04

Configurando políticas de escalonamento

Transcrição

[00:00] Nós configuramos o grupo de "auto-escalonamento", e foi feito também a criação da imagem customizada da nossa aplicação, que vai ter o Ubuntu, vai ter o Tomcat e a aplicação da Casa do Código. Agora precisamos falar que esse grupo de "auto-escalonamento", quando ele for subir os demais servidores por conta da demanda de acesso, que esses servidores deverão ser inicializados com essa imagem customizada. Então vamos fazer isso agora? Vamos voltar para o nosso "painel de console" da Amazon.

[00:30] Temos que pesquisar novamente o nosso serviço do EC2. Vamos aqui e colocamos EC2, e no menu lateral esquerdo vamos pesquisar pela aba "auto scaling", e temos que ir na aba "launch configuration", para falar como esse demais servidores deverão ser inicializados. Então precisa criar essa nova configuração com a imagem customizada.

[00:58] O processo lembra pouco a criação de um servidor na Amazon. Mas agora não queremos mais essas imagens bases, queremos utilizar a imagem que criamos na etapa anterior, para poder atualizar esses servidores.

[01:13] Temos que ir nessa aba "My AMIs", as minhas imagens. Então vemos que está listada as imagens customizadas na casa do código. Então eu quero que os demais servidores sejam inicializados com essa imagem customizada. Vamos selecionar essa imagem, e falamos que esses outros servidores deverão ter essas características da "T2 micro" que está dentro do uso gratuito da Amazon. Mantemos essa máquina mesmo, e colocamos "next" para ver as demais configurações que temos que fazer.

[01:45] Então aqui, temos que falar qual que é o nome desse grupo, dessa configuração que vamos levar para o nosso grupo de "auto-escalonamento". Pode ser o nome que desejar, vou colocar, por exemplo "imagem-customizada-casadocodigo", e agora isso é importante: quando for subir esses servidores, precisamos acessá-los pela internet.

[02:14] Tenho que falar que essas instâncias deverão ter um endereço IP público. A Amazon vai ter que entregar um endereço IP público para esses servidores para podermos acessar. Vamos na aba "advanced details", e depois aqui para falar para a Amazon assinalar um enderço IP público para todos esses servidores que farão parte desse grupo do "auto-escalonamento". Selecionamos aqui, essa segunda opção, para a Amazon fazer a entrega desse endereço IP público para os servidores.

[02:42] Feito isso, colocamos "next", para verificar o volume, que é o volume dentro do uso gratuito, nós podemos manter esse mesmo, e falta falar qual é o grupo de segurança que estaremos vinculando para esses servidores que farão parte do grupo de "auto-escalonamento".

[02:59] Nesse caso já criamos um grupo de segurança quando fizemos a configuração deste único servidor que está com aplicação da Casa do Código. Então em vez de criar um outro grupo de segurança, eu posso vir aqui e já selecionar o grupo de segurança existente para os servidores que tem aplicação da Casa do Código que tinha dado o nome de "SG-EC2".

[03:22] Podemos vir aqui e revisar que estamos liberando a porta "8080" e a porta "22", para poder acessar remotamente e fazer as configurações, então aqui estamos ok, depois vem em "review", para fazer uma revisão das configurações, e observamos que estamos usando a imagem customizada da Casa do Código, nós estamos aqui escolhendo a máquina "T2 micro" e só falta colocar aqui para subir essa configuração.

[03:49] Agora, temos que especificar qual é a chave que a gente vamos estar utilizando para poder acessar esses servidores. Já tínhamos criado essa chave, com o nome "chave_EC2_1". Temos essa chave, e vou falar que os servidores

que farão parte desse grupo de "auto-escalonamento", serão acessados com essa chave aqui.

[04:11] Então posso colocar que eu falo para Amazon que eu tenho essa chave salvo no meu computador, porque se não tiver essa chave, não conseguimos ser autenticados, e não conseguimos fazer as configurações desse servidores. É importante garantir que temos a posse dessa chave. Eu tenho ela aqui, e falo para Amazon que tenho, e posso falar que eu quero criar essa configuração dessa forma. Os demais servidores que vão fazer parte desse grupo de "auto-escalonamento", serão configurados dessa forma. Com essa imagem customizada, com esse grupo de segurança que nós colocamos aqui, com tudo isso que fizemos nessa etapa.

[04:49] Agora o que temos que fazer? Temos que voltar e falar que esse grupo de "auto-escalonamento" vai ter que usar essa configuração que nós acabamos de criar. Vamos em "close", e vamos voltar para o nosso menu lateral esquerdo, e agora uma vez que criamos essa configuração da imagem customizada da Casa do Código, temos que voltar aqui para o grupo do auto- escalonamento, e nesse grupo que nós havíamos criado, eu tenho que falar que temos que usar a configuração que nós acabamos de criar.

