Laboratório 6: Ajuste a escala e o balanceamento de carga da arquitetura

**Versão 4.6.6 (TESS3) + alteração personalizada**

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

O **Elastic Load Balancing** distribui automaticamente o tráfego de entrada das aplicações entre várias instâncias do Amazon EC2. Ele permite obter tolerância a falhas nas aplicações ao disponibilizar de forma transparente a capacidade necessária de balanceamento de carga para rotear o tráfego das aplicações.

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 nova imagem de máquina da Amazon (AMI) por meio de uma instância em execução (Web Server 1);
* Criar um grupo de destino e um balanceador de carga;
* 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 **30 minutos**.

**Cenário**

Você começa com a seguinte infraestrutura:

Interface gráfica do usuário, Aplicativo

Descrição gerada automaticamente

O estado final da infraestrutura é:

Uma imagem contendo Logotipo

Descrição gerada automaticamente

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 10 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.

**Tarefa 1**: Criar uma AMI (Amazon Machine Image) para o Auto Scaling

Nesta tarefa, você criará uma AMI pelo ***Servidor da Web 1*** existente.

Isso salvará o conteúdo do disco de inicialização para que novas instâncias possam ser executadas com conteúdo idêntico.

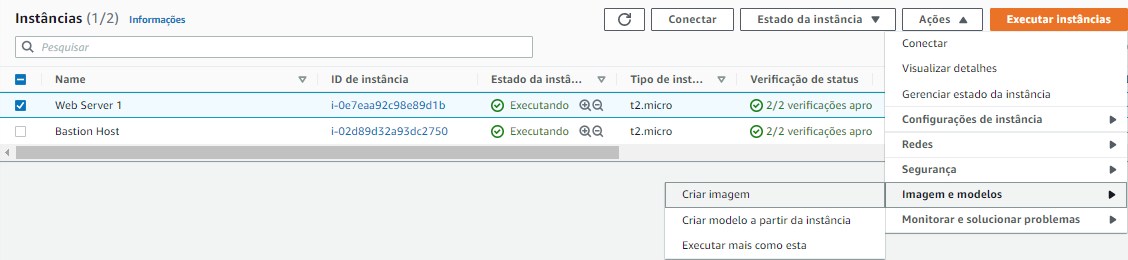
1. No **Console de Gerenciamento da AWS**, no menu  **Services**  (Serviços), clique em **EC2**.
2. No painel de navegação esquerdo, clique em **Instances** (Instâncias).

Primeiro, você confirmará que a instância está em execução.

1. Aguarde até que as **Status Checks** (Verificações de status) do ***Servidor da Web 1*** exiba *2/2 checks passed* (2/2 verificações aprovadas).
2. Clique em Refresh (Atualizar)  para atualizar.

Agora, você criará uma AMI com base nessa instância.

1. Selecione  **Web Server 1**.
2. No menu **Ações**, clique em **Imagem e Modelos** > **Criar Imagem** e configure:



* + **Nome da imagem**:  ***AMI do Servidor******Web***
  + **Descrição da imagem**:  AMI do servidor da Web para o LAB
  + Não Reinicializar: Desabilitado

Interface gráfica do usuário, Texto, Aplicativo, Email

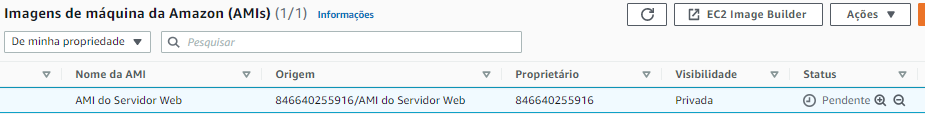
Descrição gerada automaticamente

1. Clique em **Criar imagem**

No menu à esquerda escolha AMIs:

Interface gráfica do usuário, Texto, Aplicativo, Site

Descrição gerada automaticamente



Aguarde o Status mudar de Pendente para Disponível.

Você usará essa AMI ao iniciar o Grupo de Auto Scaling posteriormente no laboratório.

