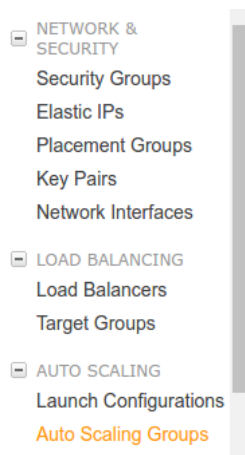


Configurando o Autoscaling

Transcrição

Para começar a utilizar o Auto Scaling, vamos selecionar do menu a esquerda a opção *Auto Scaling Groups*:



Criando uma launch configuration

O primeiro passo para implementar o serviço de Auto Scaling é criar uma launch configuration. A Launch Configuration é o template que as nossas máquinas que serão automaticamente criadas terão.

Create Auto Scaling Group

To create an Auto Scaling group, you will first need to choose a template that your Auto Scaling group will use when it launches instances for you, called a launch configuration. Choose a launch configuration or create a new one, and then apply it to your group.

Later, if you want to use a different template, you can create another launch configuration and apply it to this group, even if you already have instances running in it. Using this method, you can update the software that your group uses when it launches new instances.

Crie no botão *Create launch configuration*

O processo agora é bastante próximo ao da criação de nossa instância, logo devemos começar escolhendo a imagem que criamos `catalogo-v1` :

Em seguida vamos escolher o tipo de instância, e como queremos nos manter sem custo vamos escolher a `t2.micro` :

Create Launch Configuration

Amazon EC2 provides a wide selection of instance types optimized to fit different use cases. Instances are virtual servers that can run applications. They have varying combinations of CPU, memory, storage, and networking capacity, and give you the flexibility to choose the appropriate mix of resources for your applications. [Learn more](#) about instance types and how they can meet your computing needs.

Filter by: All instance types Current generation [Show/Hide Columns](#)

Em seguida vamos dar um nome a nossa launch configuration. Vamos chamá-la de `catalogo-lc`.

Create Launch Configuration

Name ⓘ

As outras opções desta tela podem ficar de modo padrão.

Agora escolheremos o storage, que é o dispositivo de armazenamento, que também será a configuração padrão:

Create Launch Configuration

Your instance will be launched with the following storage device settings. You can attach additional EBS volumes and instance store volumes to your instance, or edit the settings of the root volume. You can also attach additional EBS volumes after launching an instance, but not instance store volumes. <https://docs.aws.amazon.com/console/ec2/launchinstance/storage> about storage options in Amazon EC2.

Volume Type ⓘ	Device ⓘ	Snapshot ⓘ	Size (GiB) ⓘ	Volume Type ⓘ	IOPS ⓘ	Throughput ⓘ	Delete on Termination ⓘ	Encrypted ⓘ
Root	/dev/sda1	snap-0aa38b61171d7d5bf	<input type="text" value="8"/>	General Purpose (SSD) ▼	100 / 3000	N/A	<input checked="" type="checkbox"/>	No
<button>Add New Volume</button>								

Agora vamos configurar os security-groups, que são as portas de nossa aplicação, algo muito importante já que queremos que elas se comuniquem com o mundo externo, logo vamos abrir a nossa famosa porta TCP **8080**:

Por último, temos que escolher qual par de chaves, ou seja qual arquivo `.pem` será utilizado para o acesso aos nossas máquinas. Vamos utilizar o nosso `catalogo.pem` que já temos:

Não esqueça de marcar o checkbox para que ele libere a opção de prosseguir.

Agora que nossa configuração das máquinas que serão levantadas automaticamente estão prontas, vamos configurar quando isto deve acontecer.

Criando o Auto Scaling Group

Assim que você acabar de criar o seu launch configuration, você deve ser redirecionado para a tela de criação do Auto Scaling Group:

Aqui vamos preencher o campo nome com `catalogo-asg`, vamos dizer que queremos começar com 2 instâncias, não vamos alterar a *Network* padrão e vamos informar que queremos utilizar as subredes 1c e 1a.

Abaixo do formulário de informações simples, você deve ver que existem as opções avançadas:

Antes de avançar, clique para expandir este menu e vamos configurá-lo também:

Como queremos utilizar o nosso load balancer que criamos em um capítulo anterior, vamos marcar o checkbox e especificar no Target Groups que queremos escalar `catalogo-v1`.

No próximo passo é aonde configuramos a lógica que será utilizada no processo de Auto Scaling. Para isto, vamos selecionar a segunda opção e vamos começar a definir este comportamento:

Nosso Auto Scaling Group vai fazer com que nossas máquinas variem entre duas e 4 máquinas.

Agora vamos configurar quais regras vão fazer novas máquinas serem criadas ou destruídas.

Vamos criar um alarme para quando a média de utilização do processador ficar acima de 70%, por mais de 5 minutos, o nosso Auto Scaling Group tome alguma ação. Repare que também temos um checkbox que podemos ativar se desejamos receber por email uma notificação quando este alarme for ativado:

Agora vamos escolher o que deve acontecer quando este alarme soar. Queremos criar uma nova instância e podemos indicar também que esta instância leva um tempo determinado depois subir para que os serviços sejam iniciados, como é o caso do nosso servidor TomCat:

E claro, precisamos configurar também quando o Auto Scaling Group deve reduzir a quantidades de máquinas. Faremos isto nos alarmes de *Decrease Group Size*, que farão a lógica oposta do alarme anterior:

Repare que trocamos o sinal para o `<=`, pois queremos que as máquinas sejam reduzidas quando o uso de CPU estiver abaixo de 30%. Também não podemos esquecer de alterar o nome do alarme, para que ele não sobrescreva o anterior.

A ação que desejamos ter é que ele reduza as máquinas em 1 quando o alarme soar:

Podemos seguir para a tela final de revisão e finalizar a criação de nosso Auto Scaling Group.

Verificando o funcionamento do Auto Scaling Group

No final da criação do seu auto-scaling group, você será redirecionado para a tela dos seus auto-scaling groups e você deve ver a tela abaixo:

Se você voltar para o Menu de Instâncias, você verá que possui agora duas máquinas sem nome criadas:

Estas são as instâncias que foram criadas pelo Auto Scaling Group, então podemos ver tudo ocorreu como esperado.