Editar.
4. Na página Editar capacidade de leitura/gravação, em Capacidade de leitura, Unidades de capacidade provisionada, aumente a capacidade de leitura provisionada da tabela.
5. (Opcional) Se você quiser que seus índices secundários globais usem as mesmas configurações de capacidade de leitura da tabela base, marque a caixa de seleção Usar as mesmas configurações de capacidade de leitura para todos os índices secundários globais.
6. Para capacidade de gravação, unidades de capacidade provisionada, aumente a capacidade de gravação provisionada da tabela.

7. (Opcional) Se você quiser que seus índices secundários globais usem as mesmas configurações de capacidade de gravação da tabela base, marque a caixa de seleção Usar as mesmas configurações de capacidade de gravação para todos os índices secundários globais.
8. Se você não marcou as caixas de seleção nas etapas 5 ou 7, role a página para baixo para atualizar a capacidade de leitura e gravação de qualquer índice secundário global.
9. Escolha Salvar alterações para continuar.

Para aumentar a contagem de tarefas em execução para seu serviço Amazon ECS

1. Abra o console na https://console.aws.amazon.com/ecs/v2.
2. No painel de navegação, escolha Clusters e, em seguida, escolha o nome do seu cluster na lista.
3. Na seção Serviços, marque a caixa de seleção ao lado do serviço e escolha Atualizar.
4. Em Tarefas desejadas, insira o número de tarefas que você deseja executar no serviço.
5. Selecione Atualizar.

Para aumentar a capacidade de uma frota spot

1. Abra o EC2 console da Amazon em https://console.aws.amazon.com/ec2/.
2. No painel de navegação, escolha Solicitações spot e selecione sua solicitação de frota spot.
3. Escolha Actions (Ações) e Modify target capacity (Modificar capacidade de destino).
4. Em Modificar capacidade de destino, insira a nova capacidade de destino e a parte da instância sob demanda.
5. Selecione Enviar.

Etapa 6: reativar o escalonamento dinâmico

Reative o escalonamento dinâmico criando políticas de escalabilidade de rastreamento de metas.

Ao criar uma política de escalabilidade de rastreamento de metas para um grupo de Auto Scaling, você a adiciona diretamente ao grupo. Ao criar uma política de escalabilidade de rastreamento de metas para outros recursos escaláveis, primeiro registre o recurso como uma meta escalável e, em seguida, adicione uma política de escalabilidade de rastreamento de metas à meta escalável.

Tópicos

Crie políticas de escalabilidade de rastreamento de metas para grupos de Auto Scaling

Para criar políticas de escalabilidade de rastreamento de metas para grupos de Auto Scaling

1. Em um arquivo JSON, crie um `PredefinedMetricSpecification` ou `CustomizedMetricSpecification` use as configurações equivalentes do plano de escalabilidade.

   Veja a seguir exemplos de uma configuração de rastreamento de alvos. Nesses exemplos, substitua cada um *`user input placeholder`* por suas próprias informações.

   With predefined metrics

   ```
   {
     "TargetValue": 50.0,
     "PredefinedMetricSpecification":
       {
         "PredefinedMetricType": "ASGAverageCPUUtilization"
       }
   }
   ```

   Para obter mais informações, consulte PredefinedMetricSpecificationa Amazon EC2 Auto Scaling API Reference.

   With custom metrics

   ```
   {
     "TargetValue": 100.0,
     "CustomizedMetricSpecification": {
       "MetricName": "MyBacklogPerInstance",
       "Namespace": "MyNamespace",
       "Dimensions": [{
         "Name": "MyOptionalMetricDimensionName",
         "Value": "MyOptionalMetricDimensionValue"
       }],
       "Statistic": "Average",
   ```

```
    "Unit": "None"
  }
}
```

Para obter mais informações, consulte CustomizedMetricSpecificationa Amazon EC2 Auto Scaling API Reference.

2. Para criar sua política de escalabilidade, use o put-scaling-policycomando junto com o arquivo JSON que você criou na etapa anterior. No exemplo a seguir, substitua cada *user input placeholder* por suas próprias informações.

```
aws autoscaling put-scaling-policy --policy-name my-target-tracking-scaling-policy
 \
  --auto-scaling-group-name my-asg --policy-type TargetTrackingScaling \
  --target-tracking-configuration file://config.json
```

3. Repita esse processo para cada política de escalabilidade baseada em plano de escalabilidade que você está migrando para uma política de escalabilidade de rastreamento de metas baseada no Amazon EC2 Auto Scaling.

Crie políticas de escalabilidade de rastreamento de metas para outros recursos escaláveis

Em seguida, crie políticas de escalabilidade de rastreamento de metas para outros recursos escaláveis executando as seguintes tarefas de configuração.

- Registre uma meta escalável para escalonamento automático com o serviço Application Auto Scaling.
- Adicione uma política de escalabilidade de monitoramento do objetivo ao destino escalável.

Para criar políticas de escalabilidade de rastreamento de metas para outros recursos escaláveis

1. Use o register-scalable-targetcomando para registrar o recurso como um alvo escalável e definir os limites de escalabilidade para a política de escalabilidade.