**Tarefa 2**: Criar um Modelo de configuração de execução e um grupo de Auto Scaling

**Etapa 1 – Criar Modelo de Execução**

**Entre no console AWS, EC2 no menu à esquerda clique em Modelos de execução**

Interface gráfica do usuário, Aplicativo

Descrição gerada automaticamente

Dentro de **Modelos de execução**

Interface gráfica do usuário, Texto

Descrição gerada automaticamente

Clique em **Criar modelo de execução**

Interface gráfica do usuário, Texto, Aplicativo, Email

Descrição gerada automaticamente

Preencha o nome: **ModeloServidorWebAPP**

Descrição o nome: Modelo para ser usado no Grupo de Auto Scalling

Marque: “Fornecer orientação para me ajudar a configurar um modelo...”

Não vamos preencher **Tags de modelo**

Não vamos mudar **Modelo de origem**

Dentro de conteúdo do modelo, marque “**Minhas AMIs**”

Interface gráfica do usuário, Texto, Aplicativo, Email

Descrição gerada automaticamente

Selecione a **AMI do Servidor Web APP**

Dentro de “**Tipo de instância**”, selecione: t2.micro

Interface gráfica do usuário, Texto, Aplicativo, Email

Descrição gerada automaticamente

Dentro de “**Par de chaves**”, não selecione nada, vamos aumentar nossa segurança.

Dentro de “**Configurações de rede**”, selecione: Sub-Rede-Publica-1

Interface gráfica do usuário, Texto, Aplicativo, Email

Descrição gerada automaticamente

Dentro de “**Firewall**”, “**Selecionar grupo de segurança existente**”, selecione: GrupoSec-ServidoresLinux.

Dentro de “**Detalhes avançados**”, role até:

“**Tipo de nome host**”, selecione: Nome do IP.

“**Nome do host DNS**”, selecione: Habilitar solicitações de DNS IPv4 (registro A).

Interface gráfica do usuário, Texto, Aplicativo, Email

Descrição gerada automaticamente

No menu do lado direito “**Resumo**”

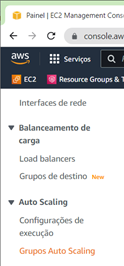
Interface gráfica do usuário, Texto, Aplicativo

Descrição gerada automaticamente

Clique em “**Criar um modelo de execução**

**Tarefa 3 – Criar Grupo do Auto Scanlling**

**Entre no console AWS, EC2 no menu à esquerda clique em Grupos do Auto Scalling**

****

Você será redirecionado para essa página:

Interface gráfica do usuário, Texto

Descrição gerada automaticamente **Clique em “Criar o grupo de Auto Scalling”:**

**Etapa 1:** Escolha o nome e o modelo de execução:

Interface gráfica do usuário, Texto, Aplicativo, Email

Descrição gerada automaticamente

* Informe o nome: **GrupoAutoScalling**
* Escolha em “**Modelo de execução**”: **ModeloServidorWebAPP**

Interface gráfica do usuário, Texto, Aplicativo, Email

Descrição gerada automaticamente

Manter versão “Default(1) e Clique em **Próximo**.

**Etapa 2: Selecione sua VPC e AZ’s:**

Interface gráfica do usuário, Texto, Aplicativo, Email

Descrição gerada automaticamente

Usar a VPC **Lab-VPC** e as AZs **Sub-Rede-Publica-1** e **Sub-Rede-Publica-2**.

Interface gráfica do usuário, Texto, Aplicativo, Email

Descrição gerada automaticamente

Clique em **Próximo**.

**Etapa 3:**

**Interface gráfica do usuário, Texto, Aplicativo, Email

Descrição gerada automaticamente**

* Aqui, nas configurações avançadas, você pode anexar diretamente com um load balancer (balanceador de carga) existente.
* Porém, como esse é o conteúdo não contemplas o ELB, não estaremos fazendo por enquanto.

