Professional Data Engineer certification

Objectifs

- Concevoir et construire des systèmes de traitement de données sur Google Cloud Platform
- Traiter des lots de données de diffusion en implémentant des pipelines de données à mise à l'échelle automatique sur Cloud Dataflow
- Dériver les perspectives commerciales d'extrêmement grandes ensembles de données utilisant Google BigQuery
- Former, évaluer et prédire en utilisant des modèles d'apprentissage automatique en utilisant Tensorflow et Cloud ML
- Tirer parti des données non structurées à l'aide des API Spark et ML sur Cloud Dataproc
- Activer les informations instantanées à partir des données en continu

Prérequis

- Avoir suivi la formation Google Cloud Platform sur les fondamentaux du Big Data et du Machine Learning (GO8325) ou posséder une expérience équivalente
- Connaître un langage de requête commun tel que SQL
- Avoir une expérience de la modélisation, de l'extraction, de la transformation et du chargement de données
- Savoir développer des applications en utilisant un langage de programmation commun tel que Python
- Etre familiarisé avec le machine learning et/ ou les statistiques

Programme

Analyse de données sans serveur avec BigQuery

- Qu'est-ce que BigQuery
- Capacités avancées
- Performance et prix

Canaux de données Autoscaling sans serveur avec Dataflow

Premiers pas avec le Machine Learning

- Qu'est-ce que le machine Learning (ML)
- ML efficace: concepts, types
- Évaluation de ML
- Ensembles de données ML: généralisation

Construire des modèles ML avec Tensorflow

- Premiers pas avec TensorFlow
- Graphes et boucles TensorFlow + lab

- Surveillance de la formation au ML

Mise à l'échelle des modèles ML avec CloudML

- Pourquoi Cloud ML?
- Emballer un modèle TensorFlow
- Formation de bout en bout

Ingénierie des caractéristiques

- Créer de bonnes fonctionnalités
- Transformer les entrées
- Caractéristiques synthétiques
- Pré-traitement avec Cloud ML

Architectures ML

- Large et profond
- L'analyse d'image
- Embeddings et séquences
- Systèmes de recommandation

Présentation de Google Cloud Dataproc

- Présentation de Google Cloud Dataproc
- Créer et gérer des clusters
- Définition des noeuds maître et de travail
- Tirer parti des types de machines personnalisées et des nœuds de travail préemptifs
- Création de clusters avec la console Web
- Clusters de script avec l'interface de ligne de commande
- Utilisation de l'API REST Dataproc
- Tarification Dataproc
- Mise à l'échelle et suppression des clusters

Exécution de Jobs Dataproc

- Contrôle des versions d'application
- Soumettre des emplois
- Accès à HDFS et GCS
- Hadoop
- Spark et PySparkCochon et ruche
- Journalisation et surveillance des travaux
- Accès aux nœuds maîtres et travailleurs avec SSH
- Travailler avec PySpark REPL (interpréteur de ligne de commande)

Intégration de Dataproc à Google Cloud Platform

- Actions d'initialisation
- Programmation des ordinateurs portables Jupyter / Datalab
- Accéder à Google Cloud Storage

- Exploiter les données relationnelles avec Google Cloud SQL
- Lire et écrire des données en streaming avec Google BigTable
- Interrogation de données à partir de Google BigQuery
- Création d'appels API Google à partir de blocs-notes

Comprendre les données non structurées avec les API d'apprentissage machine de Google

- API Google Machine Learning
- Cas d'utilisation de ML communs
- API Vision
- API de langage naturelTraduire
- API Speech

Nécessité d'une analyse en continu en temps réel

- Qu'est-ce que Streaming Analytics?
- Cas d'utilisation
- Batch vs Streaming (en temps réel)
- Terminologies connexes
- Les produits GCP qui aident à construire pour la haute disponibilité, la résilience, le haut débit, les analyses en temps réel (revue de Pub / Sub et Dataflow)

Architecture des pipelines en streaming

- Architectures de streaming et considérations
- Choisir les bons composants Lumières
- Agrégation en continu
- Evénements, déclencheurs

Diffuser des données et des événements dans PubSub

- Sujets et abonnements
- Publication d'événements dans Pub / Sub
- Options d'abonnement: Push vs Pull
- Alertes

Construire un pipeline de traitement de flux

- Pipelines, PCollections et Transforms
- Windows, événements et déclencheurs
- Statistiques d'agrégation
- Analyse en continu avec BigQuery
- Alertes à faible volume

Haut débit et faible latence avec Bigtable

- Considérations de latence
- Qu'est-ce que Bigtable
- Concevoir des clés de ligne

- Considérations de performance

Haut débit et faible latence avec BigTable

- Qu'est-ce que Google Data Studio?
- Des données aux décisions