

## Data Engineering on Google Cloud

Course Code: GCPDE

Duration: 4 days

Instructor-led Training (ILT) | Virtual Instructor-led Training (VILT)

### OVERVIEW

Get hands-on experience with designing and building data processing systems on Google Cloud. This course uses lectures, demos, and hand-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.

### SKILLS COVERED

- Design and build data processing systems on Google Cloud Platform.
- Leverage unstructured data using Spark and ML APIs on Cloud Dataproc.
- 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.
- Enable instant insights from streaming data

### WHO SHOULD ATTEND

This class is intended for experienced developers who are responsible for managing big data transformations including:

- 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

To get the most out of this course, participants should have:

- Completed Google Cloud Fundamentals- Big Data and 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 basic statistics

### MODULES

#### Module 1: Introduction to Data Engineering

- Explore the role of a data engineer.
- Analyze data engineering challenges.
- Intro to BigQuery.
- Data Lakes and Data Warehouses.
- Demo: Federated Queries with BigQuery.
- Transactional Databases vs Data Warehouses.
- Website Demo: Finding PII in your dataset with DLP API.
- Partner effectively with other data teams.
- Manage data access and governance.
- Build production-ready pipelines.
- Review GCP customer case study.

#### Lab: Analyzing Data with BigQuery.

**Module 2: Building a Data Lake**

- Introduction to Data Lakes.
- Data Storage and ETL options on GCP.
- Building a Data Lake using Cloud Storage.
- Optional Demo: Optimizing cost with Google Cloud Storage classes and Cloud Functions.
- Securing Cloud Storage.
- Storing All Sorts of Data Types.
- Video Demo: Running federated queries on Parquet and ORC files in BigQuery.

Cloud SQL as a relational Data Lake.

**Lab: Loading Taxi Data into Cloud SQL.****Module 3: Building a Data Warehouse**

- The modern data warehouse.
- Intro to BigQuery.
- Demo: Query TB+ of data in seconds.
- Getting Started.
- Loading Data.
- Video Demo: Querying Cloud SQL from BigQuery.

**Lab: Loading Data into BigQuery.**

- Exploring Schemas.
- Demo: Exploring BigQuery Public Datasets with SQL using INFORMATION\_SCHEMA.
- Schema Design.
- Nested and Repeated Fields.
- Demo: Nested and repeated fields in BigQuery.

**Lab: Working with JSON and Array data in BigQuery.**

- Optimizing with Partitioning and Clustering.

- Demo: Partitioned and Clustered Tables in BigQuery.
- Preview: Transforming Batch and Streaming Data.

**Module 4: Introduction to Building Batch Data Pipelines**

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

**Module 5: Executing Spark on Cloud Dataproc**

- The Hadoop ecosystem.
- Running Hadoop on Cloud Dataproc.
- GCS instead of HDFS.
- Optimizing Dataproc.

**Lab: Running Apache Spark jobs on Cloud Dataproc.****Module 6: Serverless Data Processing with Cloud Dataflow**

- Cloud Dataflow.
- Why customers value Dataflow.
- Dataflow Pipelines.

**Lab: A Simple Dataflow Pipeline (Python/Java).****Lab: MapReduce in Dataflow (Python/Java).****Lab: Side Inputs (Python/Java).**

- Dataflow Templates.
- Dataflow SQL.

**Module 7: Manage Data Pipelines with Cloud****Data Fusion and Cloud Composer**

- Building Batch Data Pipelines visually with Cloud Data Fusion.
- Components.
- UI Overview.
- Building a Pipeline.
- Exploring Data using Wrangler.

**Lab: Building and executing a pipeline graph in Cloud Data Fusion.**

- Orchestrating work between GCP services with Cloud Composer.
- Apache Airflow Environment.
- DAGs and Operators.
- Workflow Scheduling.
- Optional Long Demo: Event-triggered Loading of data with Cloud Composer, Cloud Functions, Cloud Storage, and BigQuery.
- Monitoring and Logging.

**Lab: An Introduction to Cloud Composer.****Module 8: Introduction to Processing Streaming Data**

Processing Streaming Data.

**Module 9: Serverless Messaging with Cloud Pub/Sub**

Cloud Pub/Sub.

**Lab: Publish Streaming Data into Pub/Sub.****Module 10: Cloud Dataflow Streaming Features**

- Cloud Dataflow Streaming Features.

**Lab: Streaming Data Pipelines.****Module 11: High-Throughput BigQuery and Bigtable Streaming Features**

- BigQuery Streaming Features.

**Lab: Streaming Analytics and Dashboards.**

- Cloud Bigtable.

**Lab: Streaming Data Pipelines into Bigtable.****Module 12: Advanced BigQuery Functionality and Performance**

- Analytic Window Functions.
- Using With Clauses.
- GIS Functions.
- Demo: Mapping Fastest Growing Zip Codes with BigQuery GeoViz.
- Performance Considerations.

**Lab: Optimizing your BigQuery Queries for Performance.**

- Optional Lab: Creating Date-Partitioned Tables in BigQuery.

**Module 13: Introduction to Analytics and AI**

- What is AI?.
- From Ad-hoc Data Analysis to Data Driven Decisions.
- Options for ML models on GCP.

**Module 14: Prebuilt ML model APIs for Unstructured Data**

- Unstructured Data is Hard.
- ML APIs for Enriching Data.

**Lab: Using the Natural Language API to Classify Unstructured Text.****Module 15: Big Data Analytics with Cloud AI Platform Notebooks**

- Whats a Notebook.
- BigQuery Magic and Ties to Pandas.

**Lab: BigQuery in Jupyter Labs on AI Platform.****Module 16: Production ML Pipelines with Kubeflow**

- Ways to do ML on GCP.
- Kubeflow.
- AI Hub.

**Lab: Running AI models on Kubeflow.****Module 17: Custom Model building with SQL in BigQuery ML**

- BigQuery ML for Quick Model Building.
- Demo: Train a model with BigQuery ML to predict NYC taxi fares.
- Supported Models.

**Lab Option 1: Predict Bike Trip Duration with a Regression Model in BQML.****Lab Option 2: Movie Recommendations in BigQuery ML.****Module 18: Custom Model building with Cloud AutoML**

- Why Auto ML?
- Auto ML Vision.
- Auto ML NLP.
- Auto ML Tables.