

Big Data with Google Cloud — 32-Hour Curriculum

Day 1:

Introduction to Big Data & GCP

- **Topics:**
 - Big Data concepts (Volume, Velocity, Variety, Veracity, Value)
 - Cloud vs On-premise for Big Data
 - Overview of Google Cloud Platform (GCP) services
 - Projects, IAM, Billing, VPC overview
- **Hands-on:**
 - Create a GCP project
 - Setup service accounts
 - Create a Cloud Storage bucket

Cloud Storage Deep Dive

- **Topics:**
 - Bucket types (Standard, Nearline, Coldline, Archive)
 - Lifecycle rules and versioning
 - Data lake design patterns
- **Hands-on:**
 - Create buckets with different storage classes
 - Apply lifecycle rules
 - Upload and manage files

Cloud Bigtable

- **Topics:**
 - Wide-column NoSQL concepts
 - Schema design for time-series or IoT data
 - Bigtable vs BigQuery vs Firestore
 - **Hands-on:**
 - Create a Bigtable instance
 - Insert and query data using cbt or Python client
-

Day 2:

Dataproc Introduction

- **Topics:**
 - Managed Hadoop and Spark
 - Cluster types, autoscaling, preemptible VMs
 - Workflow templates
- **Hands-on:**
 - Create a Dataproc cluster
 - Explore cluster nodes and HDFS

Spark on Dataproc

- **Topics:**
 - RDDs, DataFrames, Spark SQL
 - Spark job submission on Dataproc
- **Hands-on:**
 - PySpark ETL: CSV → Parquet
 - Run Spark jobs from CLI

Hive on Dataproc

- **Topics:**
 - Hive tables on GCS
 - External vs managed tables
 - Querying large datasets
 - **Hands-on:**
 - Create Hive tables
 - Load data from GCS and query
-

Day 3:

Dataflow (Batch)

- **Topics:**
 - Apache Beam concepts
 - Batch pipelines: transforms, I/O
- **Hands-on:**
 - Build a batch pipeline: GCS → BigQuery
 - Test with small CSV dataset

Dataflow (Streaming)

- **Topics:**
 - Streaming concepts: windowing, triggers, watermarks
 - Event-time vs processing-time
- **Hands-on:**
 - Pub/Sub → Dataflow → BigQuery pipeline
 - Simulate streaming events

BigQuery Fundamentals

- **Topics:**
 - Architecture, storage vs compute
 - Datasets, tables, external tables
- **Hands-on:**
 - Create datasets and tables
 - Load CSV/JSON from GCS
 - Simple SQL queries

BigQuery Advanced

- **Topics:**
 - Partitioned & clustered tables
 - Materialized views
 - Optimizing queries

- **Hands-on:**
 - Partition table by date
 - Cluster table by column
 - Run queries on public dataset

Day 4:

BigQuery ML

- **Topics:**
 - Linear regression, classification
 - BigQuery ML model creation and evaluation
 - **Hands-on:**
 - Train model on sales dataset
 - Predict future values
-

Cloud Composer

- **Topics:**
 - Managed Apache Airflow
 - DAGs, scheduling, cross-service orchestration
 - **Hands-on:**
 - Create DAG for ETL pipeline: GCS → Dataproc → BigQuery
 - Setup retries, logging, and alerts
-

Pub/Sub Deep Dive

- **Topics:**
 - Pub/Sub architecture (topics, subscriptions, ack)
 - Push vs pull subscribers
 - **Hands-on:**
 - Create topic and subscription
 - Publish/consume messages
 - Simulate streaming data
-

Streaming Use Cases

- **Topics:**
 - Real-time analytics
 - Fraud detection or IoT sensor analytics
- **Hands-on:**
 - End-to-end pipeline: Pub/Sub → Dataflow → BigQuery

Day 5:

Data Governance & Security

- **Topics:**
 - IAM for data services
 - VPC Service Controls
 - Encryption (CMEK, CSEK)
 - Cloud DLP
- **Hands-on:**
 - Column-level security in BigQuery
 - Scan dataset for PII with DLP
 - Data catalog tagging

Vertex AI

- **Topics:**
 - ML model training & deployment
 - AutoML, Feature Store
 - Vertex AI Pipelines
- **Hands-on:**
 - Train AutoML model on sales dataset
 - Deploy prediction endpoint
 - Batch predictions

Mini Project – 3 hours