## **Aula 12 – Sobrecarga do Auto-Scaling**

Este módulo aborda os seguintes tópicos:

* **Objetivo 4 -** Aplicando Auto-Scaling em ambiente com muitos acessos;

## **Referência Bibliográfica**

Auto Scaling - <https://docs.aws.amazon.com/pt_br/autoscaling/ec2/userguide/what-is-amazon-ec2-auto-scaling.html>

Multi-AZ - <https://aws.amazon.com/pt/rds/features/multi-az/>

## **Objetivo 4 - Aplicando Auto-Scaling e ambiente com muitos acessos**

Este laboratório orienta você a usar os serviços Elastic Load Balancing (ELB) e Auto Scaling para balancear cargas e dimensionar automaticamente a infraestrutura.

O **Auto Scaling** ajuda a manter a disponibilidade da aplicação e permite aumentar ou reduzir a capacidade do Amazon EC2 de forma automática, de acordo com condições definidas por você. Você pode usar o Auto Scaling para ajudar a garantir que está executando o número desejado de instâncias do Amazon EC2. O Auto Scaling também pode aumentar automaticamente o número de instâncias do Amazon EC2 durante picos de demanda para manter a performance, e diminuir a capacidade durante períodos ociosos para reduzir os custos. O Auto Scaling é ideal para aplicações com padrões de demanda estáveis ou que passam por variações de utilização horárias, diárias ou semanais.

**Objetivos**

Depois de concluir este laboratório, você será capaz de:

* Criar uma imagem de máquina da Amazon (AMI) por meio de uma instância em execução (Servidor Web APP Auto);
* Criar uma configuração de execução e um grupo de Auto Scaling;
* Ajustar a escala automaticamente de novas instâncias em uma sub-rede privada;
* Criar alarmes do Amazon CloudWatch e monitorar a performance da sua infraestrutura.

**Duração**

Este laboratório tem duração aproximada de **10 minutos**.

**Cenário**

Você começa com a seguinte infraestrutura:

Uma imagem contendo Tabela

Descrição gerada automaticamenteÍcone

Descrição gerada automaticamente com confiança média

O estado final da infraestrutura é:

Uma imagem contendo Calendário

Descrição gerada automaticamenteÍcone

Descrição gerada automaticamente com confiança médiaÍcone

Descrição gerada automaticamente com confiança média

**Acessar o Console de Gerenciamento da AWS**

1. Na parte superior destas instruções, clique em **Start Lab** (Iniciar laboratório) para iniciar o laboratório.

Um painel Start Lab (Iniciar laboratório) é aberto exibindo o status do laboratório.

1. Aguarde até ver a mensagem “**Lab status: in creation** (Status do laboratório: em criação)” e clique no **X** para fechar o painel Start Lab (Iniciar laboratório).

**Observação**: pode levar aproximadamente 20 minutos ou mais para que o status do laboratório mude para “ready” (pronto).

1. Na parte superior destas instruções, clique em AWS

Isso abrirá o Console de Gerenciamento da AWS em uma nova guia do navegador. O sistema fará o seu login automaticamente.

**Dica**: se uma nova guia do navegador não for aberta normalmente, um banner ou um ícone na parte superior do navegador indicará que o navegador está impedindo que o site abra janelas pop-up. Clique no banner ou ícone e escolha "Allow pop ups" (Permitir pop-ups).

1. Organize a guia do Console de Gerenciamento da AWS para que ela seja exibida com estas instruções. Em um cenário ideal, você poderá ver as duas guias do navegador ao mesmo tempo, para facilitar o acompanhamento das etapas do laboratório.

**Já passamos pelos Objetivos:**

**Objetivo 1**: Criar uma Instancia EC2 Amazon Linux com APP

**Objetivo 2**: Criar um Modelo AMI (Amazon Machine Image) para o Auto Scaling

**Objetivo 3**: Criamos o Auto Scalling

Nesta tarefa, você criou regras de funcionamento do Auto Scalling onde identificou que 2 instâncias EC2 em zonas de disponibilidade diferente foram criadas.

