

Data Engineering on Google Cloud Platform

CPB400 - Version: 1

 4 days Course

Description:

This four-day instructor-led class provides participants a hands-on introduction to designing and building data processing systems on Google Cloud Platform. Through a combination of presentations, demos, and hand-on labs, participants will learn how to design data processing systems, build end-to-end data pipelines, analyze data and carry out machine learning. The course covers structured, unstructured, and streaming data.

Intended audience:

Extracting, Loading, Transforming, cleaning, and validating data
Designing pipelines and architectures for data processing
Creating and maintaining machine learning and statistical models
Querying datasets, visualizing query results and creating reports

Prerequisites:

Completed Google Cloud Fundamentals: Big Data & Machine Learning course OR have equivalent experience
Basic proficiency with common query language such as SQL
Experience with data modeling, extract, transform, load activities
Developing applications using a common programming language such as Python
Familiarity with Machine Learning and/or statistics

Objectives:

Design and build data processing systems on Google Cloud Platform
Process batch and streaming data by implementing autoscaling data pipelines on Cloud Dataflow

Derive business insights from extremely large datasets using Google BigQuery
Train, evaluate and predict using machine learning models using Tensorflow and
Cloud ML

Leverage unstructured data using Spark and ML APIs on Cloud Dataproc

Enable instant insights from streaming data

Topics:

Module 1: Google Cloud Dataproc Overview

- Creating and managing clusters.
- Leveraging custom machine types and preemptible worker nodes.
- Scaling and deleting Clusters.
- Lab: Creating Hadoop Clusters with Google Cloud Dataproc.

Module 2: Running Dataproc Jobs

- Running Pig and Hive jobs.
- Separation of storage and compute.
- Lab: Running Hadoop and Spark Jobs with Dataproc.
- Lab: Submit and monitor jobs.

Module 3: Integrating Dataproc with Google Cloud Platform

- Customize cluster with initialization actions.
- BigQuery Support.
- Lab: Leveraging Google Cloud Platform Services.

Module 4: Making Sense of Unstructured Data with Google's Machine Learning APIs

- Google's Machine Learning APIs.
- Common ML Use Cases.
- Invoking ML APIs.
- Lab: Adding Machine Learning Capabilities to Big Data Analysis.

Module 5: Serverless data analysis with BigQuery

- What is BigQuery.
- Queries and Functions.
- Lab: Writing queries in BigQuery.
- Loading data into BigQuery.
- Exporting data from BigQuery.
- Lab: Loading and exporting data.
- Nested and repeated fields.
- Querying multiple tables.
- Lab: Complex queries.
- Performance and pricing.

Module 6: Serverless, autoscaling data pipelines with Dataflow

- The Beam programming model.
- Data pipelines in Beam Python.
- Data pipelines in Beam Java.
- Lab: Writing a Dataflow pipeline.
- Scalable Big Data processing using Beam.
- Lab: MapReduce in Dataflow.
- Incorporating additional data.
- Lab: Side inputs.
- Handling stream data.
- GCP Reference architecture.

Module 7: Getting started with Machine Learning

- What is machine learning (ML).
- Effective ML: concepts, types.
- ML datasets: generalization.
- Lab: Explore and create ML datasets.

Module 8: Building ML models with Tensorflow

- Getting started with TensorFlow.
- Lab: Using tf.learn.
- TensorFlow graphs and loops + lab.
- Lab: Using low-level TensorFlow + early stopping.
- Monitoring ML training.
- Lab: Charts and graphs of TensorFlow training

Module 9: Scaling ML models with CloudML

- Why Cloud ML?
- Packaging up a TensorFlow model.
- End-to-end training.
- Lab: Run a ML model locally and on cloud.

Module 10: Feature Engineering

- Creating good features.
- Transforming inputs.
- Synthetic features.
- Preprocessing with Cloud ML.
- Lab: Feature engineering.

Module 11: Architecture of streaming analytics pipelines

- Stream data processing: Challenges.
- Handling variable data volumes.
- Dealing with unordered/late data.
- Lab: Designing streaming pipeline.

Module 12: Ingesting Variable Volumes

- What is Cloud Pub/Sub?
- How it works: Topics and Subscriptions.
- Lab: Simulator.

Module 13: Implementing streaming pipelines

- Challenges in stream processing.
- Handle late data: watermarks, triggers, accumulation.
- Lab: Stream data processing pipeline for live traffic data.

Module 14: Streaming analytics and dashboards

- Streaming analytics: from data to decisions.
- Querying streaming data with BigQuery.
- What is Google Data Studio?
- Lab: build a real-time dashboard to visualize processed data.

Module 15: High throughput and low-latency with Bigtable

- What is Cloud Spanner?
- Designing Bigtable schema.
- Ingesting into Bigtable.
- Lab: streaming into Bigtable.