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;
* Criar 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 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** (Servidor da Web 1).
2. No menu **Actions** (Ações), clique em **Image** (Imagem) > **Create Image** (Criar imagem) e configure:
   * **Image name** (Nome da imagem):
   * ***Web Server AMI***(AMI do servidor da Web)
   * **Image description** (Descrição da imagem):
   * Lab AMI for Web Server (AMI do laboratório para servidor da Web)
   * 
3. Clique em   (Criar imagem)

A tela de confirmação exibe o **AMI ID** (ID da AMI) para a nova AMI.

1. Clique em **Close** (Fechar)

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

**Tarefa 2**: 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)**.
2. Clique em **Create Load Balancer** (Criar balanceador de carga)

Vários tipos diferentes de balanceador de carga são exibidos.

Você usará um *Application Load Balancer* que opera no nível de solicitação (camada 7), roteando o tráfego para os destinos (instâncias EC2, contêineres, endereços IP e funções do Lambda) com base no conteúdo da solicitação. Para obter mais informações, consulte: [Comparação de balanceadores de carga](https://aws.amazon.com/elasticloadbalancing/features/#compare)

1. Em **Application Load Balancer**, clique em **Create** (Criar) e configure:

* **Name** (Nome): **LabELB**
* **VPC:** ***Lab VPC*** (na seção **Availability Zones** (Zonas de disponibilidade))
* **Availability Zones** (Zonas de disponibilidade): selecione  ambas para ver as sub-redes disponíveis.
* Selecione **Public Subnet 1** (Sub-rede pública 1) e **Public Subnet 2** (Sub-rede pública 2)

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

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

Descrição gerada automaticamente

1. Clique em **Next: Configure Security Settings** (Próximo: definir configurações de segurança).

 Você pode ignorar o aviso *"Improve your load balancer's security." (Melhore a segurança do seu balanceador de carga)*.

1. Clique em **Next: Configure Security Groups** (Próximo: configurar grupos de segurança).

Um *Web Security Group (Grupo de segurança da Web)* já foi criado para você, o que permite o acesso HTTP.

1. Selecione  **Web Security Group** (Grupo de segurança da Web) e desmarque  **default** (padrão).
2. Clique em **Next: Configure Routing** (Próximo: configurar roteamento).

O roteamento configura para onde enviar as solicitações que chegam no balanceador de carga. Você criará um *Grupo de destino* que será usado pelo Auto Scaling.

1. Em **Name** (Nome), insira: **LabGroupRouting**
2. Clique em **Next: Register Targets** (Próximo: registrar destinos).

O Auto Scaling registrará automaticamente instâncias como destinos posteriormente no laboratório.

1. Clique em **Next: Review** (Próximo: revisão).
2. Clique em **Create** (Criar) e em **Close** (Fechar)

O balanceador de carga mostrará um estado de *provisioning (provisionamento)*. Não há necessidade de esperar até que ele esteja pronto. Siga para a próxima etapa.

**Tarefa 3**: Criar uma configuração de execução e um grupo de Auto Scaling

Nesta tarefa, você criará uma *configuração de execução* para seu grupo de Auto Scaling.

Uma configuração de execução é um modelo que um grupo de Auto Scaling usa para iniciar instâncias do EC2.

Ao criar uma configuração de execução, você específica informações para as instâncias, como a AMI, o tipo de instância, um par de chaves, grupo de segurança e discos.

1. No painel de navegação à esquerda, clique em **Launch Configurations** (Configurações de execução).
2. Clique em **Create launch configuration** (Criar configuração de execução)
3. Defina estas configurações:

* **Launch configuration name** (Nome da configuração de execução): LabConfig
* **Amazon Machine Image (AMI)** (Imagem de máquina da Amazon (AMI)): escolha ***Web Server AMI*** *(AMI do servidor da Web)*
* **Instance type** (Tipo de instância):
  + Selecione **Choose instance type** (Escolher tipo de instância)
  + Selecione *t3.micro*
  + Selecione **Choose** (Escolher)

**Observação:** se você tiver executado o laboratório na região us-east-1, selecione o tipo de instância **t2.micro**.

Para localizar a região, procure no canto superior direito do console do Amazon EC2.

**Observação:** se receber a mensagem de erro “Something went wrong. Please refresh and try again (Algo deu errado. Atualize e tente novamente)”, você poderá ignorá-la e continuar com o exercício.

* **Configuração adicional**
  + **Monitoring** (Monitoramento): Selecione  *Enable EC2 instance detailed monitoring within CloudWatch (Habilitar monitoramento detalhado de instância do EC2 no CloudWatch)*

Interface gráfica do usuário, Texto, Aplicativo, Email

Descrição gerada automaticamente

Isso permite que o Auto Scaling reaja rapidamente a alterações na utilização.

1. Em **Security groups** (Grupos de segurança), você definirá a configuração de execução para usar o *Web Security Group (Grupo de segurança da Web)* que criou anteriormente.

* Escolha **Select an existing security group** (Selecionar um grupo de segurança existente)
* Selecione  **Web Security Group** (Grupo de segurança da Web)

Interface gráfica do usuário, Texto, Aplicativo, Email

Descrição gerada automaticamente

1. Em **Key pair** (par de chaves), configure:

* **Key pair options** (Opções de par de chaves): *Choose an existing key pair (Escolher um par de chaves existente)*
* **Existing key pair** (par de chaves existente): **vockey**
* Selecione  **I acknowledge...** (Eu reconheço...)
* Clique em **Create launch configuration (Criar configuração de execução)**

Agora, você criará um grupo de Auto Scaling que usa essa configuração de execução.