## **Objetivo 4 - Aplicando Auto-Scaling e ambiente com muitos acessos**

**Etapa 1** - Vamos criar uma polícia de escalabilidade automática

No Console **AWS**, em **EC2** clique em **Grupos do Auto Scaling:**

Tela de computador com texto preto sobre fundo branco

Descrição gerada automaticamente

Clique em **GrupoAutoScaling**

Interface gráfica do usuário, Aplicativo, Teams

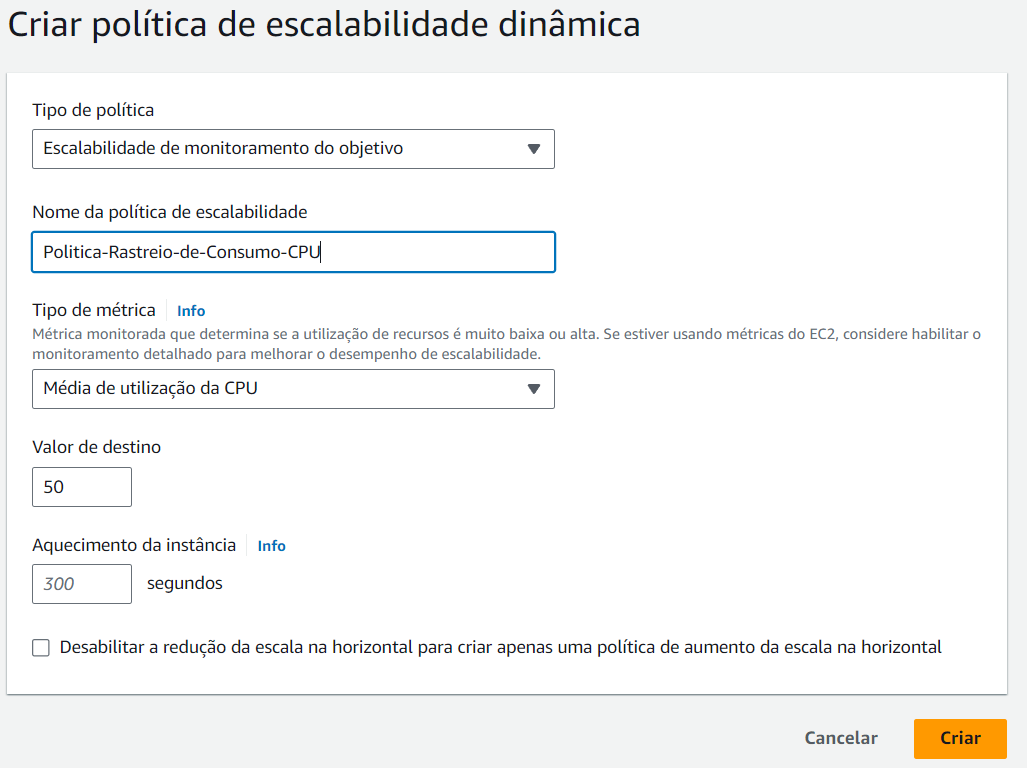
Descrição gerada automaticamente

Dentro do **GrupoAutoScaling** escolha a aba “**Escalabilidade automática**”

Interface gráfica do usuário, Texto, Aplicativo

Descrição gerada automaticamente

Clique em “**Criar política de escalabilidade dinâmica**”



Dentro de “Criar política de escalabilidade dinâmica

* Escolar o “Tipo de política” escolha: **Política de escalabilidade de monitoramento do objetivo**;
* Nome da política: **Politica-Rastreio-de-Consumo-CPU**
* Tipo de métrica: **Média de Utilização média da CPU***;*
* Valor de destino: **50**

Isso informa ao Auto Scaling para manter uma utilização de CPU *média* em *todas as instâncias* de 50%.

O Auto Scaling adicionará ou removerá automaticamente a capacidade conforme necessário para manter a métrica no valor de destino especificado ou próximo a ele.

Acesse a guia do navegador com a aplicação Web APP de **cada instância**.

Interface gráfica do usuário, Texto, Aplicativo, chat ou mensagem de texto

Descrição gerada automaticamente

Clique em **Load Test** (Carregar teste) ao lado do logotipo da AWS.

