01

Qual é a quantidade certa de máquinas que precisamos?

Transcrição

Agora que você já sabe como escalar sua aplicação e como balancear a carga entre diversas máquinas, precisamos responder algumas questões importantes:

Qual é o melhor momento para escalar nossa aplicação?

Qual o melhor momento para criar uma nova máquina **para ajudar** uma máquina anterior que está começando a sobrecarregar?

Qual o momento ideal de **desligar** uma máquina que não está sendo muito utilizada, para que não precisemos mais ficar pagando por ela ?

Analisando os dados de acesso

Essas questões podem ser resolvidas de diversas formas, uma delas é **analisando os dados de acesso da sua aplicação** e chegando a conclusão que, por exemplo, a sua aplicação tem um pico de usuários ao meio dia , e quase nenhum usuário de madrugada, logo faz muito sentido aumentar a quantidade de máquinas no meio da tarde e no início da madrugada reduzir o número de máquinas para diminuir os custos.

Este tipo de análise é bom quando temos uma aplicação tem que um fluxo de acesso previsível. Agora quando não esta possibilidade, precisamos recorrer a outras táticas e recursos para nos ajudar.

O que veremos neste capítulo é um serviço da Amazon chamado de **Auto Scaling**, que leva em consideração determinadas métricas da sua aplicação para criar ou destruir novas máquinas.