Interface gráfica do usuário, Texto, Aplicativo

Descrição gerada automaticamente

Habilitar coleta de métricas de grupo no CloudWatch e clique em **Próximo**.

**Etapa 4:**

Interface gráfica do usuário, Texto, Aplicativo

Descrição gerada automaticamente

* + Capacidade mínima: **1**
  + Capacidade desejada: **2**
  + Capacidade máxima: **4**
* Política de escalabilidade – opcional, manter marcado “Nenhum”.

Interface gráfica do usuário, Texto, Aplicativo, Email

Descrição gerada automaticamente

* + Nome da Política de escalabilidade: **Politica-Restreio-Consumo-CPU**
  + Tipo de Métrica: **Média de utilização de CPU**
  + Valor de destino: **50**
  + Desmarcar: **Desabilitar a redução para criar apenas uma politica de expansão**

Interface gráfica do usuário, Texto, Aplicativo, Email

Descrição gerada automaticamente

Clique em **Próximo**.

**Etapa 5:**

**Interface gráfica do usuário, Texto, Aplicativo, Email

Descrição gerada automaticamente**

* Você também pode enviar notificações para tópicos do SNS (Simple Notification Service) mas neste caso não vamos usar.

**Etapa 6 - Tags:**

Interface gráfica do usuário, Texto, Aplicativo, Email

Descrição gerada automaticamente

Adicionar Tag: **WebAutoScalling**

Adicionar tags às suas configurações na AWS são de extrema ajuda em casos de grandes empresas, podendo te poupar muito tempo na procura por erros e mais.

**Etapa 7 - Revisão:**

Aqui, você irá revisar suas configurações, mostrando as 6 etapas anteriores.

Interface gráfica do usuário, Texto, Aplicativo, Email

Descrição gerada automaticamente

Interface gráfica do usuário, Texto, Aplicativo, Email

Descrição gerada automaticamente

Interface gráfica do usuário, Texto, Aplicativo, Email

Descrição gerada automaticamente

Interface gráfica do usuário, Texto, Aplicativo, Email

Descrição gerada automaticamente

Interface gráfica do usuário, Texto, Aplicativo, Email, Teams

Descrição gerada automaticamente

Clique em “**Criar o grupo do Auto Scaling**”.

Grupo criado com sucesso

Interface gráfica do usuário, Aplicativo, Teams

Descrição gerada automaticamente

Já esta “Atualizando capacidades”

Veja em **EC2** e **Instâncias** que as máquinas começaram a ser criadas:

Interface gráfica do usuário, Texto, Aplicativo

Descrição gerada automaticamente

**Tarefa 4**: Criar grupo de Destino para o balanceador de carga

Vamos criar um balanceador de carga que pode balancear o tráfego entre várias instâncias EC2 e zonas de disponibilidade.

1. No painel de navegação à esquerda, clique em **Grupo de Destino**

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

Descrição gerada automaticamente

Criar em **Criar grupo de destino**

Interface gráfica do usuário, Texto, Aplicativo

Descrição gerada automaticamente

1. Selecione o “**Instâncias**”.

Interface gráfica do usuário, Texto, Aplicativo, Email

Descrição gerada automaticamente

Em **Nome**, insira: **ALBGrupoDestino**

Manter porta 80 = HTTP

Selecione o “Lab VPC”.

Interface gráfica do usuário, Texto, Aplicativo, Email

Descrição gerada automaticamente

Não alterar nada nos demais itens:



Selecione **Próximo**

1. Masque as duas “**Instâncias**”

Interface gráfica do usuário, Texto, Aplicativo, Email

Descrição gerada automaticamente

1. Clique em **Criar Grupo de destino**

Veja que o Grupo de Destino foi criado com sucesso, mas não foi atrelado a um Load Balancer:

Interface gráfica do usuário, Texto, Aplicativo

Descrição gerada automaticamente

**Tarefa 5**: Criar um balanceador de carga

Nesta tarefa, você criará um balanceador de carga que pode balancear o tráfego entre várias instâncias EC2 e zonas de disponibilidade.