A máquina irá começar a consumir muito processamento chegando em 100%.

Veja se a instância esta com IP Válido e DNS e faça acesso em cada uma pelo endereço:

<http://ec2-107-22-24-147.compute-1.amazonaws.com/index.php>

Veja que as instancias estão sendo criadas em média de 5 minutos.

Interface gráfica do usuário, Aplicativo

Descrição gerada automaticamente

Você deve ver duas novas instâncias chamadas de **WebAutoScalling**

Elas foram iniciadas pelo Auto Scaling.

Se as instâncias ou nomes não forem exibidos, aguarde 30 segundos e clique em atualizar no canto superior direito.

Primeiro, você confirmará que as novas instâncias foram aprovadas na verificação de integridade.

**Tarefa 6**: Testar o Auto Scaling

Você criou um grupo de Auto Scaling com um mínimo de duas instâncias e um máximo de seis instâncias. Atualmente, duas instâncias estão em execução porque o tamanho mínimo é duas e o grupo não está atualmente sob nenhuma carga. Agora, você aumentará a carga para fazer com que o Auto Scaling acrescente instâncias adicionais.

1. Volte para o Console de Gerenciamento da AWS, mas não feche a guia da aplicação. Você retornará a ele em breve.
2. No menu **Services** (Serviços), clique em **CloudWatch**.
3. No painel de navegação à esquerda, clique em **Alarms** (Alarmes) (*não em* **ALARM** (ALARME)).

Verifique no serviço do CloudWatch que aletas foram gerados.

Tela de computador com texto preto sobre fundo branco

Descrição gerada automaticamente

Clique na guia **Alarmes**.

Interface gráfica do usuário, Texto, Aplicativo, Email

Descrição gerada automaticamente

1. Aguarde até que o **Status** de ambas as instâncias mude para **Healthy** *(íntegro)*.
2. Clique em atualizar no canto superior direito para verificar se há atualizações.

O status **Healthy** *(íntegro)* indica que a instância passou na verificação de integridade.

Alarmes serão exibidos. Eles foram criados automaticamente pelo grupo de Auto Scaling.

Eles manterão automaticamente a carga média da CPU próxima a 50%, permanecendo também dentro da limitação de ter duas a seis instâncias.

​

* No menu **Services** (Serviços), clique em **CloudWatch**.
* No painel de navegação à esquerda, clique em **All Alarms** (Todos os Alarmes) (*não em* **ALARM** (ALARME)) e confirme se vê dois alarmes.

1. Clique no alarme **OK**, que tem *AlarmHigh (Alarme alto)* no nome.

Se nenhum alarme estiver mostrando **OK**, aguarde um minuto e clique em Refresh (Atualizar)  no canto superior direito até que o status do alarme mude.

O **OK** indica que o alarme *não* foi acionado.

Interface gráfica do usuário, Aplicativo

Descrição gerada automaticamente

Em menos de 5 minutos, o alarme **AlarmLow** (Alarme baixo) deverá mudar para **OK** e o status do alarme **AlarmHigh** (Alarme alto) deverá mudar para *ALARM* (ALARME).

 Você pode clicar em Refresh (Atualizar)  no canto superior direito a cada 60 segundos para atualizar a exibição.

Você deve ver o gráfico **AlarmHigh**(Alarme alto) indicando uma porcentagem crescente de CPU. Depois de passar da linha de 60% por mais de 3 minutos, o Auto Scaling acionará a adição de novas instâncias.

1. Aguarde até que o alarme **AlarmHigh** (Alarme alto) entre no estado *ALARM (ALARME)*.

Agora você pode visualizar as instâncias adicionais que foram executadas.

1. No menu **Services** (Serviços), clique em **EC2**.
2. No painel de navegação esquerdo, clique em **Instances** (Instâncias).

Agora deve haver mais de duas instâncias rotuladas **Lab Instance** (Instância do laboratório) em execução. As novas instâncias foram criadas pelo Auto Scaling em resposta ao alarme.

**Remover Grupos de Auto Scaling, AMIs e Instâncias**