Mastering GCP Data Engineering

BigQuery | Dataflow | Dataproc | Pub/Sub | Composer

Databricks on GCP | Vertex AI | Kafka | Terraform

16	158+	214+	4
Modules	Hours	Topics	Capstone Projects

Technologies: BigQuery · Dataflow · Dataproc · Pub/Sub · Cloud Composer · GCS · Cloud SQL · Spanner · Bigtable · Databricks · Delta Lake · Apache Kafka · Vertex AI · Gemini · Terraform · Cloud Functions · Cloud Run · Dataplex · BigLake · Looker Studio

www.databrickstraining.in | WhatsApp: **+91-8500002025**

Aligned with Google Cloud Professional Data Engineer Certification

Course Modules

01	Introduction to GCP & Cloud Fundamentals	6 Hours
02	Google Cloud Storage (GCS)	6 Hours
03	BigQuery — Cloud Data Warehouse	16 Hours
04	Cloud Dataflow — Serverless ETL	12 Hours
05	Cloud Dataproc — Managed Spark & Hadoop	12 Hours
06	Cloud Pub/Sub & Event-Driven Pipelines	8 Hours
07	Cloud Composer — Orchestration with Airflow	10 Hours
08	Cloud SQL, Spanner & Bigtable	10 Hours
09	Data Catalog, Dataplex & Governance	8 Hours
10	Databricks on GCP	14 Hours
11	Apache Kafka & Streaming on GCP	10 Hours
12	Vertex AI & GenAI on GCP	10 Hours
13	Cloud Functions, Cloud Run & Serverless	6 Hours
14	Terraform & Infrastructure as Code on GCP	6 Hours
15	GCP Data Engineering Capstone Projects	16 Hours
16	GCP Data Engineer Certification Prep	8 Hours

Total: 158+ Hours | 214+ Topics | 16 Modules

MODULE 01

6 Hours

Introduction to GCP & Cloud Fundamentals

Foundational understanding of Google Cloud Platform — global infrastructure, IAM, billing, and the GCP console.

GCP Overview

- What is Cloud Computing? — IaaS, PaaS, SaaS
- GCP Global Infrastructure — Regions, Zones, Edge Network
- GCP Console, Cloud Shell & gcloud CLI
- GCP vs AWS vs Azure — Service Mapping Comparison
- GCP Free Tier & Cost Optimization Strategies
- Setting Up a GCP Project & Billing Account

IAM & Security

- IAM — Users, Service Accounts, Roles & Policies
- Custom IAM Roles & Conditional Policies
- Organization Policies & Resource Hierarchy
- VPC Service Controls & Data Perimeters
- Secret Manager for Credentials Management
- Cloud Audit Logs & Compliance

MODULE 02

Google Cloud Storage (GCS)

6 Hours

GCP's object storage service — storage classes, lifecycle management, access control, and integration patterns.

GCS Fundamentals

- GCS Architecture — Buckets, Objects, Metadata
- Storage Classes — Standard, Nearline, Coldline, Archive
- Lifecycle Management & Auto-Tiering
- gsutil Commands & Parallel Uploads
- GCS FUSE — Mount Buckets as File System
- Cross-Region & Dual-Region Replication

GCS Security & Integration

- IAM vs ACL-Based Access Control
- Signed URLs & Signed Policy Documents
- Customer-Managed Encryption Keys (CMEK)
- GCS Event Notifications → Pub/Sub → Cloud Functions
- GCS Transfer Service — On-Prem & Cross-Cloud Migrations
- Best Practices for Data Lake on GCS

MODULE 03

16 Hours

BigQuery — Cloud Data Warehouse

Google's serverless, petabyte-scale data warehouse — SQL analytics, ML, cost optimization, and advanced features.

BigQuery Architecture

- Dremel Engine, Colossus Storage & Jupiter Network
- Slots, Reservations & On-Demand vs Flat-Rate Pricing
- Datasets, Tables, Views & Materialized Views
- BigQuery Storage API for High-Throughput Reads
- Information Schema & JOBS Table Monitoring

BigQuery SQL & Data Loading

- Standard SQL — Window Functions, CTEs, Pivots, UDFs
- Partitioned Tables — Time, Range & Integer Partitioning
- Clustered Tables for Query Cost Optimization
- Loading Data — Batch (GCS, Local), Streaming Inserts
- BigQuery Data Transfer Service (DTS)
- Federated Queries — GCS, Cloud SQL, Bigtable, Drive
- Scheduled Queries & Parameterized Queries