1. No painel de navegação à esquerda, clique em **Load Balancers (Balanceadores de carga)**.

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

Descrição gerada automaticamente

1. Clique em **Criar Load Balancer**

Vários tipos diferentes de balanceador de carga são exibidos.

Vamos usar o **Application Load Balancer**, roteando o tráfego para os destinos (instâncias EC2, contêineres, endereços IP) ou para Camada 7 (aplicação).

Interface gráfica do usuário

Descrição gerada automaticamente

Clique em Criar:

Para obter mais informações, consulte: [Comparação de balanceadores de carga](https://aws.amazon.com/elasticloadbalancing/features/#compare)

* **Nome do Load balancer**: **LabALB**
* **Esquema**: **Voltado para a Internet**
* **Tipo de endereço IP**: **IPv4**

Interface gráfica do usuário, Texto, Aplicativo, Email

Descrição gerada automaticamente

Interface gráfica do usuário, Texto, Aplicativo, Email

Descrição gerada automaticamente

* **VPC:** **Lab VPC**
* **Mapeamentos**: marcar ambas as sub-redes disponíveis.
* Selecione **Sub-rede-Publica-1** e **Sub-rede-Publica-2**

Isso configura o balanceador de carga para operar em duas zonas de disponibilidade.

Selecione  **GrupoSecServidoresLinux** (Grupo de segurança da Web) e desmarque  **default** (padrão).

Interface gráfica do usuário, Texto, Aplicativo, Email

Descrição gerada automaticamente

Marque o Grupo de Destino criado: **ALBGrupoDestino**

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

Descrição gerada automaticamente

O roteamento configura para onde enviar as solicitações que chegam no balanceador de carga.

Não adicionar, Acelerador

Interface gráfica do usuário, Texto, Aplicativo, Email

Descrição gerada automaticamente

Clique em “**Criar load balancer**”

O Load Balancer foi **LabALB** foi criado com sucesso:

Interface gráfica do usuário, Aplicativo

Descrição gerada automaticamente

O balanceador de carga mostrará um estado de “**Provisionando***”*. Não há necessidade de esperar ficar “**Ativo”**.

Siga para a próxima etapa.

NÃO ADIANTA FAZER O TESTE DO LOAD BALANCER SEM TER AS INSTÂNCIAS DO AUTO-SCALING CRIADAS

**Tarefa 6**: Ajusto do Grupo Auto Scalling

**Entre no** **console AWS, EC2 no menu à esquerda clique em Grupos do Auto Scalling**

**Interface gráfica do usuário, Texto, Aplicativo

Descrição gerada automaticamente**

Selecione o **GrupoAutoScalling** clieque em Editar

Interface gráfica do usuário, Texto, Aplicativo

Descrição gerada automaticamente

Na opção “Balanceamento de carga – opcional”, marque Grupo de Destino e escolha “**ALBGrupoDestino**”

Interface gráfica do usuário, Texto, Aplicativo, Email

Descrição gerada automaticamente

Role a até em baixo e clique em **Atualizar**

**Tarefa 7**: Verificar se o balanceamento de carga está funcionando

Nesta tarefa, você verificará se o balanceamento de carga está funcionando corretamente.

No painel de navegação esquerdo, clique em **Instâncias**.

Você deve ver duas novas instâncias chamadas de **WebAutoScalling** (Instância do laboratório).

Interface gráfica do usuário, Texto, Aplicativo

Descrição gerada automaticamente

Elas foram iniciadas pelo Auto Scaling.

Se as instâncias ou nomes não forem exibidos, aguarde 30 segundos, clique em atualizar.

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

1. No painel de navegação à esquerda, clique em **Grupos de destino**), na seção *Load Balancing (Balanceamento de carga)*.
2. Escolha **ALBGrupoDestino**

Interface gráfica do usuário, Texto, Aplicativo

Descrição gerada automaticamente

1. Clique **Destinos**.

Dois destinos **WebAutoScalling** devem ser listados para esse grupo de destino.

