Data Engineering on Google Cloud

Duration: 4 Days Course Code: GO5975

Overview:

Get hands-on experience 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, and analyze data.
This course covers structured, unstructured, and streaming data.
It is a combination of four sequences:
- Introduction to Data Engineering on Google Cloud
- Build Data Lakes and Data Warehouses with Google Cloud
- Build Batch Data Pipelines on Google Cloud
- Build Streaming Data Pipelines on Google Cloud

Updated 20/02/2026

Target Audience:

- Data engineers
- Database administrators
- System administrators

Objectives:

- In this course participants will learn:
- Design scalable data processing systems in Google Cloud.
- Differentiate data architectures and implement data lakehouse and pipeline concepts.
- Build and manage robust streaming and batch data pipelines.
- Utilize AI/ML tools to optimize performance and gain process and data insights.

Prerequisites:

- Understanding of data engineering principles, including ETL/ELT processes, data modeling, and common data formats (Avro, Parquet, JSON).
- Familiarity with data architecture concepts, specifically Data Warehouses and Data Lakes.
- Proficiency in SQL for data querying.
- Proficiency in a common programming language (Python recommended).
- Familiarity with using Command Line Interfaces (CLI).
- Familiarity with core Google Cloud concepts and services (Compute, Storage, and Identity management).

Testing and Certification

This course is part of the Learning path towards Google Certification Professional Data Engineer

Follow-on-Courses:

No recommendation

Content:

Sequence 1: Introduction to Data Engineering on Google Cloud

Module 1: Data Engineering Tasks and Components

- 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.

Module 2: Data replication and migration

- 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.

Module 3: The extract and load data pipeline pattern

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

Module 4: The extract, load, and transform data pipeline pattern

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

Module 5: The extract, transform, and load data pipeline pattern

Module 7: Introduction to modern data engineering and Google Cloud

- Compare and contrast data lake, data warehouse, and data lakehouse architectures.
- Evaluate the benefits of the lakehouse approach
- Lab: Using BigQuery to Do Analysis

Module 8: Building a data lakehouse with Cloud Storage, open formats, and BigQuery

- Discuss data storage options, including Cloud Storage for files, open table formats like Apache Iceberg, BigQuery for analytic data, and AlloyDB for operational data.
- Understand the role of AlloyDB for operational data use cases.
- Lab: Federated Query with BigQuery

Module 9: Modernizing Data Warehouses with BigQuery and BigLake

- Explain why BigQuery is a scalable data warehousing solution on Google Cloud.
- Discuss the core concepts of BigQuery.
- Understand BigLake's role in creating a unified lakehouse architecture and its integration with BigQuery for external data.
- Learn how BigQuery natively interacts with Apache Iceberg tables via BigLake.
- Lab: Querying External Data and Iceberg Tables

Module 10: Advanced lakehouse patterns and data governance

- Implement robust data governance and security practices across the unified data platform, including sensitive data protection and metadata management.
- Explore advanced analytics and machine learning directly on lakehouse data.

Module 11: Labs and best practices

- Reinforce the core principles of Google Cloud's data platform.
- Lab: Getting Started with BigQuery ML
- Lab: Vector Search with BigQuery

Sequence 3: Build Batch Data Pipelines on Google Cloud

Module 12: When to choose batch data pipelines

- Explain the critical role of a data engineer in developing and maintaining batch data

Module 14: Control Data Quality in Batch Data Pipelines

- Develop data validation rules and cleansing logic to ensure data quality within batch pipelines.
- Implement strategies for managing schema evolution and performing data deduplication in large datasets.
- Lab: Validate Data Quality in a Batch Pipeline with Serverless for Apache Spark (optional)

Module 15: Orchestrate and Monitor Batch Data Pipelines

- Orchestrate complex batch data pipeline workflows for efficient scheduling and lineage tracking.
- Implement robust error handling, monitoring, and observability for batch data pipelines.
- Lab: Building Batch Pipelines in Cloud Data Fusion

Sequence 4: Build Streaming Data Pipelines on Google Cloud

Module 16: Concept of streaming data pipelines

- Introduce the sequence learning objectives, and the scenario that will be used to bring hands on learning to building streaming data pipelines.
- Describe the concept of streaming data pipelines, challenges associated with it, and the role of these pipelines within the data engineering process.

Module 17: Streaming use cases and reference architectures

- Understand various streaming use cases and their applications, including Streaming ETL, Streaming AI/ML, Streaming Application, and Reverse ETL.
- Identify and describe common sample architectures for streaming data, including Streaming ETL, Streaming AI/ML, Streaming Application, and Reverse ETL.

Module 18: Product deep dives

- Pub/Sub and Managed Service for Apache Kafka: Define messaging concepts, know when to use Pub/Sub or Managed Service for Apache Kafka.
- Dataflow: Describe the service and challenges with streaming data, build and deploy a streaming pipeline.
- BigQuery: Explore various data ingestion

- 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 to use Dataproc Serverless for Spark for ETL.
- Explain streaming data processing options.
- Explain the role Bigtable plays in data pipelines.
- Lab: Use Dataproc Serverless for Spark to Load BigQuery
- Lab: Creating a Streaming Data Pipeline for a Real-Time Dashboard with Dataflow

Module 6: Automation techniques

- 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.
- Lab: Use Cloud Run Functions to Load BigQuery

Sequence 2: Build Data Lakes and Data Warehouses with Google Cloud

pipelines.
- Describe the core components and typical lifecycle of batch data pipelines from ingestion to downstream consumption.
- Analyze common challenges in batch data processing, such as data volume, quality, complexity, and reliability, and identify key Google Cloud services that can address them.

Module 13: Design and Build Scalable Batch Data Pipelines

- Design scalable batch data pipelines for high-volume data ingestion and transformation.
- Optimize batch jobs for high throughput and cost-efficiency using various resource management and performance tuning techniques.
- Lab: Build a Simple Batch Data Pipeline with Serverless for Apache Spark (optional)
- Lab: Build a Simple Batch Data Pipeline with Dataflow Job Builder UI (optional)

methods, use BigQuery continuous queries, BigQuery ETL, and reverse ETL, configure Pub/Sub to BigQuery streaming, architecting BigQuery streaming pipelines.
- Bigtable: Describe the big picture of data movement and interaction, establish a streaming pipeline from Dataflow to Bigtable, analyze the Bigtable continuous data stream for trends using BigQuery, synchronize the trends analysis back into the user-facing application.
- Lab: Stream data with pipelines - Esports use case (optional)
- Lab: Use Apache Beam and Bigtable to enrich esports downloadable content (DLC) data
- Lab: Stream e-sports data with Pub/Sub and BigQuery
- Lab: Monitor e-sports chat with Streamlit

Module 19: Key takeaways

Additional Information:

Official course book provided to participants.

Further Information:

For More information, or to book your course, please call us on Head Office 01189 123456 / Northern Office 0113 242 5931

info@globalknowledge.co.uk

www.globalknowledge.com/en-gb/

Global Knowledge, Mulberry Business Park, Fishponds Road, Wokingham Berkshire RG41 2GY UK