   No exemplo a seguir, substitua cada *user input placeholder* por suas próprias informações. Para as opções de comando, forneça as seguintes informações:

- `--service-namespace`— Um namespace para o serviço de destino (por exemplo,**`ecs`**). Para obter namespaces de serviço, consulte a RegisterScalableTargetreferência.
- `--scalable-dimension`— Uma dimensão escalável associada ao recurso de destino (por exemplo,**`ecs:service:DesiredCount`**). Para obter dimensões escaláveis, consulte a RegisterScalableTargetreferência.
- `--resource-id`— Um ID de recurso para o recurso de destino (por exemplo,**`service/`*`my-cluster`*`/`*`my-service`***). Para obter informações sobre a sintaxe e exemplos de recursos específicos IDs, consulte a RegisterScalableTargetreferência.

```
aws application-autoscaling register-scalable-target --service-namespace namespace
 \
  --scalable-dimension dimension \
  --resource-id identifier \
  --min-capacity 1 --max-capacity 10
```

Se obtiver êxito, esse comando retornará o ARN do destino escalável.

```
{
    "ScalableTargetARN": "arn:aws:application-autoscaling:region:account-
id:scalable-target/1234abcd56ab78cd901ef1234567890ab123"
}
```

2. Em um arquivo JSON, crie um `PredefinedMetricSpecification` ou `CustomizedMetricSpecification` use as configurações equivalentes do plano de escalabilidade.

   Veja a seguir exemplos de uma configuração de rastreamento de alvos.

   With predefined metrics

   ```
   {
     "TargetValue": 70.0,
     "PredefinedMetricSpecification":
       {
         "PredefinedMetricType": "ECSServiceAverageCPUUtilization"
       }
   }
   ```

Para obter mais informações, consulte a Referência PredefinedMetricSpecificationda API Application Auto Scaling.

With custom metrics

```
{
  "TargetValue": 70.0,
  "CustomizedMetricSpecification": {
    "MetricName": "MyUtilizationMetric",
    "Namespace": "MyNamespace",
    "Dimensions": [{
      "Name": "MyOptionalMetricDimensionName",
      "Value": "MyOptionalMetricDimensionValue"
    }],
    "Statistic": "Average",
    "Unit": "Percent"
  }
}
```

Para obter mais informações, consulte a Referência CustomizedMetricSpecificationda API Application Auto Scaling.

3. Para criar sua política de escalabilidade, use o put-scaling-policycomando junto com o arquivo JSON que você criou na etapa anterior.

```
aws application-autoscaling put-scaling-policy --service-namespace namespace \
  --scalable-dimension dimension \
  --resource-id identifier \
  --policy-name my-target-tracking-scaling-policy --policy-
type TargetTrackingScaling \
  --target-tracking-scaling-policy-configuration file://config.json
```

4. Repita esse processo para cada política de escalabilidade baseada em plano de escalabilidade que você está migrando para uma política de escalabilidade de rastreamento de metas baseada em Application Auto Scaling.

Etapa 7: reativar a escala preditiva

Se você não usa a escala preditiva, pule esta etapa.

Reative a escala preditiva trocando a escala preditiva por previsão e escala.

Para fazer essa alteração, atualize os arquivos JSON que você criou Etapa 2: criar políticas de escalabilidade preditiva e altere o valor da `Mode` opção para `ForecastAndScale` como no exemplo a seguir:

```
"Mode":"ForecastAndScale"
```

Em seguida, atualize cada política de escalabilidade preditiva com o put-scaling-policycomando. Neste exemplo, substitua cada um *`user input placeholder`* por suas próprias informações.

```
aws autoscaling put-scaling-policy --policy-name my-predictive-scaling-policy \
  --auto-scaling-group-name my-asg --policy-type PredictiveScaling \
  --predictive-scaling-configuration file://my-predictive-scaling-config.json
```

Como alternativa, você pode fazer essa alteração no console do Amazon EC2 Auto Scaling ativando a escala com base na configuração de previsão. Para obter mais informações, consulte Escalabilidade preditiva para Amazon Auto EC2 Scaling no Guia do usuário do Amazon Auto EC2 Scaling.

Referência do Amazon EC2 Auto Scaling para migrar políticas de escalabilidade de rastreamento de metas

Para fins de referência, a tabela a seguir lista todas as propriedades de configuração de rastreamento de destino no plano de escalabilidade com suas propriedades correspondentes na operação da API Amazon EC2 Auto `PutScalingPolicy` Scaling.

Propriedade de origem do plano de escalabilidade	Propriedade alvo do Amazon EC2 Auto Scaling
`PolicyName`	`PolicyName`
`PolicyType`	`PolicyType`
`TargetTrackingConfiguration.CustomizedScalingMetricSpecification.Dimensions.Name`	`TargetTrackingConfiguration.CustomizedMetricSpecification.Dimensions.Name`

