## **Aula 12 – Escalabilidade com Auto-Scaling**

Este módulo aborda os seguintes tópicos:

* **Objetivo 1 -** Auto Scaling - Criando instância para o Auto Scaling;
* **Objetivo 2 -** Criando Instância modelo AMI;
* **Objetivo 3 -** Auto Scaling Group;

## **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/>

## **Auto Scaling - Criando instância Servidor Web APP para o Auto Scalling**

**Benefícios do Multi-AZ**

Multi-AZ é importante para implementações de serviços na AWS pois com múltiplas Zonas de Disponibilidade (AZ) temos uma série de benefícios. Entre eles são:

* Resiliência;
* Failover automático;
* Maior disponibilidade;
* Durabilidade aperfeiçoada;
* Proteção do desempenho do seu banco de dados;

**O que é Resiliência e Failover automático?**

Resiliência é a capacidade de uma carga de trabalho se recuperar de interrupções de infraestrutura ou serviço, adquirir dinamicamente recursos computacionais para atender à demanda e mitigar interrupções, como configurações incorretas ou problemas temporários de rede.

Se um volume de armazenamento na instância primária falhar em uma implantação Multi-AZ do Amazon RDS, ele iniciará automaticamente um failover (instância que estava em standby) para a espera atualizada (ou para uma réplica, no caso do Amazon Aurora).

**O que é maior disponibilidade?**

Você se beneficia da disponibilidade otimizada do servidor ao executar implantações Multi-AZ. Se ocorrer uma falha em uma Zona de disponibilidade ou em uma instância de DB, por exemplo, o impacto na disponibilidade será limitado ao tempo necessário para a conclusão do failover automático, que geralmente é inferior a um minuto para o Amazon Aurora (chegando a 30 segundos com o uso do MariaDB Connector/J) e de um a dois minutos para outros mecanismos de banco de dados (consulte as perguntas frequentes do RDS para obter detalhes).

**O que é Proteção do desempenho do seu banco de dados?**

Diferentemente das implantações Single-AZ, a atividade de E/S não é suspensa em sua instância principal durante o backup nas implantações Multi-AZ de mecanismos MySQL, MariaDB, Oracle e PostgreSQL, pois o backup é feito por meio da instância de espera.

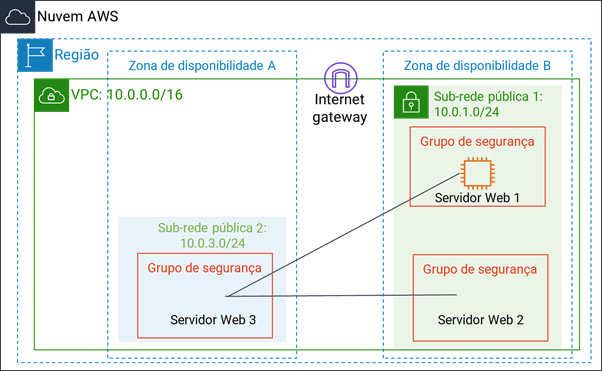
**Mais sobre implantações Multi-AZ**

Implantações Multi-AZ podem ter uma ou duas instâncias de banco de dados, por exemplo, em espera.

Uma implantação de instância de banco de dados Multi-AZ tem uma instância de banco de dados em espera (standby) que fornece suporte para failover, mas não serve tráfego de leitura.

Quando a implantação tem duas instâncias de banco de dados em espera, ela é chamada de implantação de cluster de banco de dados Multi-AZ. Uma implantação de cluster de banco de dados Multi-AZ tem instâncias de banco de dados em espera que fornecem suporte para failover e podem servir tráfego de leitura.

**Exemplo de topologia MultiAZ:**

Ícone

Descrição gerada automaticamente com confiança médiaÍcone

Descrição gerada automaticamente com confiança média

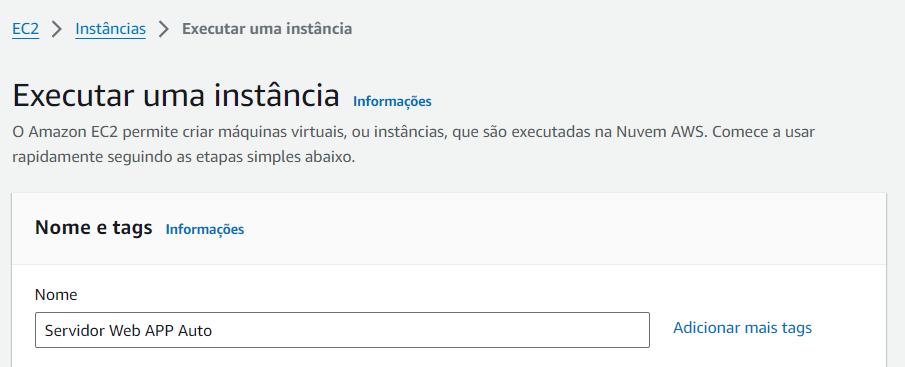
Aqui temos um exemplo de uma topologia com multi-AZ com uma alta disponibilidade e resiliência.

Nesta topologia temos 3 instâncias web, estando uma em uma Zona de Disponibilidade A e duas em uma Zona de Disponibilidade B. Caso uma das zonas caia, perca conectividade ou tenha algum problema parecido.