BigQuery Advanced

- BigQuery ML — Train ML Models with SQL (Linear Reg, XGBoost, DNN)
- BigQuery BI Engine — In-Memory Acceleration
- BigQuery Omni — Query S3 & Azure Blob Natively
- Authorized Views & Row-Level Security (RLS)
- Column-Level Security & Data Masking Policies
- BigQuery Remote Functions & Cloud Functions UDFs
- BigQuery Editions & Cost Optimization Best Practices
- Change Data Capture (CDC) with BigQuery

MODULE 04

12 Hours

Cloud Dataflow — Serverless ETL

Apache Beam-based serverless data processing — batch & streaming pipelines, templates, and auto-scaling.

Apache Beam Fundamentals

- Apache Beam Programming Model — PCollections & PTransforms
- ParDo, Map, FlatMap, Filter, GroupByKey, CoGroupByKey
- Beam Schemas & Row-Based Processing
- Side Inputs, Side Outputs & Multi-Output Transforms
- Beam SQL & DataFrames API
- Testing Beam Pipelines with TestPipeline

Dataflow Pipelines

- Batch Pipelines — GCS → Transform → BigQuery
- Streaming Pipelines — Pub/Sub → Dataflow → BigQuery
- Windowing — Fixed, Sliding, Session & Global Windows
- Watermarks, Triggers & Allowed Lateness
- Dataflow Flex Templates — Reusable Parameterized Pipelines
- Dataflow Prime & Auto-Scaling Strategies
- Streaming Inserts vs Storage Write API
- Monitoring with Cloud Monitoring & Dataflow UI

MODULE 05

12 Hours

Cloud Dataproc — Managed Spark & Hadoop

Managed Apache Spark, Hadoop, Hive & Presto on GCP — cluster management, PySpark, and BigQuery integration.

Dataproc Cluster Management

- Dataproc Architecture — Master, Worker & Preemptible Nodes
- Cluster Creation with gcloud, Terraform & REST API
- Initialization Actions & Custom Images
- Dataproc Autoscaling Policies
- Dataproc on GKE — Kubernetes-Based Spark
- Dataproc Serverless — Zero Cluster Management

Spark on Dataproc

- PySpark Jobs — Submit, Monitor & Debug
- Spark SQL on Dataproc
- Hive on Dataproc with Dataproc Metastore
- BigQuery Connector for Spark — Read/Write BQ Tables
- GCS Connector — Use GCS as HDFS Replacement
- Spark Structured Streaming on Dataproc
- Jupyter & Zeppelin Notebooks on Dataproc
- Dataproc vs EMR vs Databricks Cost Comparison

MODULE 06

8 Hours

Cloud Pub/Sub & Event-Driven Pipelines

Fully managed messaging service — topics, subscriptions, delivery modes, and real-time event-driven architectures.

Pub/Sub Core Concepts

- Pub/Sub Architecture — Topics, Subscriptions, Messages
- Push vs Pull Delivery Modes
- Exactly-Once Delivery & Message Ordering
- Dead-Letter Topics & Retry Policies
- Pub/Sub Lite — Low-Cost Zonal Messaging
- Pub/Sub Schemas & Schema Registry

Event-Driven Architecture

- Pub/Sub → Dataflow Streaming Pipeline
- Pub/Sub → BigQuery Direct Subscription
- Pub/Sub → Cloud Functions (Serverless Processing)
- Eventarc — Unified Event Routing
- Pub/Sub vs Kafka — When to Use Which
- Cloud Tasks vs Pub/Sub — Use Case Differences

MODULE 07

10 Hours

Cloud Composer — Orchestration with Airflow

Managed Apache Airflow on GCP — DAG design, GCP operators, scheduling, and production best practices.

Cloud Composer Setup & DAGs

- Cloud Composer 2 Architecture & Environment Setup
- DAG Design Patterns & Best Practices
- GCP Operators — BigQuery, GCS, Dataflow, Dataproc
- Sensors — GCS, BigQuery, Pub/Sub Sensors
- TaskGroups, SubDAGs & Dynamic DAG Generation
- XComs, Variables & Connections Management

Production Airflow on GCP

- Composer 2 Auto-Scaling & Workload Identity
- CI/CD for DAGs — GitHub → Cloud Build → Composer
- Monitoring with Cloud Monitoring & Alerting
- Airflow REST API & Triggering DAGs Externally
- Cost Optimization — Node Pools & Scheduling
- Airflow on GKE vs Cloud Composer Comparison

MODULE 08

10 Hours

Cloud SQL, Spanner & Bigtable

Managed relational and NoSQL databases — Cloud SQL, Cloud Spanner, Bigtable, Firestore & Memorystore.