Propriedade de origem do plano de escalabilidade	Propriedade alvo do Amazon EC2 Auto Scaling
TargetTrackingConfiguration.CustomizedScalingMetricSpecification.Dimensions.Value	TargetTrackingConfiguration.CustomizedMetricSpecification.Dimensions.Value
TargetTrackingConfiguration.CustomizedScalingMetricSpecification.MetricName	TargetTrackingConfiguration.CustomizedMetricSpecification.MetricName
TargetTrackingConfiguration.CustomizedScalingMetricSpecification.Namespace	TargetTrackingConfiguration.CustomizedMetricSpecification.Namespace
TargetTrackingConfiguration.CustomizedScalingMetricSpecification.Statistic	TargetTrackingConfiguration.CustomizedMetricSpecification.Statistic
TargetTrackingConfiguration.CustomizedScalingMetricSpecification.Unit	TargetTrackingConfiguration.CustomizedMetricSpecification.Unit
TargetTrackingConfiguration.DisableScaleIn	TargetTrackingConfiguration.DisableScaleIn
TargetTrackingConfiguration.EstimatedInstanceWarmup	TargetTrackingConfiguration.EstimatedInstanceWarmup [1]
TargetTrackingConfiguration.PredefinedScalingMetricSpecification.PredefinedScalingMetricType	TargetTrackingConfiguration.PredefinedMetricSpecification.PredefinedMetricType
TargetTrackingConfiguration.PredefinedScalingMetricSpecification.ResourceLabel	TargetTrackingConfiguration.PredefinedMetricSpecification.ResourceLabel

Propriedade de origem do plano de escalabilidade	Propriedade alvo do Amazon EC2 Auto Scaling
`TargetTrackingConfiguration.ScaleInCooldown`	`Not available`
`TargetTrackingConfiguration.ScaleOutCooldown`	`Not available`
`TargetTrackingConfiguration.TargetValue`	`TargetTrackingConfiguration.TargetValue`

[1] O aquecimento de instâncias é um recurso para grupos de Auto Scaling que ajuda a garantir que as instâncias recém-lançadas estejam prontas para receber tráfego antes de contribuir com seus dados de uso para a métrica de escalabilidade. Enquanto as instâncias ainda estão se aquecendo, o Amazon EC2 Auto Scaling retarda o processo de adição ou remoção de instâncias do grupo. Em vez de especificar um tempo de aquecimento para uma política de escalabilidade, recomendamos que você use a configuração padrão de aquecimento da instância do seu grupo de Auto Scaling para garantir que todas as execuções de instâncias usem o mesmo tempo de aquecimento da instância. Para obter mais informações, consulte Definir o aquecimento padrão da instância para um grupo de Auto Scaling no Guia do usuário do Amazon Auto EC2 Scaling.

Referência do Application Auto Scaling para migrar políticas de escalabilidade de rastreamento de metas

Para fins de referência, a tabela a seguir lista todas as propriedades de configuração de rastreamento de destino no plano de escalabilidade com suas propriedades correspondentes na operação da API Application Auto `PutScalingPolicy` Scaling.

Propriedade de origem do plano de escalabilidade	Propriedade de destino do Application Auto Scaling
`PolicyName`	`PolicyName`
`PolicyType`	`PolicyType`

Propriedade de origem do plano de escalabilidade	Propriedade de destino do Application Auto Scaling
`TargetTrackingConfiguration.CustomizedScalingMetricSpecification.Dimensions.Name`	`TargetTrackingScalingPolicyConfiguration.CustomizedMetricSpecification.Dimensions.Name`
`TargetTrackingConfiguration.CustomizedScalingMetricSpecification.Dimensions.Value`	`TargetTrackingScalingPolicyConfiguration.CustomizedMetricSpecification.Dimensions.Value`
`TargetTrackingConfiguration.CustomizedScalingMetricSpecification.MetricName`	`TargetTrackingScalingPolicyConfiguration.CustomizedMetricSpecification.MetricName`
`TargetTrackingConfiguration.CustomizedScalingMetricSpecification.Namespace`	`TargetTrackingScalingPolicyConfiguration.CustomizedMetricSpecification.Namespace`
`TargetTrackingConfiguration.CustomizedScalingMetricSpecification.Statistic`	`TargetTrackingScalingPolicyConfiguration.CustomizedMetricSpecification.Statistic`
`TargetTrackingConfiguration.CustomizedScalingMetricSpecification.Unit`	`TargetTrackingScalingPolicyConfiguration.CustomizedMetricSpecification.Unit`
`TargetTrackingConfiguration.DisableScaleIn`	`TargetTrackingScalingPolicyConfiguration.DisableScaleIn`
`TargetTrackingConfiguration.EstimatedInstanceWarmup`	`Not available`

Propriedade de origem do plano de escalabilidade	Propriedade de destino do Application Auto Scaling
`TargetTrackingConfiguration.PredefinedScalingMetricSpecification.PredefinedScalingMetricType`	`TargetTrackingScalingPolicyConfiguration.PredefinedMetricSpecification.PredefinedMetricType`
`TargetTrackingConfiguration.PredefinedScalingMetricSpecification.ResourceLabel`	`TargetTrackingScalingPolicyConfiguration.PredefinedMetricSpecification.ResourceLabel`
`TargetTrackingConfiguration.ScaleInCooldown` [1]	`TargetTrackingScalingPolicyConfiguration.ScaleInCooldown`
`TargetTrackingConfiguration.ScaleOutCooldown` [1]	`TargetTrackingScalingPolicyConfiguration.ScaleOutCooldown`
`TargetTrackingConfiguration.TargetValue`	`TargetTrackingScalingPolicyConfiguration.TargetValue`

[1] O Application Auto Scaling usa períodos de espera para diminuir a escalabilidade quando seu recurso escalável está se expandindo (aumentando a capacidade) e aumentando a escala (reduzindo a capacidade). Para obter mais informações, consulte Definir períodos de espera no Guia do usuário do Application Auto Scaling.

