[sfeir] Institute

Google Cloud | GCP100DE

Introduction to Data Engineering on Google Cloud

The best course to start your Data Engineering journey on Google Cloud.

1 jour / 7h

Course overview

In this course, you learn about data engineering on Google Cloud, the roles and responsibilities of data engineers, and how those map to offerings provided by Google Cloud. You also learn about ways to address data engineering challenges.

Learning outcomes

- Understand the role of a data engineer.
- Identify data engineering tasks and core components used on Google Cloud.
- Understand how to create and deploy data pipelines of varying patterns on Google Cloud.
- Identify and utilize various automation techniques on Google Cloud.

Target audience

- Data engineers
- Database administrators
- System administrators

Prerequisites

- Prior Google Cloud experience at the fundamental level using Cloud Shell and accessing products from the Google Cloud console.
- Basic proficiency with a common query language such as SQL.
- Experience with data modeling and ETL (extract, transform, load) activities.
- Experience developing applications using a common programming language such as Python

Course Outline

Module 01: Data Engineering Tasks and Components

Topics:

- The role of a data engineer
- Data sources versus data sinks
- Data formats
- Storage solution options on Google Cloud
- Metadata management options on Google Cloud
- Sharing datasets using Analytics Hub

Objectives:

- Explain the role of a data engineer.
- Understand the differences between a data source and a data sink.
- Explain the different types of data formats.
- Explain the storage solution options on Google Cloud.
- Learn about the metadata management options on Google Cloud.
- Understand how to share datasets with ease using Analytics Hub.
- Understand how to load data into BigQuery using the Google Cloud console or the gcloud CLI.

Activities:

- Lab: Loading Data into BigQuery
- Quiz

Module 02: Data Replication and Migration

Topics:

- Replication and migration architecture
- The gcloud command-line tool
- Moving datasets
- Datastream

Objectives:

- Explain the baseline Google Cloud data replication and migration architecture.
- Understand the options and use cases for the gcloud command-line tool.
- Explain the functionality and use cases for Storage Transfer Service.
- Explain the functionality and use cases for Transfer Appliance.
- Understand the features and deployment of Datastream.

Activities:

- Lab: Datastream: PostgreSQL Replication to BigQuery (optional for ILT)
- Quiz

Module 03: The Extract and Load Data Pipeline Pattern

Topics:

- Extract and load architecture
- The bq command-line tool

Objectives:

- Explain the baseline extract and load architecture diagram.
- Understand the options of the bq command-line tool.
- Explain the functionality and use cases for BigQuery Data Transfer Service.
- Explain the functionality and use cases for BigLake as a non-extract-load pattern

Activities:

- Lab: BigLake: Qwik Start
- Quiz

Module 04: The Extract, Load, and Transform Data Pipeline Pattern

Topics:

- Extract, load, and transform (ELT) architecture
- SQL scripting and scheduling with BigQuery
- Dataform

Objectives:

- Explain the baseline extract, load, and transform architecture diagram.
- Understand a common ELT pipeline on Google Cloud.
- Learn about BigQuery's SQL scripting and scheduling capabilities.
- Explain the functionality and use cases for Dataform.

Activities:

- Lab: Create and Execute a SQL Workflow in Dataform
- Quiz

Module 05: The Extract, Transform, and Load Data Pipeline Pattern

Topics:

- Extract, transform, and load (ETL) architecture
- Google Cloud GUI tools for ETL data pipelines
- Batch data processing using Dataproc
- Streaming data processing options
- Bigtable and data pipelines

Objectives:

- Explain the baseline extract, transform, and load architecture diagram.
- Learn about the GUI tools on Google Cloud used for ETL data pipelines.
- Explain batch data processing using Dataproc.
- Learn how to use Dataproc Serverless for Spark for ETL.
- Explain streaming data processing options.
- Explain the role Bigtable plays in data pipelines.

Activities:

- Lab: Use Dataproc Serverless for Spark to Load BigQuery (optional for ILT)
- Lab: Creating a Streaming Data Pipeline for a Real-Time Dashboard with Dataflow
- Quiz

Module 06: Automation Techniques

Topics:

- Automation patterns and options for pipelines
- Cloud Scheduler and Workflows
- Cloud Composer
- Cloud Run Functions
- Eventarc

Objectives:

- Explain the automation patterns and options available for pipelines.
- Learn about Cloud Scheduler and Workflows.
- Learn about Cloud Composer.
- Learn about Cloud Run functions.
- Explain the functionality and automation use cases for Eventarc.

Activities:

- Lab: Use Cloud Run Functions to Load BigQuery (optional for ILT)
- Quiz