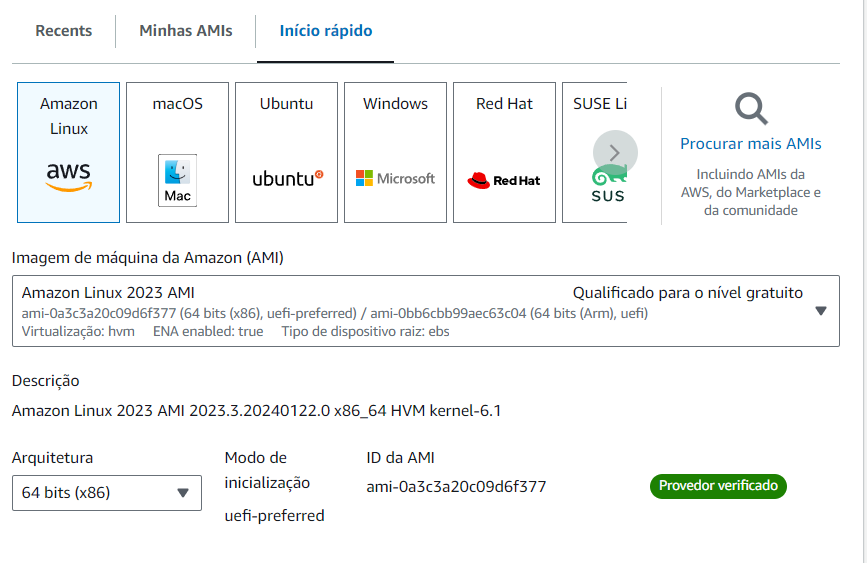
**Objetivo 1: Criar Instância com WebServer APP para o Auto Scalling**

## 

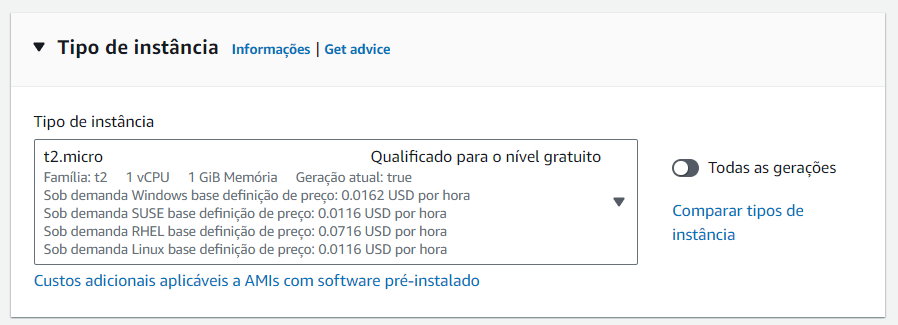
1. No Console de Gerenciamento da AWS no menu **Serviços**, clique em **EC2**.
2. No **Painel EC2,** clique em Executar instância.
3. Em **Nome e tags**, nomeie a instância como Servidor Web APP Auto



Seleciona a AMI Amazon Linux que é qualificada para o nível gratuito, mantenha 64 bits (x86) selecionado.

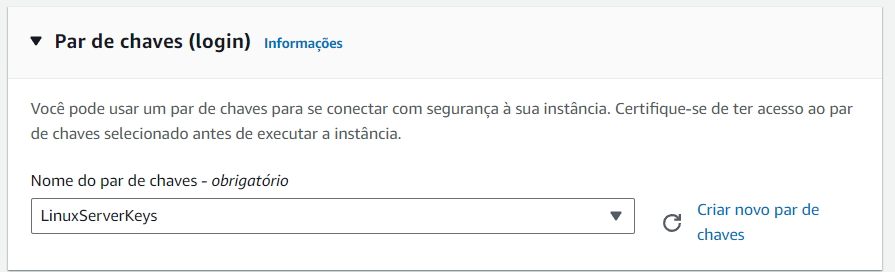


No Tipo de instância, mantenha o tipo da instância, **t2.micro** é neste momento que estamos escolhendo um processador e memória mais potentes conforme solicitação da empresa.

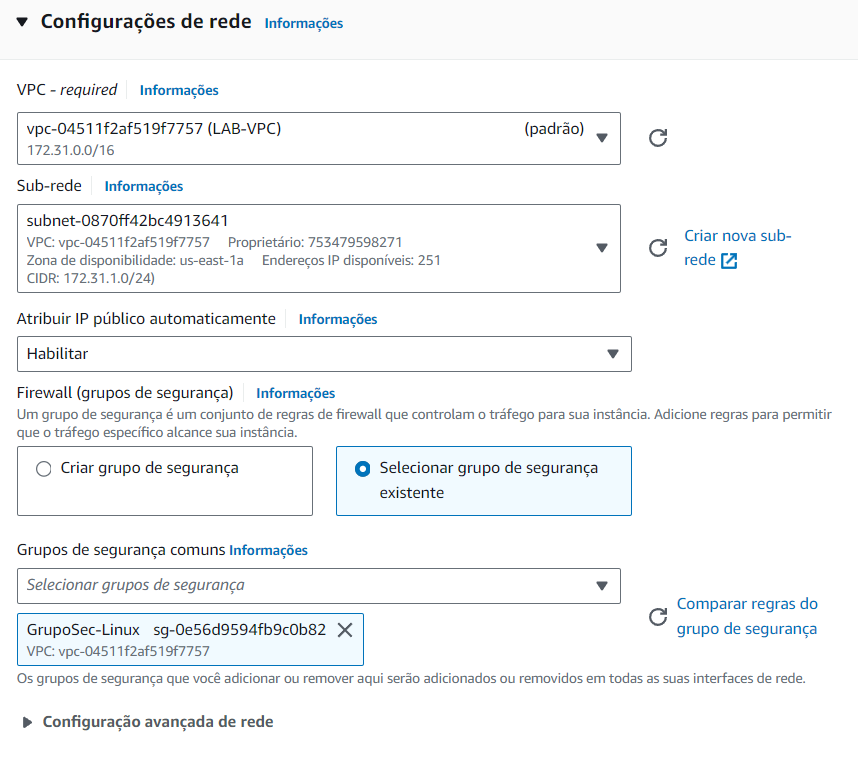


Segue relação e detalhamento de cada família de máquinas: <https://aws.amazon.com/pt/ec2/instance-types/>