Mais informações

Para saber como criar novas políticas de escalabilidade preditiva a partir do console, consulte o tópico a seguir:

- Amazon EC2 Auto Scaling — Crie uma política de escalabilidade preditiva no Guia do usuário do Amazon Auto Scaling. EC2

Para saber como criar novas políticas de escalabilidade de rastreamento de metas usando o console, consulte os tópicos a seguir:

- Amazon Aurora — Usando o Amazon Aurora Auto Scaling com réplicas do Aurora no Guia do usuário do Amazon RDS.
- DynamoDB — Uso do com o AWS Management Console DynamoDB auto scaling no Amazon DynamoDB Developer Guide.
- Amazon EC2 Auto Scaling — Crie uma política de escalabilidade de rastreamento de metas no Guia do usuário do Amazon EC2 Auto Scaling.
- Amazon ECS — Atualização de um serviço usando o console no Amazon Elastic Container Service Developer Guide.
- Spot Fleet — Dimensione o Spot Fleet usando uma política de rastreamento de alvos no Guia EC2 do Usuário da Amazon.

Segurança do plano de escalabilidade

A segurança na nuvem AWS é a maior prioridade. Como AWS cliente, você se beneficia de uma arquitetura de data center e rede criada para atender aos requisitos das organizações mais sensíveis à segurança.

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

- Segurança da nuvem — AWS é responsável por proteger a infraestrutura que executa AWS os serviços na AWS nuvem. AWS também fornece serviços que você pode usar com segurança. Auditores terceirizados testam e verificam regularmente a eficácia de nossa segurança como parte dos AWS programas de de . Para saber mais sobre os programas de conformidade que se aplicam a AWS Auto Scaling, consulte AWS serviços no escopo por programa de conformidade AWS .
- Segurança na nuvem — Sua responsabilidade é determinada pelo AWS serviço que você usa. Você também é responsável por outros fatores, incluindo a confidencialidade de seus dados, os requisitos da sua empresa e as leis e normas aplicáveis.

Esta documentação ajuda 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

Acesse planos de escalabilidade usando endpoints VPC de interface

Você pode usar AWS PrivateLink para criar uma conexão privada entre sua VPC e. AWS Auto Scaling Você pode acessar AWS Auto Scaling como se estivesse em sua VPC, sem o uso de um

gateway de internet, dispositivo NAT, conexão VPN ou conexão. AWS Direct Connect As instâncias na sua VPC não precisam de endereços IP públicos para acessar o AWS Auto Scaling.

Estabeleça essa conectividade privada criando um endpoint de interface, habilitado pelo AWS PrivateLink. Criaremos um endpoint de interface de rede em cada sub-rede que você habilitar para o endpoint de interface. Estas são interfaces de rede gerenciadas pelo solicitante que servem como ponto de entrada para o tráfego destinado ao AWS Auto Scaling.

Para obter mais informações, consulte Acesso Serviços da AWS por meio AWS PrivateLink do AWS PrivateLink Guia.

Tópicos

- Criar um endpoint de interface da VPC para planos de escalabilidade
- Criar uma política de endpoint da VPC para planos de escalabilidade
- Migração de endpoints

Criar um endpoint de interface da VPC para planos de escalabilidade

Crie um endpoint para planos de AWS Auto Scaling escalabilidade usando o seguinte nome de serviço:

```
com.amazonaws.region.autoscaling-plans
```

Para obter mais informações, consulte Acessar um AWS serviço usando uma interface VPC endpoint no Guia.AWS PrivateLink

Você não precisa alterar nenhuma outra configuração. AWS Auto Scaling A API chama outros Serviços da AWS usando endpoints de serviço ou endpoints VPC de interface privada, os que estiverem em uso.

Criar uma política de endpoint da VPC para planos de escalabilidade

Você pode anexar uma política ao seu VPC endpoint para controlar o acesso à API. 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 VPC endpoint policies (Políticas de endpoint da VPC) no AWS PrivateLink Guide (Guia do).

Migração de endpoints

Em 22 de novembro de 2019, apresentamos `autoscaling-plans.region.amazonaws.com` o novo nome de host e endpoint DNS padrão para chamadas para a API. AWS Auto Scaling O novo endpoint é compatível com a versão mais recente do AWS CLI e. SDKs Se você ainda não tiver feito isso, instale a versão mais recente AWS CLI e SDKs use o novo endpoint. Para atualizar o AWS CLI, consulte Instalando ou atualizando o AWS CLI no Guia AWS Command Line Interface do Usuário. Para obter informações sobre o AWS SDKs, consulte Tools for Amazon Web Services.

> **Important**
>
> Para compatibilidade com versões anteriores, o `autoscaling.region.amazonaws.com` endpoint existente continuará sendo suportado para chamadas para a AWS Auto Scaling API. Para configurar o *`autoscaling.region.amazonaws.com`* endpoint como um endpoint VPC de interface privada, consulte Amazon Auto EC2 Scaling e faça a interface de VPC endpoints no Guia do usuário do Amazon Auto Scaling. EC2

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

Para a versão atual do AWS Auto Scaling, suas chamadas para a AWS Auto Scaling API vão automaticamente para o `autoscaling-plans.region.amazonaws.com` endpoint 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 o APIs, consulte a documentação do SDK de interesse para saber como direcionar as solicitações para um endpoint específico. Para obter mais informações, consulte Ferramentas para a Amazon Web Services.

