



Data Engineering on Google Cloud

Course description

Learn to Design, Build, Automate, Secure, and Optimize Batch and Streaming Data Processing Systems using Google Cloud's Modern Data Engineering Platform.

The course covers structured, unstructured, and streaming data processing using Google Cloud data engineering products and modern data architecture principles.

- Course level: Intermediate
- Duration: 4 days

Activities

This course includes instructor-led sessions, on-demand learning, hands-on labs, architecture demonstrations, data pipeline workshops, streaming analytics exercises, and governance and monitoring labs to reinforce learning.

Course objectives

In this course, you will learn to:

- Design scalable and secure data processing systems on Google Cloud.
- Build and manage batch and streaming data pipelines.
- Implement modern data lakehouse architectures.

- Automate orchestration and workflow management.
- Use BigQuery, Dataflow, Dataproc, Pub/Sub, and Bigtable effectively.
- Apply AI/ML and analytics capabilities to data engineering workflows.
- Implement monitoring, governance, and observability practices.
- Build operational and analytical data platforms using Google Cloud services.

Key features

- Modern Data Engineering Foundations
- Lakehouse and Data Warehouse Architectures
- Batch and Streaming Data Pipelines
- Automation and Orchestration
- Data Quality, Governance, and Security
- AI/ML and Advanced Analytics Integration
- Observability and Monitoring
- Real-World Hands-On Labs and Use Cases

Intended audience

This course is intended for:

- Data Engineers
- Data Analysts
- Data Architects

Prerequisites

We recommend that attendees of this course have:

- Google Cloud Fundamentals: Core Infrastructure
- Introduction to Data Engineering on Google Cloud
- Introduction to AI and Machine Learning on Google Cloud

Products covered

- AlloyDB
- BigLake

- BigQuery
- Bigtable
- Cloud Composer
- Cloud Data Fusion
- Cloud Logging
- Cloud Monitoring
- Dataflow
- Dataform
- Dataplex Universal Catalog
- Dataproc
- Managed Service for Apache Kafka
- Pub/Sub
- Serverless for Apache Spark
- Vertex AI

Course outline

Module 01 - Data Engineering Tasks and Components

- Role of a Data Engineer
- Data Sources and Data Sinks
- Data Formats
- Storage Solutions
- Metadata Management
- Analytics Hub

Module 02 - Data Replication and Migration

- Replication and Migration Architecture
- gcloud CLI
- Storage Transfer Service
- Transfer Appliance
- Datastream

Module 03 - Extract and Load Data Pipeline Pattern

- Extract and Load Architecture

- bq Command-Line Tool
- BigQuery Data Transfer Service
- BigLake

Module 04 - Extract, Load, and Transform (ELT) Pipeline Pattern

- ELT Architecture
- BigQuery SQL Scripting
- Scheduling
- Dataform

Module 05 - Extract, Transform, and Load (ETL) Pipeline Pattern

- ETL Architecture
- GUI Tools for ETL
- Dataproc
- Dataproc Serverless for Spark
- Streaming Processing
- Bigtable Pipelines

Module 06 - Automation Techniques

- Automation Patterns
- Cloud Scheduler
- Workflows
- Cloud Composer
- Cloud Run Functions
- Eventarc

Module 07 - Introduction to Modern Data Engineering

- Data Lakes
- Data Warehouses
- Data Lakehouse
- Architecture Selection

Module 08 - Building a Data Lakehouse

- Cloud Storage Foundation
- Apache Iceberg

- BigQuery
- AlloyDB
- Federated Queries

Module 09 - Modernizing Data Warehouses

- BigQuery Fundamentals
- Partitioning and Clustering
- BigLake
- External Tables

Module 10 - Advanced Lakehouse Patterns and Governance

- Governance and Security
- Data Loss Prevention
- Analytics and Machine Learning
- Migration Strategies

Module 11 - When to Choose Batch Data Pipelines

- Batch Pipeline Use Cases
- Processing Challenges

Module 12 - Design and Build Scalable Batch Data Pipelines

- Batch Pipeline Design
- Large-Scale Transformations
- Dataflow
- Serverless Spark
- Orchestration

Module 13 - Control Data Quality in Batch Data Pipelines

- Validation and Cleansing
- Error Analysis
- Schema Evolution
- Deduplication

Module 14 - Orchestrate and Monitor Batch Data Pipelines

- Cloud Composer
- Observability

- Alerts and Troubleshooting
- Visual Pipeline Management

Module 15 - Course Introduction

- Streaming Pipeline Concepts
- Business Use Case
- Challenges and Mission

Module 16 - Streaming Use Cases and Reference Architectures

- Streaming ETL
- Streaming AI/ML
- Streaming Applications
- Reverse ETL

Module 17 - Product Deep Dives: Messaging and Processing

- Pub/Sub
- Managed Service for Apache Kafka
- Dataflow
- Streaming Processing

Module 18 - Product Deep Dives: Analytics and Operational Data

- BigQuery Streaming
- Continuous Queries
- Reverse ETL
- Bigtable Integration

Module 19 - Key Takeaways

- Course Summary
- Next Steps

Certification details

CloudThat Course Completion Certificate will be awarded to all learners who complete the training.

FAQs

Who should attend this course?

This course is designed for data engineers, analysts, and architects building modern data platforms on Google Cloud.

Is this course hands-on?

Yes. The course includes multiple labs, architecture workshops, pipeline exercises, and real-world streaming scenarios.

What technologies are covered?

The course covers BigQuery, Dataflow, Dataproc, Pub/Sub, Bigtable, BigLake, Vertex AI, and several Google Cloud data services.

Will streaming pipelines be covered?

Yes. Streaming ETL, streaming analytics, reverse ETL, Pub/Sub, Kafka, and Dataflow are major focus areas.

Does the course include lakehouse architectures?

Yes. Learners will explore modern lakehouse architectures using BigLake, BigQuery, and Apache Iceberg.

What is the delivery format?

The course is available in Instructor-Led Training (ILT) and On-Demand formats.

Is programming knowledge required?

Yes. SQL proficiency and familiarity with Python are recommended.

Will I receive a certificate?

Yes. A CloudThat Course Completion Certificate will be awarded after successful completion of the training.