



Data Engineering on Google Cloud Platform

Duration: 4 Days **Course Code: GO5975**

Overview:

Learn how to design and build data processing systems.

This four-day instructor-led class provides you with a hands-on introduction to designing and building data processing systems on Google Cloud Platform. Through a combination of presentations, demos, and hand-on labs, you will learn how to design data processing systems, build end-to-end data pipelines, analyze data and carry out machine learning. The course covers structured, unstructured, and streaming data.

Target Audience:

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

Objectives:

- In this course you will learn:
 - Design and build data processing systems on Google Cloud Platform
 - 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
 - Leverage unstructured data using Spark and ML APIs on Cloud Dataproc
 - Enable instant insights from streaming data
-

Prerequisites:

- Completed Google Cloud Fundamentals- Big Data and Machine Learning course #8325 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 Python
 - Familiarity with Machine Learning and/or statistics
-

Content:

1. Serverless Data Analysis with BigQuery

- What is BigQuery
- Advanced Capabilities
- Performance and pricing

2. Serverless, Autoscaling Data Pipelines with Dataflow

3. Getting Started with Machine Learning

- What is machine learning (ML)
- Effective ML: concepts, types
- Evaluating ML
- ML datasets: generalization

4. Building ML Models with Tensorflow

- Getting started with TensorFlow
- TensorFlow graphs and loops + lab
- Monitoring ML training

5. Scaling ML Models with CloudML

- Why Cloud ML?
- Packaging up a TensorFlow model
- End-to-end training

6. Feature Engineering

- Creating good features
- Transforming inputs
- Synthetic features
- Preprocessing with Cloud ML

7. ML Architectures

- Wide and deep
- Image analysis
- Embeddings and sequences
- Recommendation systems

8. Google Cloud Dataproc Overview

- Introducing Google Cloud Dataproc
- Creating and managing clusters
- Defining master and worker nodes
- Leveraging custom machine types and preemptible worker nodes
- Creating clusters with the Web Console
- Scripting clusters with the CLI
- Using the Dataproc REST API
- Dataproc pricing
- Scaling and deleting Clusters

9. Running Dataproc Jobs

- Controlling application versions
- Submitting jobs
- Accessing HDFS and GCS
- Hadoop
- Spark and PySpark
- Pig and Hive
- Logging and monitoring jobs
- Accessing onto master and worker nodes with SSH
- Working with PySpark REPL (command-line interpreter)

10. Integrating Dataproc with Google Cloud Platform

- Initialization actions
- Programming Jupyter/Datalab notebooks
- Accessing Google Cloud Storage
- Leveraging relational data with Google Cloud SQL
- Reading and writing streaming Data with Google BigTable
- Querying Data from Google BigQuery
- Making Google API Calls from notebooks

11. Making Sense of Unstructured Data with Google's Machine Learning APIs

- Google's Machine Learning APIs
- Common ML Use Cases
- Vision API
- Natural Language API
- Translate
- Speech API

12. Need for Real-Time Streaming Analytics

- What is Streaming Analytics?
- Use-cases
- Batch vs. Streaming (Real-time)
- Related terminologies
- GCP products that help build for high

13. Architecture of Streaming Pipelines

- Streaming architectures and considerations
- Choosing the right components
- Windowing
- Streaming aggregation
- Events, triggers

14. Stream Data and Events into PubSub

- Topics and Subscriptions
- Publishing events into Pub/Sub
- Subscribing options: Push vs Pull
- Alerts

15. Build a Stream Processing Pipeline

- Pipelines, PCollections and Transforms
- Windows, Events, and Triggers
- Aggregation statistics
- Streaming analytics with BigQuery
- Low-volume alerts

16. High Throughput and Low-Latency with Bigtable

- Latency considerations
- What is Bigtable
- Designing row keys
- Performance considerations

17. High Throughput and Low-Latency with Bigtable

- What is Google Data Studio?
- From data to decisions

availability, resiliency, high-throughput,
real-timestreaming analytics (review of
Pub/Sub and Dataflow)

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.co.uk

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