Proteção de dados para planos de escalabilidade

O modelo de responsabilidade AWS compartilhada modelo se aplica à proteção de dados em AWS Auto Scaling. Conforme descrito neste modelo, AWS é responsável por proteger a infraestrutura global que executa todos os Nuvem AWS. Você é responsável por manter o controle sobre o conteúdo hospedado nessa infraestrutura. Você também é responsável pelas tarefas de configuração e gerenciamento de segurança dos Serviços da AWS que usa. Para obter mais informações sobre a privacidade de dados, consulte as Data Privacy FAQ. Para obter mais informações sobre a proteção de dados na Europa, consulte a postagem do blog AWS Shared Responsibility Model and RGPD no Blog de segurança da AWS .

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

- Use uma autenticação multifator (MFA) com cada conta.
- Use SSL/TLS para se comunicar com AWS os recursos. Exigimos TLS 1.2 e recomendamos TLS 1.3.

- Configure a API e o registro de atividades do usuário com AWS CloudTrail. Para obter informações sobre o uso de CloudTrail trilhas para capturar AWS atividades, consulte Como trabalhar com CloudTrail trilhas no Guia AWS CloudTrail do usuário.
- Use soluções de AWS criptografia, juntamente com todos os controles de segurança padrão Serviços da AWS.
- Use serviços gerenciados de segurança avançada, como o Amazon Macie, que ajuda a localizar e proteger dados sigilosos armazenados no Amazon S3.
- Se você precisar de módulos criptográficos validados pelo FIPS 140-3 ao acessar AWS por meio de uma interface de linha de comando ou de uma API, use um endpoint FIPS. Para obter mais informações sobre os endpoints FIPS disponíveis, consulte Federal Information Processing Standard (FIPS) 140-3.

É altamente recomendável que nunca sejam colocadas informações confidenciais ou sigilosas, como endereços de e-mail de clientes, em tags ou campos de formato livre, como um campo Nome. Isso inclui quando você trabalha com AWS Auto Scaling ou Serviços da AWS usa o console, a API ou AWS SDKs. AWS CLI Quaisquer dados inseridos em tags ou em campos de texto de formato livre usados para nomes podem ser usados para logs de faturamento ou de diagnóstico. Se você fornecer um URL para um servidor externo, é fortemente recomendável que não sejam incluídas informações de credenciais no URL para validar a solicitação nesse servidor.

Gerenciamento de identidade e acesso para planos de escalabilidade

AWS Identity and Access Management (IAM) é uma ferramenta AWS service (Serviço da AWS) que ajuda o administrador a controlar com segurança o acesso aos AWS recursos. Os administradores do IAM controlam quem pode ser autenticado (conectado) e autorizado (tem permissões) a usar AWS Auto Scaling os recursos. O IAM é um AWS service (Serviço da AWS) que você pode usar sem custo adicional.

Para concluir a documentação do IAM, consulte o Guia do usuário do IAM.

Controle de acesso

É possível ter credenciais válidas para autenticar suas solicitações. No entanto, a menos que 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 AWS Auto Scaling escalabilidade, você deve entender quais recursos do IAM estão disponíveis para uso com os 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 Guia do usuário do IAM.

Ações

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

O elemento `Action` de uma política JSON descreve as ações que podem ser usadas para permitir ou negar acesso em uma política. As ações de política geralmente têm o mesmo nome da operação

de AWS API associada. Existem algumas exceções, como ações somente de permissão, que não têm uma operação de API correspondente. Algumas operações também exigem várias ações em uma política. Essas ações adicionais são chamadas de ações dependentes.

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

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 de recursos da Amazon (ARNs) para você usar 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 `Condition`) permite que você especifique condições nas quais uma instrução estiver 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 as chaves de condição AWS globais, consulte as chaves de contexto de condição AWS global no Guia 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.

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 oferecem suporte a listas de controle de acesso (ACLs).

Autorização baseada em tags do

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 para esse AWS service (Serviço da AWS).

Perfis do IAM

Um perfil 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. Obtenha credenciais de segurança temporárias chamando operações de API AWS STS tais como AssumeRole ou GetFederationToken.

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

Funções vinculadas a serviço para planos de escalabilidade

AWS Auto Scaling usa funções vinculadas a serviços para obter as permissões necessárias para chamar outros AWS serviços 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.

AWS Auto Scaling usa alguns tipos de funções vinculadas a serviços para chamar outras pessoas Serviços da AWS em seu nome quando você trabalha com um plano de escalabilidade:

- Função vinculada ao serviço de escalabilidade preditiva — permite AWS Auto Scaling acessar dados métricos históricos de. 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 obter mais informações, consulte Função vinculada ao serviço de escalabilidade preditiva.
- Função vinculada ao serviço Amazon EC2 Auto Scaling — Permite AWS Auto Scaling acessar e gerenciar políticas de escalabilidade de rastreamento de metas para grupos de Auto Scaling. Para obter mais informações, consulte Funções vinculadas a serviços para o Amazon Auto EC2 Scaling no Guia do usuário do Amazon Auto EC2 Scaling.
- Função vinculada ao serviço do Application Auto Scaling — Permite acessar e gerenciar políticas de escalabilidade de rastreamento de metas AWS Auto Scaling para outros recursos escaláveis. Há uma função vinculada ao serviço para cada serviço. Para ter 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 Perfis.
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.