[05:20] Para isso, temos que vir nessa aba "detalhes", "details", e editar esses detalhes. Então agora a configuração que deverá ser utilizada por esse grupo de "auto-escalonamento", é o da imagem customizada na Casa do Código. Então essa configuração já vai ter tudo que fizemos, vai ter imagem customizada para casa do código, vai ter aquelas configurações do grupo de segurança que nós acabamos de fazer agora na etapa anterior no "launch configuration".

[05:47] Agora temos a opção de especificar como é que vai ser realizado esse escalonamento. Logo aqui embaixo temos essas configurações da quantidade de servidores vamos ter. Então temos que nós desejamos ter um servidor, e aqui o mínimo está zero. Vamos colocar como "sempre quer ter um servidor" e o mínimo para também ser "um", para sempre garantir que no mínimo vai ter um servidor rodando.

[06:18] Conforme a demanda de acessos aumentar, queremos escalonar a nossa aplicação para poder distribuir esses acessos nos demais servidores. Então vamos colocar que queremos ter no máximo "dois" servidores. Então no mínimo "um" e no máximo "dois", para quando tiver essa grande demanda de acessos. Tendo essa grande demanda de acessos, eu vou pedir para a Amazon criar esse segundo servidor, coloco o máximo "dois" servidores.

[06:41] O servidor que temos no momento, que está com aplicação da Casa do Código, tínhamos configurado para atuar nessa localidade "us-east-1b". Então o que podemos fazer? Para garantir sempre a disponibilidade dos nossos serviços para os usuários, o que pode acontecer como nós vimos? Se uma localidade tiver algum problema, tiver uma queda de energia, por exemplo, o nosso servidor ficaria indisponível.

[07:09] Então para garantir sempre uma disponibilidade para esses nossos usuários, vamos colar uma outra localidade para subir esse segundo servidor, para caso tenha essa grande demanda de acessos, cada servidor vai ficar numa localidade diferente, tendo uma garantia maior que caso tenha algum problema em alguma localidade teremos a outra que pode assumir.

[07:30] Vou colocar, por exemplo, que quero criar esse segundo servidor, numa localidade por exemplo "a localidade C". Então coloco esse segundo servidor nessa localidade dentro da Virgínia do Norte que é a "us-east-c".

[07:47] Com isso definimos qual a configuração que vai ser utilizada para subir esses demais servidores, e a quantidade de servidores que queremos ter. Queremos ter um, mais conforme tiver uma grande demanda de acessos, vamos colocar para ter dois. E estamos colocando cada servidor pra ficar em uma localidade diferente. Se tiver um problema em uma localidade, teremos uma outra que poderá assumir.

[08:12] Mas quando é que vou criar esse segundo servidor? Quais são as políticas que vai definir para que seja criado esse número máximo de dois servidores. Temos que fazer essa definição. Vamos salvar essa configuração que nós

fizemos agora, e vamos nessa aba "scaling policies" para falar quais são as políticas que vamos adotar para dizer que um servidor já consumiu muitos recursos e deve ir para o número máximo de dois servidores.

[08:43] Então temos que vir aqui nessa aba "scaling polices" para adicionar essa política. Então vamos colocar o nome dessa política como sendo "aumentando-servidores". E agora temos que selecionar qual que vai ser a nossa política. Eu coloco "create a scaling polices with steps", e criamos um alarme. Logo na primeira opção, temos a oportunidade de falar para qual e-mail mandar. Como não quero mandar para nenhum e-mail, eu vou não vou selecionar esse checkbox. Estamos preocupados somente nessa criação de um novo servidor.

[09:29] O que vamos fazer aqui? Sempre que tiver um consumo médio da utilização de CPU desse servidor que foram maior aqui do que, por exemplo, vou colocar um valor baixo de 0,5%, só para poder ver a criação dessa instância, porque se eu colocar um valor muito alto vai demorar para consumir esses recursos, e vai demorar para criar esse segundo servidor. Então vou colocar um valor baixo de 0,5% de utilização do servidor, para fazer essa simulação e ver se o nosso grupo de "auto-escalonamento" está trabalhando como imaginamos.

[10:12] Vamos falar que sempre que tiver um consumo médio da utilização de CPU do servidor que for maior ou igual a 0,5%, por um período de um minuto, vou falar para aumentar e escalonar para o nosso número máximo de servidores que são dois. Então, vamos colocar para criar esse alarme, e vamos fazer "take the action", fazer ação de adicionar uma Instância.

[10:34] Então com isso o que estamos fazendo? Estamos falando que caso esse servidor que eu tenho hoje ultrapasse o 0,5% de utilização da CPU, vamos adicionar uma instância chegando no número máximo que nós definimos de dois servidores. Então com isso podemos criar essa política. E agora falta o que? Falta testar esse nosso grupo de "auto-escalonamento".

