



Introduction to Data Engineering on Google Cloud

Course description

In this course, you will learn the fundamental concepts, tools, and best practices for data engineering on the Google Cloud Platform (GCP). You will gain a solid understanding of how to design, build, and manage robust, scalable, and cost-effective data pipelines using GCP's comprehensive suite of data services. This course covers everything from data ingestion and storage to processing, transformation, orchestration, and monitoring of data workflows, equipping you with the practical skills needed to implement real-world data solutions on Google Cloud.

- Course level: Fundamental
- Duration: 2 days

Activities

This course includes presentations, hands-on labs, demonstrations, and interactive discussions.

Course objectives

In this course, you will learn to:

- Summarize the core principles and challenges of data engineering.
- Identify and select appropriate Google Cloud services for various data engineering use cases (storage, processing, analytics).

- Design and implement scalable data ingestion and storage solutions on GCP.
- Build batch and streaming data pipelines using services like Cloud Dataflow and Cloud Dataproc.
- Orchestrate complex data workflows with Cloud Composer (Apache Airflow).
- Understand and apply best practices for data governance, security, and monitoring on Google Cloud.
- Prepare and transform data effectively for downstream analytics and machine learning applications.

Intended audience

This course is intended for those interested in building and managing data solutions on Google Cloud, including:

- Aspiring Data Engineers
- Data Analysts and Scientists looking to expand into data pipeline development
- Software Developers transitioning to data-intensive roles
- IT Professionals responsible for data infrastructure
- Solution Architects evaluating Google Cloud for data initiatives

Prerequisites

We recommend that attendees of this course have:

- Basic understanding of cloud computing concepts.
- Familiarity with SQL.
- Foundational knowledge of programming (preferably Python) is beneficial but not strictly required for core concepts.

Course outline

Module 1: Introduction to Data Engineering and Google Cloud

- What is Data Engineering? Role, responsibilities, and key challenges
- The modern data landscape and the need for scalable solutions
- Overview of Google Cloud Platform for data engineering
- Key GCP services for data: Storage, compute, analytics, and AI/ML
- Understanding infrastructure-as-code and managed services on GCP

Module 2: Data Storage Solutions on Google Cloud

- Cloud Storage (GCS): Object storage for structured, semi-structured, and unstructured data
- Storage classes, lifecycle management, and security
- BigQuery: Serverless, highly scalable, and cost-effective data warehouse
- Architecture, datasets, tables, views, and querying
- Cloud SQL: Managed relational database service (PostgreSQL, MySQL, SQL Server)
- Cloud Spanner: Globally distributed, strongly consistent relational database
- Cloud Firestore & Datastore: NoSQL document databases for application data
- Hands-on Lab: Storing and Managing Data in Cloud Storage and BigQuery

Module 3: Batch Data Processing and Transformation

- Cloud Dataflow (Apache Beam): Unified programming model for batch and streaming data processing
- Concepts: PCollection, Transforms, Pipelines
- Dataflow templates and custom pipelines
- Cloud Dataproc: Managed Apache Hadoop, Spark, Flink, and Presto service
- Cluster management, job submission, and integration with GCS
- Cloud Data Fusion: Fully managed, cloud-native data integration service (powered by CDAP)
- ETL/ELT pipeline design and execution
- Cloud Functions: Serverless event-driven compute for small data tasks
- Hands-on Lab: Building and Executing a Batch Dataflow Pipeline

Module 4: Real-time Data Ingestion and Streaming Analytics

- Cloud Pub/Sub: Asynchronous messaging service for real-time data ingestion
- Topics, subscriptions, push/pull delivery
- Cloud Dataflow Streaming: Applying Apache Beam to real-time data
- Windowing, watermarks, and stateful processing
- BigQuery Streaming Inserts: Ingesting real-time data directly into BigQuery
- Cloud IoT Core: Connecting and managing IoT devices (overview)

- Hands-on Lab: Ingesting Real-time Data with Pub/Sub and Processing with Dataflow

Module 5: Orchestration and Workflow Management

- Cloud Composer (Apache Airflow): Fully managed workflow orchestration service
 - DAGs (Directed Acyclic Graphs), operators, sensors
 - Scheduling, monitoring, and managing complex data pipelines
- Cloud Workflows: Orchestrating serverless services
- Scheduled Queries in BigQuery: Automating routine data tasks
- Hands-on Lab: Orchestrating a Multi-step Data Pipeline with Cloud Composer

Module 6: Data Governance, Security, and Monitoring

- Cloud IAM: Managing access and permissions for data services
 - Service accounts, roles, and principle of least privilege
- Data Loss Prevention (DLP) API: Discovering, classifying, and protecting sensitive data
- Cloud Logging and Monitoring (Operations Suite): Observing data pipelines and services
 - Metrics, logs, dashboards, and alerts
- Cloud Data Catalog: Metadata management service for data discovery and governance
- Security best practices for data at rest and in transit on GCP
- Hands-on Lab: Implementing IAM for Data Services and Monitoring a Data Pipeline

Module 7: Integrating Data Engineering with Machine Learning and AI

- The role of data engineering in the ML lifecycle (MLOps)
- Vertex AI: Unified platform for building, deploying, and managing ML models
 - Data preparation for Vertex AI Training
- BigQuery ML: Creating and executing machine learning models directly within BigQuery
 - Common ML models (linear regression, logistic regression, k-means)
- Data Labeling Service: Preparing datasets for supervised learning (overview)

- Demo: Preparing Data for ML with BigQuery ML and Exporting for Vertex AI

Module 8: Course Wrap-up

- Review of core Google Cloud data engineering concepts and services
- Best practices for designing and deploying robust data pipelines
- Pathways for further learning and advanced certifications
- Q&A session and discussion of real-world use cases
- Course summary

Your course overview ends here.
But your cloud journey doesn't!

Unlock advanced training and certification pathways with **CloudThat**.



About CloudThat

Established in 2012, CloudThat is an award-winning company and the first in India to offer cloud training and consulting services for individuals and enterprises worldwide. Recently, it won Google Cloud's New Training Partner of the Year Award for 2025, becoming the first company in the world in 2025 to hold awards from all three major cloud giants: AWS, Microsoft, and Google. CloudThat notably won consecutive AWS Training Partner of the Year (APJ) awards in 2023 and 2024 and the Microsoft Training Services Partner of the Year Award in 2024, bringing its total award count to an impressive 12 awards in the last 8 years.

CloudThat specializes in Cloud Migration, Data Platforms, DevOps, IoT, and cutting-edge technologies like Gen AI & AI/ML. As a Microsoft Solutions Partner, AWS Premier Tier Training Partner, Google Cloud Platform Partner, and collaborator with leading organizations like HPE, Oracle, Nvidia and Databricks, CloudThat has trained over 850,000 professionals across 600+ cloud certifications, empowering students and professionals worldwide to advance their skills and careers.

To know more about our cloud certification training, email at sales@cloudthat.com or call us at [+91-8880002200](tel:+91-8880002200) or [+1-8555588830](tel:+1-8555588830) for those from outside India.