Perfis de serviço

AWS Auto Scaling não tem funções de serviço para planos de escalabilidade.

Função vinculada ao serviço de escalabilidade preditiva

AWS Auto Scaling usa funções vinculadas ao serviço para obter as permissões necessárias para ligar para outras pessoas AWS em seu nome quando você trabalha com um plano de escalabilidade. Para obter mais informações, consulte Funções vinculadas a serviço para planos de escalabilidade.

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

AWS Auto Scaling usa a função vinculada ao serviço nomeada AWSServiceRoleForAutoScalingPlans_EC2AutoScaling para chamar outros AWS serviços em seu nome quando você ativa a escalabilidade preditiva.

AWSServiceRoleForAutoScalingPlans_EC2AutoScalingconfia no `autoscaling-plans.amazonaws.com` serviço para assumir a função.

Esse perfil vinculado ao serviço usa a política gerenciada AWSAutoScalingPlansEC2AutoScalingPolicy. Para visualizar as permissões para esta política, consulte AWSAutoScalingPlansEC2AutoScalingPolicy na Referência de políticas gerenciadas pela AWS .

Criar a função vinculada ao serviço (automática)

Você não precisa criar manualmente a AutoScaling função AWSServiceRoleForAutoScalingPlans_EC2. AWS cria essa função para você quando você cria um plano de escalabilidade em sua conta e ativa a escalabilidade preditiva.

AWS Para 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 nível vinculado 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 ter mais informações, consulte Criar um perfil vinculado ao serviço no Guia do usuário do IAM.

Para criar uma função vinculada a serviço (AWS CLI)

Use o create-service-linked-rolecomando a seguir para criar a função vinculada ao serviço.

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

Editar a função vinculada ao serviço

Você pode editar a descrição de AWSServiceRoleForAutoScalingPlans_EC2 AutoScaling usando o IAM. Para obter mais informações, consulte Editar uma descrição de perfil vinculado ao serviço no Guia do usuário do IAM.

Excluir a função vinculada ao serviço

Se você não precisar mais usar planos de escalabilidade, recomendamos que você exclua AWSServiceRoleForAutoScalingPlans_EC2 AutoScaling.

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 AWSServiceRoleForAutoScalingPlans_EC2AutoScaling vinculada ao serviço, AWS Auto Scaling criará a função novamente se você criar um plano de escalabilidade com a escalabilidade preditiva ativada.

Regiões do compatíveis

AWS Auto Scaling suporta o uso de funções vinculadas a serviços em todos os Regiões da AWS planos de escalabilidade 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 e atribuir políticas do IAM que concedam a uma identidade do IAM (como um usuário ou perfil) permissão para trabalhar com planos de escalabilidade.

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.

Tópicos

Práticas recomendadas de política

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

- Comece com as políticas AWS gerenciadas e passe para as permissões de privilégios mínimos — Para começar a conceder permissões aos seus usuários e cargas de trabalho, use as políticas AWS gerenciadas que concedem permissões para muitos casos de uso comuns. Eles estão disponíveis no seu Conta da AWS. Recomendamos que você reduza ainda mais as permissões definindo políticas gerenciadas pelo AWS cliente que sejam específicas para seus casos de uso. Para obter mais informações, consulte Políticas gerenciadas pela AWS ou Políticas gerenciadas pela AWS para funções de trabalho no Guia do usuário do IAM.
- Aplique permissões de privilégio mínimo: ao definir permissões com as políticas do IAM, conceda apenas as permissões necessárias para executar uma tarefa. Você faz isso definindo as ações que podem ser executadas em recursos específicos sob condições específicas, também

conhecidas como permissões de privilégio mínimo. Para obter mais informações sobre como usar o IAM para aplicar permissões, consulte Políticas e permissões no IAM no Guia do usuário do IAM.

- Use condições nas políticas do IAM para restringir ainda mais o acesso: você pode adicionar uma condição às políticas para limitar o acesso a ações e recursos. Por exemplo, você pode escrever uma condição de política para especificar que todas as solicitações devem ser enviadas usando SSL. Você também pode usar condições para conceder acesso às ações de serviço se elas forem usadas por meio de uma ação específica AWS service (Serviço da AWS), como AWS CloudFormation. Para obter mais informações, consulte Elementos da política JSON do IAM: condição no Guia do usuário do IAM.
- Use o IAM Access Analyzer para validar suas políticas do IAM a fim de garantir permissões seguras e funcionais: o IAM Access Analyzer valida as políticas novas e existentes para que elas sigam a linguagem de política do IAM (JSON) e as práticas recomendadas do IAM. O IAM Access Analyzer oferece mais de cem verificações de política e recomendações práticas para ajudar a criar políticas seguras e funcionais. Para obter mais informações, consulte Validação de políticas do IAM Access Analyzer no Guia do Usuário do IAM.
- Exigir autenticação multifator (MFA) — Se você tiver um cenário que exija usuários do IAM ou um usuário root, ative Conta da AWS a MFA para obter segurança adicional. Para exigir MFA quando as operações de API forem chamadas, adicione condições de MFA às suas políticas. Para obter mais informações, consulte Configuração de acesso à API protegido por MFA no Guia do Usuário do IAM.