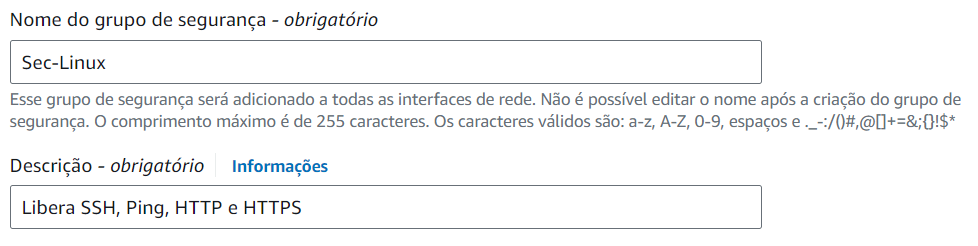
Agora devemos escolher um Par de chaves para fazer a conexão à instância.



Na etapa de Configurações de rede, podemos fazer configurações detalhadas de nossa instância para este caso, usaremos a VPC padrão, onde nossa instância **Amazon EC2**, que será atribuído automaticamente um endereço IP público à nossa máquina virtual incluindo DNS.



Se não tiver grupo de segurança criar:



1. Estamos usando **VPC-ANEIS** padrão da AWS “vpc-05c689....”;
2. Alterar **Atribuir IP público automaticamente** para: Habilitar;

c) Clicamos em **Selecionar grupo de segurança existente** e selecionamos o grupo **Sec-Linux**;

Role para baixo para ver a seção **Detalhes avançados**.

* Expanda **Detalhes avançados**. Aparecerá no último campo **Dados do Usuário**.
* Copie os comandos a seguir e cole-os no campo **Dados do Usuário**. Mantendo a opção “Como texto” selecionada.

Clique em “Próximo: Configure os detalhes da instância”;

#!/bin/bash -ex

# Updated Amazon Linux 2

sudo su -

yum -y update

# Instalar Apache2 - PHP - Banco de Dados MySQL

yum -y install httpd

yum -y install php

# Instalar Apache2 - MariaDB

yum install -y mariadb105-server

yum install -y php8.3

# Habilitar Apache2

/usr/bin/systemctl enable httpd

# Iniciar Apache2

/usr/bin/systemctl start httpd

# Habilitar Web-APP

cd /var/www/html

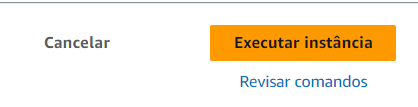
wget https://aws-tc-largeobjects.s3.amazonaws.com/CUR-TF-100-ACCLFO-2/lab5-rds/lab-app-php7.zip

unzip lab-app-php7.zip -d /var/www/html/

chown apache:root /var/www/html/rds.conf.php

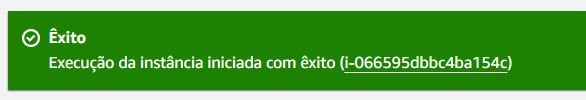
**Bootstrapping**, código fornecido que é executado quando um computador é inicializado.

Neste caso, para o servidor se tornar um servidor Web com APP para Sobrecarregar a máquina virtual e fazer com que o Auto Scaling comece a funcionar.



Clique em 

Você será direcionado para essa tela:

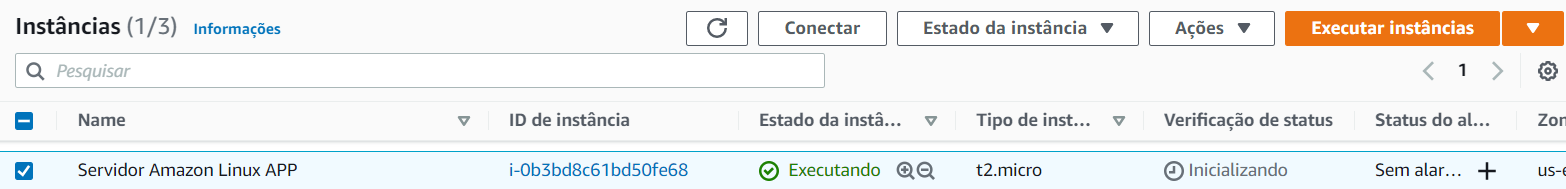


Clique no ID da instância.

**Etapa 8 –** Acesso APP do Servidor Web.

Temos as colunas:

* Name “Servidor WEB”;
* Estado da Instância (**Executando**);
* Tipo da Instância: ex.: **t2.micro**;
* Verificação de Status (Inicializando, 2/2 verificações aprovadas).



Assim que iniciar, faça o teste de acesso WEB.

Interface gráfica do usuário, Texto, Aplicativo

Descrição gerada automaticamente

Nesta mesma aba que já está marcada colete a informação: “Nome IPv4 público”: ex.: ec2-35-175-207-67.compute-1.amazonaws.com

Complete com: ec2-35-175-207-67.compute-1.amazonaws.com/index.php

Será exibida essa tela:

Interface gráfica do usuário, Texto, Aplicativo, Email

Descrição gerada automaticamente

**Objetivo 1 concluído**, vamos para o próximo laboratório sequencial.

## **Objetivo 2 - Criando Instância modelo AMI para o Auto Scaling**

Nesta tarefa, você criará uma AMI pelo **Servidor Web APP Auto** existente.