Aguarde até que o **Status** de ambas as instâncias mude para ***Healthy*** *(íntegro)*. Clique em atualizar para verificar se há atualizações.

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

Isso significa que o balanceador de carga enviará tráfego para a instância.

Agora você pode acessar o grupo de **Auto Scaling** por meio do balanceador de carga.

1. No painel de navegação à esquerda, clique em **Load Balancers**.

Interface gráfica do usuário, Texto, Aplicativo

Descrição gerada automaticamente

Selecione o **LabALB**, em Descrição copiar o **Nome do DNS**

Ele deve ser semelhante a: LabALB-443561510.us-east-1.elb.amazonaws.com

Interface gráfica do usuário, Texto, Aplicativo, Email

Descrição gerada automaticamente

Abra uma nova guia do navegador da Web, cole o nome DNS que você acabou de copiar e pressione [ENTER].

Interface gráfica do usuário, Aplicativo

Descrição gerada automaticamente

O aplicativo deve aparecer em seu navegador – OBS: o ID da Instância identifica qual você está acessando no momento.

Isso indica que o Load Balancer recebeu a solicitação, a enviou para uma das instâncias do EC2 e, em seguida, repassou o resultado.

**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)).

Dois alarmes serão exibidos. Eles foram criados automaticamente pelo grupo de Auto Scaling. Eles manterão automaticamente a carga média da CPU próxima a 60%, permanecendo também dentro da limitação de ter duas a seis instâncias.

​  **Observação**: siga estas etapas somente se você não vir os alarmes em 60 segundos.

* No menu **Services** (Serviços), clique em **EC2**.
* No painel de navegação à esquerda, clique em **Auto Scaling Groups** (Grupos de Auto Scaling) e, em seguida,
* Selecione a politica **LabGroupAutoScaling**
* clique em **Automatic** **Scaling** (Escalabilidade automática).
* Clique em **Ações e Editar** (Editar).
* Altere o **Target Value** (Valor de destino) para **50**.
* 
* Clique em   (Atualizar).
* 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.

Interface gráfica do usuário, Texto, Aplicativo, Email

Descrição gerada automaticamente

 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.

É o alarme para **CPU Utilization > 60** (Utilização de CPU > 60), que adicionará instâncias quando a CPU média estiver alta. O gráfico deve mostrar níveis muito baixos de CPU no momento.

Agora, você informará ao aplicativo para executar cálculos que devem aumentar o nível de CPU.

1. Volte para a guia do navegador com a aplicação Web.

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

Descrição gerada automaticamente

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

Isso fará com que a aplicação gere cargas elevadas. A página do navegador será atualizada automaticamente para que todas as instâncias no grupo de Auto Scaling gerem carga. Não feche esta guia.

1. Volte para a guia do navegador com o console do **CloudWatch**.

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.

**Tarefa 7**: Encerrar a Grupo Auto Scalling e as instâncias serão removidas

Nesta tarefa, No console AWS**, EC2** no menu à esquerda clique em **Grupos do Auto Scalling**

Selecione o **GrupoAutoScalling**

Interface gráfica do usuário, Texto, Aplicativo, Email

Descrição gerada automaticamente

Escola a opção **Excluir**

Interface gráfica do usuário, Texto, Aplicativo, Email

Descrição gerada automaticamente

Preencha “exclui” e clique em **Excluir**.

**Laboratório concluído**

Parabéns! Você concluiu o laboratório.

1. Clique em End Lab (Encerrar laboratório) na parte superior desta página e, em seguida, clique em **Yes** (Sim) para confirmar que você deseja encerrar o laboratório.

Um painel será exibido com a mensagem: "DELETE has be initiated... You may close this message box now." (a EXCLUSÃO foi iniciada... Você pode fechar esta caixa de mensagem agora).

1. Clique no **X** no canto superior direito para fechar o painel.

Envie um e-mail com feedback, sugestões ou correções para: [*aws-course-feedback@amazon.com*](mailto:aws-course-feedback@amazon.com)