Para obter mais informações sobre as práticas recomendadas do IAM, consulte Práticas recomendadas de segurança no IAM no Guia do usuário do IAM.

Permitir que os usuários criem planos de escalabilidade

Veja a seguir um exemplo de política baseada em permissões que concede aos usuários permissão para criar planos de escalabilidade.

JSON

```
{
    "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, os usuários finais precisam ter permissões adicionais que permitam trabalhar com recursos específicos na conta deles. Essas permissões estão listadas em Permissões adicionais necessárias.

Cada usuário do console também precisa de permissões que permitam descobrir os recursos escaláveis em sua conta e visualizar gráficos de dados CloudWatch métricos do AWS Auto Scaling console. O conjunto adicional de permissões necessárias para trabalhar com o AWS Auto Scaling console 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 baseada em permissões que concede aos usuários permissão para habilitar escalabilidade preditiva. Essas permissões estendem os recursos dos planos de escalabilidade configurados para escalar grupos do Auto Scaling.

JSON

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

```
        {
            "Effect": "Allow",
            "Action": [
              "cloudwatch:GetMetricData",
              "autoscaling:DescribeAutoScalingGroups",
              "autoscaling:DescribeScheduledActions",
              "autoscaling:BatchPutScheduledUpdateGroupAction",
              "autoscaling:BatchDeleteScheduledAction"
            ],
            "Resource": "*"
        }
    ]
}
```

Permissões adicionais necessárias

Para configurar corretamente planos de escalabilidade, os usuários finais devem receber as permissões para cada serviço de destino para o qual configurarão a escalabilidade. Para conceder as permissões mínimas necessárias para trabalhar com os serviços de destino, leia as informações nesta seção e especifique as ações relevantes no elemento `Action` de uma declaração de política do IAM.

Grupos do Auto Scaling

Para adicionar grupos de Auto Scaling a um plano de escalabilidade, os usuários devem ter as seguintes permissões do Amazon Auto EC2 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 Spot Fleets a um plano de escalabilidade, os usuários devem ter as seguintes permissões da Amazon EC2 e da 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

AWS Auto Scaling exige permissões para criar uma função vinculada ao serviço na primeira vez que qualquer usuário em seu Conta da AWS criar um plano de escalabilidade com a escalabilidade preditiva ativada. Se a função vinculada ao serviço ainda não existir, AWS Auto Scaling crie-a em sua conta. A função vinculada ao serviço concede permissões para AWS Auto Scaling que ela possa chamar 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 baseada em permissões que concede aos usuários permissão para criar planos um perfil vinculado a um serviço.

JSON

```
{
    "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 obter mais informações, consulte Função vinculada ao serviço de escalabilidade preditiva.

Validação de conformidade para planos de escalabilidade

Para saber se um AWS service (Serviço da AWS) está dentro do escopo de programas de conformidade específicos, consulte Serviços da AWS Escopo por Programa de Conformidade Serviços da AWS e escolha o programa de conformidade em que você está interessado. Para obter informações gerais, consulte Programas de AWS conformidade Programas AWS de .

Você pode baixar relatórios de auditoria de terceiros usando AWS Artifact. Para obter mais informações, consulte Baixar relatórios em AWS Artifact .

Sua responsabilidade de conformidade ao usar Serviços da AWS é determinada pela confidencialidade de seus dados, pelos objetivos de conformidade de sua empresa e pelas leis e regulamentações aplicáveis. AWS fornece os seguintes recursos para ajudar na conformidade:

- Governança e conformidade de segurança: esses guias de implementação de solução abordam considerações sobre a arquitetura e fornecem etapas para implantar recursos de segurança e conformidade.

- Referência de serviços qualificados para HIPAA: lista os serviços qualificados para HIPAA. Nem todos Serviços da AWS são elegíveis para a HIPAA.
- AWS Recursos de https://aws.amazon.com/compliance/resources/ de conformidade — Essa coleção de pastas de trabalho e guias pode ser aplicada ao seu setor e local.
- AWS Guias de conformidade do cliente — Entenda o modelo de responsabilidade compartilhada sob a ótica da conformidade. Os guias resumem as melhores práticas de proteção Serviços da AWS e mapeiam as diretrizes para controles de segurança em várias estruturas (incluindo o Instituto Nacional de Padrões e Tecnologia (NIST), o Conselho de Padrões de Segurança do Setor de Cartões de Pagamento (PCI) e a Organização Internacional de Padronização (ISO)).
- Avaliação de recursos com regras no Guia do AWS Config desenvolvedor — O AWS Config serviço avalia o quão bem suas configurações de recursos estão em conformidade com as práticas internas, as diretrizes e os regulamentos do setor.
- AWS Security Hub— Isso AWS service (Serviço da AWS) fornece uma visão abrangente do seu estado de segurança interno AWS. O Security Hub usa controles de segurança para avaliar os recursos da AWS e verificar a conformidade com os padrões e as práticas recomendadas do setor de segurança. Para obter uma lista dos serviços e controles aceitos, consulte a Referência de controles do Security Hub.
- Amazon GuardDuty — Isso AWS service (Serviço da AWS) detecta possíveis ameaças às suas cargas de trabalho Contas da AWS, contêineres e dados monitorando seu ambiente em busca de atividades suspeitas e maliciosas. GuardDuty pode ajudá-lo a atender a vários requisitos de conformidade, como o PCI DSS, atendendo aos requisitos de detecção de intrusões exigidos por determinadas estruturas de conformidade.
- AWS Audit Manager— Isso AWS service (Serviço da AWS) ajuda você a auditar continuamente seu AWS uso para simplificar a forma como você gerencia o risco e a conformidade com as regulamentações e os padrões do setor.