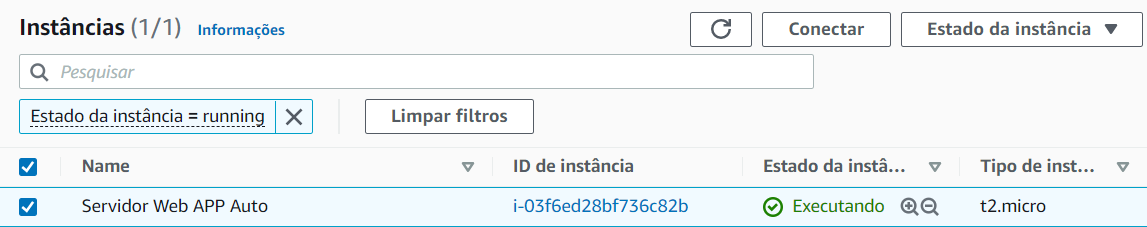
Isso salvará o conteúdo do disco de inicialização para que novas instâncias possam ser executadas com conteúdo idêntico.

No **Console de Gerenciamento da AWS**, no menu **Serviços**, clique em **EC2**.

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

1° - Confirme se a instância está em execução.

Aguarde até que as etapas de **Verificação de status** do **Servidor Web APP Auto**, estão 2/2 verificações aprovadas.



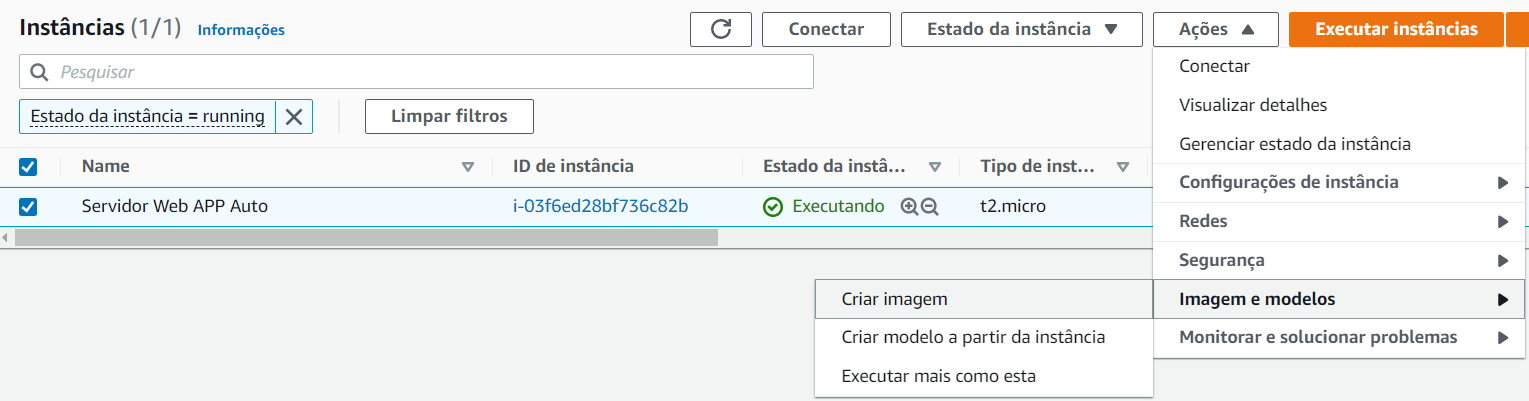


Clique em Refresh (Atualizar) até chegar ao estado desejado.

Agora, você criará uma AMI com base nessa instância.

Selecione o **Servidor Web APP Auto**.

No menu **Ações**), clique em **Imagem e Modelos** > **Criar Imagem**



Configure:

Interface gráfica do usuário, Texto, Aplicativo, Email

Descrição gerada automaticamente

* 1. **Nome da imagem**: AMI do Servidor Web APP
  2. **Descrição da imagem**: AMI do Servidor Web para o LAB
  3. **Deixe desmarcado**: Não reiniciar

Interface gráfica do usuário, Texto, Aplicativo, Email

Descrição gerada automaticamente

Clique em **Criar imagem**

No **Console de Gerenciamento da AWS**, no menu **Serviços**, clique em **EC2**.

No painel de navegação esquerdo, clique em **AMIs**

Interface gráfica do usuário, Aplicativo

Descrição gerada automaticamente

Interface gráfica do usuário, Aplicativo

Descrição gerada automaticamente

Aguarde o Status mudar de Pendente para Disponível

Demora até 5 minutos para criar.

Máquina modelo criada com sucesso:

Interface gráfica do usuário, Aplicativo

Descrição gerada automaticamente

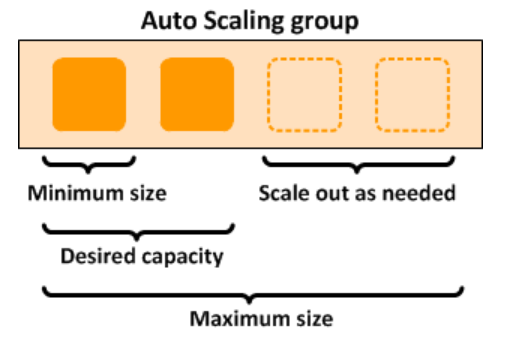
Você usará essa AMI ao iniciar o grupo de Auto Scaling posteriormente, no objetivo 3.

## **Auto Scaling Group**

É um conjunto de instâncias do Amazon EC2 que são tratadas como um agrupamento lógico para fins de escalabilidade e gerenciamento automáticos. O tamanho de um grupo de Auto Scaling depende do número de instâncias definidas como a capacidade desejada.

Você pode ajustar seu tamanho para atender à demanda manualmente ou usando a escalabilidade automática.

A seguir, você verá uma imagem ilustrativa contendo as informações necessárias ao criar um Auto Scaling Group e poder entender um pouco melhor como ele funciona na prática logo em seguida.



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

* Minimum size (tamanho mínimo): O menor número de instâncias ativas que seu Auto Scaling pode ter.
* Desired capacity (capacidade desejada): O número de instâncias recomendado.
* Scale out as needed (dimensionar conforme necessário): Espaço disponível para expansão do número de instâncias.
* Maximum size (tamanho máximo): O número máximo de instâncias que podem ser executadas ao mesmo tempo.