1. Marque a caixa de seleção para a configuração de execução ***LabConfig***.
2. No menu **Actions** (Ações), escolha *Create Auto Scaling group* (Criar grupo de Auto Scaling)
3. Insira o nome do grupo de Auto Scaling:

* **Name** (Nome): **LabGroupAutoS** (Grupo de Auto Scaling do laboratório)

1. Escolha **Next** (Próximo).
2. Na página **Network** (Rede), configure

* **Network** (Rede): *Lab VPC (VPC de laboratório)*

 Você pode ignorar a mensagem "No public IP address" (Sem endereço IP público)

* **Subnet** (Sub-rede): selecione *Private Subnet 1 (10.0.1.0/24) (Sub-rede privada 1 [10.0.1.0/24])* **e** *Private Subnet 2 (10.0.3.0/24) (Sub-rede privada 2 [10.0.3.0/24])*

Interface gráfica do usuário, Texto, Aplicativo, Email

Descrição gerada automaticamente

Isso executará instâncias do EC2 em sub-redes privadas em ambas as zonas de disponibilidade.

1. Escolha **Next** (Próximo).
2. Em **Load balancing** (Balanceamento de carga):

* Selecione  **Attach to an existing load balancing** (Habilitar existente balanceamento de carga)
* Escolha **Application Load Balancer or Network Load Balancer** (Application Load Balancer ou Network Load Balancer)
* **Escolha um grupo de destino para o balanceador de carga**: **LabGroupRouting**

1. Em **Additional settings** (Configurações adicionais), selecione  **Habilitar coleta de métricas de grupo no CloudWatch**

Essa ação captura métricas em intervalos de 1 minuto, o que permite que o Auto Scaling reaja rapidamente a mudanças nos padrões de uso.

1. Escolha **Next** (Próximo).
2. Em **Group Size** (Tamanho do grupo), configure:

* **Desired capacity** (Capacidade desejada): 2
* **Minimum capacity** (Capacidade mínima): 2
* **Maximum capacity** (Capacidade máxima): 6

Isso permitirá que o Auto Scaling adicione/remova instâncias automaticamente, mantendo sempre de 2 a 6 instâncias em execução.

1. Em **Scaling policies** (Políticas de escalabilidade), escolha *Target tracking scaling policy (política de escalabilidade de rastreamento de destino)* e configure:

* **Lab policy name** (Nome da política do laboratório): **LabScalingPolicy**
* **Metric type** (Tipo de métrica): *Average CPU Utilization (Utilização média de CPU)*
* **Target value** (Valor de destino): **60**

Isso informa ao Auto Scaling para manter uma utilização de CPU *média* em *todas as instâncias* de 60%.

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

Ele se ajusta às flutuações na métrica devido a um padrão de carga flutuante.

1. Escolha **Next** (Próximo).

O Auto Scaling pode enviar uma notificação quando ocorre um evento de escalabilidade.

Você usará as configurações padrão.

1. Escolha **Next** (Próximo).

As tags aplicadas ao grupo de Auto Scaling serão propagadas automaticamente para as instâncias executadas.

1. Escolha **Add tag** (Adicionar tag ) e configure o seguinte:

* **Key** (Chave): Name (Nome)
* **Value** (Valor): **LabInstanceAutoS** (Instância de laboratório)

1. Clique em **Next** (Próximo)
2. Revise os detalhes de seu grupo de Auto Scaling e clique em **Create Auto Scaling group** (Criar grupo de Auto Scaling).
3. Se receber a mensagem de erro **Failed to create Auto Scaling group** (Falha ao criar o grupo de Auto Scaling), clique em **Retry Failed Tasks** (Tentar tarefas com falha novamente).

O grupo de Auto Scaling mostrará inicialmente uma contagem de instâncias igual a zero, mas novas instâncias serão executadas para atingir a contagem **Desired** (Desejada) de 2 instâncias.

**Tarefa 4**: Verificar se o balanceamento de carga está funcionando

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

1. No painel de navegação esquerdo, clique em **Instances** (Instâncias).

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

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.

1. No painel de navegação à esquerda, clique em **Target Groups** (Grupos de destino), na seção *Load Balancing (Balanceamento de carga)*.
2. Escolha ***LabGroupRouting***
3. Clique na guia **Targets** (Destinos).

Dois destinos **LabInstancea** (Instância do laboratório) devem ser listados para esse grupo de destino.

1. Aguarde até que o **Status** de ambas as instâncias mude para ***healthy*** *(íntegro)*. 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 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** (Balanceadores de carga).
2. No painel inferior, copie o **DNS name** (Nome do DNS) do balanceador de carga, omitindo “(A Record)” (Registro A).

Ele deve ser semelhante a: *LabELB-1998580470.us-west-2.elb.amazonaws.com*

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

O aplicativo deve aparecer em seu navegador.

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

**Tarefa 5**: 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 **LabGroupAutoS**
* clique em **Automatic** **Scaling** (Escalabilidade automática).
* Clique em **Actions** (Ações) e em **Edit** (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 6**: Encerrar o servidor da Web 1

Nesta tarefa, você encerrará o *Servidor da Web 1*. Essa instância foi usada para criar a AMI usada por seu grupo de Auto Scaling, mas ela não é mais necessária.

1. Selecione  ***Web Server 1*** (Servidor da Web 1) (e verifique se é a única instância selecionada).
2. No menu **Actions** (Ações), clique em **Instance State** (Estado da instância) > **Terminate** (Encerrar).
3. Escolha **Yes, Terminate** (Sim, encerrar).

**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)