Google Cloud

Welcome!

Data Engineering on Google Cloud Platform

30-40 minutes. Skip student intros for large classes.

Logistics

Parking

Facilities

Food

Etiquette

Silence your phone

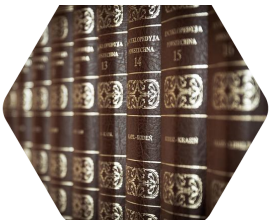
Don't record this class

Ask questions

Notes:

In a virtual class (Hangouts on Air), use the Q&A feature to ask questions. The instructor will *verbally* answer those questions.

Course Materials

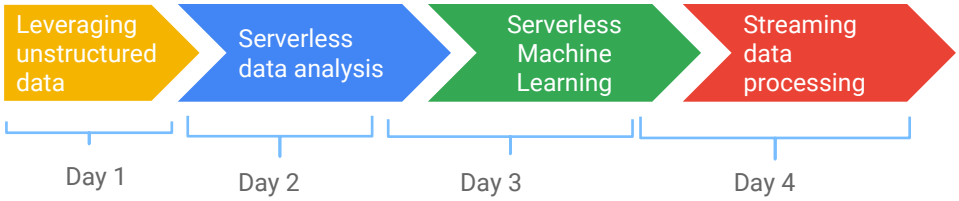


Modules
Labs

1. Log into Qwiklabs environment with the same email address that you used to register for this class. [Sign up for Qwiklabs if necessary]
2. Select “Data Engineering on Google Cloud Platform (v1.1)” class from the drop-down list
3. Modules are available in pdf format
   Available 2 years after the end of this class
4. Lab code is on GitHub:
   http://github.com/GoogleCloudPlatform/training-data-analyst

What is available in Qwiklabs after this class will reflect the most recent content.

Agenda



The ML content is approximately 1.5 days from end of Day 2 to start of Day 4. Timelines are approximate, of course.

Leveraging unstructured data

Module	Description
1. Introduction to Cloud Dataproc	Cloud Dataproc automation helps you create clusters quickly, manage them easily, and save money by turning clusters off when you don't need them.
2. Running Dataproc jobs	Run Pig And Spark jobs on Dataproc Cluster
3. Leveraging Google Cloud Platform	Integrate GCP services
4. Analyzing unstructured data	Use ML APIs like Speech, Vision, NL

Notes:

¾ day approximately

Serverless data analysis

Module	Description	Topics
5. Serverless SQL data analysis	No-ops data warehousing and analytics	• Queries • Functions • Load/Export • Nested, repeated fields • Windows • UDFs
6. Autoscaling data processing pipelines	No-ops data pipelines for reliable, scalable data processing	• Pipeline concepts • MapReduce • Side inputs • Streaming

Notes:

½ day each

Serverless Machine Learning

Module	Description
7. Getting started with Machine Learning	Uses, Challenges, Scope. Learn to sample dataset and create training, validation, and testing datasets for local development of TensorFlow models
8. Building ML models with TensorFlow	Linear Regression model in TensorFlow (train, evaluate, predict)
9. Scaling TF models with Cloud ML Engine	Packaging TensorFlow model to run local or in cloud
10. Improving ML through feature engineering	Improve ML model using feature engineering

Notes:

1½ days: these are long chapters.

Building resilient streaming systems

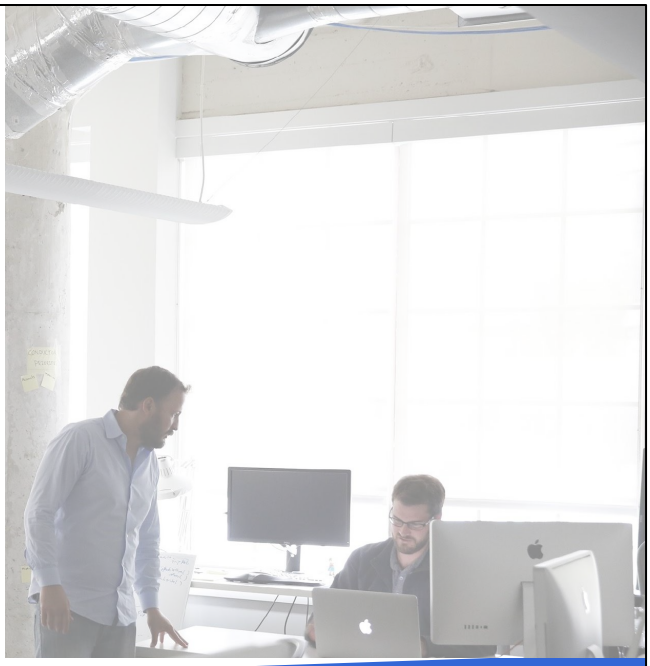
Module	Description
11. Architecture of streaming analytics pipelines	Challenges with stream processing and dealing with late, out-of-order data
12: Ingesting variable volumes	Global messaging infrastructure connecting applications and services
13: Implementing streaming pipelines	Data pipelines for streaming data processing
14: Streaming analytics and dashboards	Deriving insights, running queries against streaming data
15: Handling throughput and latency requirements	Designing for different types of resilience

Notes:

½ day each

Introductions

- Your instructor
  - Organization
  - Background
  - Course goals
- You
  - Name
  - Organization
  - Job role
  - Course goals




cloud.google.com
Image by Connie Zhou