## **Objetivo 3 – Auto Scaling Group**

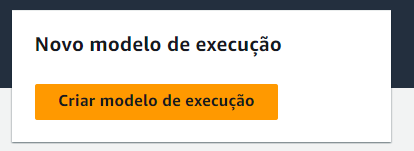
**Etapa 1 – Criar Modelo de Execução**

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

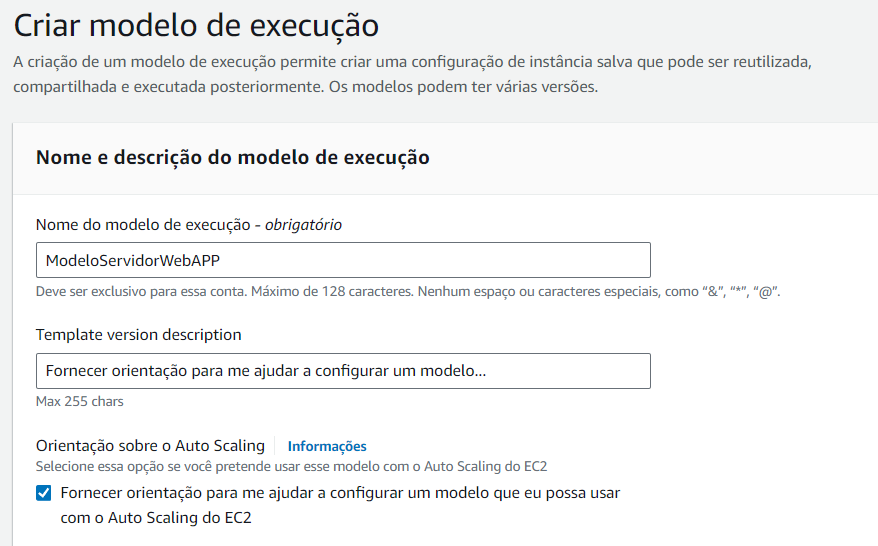
Interface gráfica do usuário, Aplicativo

Descrição gerada automaticamente

Dentro de **Modelos de execução**



Clique em **Criar modelo de execução**



Preencha o nome: ModeloServidorWebAPP

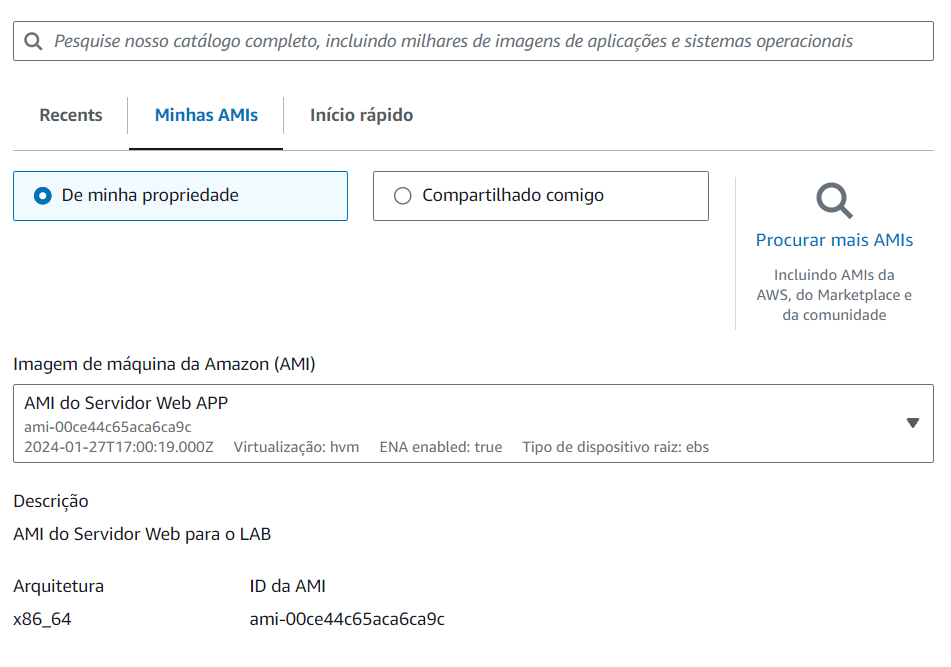
Descrição o nome: Modelo para ser usado no Grupo de Auto Scaling

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**”



Selecione a **AMI do Servidor Web APP**

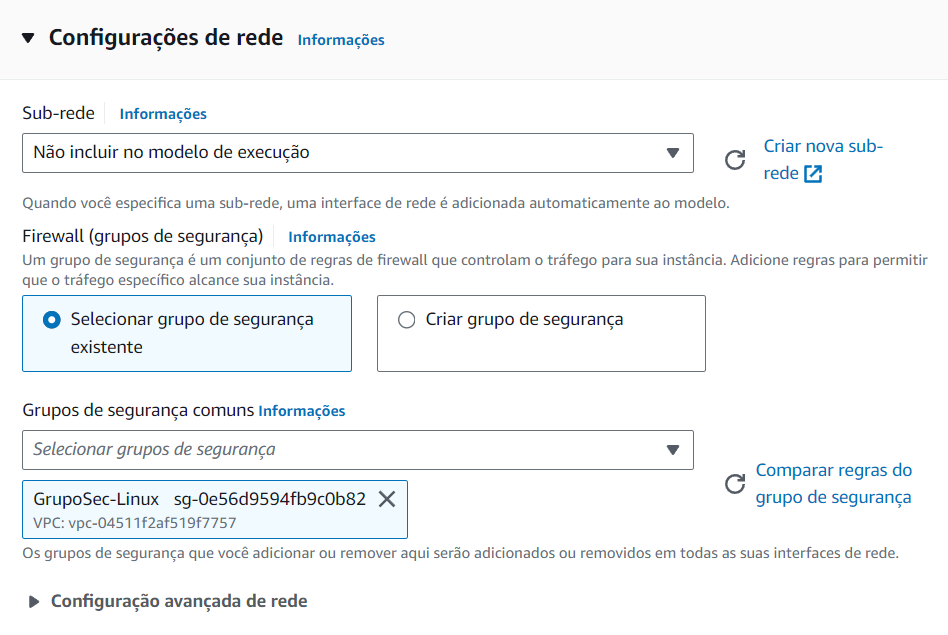
Dentro de “**Tipo de instância**”, selecione: t2.micro