[11:04] Para poder fazer esse teste e simular várias requisições de vários usuários, vamos utilizar um programa chamado ApacheBench, que vai fazer essa simulação de várias requisições de vários usuários para ver se ultrapassa esse limiar de 0,5% de utilização da CPU do servidor, e a Amazon vai criar esse segundo servidor. Então para poder utilizar o "Apache bench" no Windows, eu instalei o XAMPP, que já tem o ApacheBench para fazer esses testes. Não se preocupa que logo abaixo do vídeo, nos exercícios vai ter as configurações que você precisa fazer para poder instalar o ApacheBench.

[11:40] Vamos voltar para nossa de instâncias das nossas máquinas, dos nossos servidores, que tem a aplicação da Casa do Código, e observamos que temos esse único servidor. Então o que vou fazer? Vou vir abrir o prompt, vou voltar para acessar XAMPP, vou entrar na pasta "Apache", e vou entrar na pasta "bin". Então para poder utilizar o ApacheBench, vou colocar o número de requisição como um número alto, por exemplo, 5000 requisições.

[12:23] E vamos ter que falar qual é a URL que queremos fazer essas várias requisições. Bom, temos que fazer essas várias requisições com o único servidor que temos disponível no momento, para ver se o nosso grupo de "auto-escalonamento" está funcionando. Então, pegamos o endereço IP público desse nosso servidor, e vamos colocar em uma outra aba no "browser", teríamos a porta de comunicação do Tomcat, "8080", e a nossa aplicação da Casa do Código .

[12:48] Eu quero fazer várias requisições, nessa URL para verificar se a Amazon vai criar o segundo servidor conforme definido nas políticas de "auto-escalonamento". Então vou copiar essa URL, voltar para o "prompt", e quero fazer 5000 requisições aqui nessa URL. Vou colocar os dois aqui do lado para ver se conseguimos ver que uma segunda instância vai ser criada. Então temos essa única instância, eu vou começar rodar, e vamos ver se essa segunda instância será criada.

[13:19] Vou voltar para o "prompt", e vamos instanciar o nosso teste dessas 5.000 requisições. Então vamos esperar alguns segundos, que vai demorar um pouquinho, e faz a análise para ver se a nossa segunda instância foi criada, eu

vou pausar o vídeo e volto daqui a pouco.

[13:44] Demorou um pouco para fazer o teste com 5000 requisições, eu voltei o teste para o ApacheBench, e coloquei 2000 requisições. Então foram feitas várias requisições para esse nosso servidor. Teoricamente, deve ter ultrapassado aquele limite de 0,5% do consumo de CPU desse servidor que temos na Amazon.

[14:06] Eu vou voltar aqui e colocar uma nova atualização para ver se esse segundo servidor foi criado. E o que temos? Agora foi criado o que? Esse segundo servidor na Amazon. Em qual localidade? Na "us-east-1c" que era a segunda localidade no norte da Virginia, que pedimos para que fosse criado esse servidor.

[14:29] Agora temos dois servidores, e esses dois servidores devem estar configurados com aplicação da Casa do Código. Então vemos que está no período em que ele está inicializando esse servidor. Vou parar o vídeo para esperar esse servidor terminar de ser inicializado, para ver se conseguimos acessar a aplicação da Casa do Código nesse segundo servidor que subimos com o grupo do "auto-escalonamento". Vou pausar o vídeo, e volto daqui a pouco.

[15:01] Terminou a inicialização desse segundo servidor, e agora vamos ver se temos a aplicação da Casa do Código com o Tomcat e tudo funcionando nesse segundo servidor. Pedimos para todos os demais servidores desse grupo do "auto-escalonamento" utilizarem aquela configuração customizada com a imagem que tem a aplicação da Casa do Código.

[15:21] Vamos nesse segundo servidor que colocamos na localidade "us-east-1c" e vamos na aba "description" e vamos pegar o endereço IP público desse segundo servidor. Copiamos e colamos esse IP público do servidor, e ":8080" para acessar a porta de comunicação do Tomcat que já deve estar instalado nesse servidor, e a aplicação da Casa do Código que também já deve estar configurada.

[15:48] Vou colocar "Enter" e maravilha, temos os dois servidores com a aplicação da Casa do Código. Vamos só confirmar que o primeiro está funcionando? Vou voltar nos nossos servidores, voltar para o nosso primeiro servidor, pegar seu endereço IP público e colocar em outra aba. Agora possuímos estes dois servidores com a aplicação da Casa do Código funcionando. Agora temos que configurar esse balanceamento para esses dois servidores, para essas duas instâncias da aplicação da Casa do Código. Vamos fazer isto na próxima etapa.