Google Cloud

# Data Engineering on Google Cloud

Get hands-on experience with designing and building data processing systems on Google Cloud. This course uses lectures, demos, and hands-on labs to show you how to design data processing systems, build end-to-end data pipelines, analyze data, and implement machine learning. This course covers structured, unstructured, and streaming data.

**DURATION**
4 days

**LEVEL**
Intermediate

**FORMAT**
Instructor led
On-demand

## What you'll learn

- Design and build data processing systems on Google Cloud.
- Process batch and streaming data by implementing autoscaling data pipelines on Dataflow.
- Derive business insights from extremely large datasets using BigQuery.
- Leverage unstructured data using Spark and ML APIs on Dataproc.
- Enable instant insights from streaming data.
- Understand ML APIs and BigQuery ML, and learn to use AutoML to create powerful models without coding.

**Overview**

18 Modules · 143 Videos · 24 Labs · 21 Classrom activities

**Who this course is for**

This class is intended for developers who are responsible for:

- Extracting, loading, transforming, cleaning, and validating data.
- Designing pipelines and architectures for data processing.
- Integrating analytics and machine learning capabilities into data pipelines.
- Querying datasets, visualizing query results, and creating reports.

**Products**

- BigQuery
- Cloud Bigtable
- Cloud Storage
- Cloud SQL
- Cloud Spanner
- Dataproc
- Dataflow
- Cloud Data Fusion
- Cloud Composer
- Pub/Sub
- Vertex AI
- Cloud ML APIs

**Prerequisite**

To benefit from this course, participants should have completed “Google Cloud Big Data and Machine Learning Fundamentals” or have equivalent experience.

Participant should also have:

- Basic proficiency with a common query language such as SQL.
- Experience with data modeling and ETL (extract, transform, load) activities.
- Experience with developing applications using a common programming language such as Python.
- Familiarity with machine learning and/or statistics.

## Module 01 Introduction to Data Engineering

**Topics**

- Explore the role of a data engineer
- Analyze data engineering challenges
- Introduction to BigQuery
- Data lakes and data warehouses
- Transactional databases versus data warehouses
- Partner effectively with other data teams

- Manage data access and governance
- Build production-ready pipelines
- Review Google Cloud customer case study

**Objectives**

- Understand the role of a data engineer
- Discuss benefits of doing data engineering in the cloud
- Discuss challenges of data engineering practice and how building data pipelines in the cloud helps to address these
- Review and understand the purpose of a data lake versus a data warehouse, and when to use which

**Activities**

Lab: Using BigQuery to do Analysis

## Module 02 Building a Data Lake

**Topics**

- Introduction to data lakes
- Data storage and ETL options on Google Cloud
- Building a data lake using Cloud Storage
- Securing Cloud Storage
- Storing all sorts of data types
- Cloud SQL as a relational data lake

**Objectives**

- Understand why Cloud Storage is a great option for building a data lake on Google Cloud
- Learn how to use Cloud SQL for a relational data lake

**Activities**

Lab: Loading Taxi Data into Cloud SQL

## Module 03 Building a Data Warehouse

**Topics**

- The modern data warehouse
- Introduction to BigQuery
- Getting started with BigQuery
- Loading data
- Exploring schemas
- Schema design
- Nested and repeated fields
- Optimizing with partitioning and clustering

**Objectives**

- Discuss requirements of a modern warehouse
- Understand why BigQuery is the scalable data warehousing solution on Google Cloud
- Understand core concepts of BigQuery and review options of loading data into BigQuery

**Activities**

- Lab: Loading Data into BigQuery
- Lab: Working with JSON and Array Data in BigQuery

## Module 04 Introduction to Building Batch Data Pipelines

**Topics**

- EL, ELT, ETL
- Quality considerations
- How to carry out operations in BigQuery
- Shortcomings
- ETL to solve data quality issues

**Objectives**

- Review different methods of loading data into your data lakes and warehouses: EL, ELT, and ETL
- Discuss data quality considerations and when to use ETL instead of EL and ELT

**Activities** —

## Module 05 Executing Spark on Dataproc

**Topics**

- The Hadoop ecosystem
- Run Hadoop on Dataproc
- Cloud Storage instead of HDFS
- Optimize Dataproc

**Objectives**

- Review the parts of the Hadoop ecosystem
- Learn how to lift and shift your existing Hadoop workloads to the cloud using Dataproc
- Understand considerations around using Cloud Storage instead of HDFS for storage
- Learn how to optimize Dataproc jobs

**Activities** Lab: Running Apache Spark jobs on Dataproc

## Module 06 Serverless Data Processing with Dataflow

**Topics**

- Introduction to Dataflow
- Why customers value Dataflow
- Dataflow pipelines
- Aggregating with GroupByKey and Combine
- Side inputs and windows
- Dataflow templates
- Dataflow SQL

**Objectives**

- Understand how to decide between Dataflow and Dataproc for processing data pipelines
- Understand the features that customers value in Dataflow
- Discuss core concepts in Dataflow
- Review the use of Dataflow templates and SQL