Interface gráfica do usuário, Texto, Aplicativo, Email

Descrição gerada automaticamente

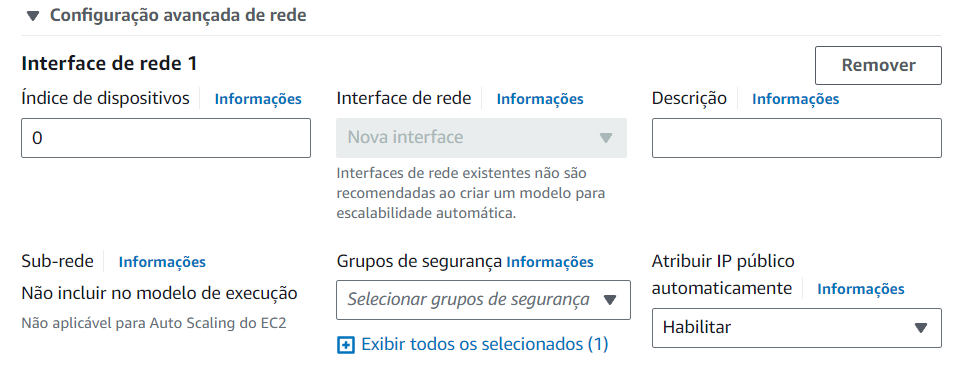
Dentro de “**Par de chaves**”, não selecione nada, vamos aumentar nossa segurança.

Dentro de “**Configurações de rede**”, selecione: Não incluir no modelo de execução

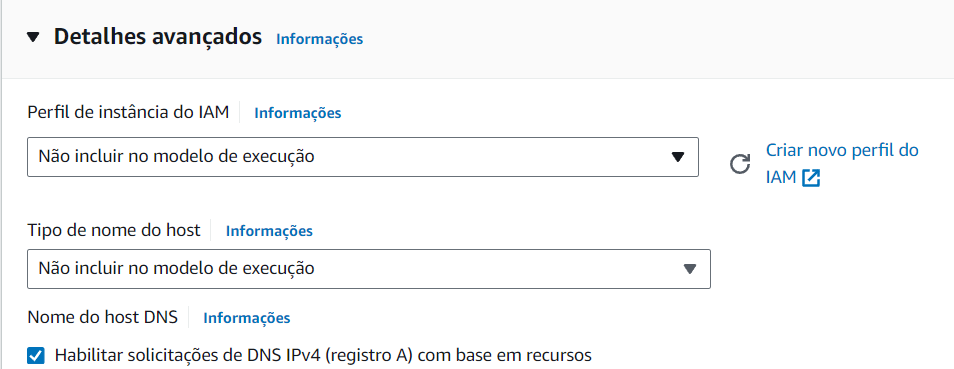


Dentro de “**Firewall**”, “**Selecionar grupo de segurança existente**”, selecione: GrupoSec-Linux.

Nas configurações avançadas de rede, em **Atribuir IP público automaticamente**, selecionamos a opção Habilitar.



Dentro de “**Detalhes avançados**”, role até “**Nome do host DNS**”, selecione: Habilitar solicitações de DNS IPv4 (registro A).



No menu do lado direito “**Resumo**”

Interface gráfica do usuário, Texto, Aplicativo

Descrição gerada automaticamente

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

**Etapa 2 – Criar Grupo do Auto Scaling**

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

Tela de computador com texto preto sobre fundo branco

Descrição gerada automaticamente

Você será redirecionado para essa página:

Interface gráfica do usuário, Texto

Descrição gerada automaticamente **Clique em ”Criar o grupo de Auto Scaling":**

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

Interface gráfica do usuário, Texto, Aplicativo, Email

Descrição gerada automaticamente

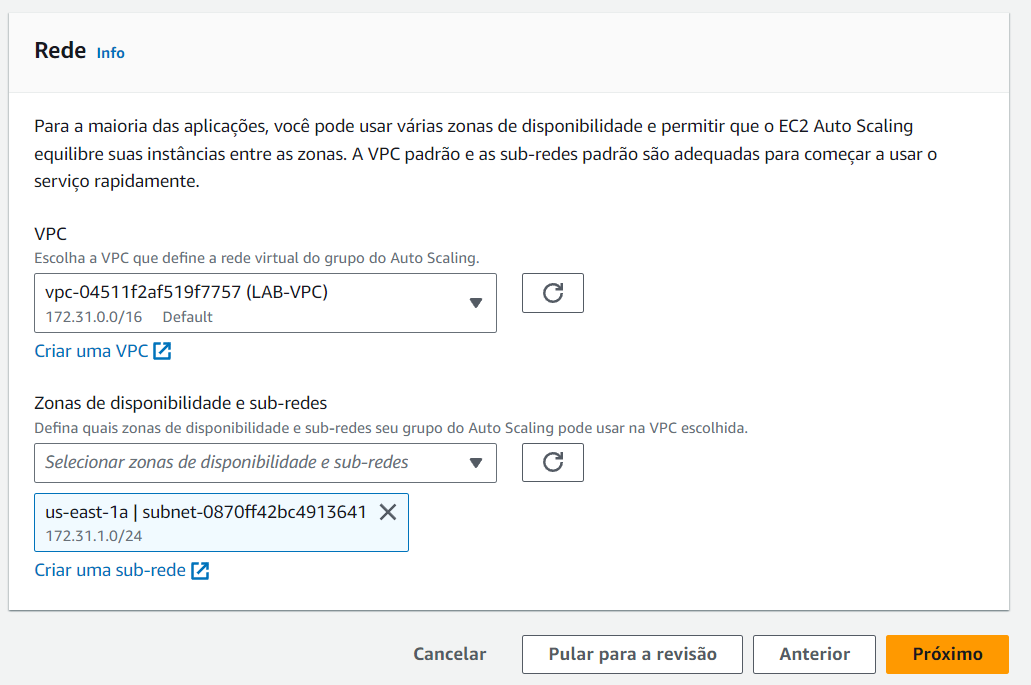
* Informe o nome: **GrupoAutoScaling**
* Escolha em “**Modelo de execução**”: **ModeloServidorWebAPP**

