coursera

Menu

Implementações de modelos do TensorFlow com escalonamento automático com TF Serving e Kubernetes

Traduzido automaticamente do Inglês

Este item inclui conteúdo que ainda não foi traduzido para o idioma de sua preferência.

Nesta tarefa, o senhor usará o TensorFlow Serving e o Google Cloud Kubernetes Engine (GKE) para configurar um sistema de serviço de alto desempenho e autoescalável para modelos do TensorFlow. Em termos mais concretos, o senhor irá:

1. Criar um cluster do GKE e implantar um modelo.

2. Baixar os arquivos do modelo em um bucket de armazenamento.

3. Crie o Kubernetes ConfigMap que aponte para o local do modelo no bucket de armazenamento.

4. Crie a implantação do Kubernetes usando uma imagem padrão do TensorFlow Serving do Docker Hub.

5. Crie o Kubernetes Service para expor a implantação por meio de um balanceador de carga.

6. Configurar o Horizontal Pod Autoscaler.

7. Testar o modelo.

Se tiver alguma dúvida sobre as tarefas deste curso, o senhor pode pedir ajuda em nossa comunidade. Se o senhor ainda não o fez, clique aqui e siga as instruções para que o senhor possa participar! As perguntas frequentes sobre todas as tarefas estão consolidadas neste tópico.

You can do it

This assignment typically takes around **1 hour, 47 minutes** to complete. Though challenging, it's a great way to build and apply your new skills.

Isso foi útil? Sim Não

Este course utiliza um aplicativo de terceiros, Implementações de modelos do TensorFlow com