**Activities**

- Lab: A Simple Dataflow Pipeline (Python/Java)
- Lab: MapReduce in Dataflow (Python/Java)
- Lab: Side inputs (Python/Java)

## Module 07 Manage Data Pipelines with Cloud Data Fusion and Cloud Composer

**Topics**

- Building batch data pipelines visually with Cloud Data Fusion
- Components
- UI overview
- Building a pipeline
- Exploring data using Wrangler
- Orchestrating work between Google Cloud services with Cloud Composer
- Apache Airflow environment
- DAGs and operators
- Workflow scheduling
- Monitoring and logging

**Objectives**

- Discuss how to manage your data pipelines with Data Fusion and Cloud Composer
- Understand Data Fusion’s visual design capabilities
- Learn how Cloud Composer can help to orchestrate the work across multiple Google Cloud services

**Activities**

- Lab: Building and Executing a Pipeline Graph in Data Fusion
- Optional Lab: An introduction to Cloud Composer

## Module 08 Introduction to Processing Streaming Data

**Topics**

Process Streaming Data

**Objectives**

- Explain streaming data processing
- Describe the challenges with streaming data
- Identify the Google Cloud products and tools that can help address streaming data challenges

**Activities**

—

## Module 09 Serverless Messaging with Pub/Sub

**Topics**

- Introduction to Pub/Sub
- Pub/Sub push versus pull
- Publishing with Pub/Sub code

**Objectives**

- Describe the Pub/Sub service
- Understand how Pub/Sub works
- Gain hands-on Pub/Sub experience with a lab that simulates real-time streaming sensor data

**Activities**

Lab: Publish Streaming Data into Pub/Sub

## Module 10 Dataflow Streaming Features

**Topics**

- Steaming data challenges
- Dataflow windowing

**Objectives**

- Understand the Dataflow service
- Build a stream processing pipeline for live traffic data
- Demonstrate how to handle late data using watermarks, triggers, and accumulation

**Activities**

Lab: Streaming Data Pipelines

## Module 11 High-Throughput BigQuery and Bigtable Streaming Features

**Topics**

- Streaming into BigQuery and visualizing results
- High-throughput streaming with Cloud Bigtable
- Optimizing Cloud Bigtable performance

**Objectives**

- Learn how to perform ad hoc analysis on streaming data using BigQuery and dashboards
- Understand how Cloud Bigtable is a low-latency solution
- Describe how to architect for Bigtable and how to ingest data into Bigtable
- Highlight performance considerations for the relevant services

**Activities**

- Lab: Streaming Analytics and Dashboards
- Lab: Streaming Data Pipelines into Bigtable

## Module 12 Advanced BigQuery Functionality and Performance

**Topics**

- Analytic window functions
- Use With clauses

- GIS functions
- Performance considerations

**Objectives**

- Review some of BigQuery’s advanced analysis capabilities
- Discuss ways to improve query performance

**Activities**

- Lab: Optimizing your BigQuery Queries for Performance
- Optional Lab: Partitioned Tables in BigQuery

## Module 13 Introduction to Analytics and AI

**Topics**

- What is AI?
- From ad-hoc data analysis to data-driven decisions
- Options for ML models on Google Cloud

**Objectives**

- Understand the proposition that ML adds value to your data
- Understand the relationship between ML, AI, and Deep Learning
- Identify ML options on Google Cloud

**Activities** —

## Module 14 Prebuilt ML Model APIs for Unstructured Data

**Topics**

- Unstructured data is hard
- ML APIs for enriching data

**Objectives**

- Discuss challenges when working with unstructured data
- Learn the applications of ready-to-use ML APIs on unstructured data

**Activities** Lab: Using the Natural Language API to Classify Unstructured Text

## Module 15 Big Data Analytics with Notebooks

**Topics**

- What’s a notebook?
- BigQuery magic and ties to Pandas

**Objectives**

- Introduce Notebooks as a tool for prototyping ML solutions
- Learn to execute BigQuery commands from Notebooks

**Activities** Lab: BigQuery in Jupyter Labs on AI Platform

## Module 16 Production ML Pipelines

**Topics**

- Ways to do ML on Google Cloud
- Vertex AI Pipelines
- AI Hub

**Objectives**

- Describe options available for building custom ML models
- Understand the use of tools like Vertex AI Pipelines

**Activities**

Lab: Running Pipelines on Vertex AI

## Module 17 Custom Model Building with SQL in BigQuery ML

**Topics**

- BigQuery ML for quick model building
- Supported models

**Objectives**

- Learn how to create ML models by using SQL syntax in BigQuery
- Demonstrate building different kinds of ML models using BigQuery ML

**Activities**

- Lab option 1: Predict Bike Trip Duration with a Regression Model in BigQuery ML
- Lab option 2: Movie Recommendations in BigQuery ML

## Module 18 Custom Model Building with AutoML

**Topics**

- Why AutoML?
- AutoML Vision
- AutoML NLP
- AutoML tables

**Objectives**

- Explore various AutoML products used in machine learning
- Learn to use AutoML to create powerful models without coding

**Activities**

—