Interface gráfica do usuário, Texto, Aplicativo, Email

Descrição gerada automaticamente

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

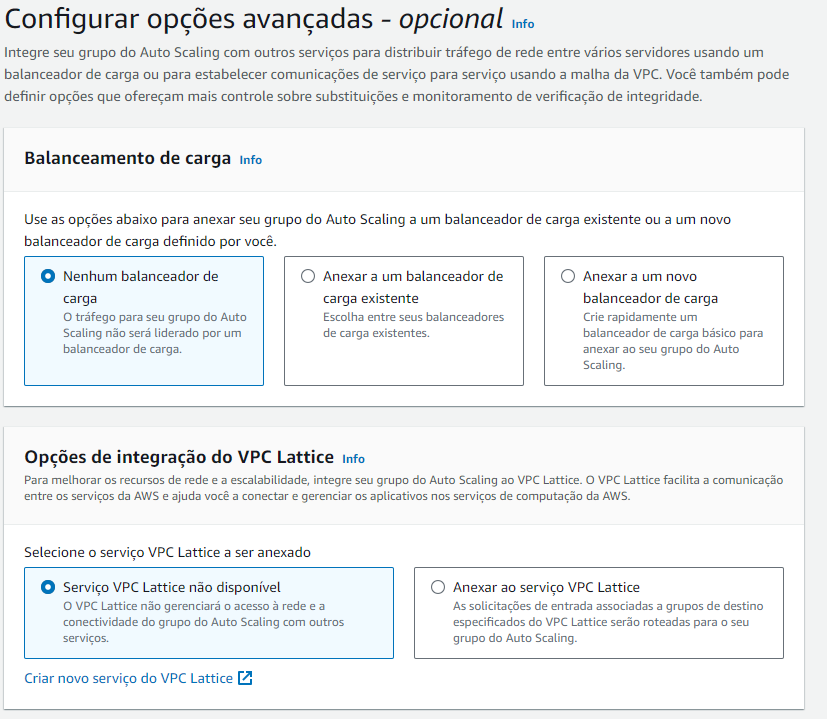
**Etapa 2: Selecione sua VPC e AZ’s:**



Usar a VPC **LAB-VPC** e a sub-rede padrão da VPC.

Clique em **Próximo**.

**Etapa 3:**

****

* Aqui, nas configurações avançadas, você pode anexar diretamente com um load balancer (balanceador de carga) existente.
* Porém, como esse conteúdo não contempla 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:**

****

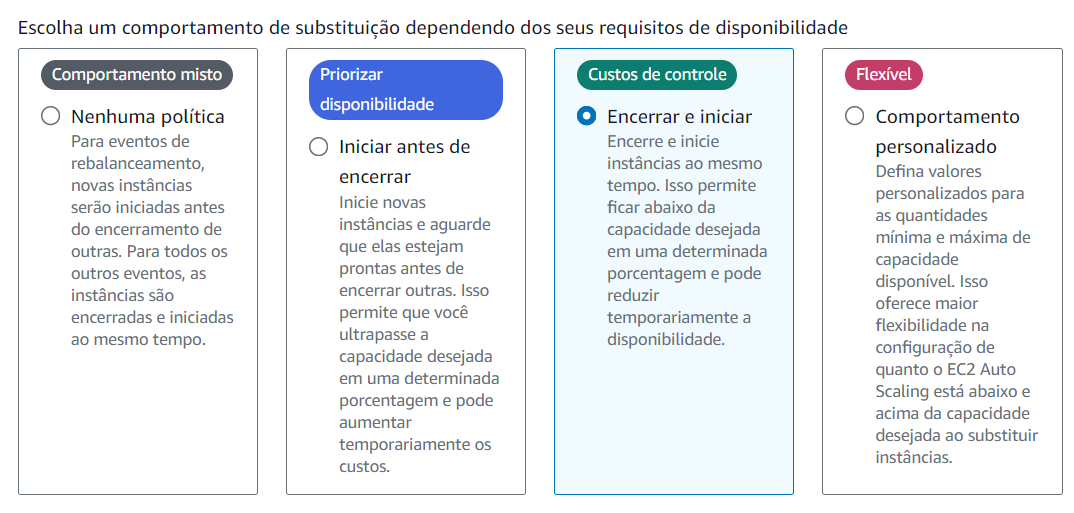
* Aqui se encontra a tela de configuração que foi representada pela imagem ilustrativa anteriormente, com as mesmas informações:

Capacidade mínima: 1

Capacidade desejada: 2

Capacidade máxima: 4

* Política de escalabilidade – opcional, manter marcado “Encerrar e Iniciar”.