Segurança de infraestrutura para planos de escalabilidade

Como serviço gerenciado, AWS Auto Scaling é protegido pela segurança de rede AWS global. Para obter informações sobre serviços AWS de segurança e como AWS proteger a infraestrutura, consulte AWS Cloud Security. Para projetar seu AWS ambiente usando as melhores práticas de segurança de infraestrutura, consulte Proteção de infraestrutura no Security Pillar AWS Well-Architected Framework.

Você usa chamadas de API AWS publicadas para acessar AWS Auto Scaling pela rede. Os clientes devem oferecer compatibilidade com:

- Transport Layer Security (TLS). Exigimos TLS 1.2 e recomendamos TLS 1.3.
- Conjuntos de criptografia com perfect forward secrecy (PFS) como DHE (Ephemeral Diffie-Hellman) ou ECDHE (Ephemeral Elliptic Curve Diffie-Hellman). A maioria dos sistemas modernos, como Java 7 e versões posteriores, comporta esses modos.

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

Você Conta da AWS tem as cotas padrão (anteriormente chamadas de limites) relacionadas aos planos de escalabilidade. A menos que especificado de outra forma, cada cota é específica da região . É possível solicitar aumentos para algumas cotas e outras cotas não podem ser aumentadas.

Para visualizar as cotas do Application Auto Scaling, abra o console do Service Quotas. No painel de navegação, escolha Serviços da AWSe selecione Planos de AWS Auto Scaling.

Para solicitar o aumento da quota, consulte Solicitar um aumento de quota no Guia do usuário do Service Quotas.

Você Conta da AWS tem as seguintes cotas relacionadas aos planos de escalabilidade.

Name	Padrão	Ajustável
Recursos escaláveis por tipo de recurso	Amazon DynamoDB: 3.000 | Grupos do Amazon Auto EC2 Scaling: 200 | Todos os outros tipos de recursos: 500	Sim
Planos de escalabilidade	100	Sim
Instruções de escalabilidade por plano de escalabilidade	500	Não
Instrução de configuração de rastreamento de destino por escalabilidade	10	Não

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, AWS Auto Scaling pode ser expandido 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 outras ofertas da Amazon Web Services, consulte Endpoints e cotas de serviços no Referência geral da Amazon Web Services.

Histórico do documento dos planos de escalabilidade

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

Alteração	Descrição	Data
Novo conteúdo para AWS Auto Scaling migrar de opções alternativas	Agora você pode migrar da escalabilidade preditiva do Amazon EC2 Auto Scaling AWS Auto Scaling para o Amazon Auto Scaling, que oferece mais funcionalidades. Para obter mais informações, consulte Migrar seu plano de escalabilidade.	5 de abril de 2024
Novo conteúdo de segurança	Lançamos um capítulo de segurança atualizado. Como parte dessa atualização, substituímos “Autenticação e controle de acesso” pelo gerenciamento de identidade e acesso do AWS Auto Scaling.	12 de março de 2020
Suporte para endpoints da Amazon VPC	Agora você pode estabelec er uma conexão privada entre sua VPC e. AWS Auto Scaling Para ver as considerações e instruções de migração, consulte Planos de escalabil idade e endpoints da VPC de interface.	22 de novembro de 2019
Support para aumentar a capacidade máxima acima da capacidade prevista	Adiciona suporte ao console para permitir que o plano de dimensionamento aumente	9 de março de 2019

	a capacidade máxima acima da capacidade da previsão por um valor de buffer especificado. Para obter mais informações, consulte Configurações de escala preditiva.	
Escalabilidade preditiva e melhorias	Agora você pode usar a escalabilidade preditiva para escalar proativamente seus grupos do Amazon Auto Scaling. EC2 Esta versão também inclui suporte para a substituição de políticas de escalabilidade criadas fora do plano de escalabilidade (como a partir de outros consoles) e controla se você ativa o recurso de escalabilidade dinâmico do plano.	20 de novembro de 2018
Suporte para a configuração de recursos personalizados	Suporte adicionado para a personalização de várias configurações para cada recurso individual ou vários recursos ao mesmo tempo.	9 de outubro de 2018
Tags como origem do aplicativo	Esta versão adiciona suporte para especificar um conjunto de tags como origem do aplicativo.	23 de abril de 2018
Novo serviço	Lançamento inicial do AWS Auto Scaling.	16 de janeiro de 2018

As traduções são geradas por tradução automática. Em caso de conflito entre o conteúdo da tradução e da versão original em inglês, a versão em inglês prevalecerá.