Então, resumo da engenharia de dados no Google Cloud Platform.

Nós cobrimos muito terreno neste curso.

Analisamos a análise de dados sem servidor,

nós olhamos como fazer dados não estruturados,

nós olhamos como construir modelos de aprendizado de máquina,

e analisamos como fazer o processamento de fluxo resiliente.

Conversamos sobre uma arquitetura de exemplo no GCP que começa com a obtenção

processamento de dados em tempo real com o pub / sub, mas também é capaz de processar dados em lote,

dados históricos, mantidos no armazenamento em nuvem e para fazer ambos,

processamento em tempo real e

processamento de dados históricos usando o mesmo código no Cloud Dataflow.

Então, se você tem um pipeline de streaming ou tem

um fluxo de dados em lote que o Dataflow pode gravar

BigQuery e BigQuery essencialmente

torna-se um lugar a partir do qual você realiza todas as suas análises.

Então, você pode se conectar ao Datalab,

você pode se conectar ao Tableau, Data Studio,

Qlik, Looker, etc, ou você pode enviá-lo para o Planilhas Google,

você pode compartilhar com seus colegas de trabalho.

Então, essa é essencialmente a arquitetura típica que você executa no GCP.

Então, o Dataflow é uma parte essencial dessa arquitetura,

ele ingerir, faz transformação e carrega,

pode fazer filtragem

pode fazer agrupamento,

ele pode fazer janelas e janelas, é claro

muito importante se você está fazendo dados ilimitados,

se você estiver fazendo dados de fluxo.

BigQuery, que é outra parte da arquitetura de referência

é onde você obtém análises em tempo real de conjuntos de dados muito grandes,

não funciona, então você só paga pelo que usa

dá-lhe durabilidade

Ele dá a replicação que lhe dá um armazenamento muito barato sobre o mesmo custo que

armazenamento em nuvem e dá-lhe

registros de auditoria imutáveis ​​para que você saiba quem acessa os dados quando,

e o mais importante é que, porque há apenas um BigQuery que é global,

Ele permite que você misture diferentes conjuntos de dados.

Então, nós olhamos isso um pouco.

Conseguimos combinar dados de duas fontes diferentes.

Pudemos analisar os dados meteorológicos em conjunto com os dados de táxi, por exemplo,

e somos capazes de obter insights basicamente analisando vários conjuntos de dados.

Esse é o principal benefício do uso do BigQuery.

Então nós também analisamos o Dataproc.

O Dataproc é uma forma de executar muitas das ferramentas do ecossistema do Hadoop,

Porco, colmeia, faísca, etc.

Você pode executá-los no GCP.

Agora, normalmente, se você está pensando em executar um cluster,

se você está fazendo isso no local você está realmente preocupado com uma variedade de coisas,

monitorando o ajuste de desempenho do cluster,

Descobrir se você está usando o cluster o tempo todo.

Agora, descobrindo como implantar e configurar o cluster,

preocupando-se com a confiabilidade do cluster,

o que acontece se uma máquina cair

o que acontece se você tiver que processar mais dados que puder.

Mas, idealmente, o que você quer focar é apenas insights e análises,

e é isso que o Dataproc ajuda você a fazer.

Em seguida, também analisamos as APIs de aprendizado de máquina como uma maneira de compreender os dados não estruturados.

Nós dissemos que se você tem imagens ou áudio ou vídeo ou texto de forma livre você pode

use as APIs de aprendizado de máquina para essencialmente extrair entidades,

rótulos, pessoas, eventos, etc,

para que você possa entender e, em seguida, pegar esses dados não estruturados, obter entidades,

obter sentimento, obter rótulos e, em seguida, se necessário,

aprendizado de máquina nesses dados.

Em seguida, analisamos como você cria modelos eficazes de aprendizado de máquina.

Nós dissemos para construir

modelos de aprendizado de máquina eficazes que você precisa para coletar todos os dados que

você pode então coletar os dados para que

Você pode fazer análises sobre os dados é extremamente importante.

Então, mas quando você tem os dados, então você quer basicamente trazer

visão humana sobre os dados usando bons recursos,

e depois também analisamos como você pode aproveitar as melhorias modernas

arquiteturas de rede neural para obter

a melhor precisão possível uma vez que você

Decidiu que é assim que você vai construir o seu ML.

Nós dissemos que quando você está fazendo aprendizado de máquina

a melhoria da precisão que você vai

obter é que você vai conseguir através do trabalho duro,

você vai conseguir através da engenharia de recursos,

através de ajuste de hyperparameter e através de muitos dados.

O que o Cloud ML Engine oferece é um ambiente no qual você pode fazer todas essas coisas.

Por fim, analisamos como fazer o processamento de fluxo resiliente no GCP.

Nós dissemos que era importante ser capaz de ingerir

volumes variáveis ​​porque você pode ter picos nos seus dados

é importante ser capaz de lidar com a latência porque a latência é um fato da vida,

e nós queremos ser capazes de derivar

insights em tempo real a partir dos dados, mesmo que os dados estejam sendo transmitidos.

Para fazer isso, olhamos

uma arquitetura que consistia em ingerir os dados com pub / sub,

processar os dados em fluxo usando o Cloud Dataflow e

streaming no BigQuery para armazenamento durável e análise interativa.

Então, quais são alguns dos nossos próximos passos?

Agora que você aprendeu um pouco sobre o Google Cloud Platform,

Encorajo-vos a ver a certificação de engenharia de dados certificada da Google.

O que você aprendeu neste curso é suficiente

mas sugiro fortemente que você desenvolva a prática de usar as tecnologias.

A certificação é mais sobre o quão bem você pode

usar as tecnologias do que saber algo em um nível teórico.

Então, o que você aprendeu

neste conjunto de cursos é necessário poder processar a certificação, não é suficiente.

Você quer basicamente desenvolver uma quantidade de conhecimento do mundo real,

Por isso, incentivo você a criar algo sobre o GCP.

Não há substituto para fazer algo sozinho.

Use todos os laboratórios e o código que temos como exemplo para você começar

mas tente fazer algo você mesmo.

Finalmente alguns recursos, existe

um grande blog de dados e aprendizado de máquina para cloud.google.com/blog/big-data/,

encorajo vivamente a segui-lo,

você tende a ter de três a cinco novas postagens a cada semana.

Muito interessante.

Eles cobrem a gama de coisas,

tudo, desde anúncios de novos produtos até aqui.

use um produto e aqui está um exemplo de alguém que usa este produto de maneira impactante.

Muitas vezes os posts são acompanhados por amostras do GitHub.

Então, você definitivamente quer seguir esse blog.

O segundo blog que vale a pena seguir é

o blog geral do Google Cloud Platform e

Isso essencialmente contém o anúncio do produto em toda a plataforma.

Então, finalmente plug sem vergonha,

há um novo livro Data Science no Google Cloud Platform escrito por

eu e isso está disponível atualmente enquanto estou gravando isso.

Está disponível no lançamento antecipado em O'Reilly,

provavelmente estará no papel em breve,

e, novamente, abrange muito do mesmo fundamento que temos

feito neste curso, mas ele faz do ponto de vista de uma aplicação.

Lembre-se eu disse que a melhor maneira de aprender

a plataforma é tentar criar um aplicativo você mesmo,

mas este livro segue minha jornada à medida que aprendi o GCP criando um aplicativo sozinho.

Então, muito obrigado a todos vocês.