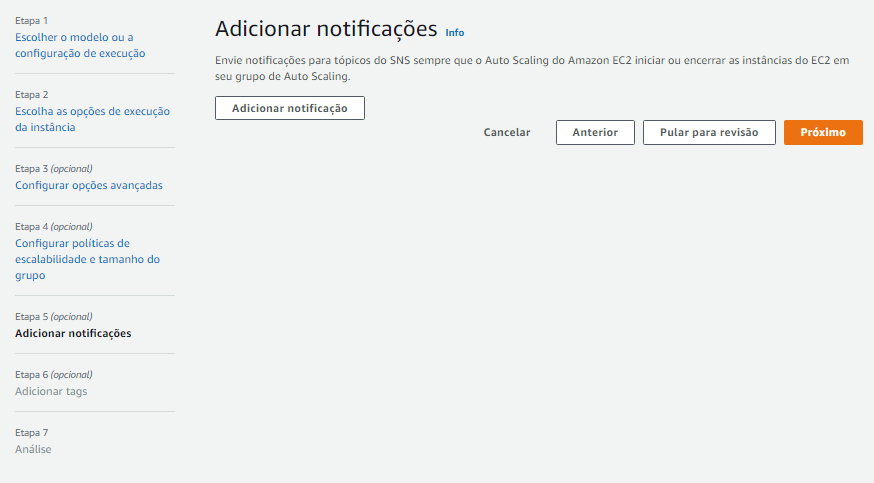


Interface gráfica do usuário, Texto, Aplicativo, Email

Descrição gerada automaticamente

Desmarcar “Habilitar a proteção contra redução de instâncias” e clique em **Próximo**.

**Etapa 5:**

****

* 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

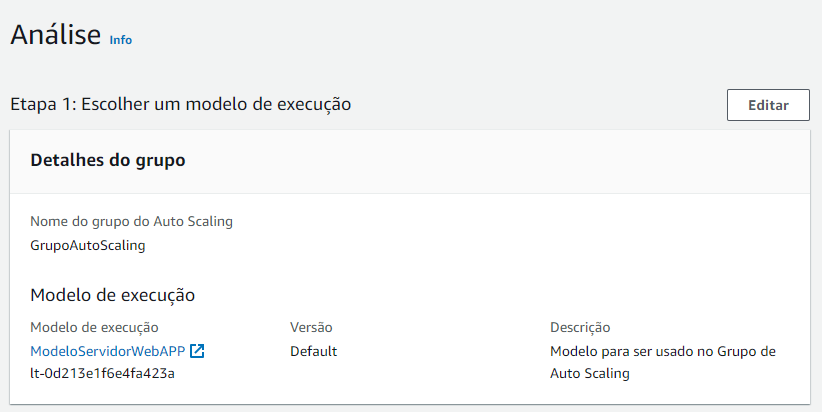
Descrição gerada automaticamente

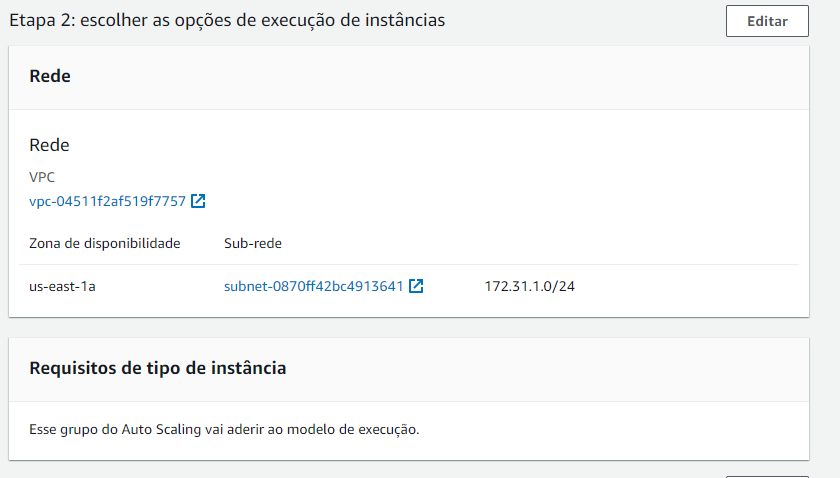
Adicionar Tag: **WebAutoScaling**

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:**

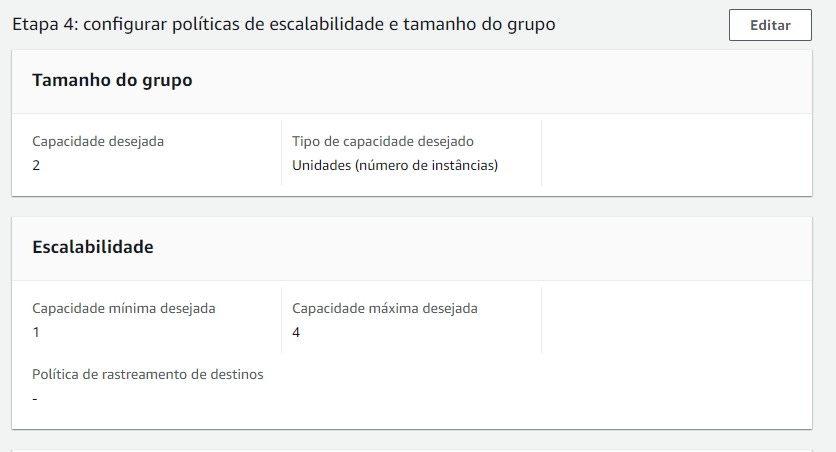
Aqui, você irá revisar suas configurações, mostrando as 6 etapas anteriores.





Interface gráfica do usuário, Texto, Aplicativo, Email

Descrição gerada automaticamente



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

Descrição gerada automaticamente

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

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