Cloud SQL & Cloud Spanner

- Cloud SQL — MySQL, PostgreSQL, SQL Server on GCP
- Cloud SQL Replicas, High Availability & Backups
- Cloud SQL Proxy & Private IP Connectivity
- Cloud Spanner — Globally Distributed SQL Database
- Spanner Schema Design & Interleaved Tables
- AlloyDB — PostgreSQL-Compatible for Analytics + OLTP

NoSQL Databases on GCP

- Cloud Bigtable — Wide-Column Store for Time-Series & IoT
- Bigtable Schema Design — Row Key Patterns
- Firestore — Document Database (Native vs Datastore Mode)
- Memorystore — Managed Redis & Memcached
- Database Migration Service (DMS) — Oracle/MySQL → GCP
- Choosing the Right GCP Database for Your Workload

MODULE 09

Data Catalog, Dataplex & Governance

8 Hours

Enterprise data governance — Data Catalog, Dataplex data mesh, BigLake, DLP & lineage tracking.

Data Catalog & Discovery

- Data Catalog — Search, Tag & Classify Data Assets
- Tag Templates & Policy Tags for Classification
- Data Catalog API — Automated Metadata Management
- Data Lineage — Track Data Flow Across Systems
- Sensitive Data Discovery with Cloud DLP

Dataplex & BigLake

- Dataplex — Data Mesh Architecture on GCP
- Dataplex Lakes, Zones & Data Quality Rules
- BigLake — Unified Access to GCS, S3 & Azure Blob
- BigLake Managed Tables & Iceberg Support
- Column-Level Security with Policy Tags
- VPC Service Controls for Data Perimeters

MODULE 10

Databricks on GCP

14 Hours

Running Databricks on Google Cloud — workspace setup, Unity Catalog, Delta Lake, and GCP-native integrations.

Databricks GCP Setup & Architecture

- Databricks on GCP Architecture — GKE-Based Clusters
- Workspace Provisioning & VPC Network Configuration
- Cluster Types — All-Purpose, Job, Serverless SQL Warehouses
- Databricks DBFS & GCS Integration
- Secrets Management with GCP Secret Manager
- Databricks PrivateLink for Secure Connectivity

Delta Lake & Lakehouse on GCP

- Delta Lake — ACID Transactions on GCS
- Delta Tables — MERGE, UPDATE, DELETE, Time Travel
- Delta Lake Optimization — ZORDER, OPTIMIZE, VACUUM
- SCD Type 1 & Type 2 with Delta Lake MERGE
- Streaming with Delta Lake — Auto Loader & Structured Streaming
- Change Data Feed (CDF) for Downstream Consumers

Unity Catalog & GCP Integration

- Unity Catalog — Centralized Governance on GCP
- External Locations with GCS Buckets
- Row-Level Security & Column Masking
- Lakehouse Federation — Query BigQuery from Databricks
- Delta Live Tables (DLT) — Declarative ETL Pipelines
- Databricks Workflows — Job Orchestration & Scheduling
- Databricks SQL Warehouses for BI & Analytics
- Databricks vs Dataproc — Feature & Cost Comparison

MODULE 11

10 Hours

Apache Kafka & Streaming on GCP

Real-time streaming — Managed Kafka, Pub/Sub integration, and end-to-end streaming architectures on GCP.

Kafka on GCP

- Kafka Architecture Recap — Brokers, Topics, Partitions
- Managed Service for Apache Kafka (Amazon MSK Alternative on GCP)
- Confluent Cloud on GCP — Fully Managed Kafka
- Kafka Connect — GCS Sink, BigQuery Sink Connectors
- Schema Registry with Confluent on GCP
- Kafka vs Pub/Sub — Architecture & Use Case Comparison

End-to-End Streaming Pipelines

- Kafka → Spark Structured Streaming → Delta Lake on GCS
- Kafka → Dataflow → BigQuery Pipeline
- Pub/Sub → Dataflow → BigQuery Real-Time Analytics
- CDC with Debezium → Kafka → BigQuery
- Streaming Data Quality & Late Data Handling
- Monitoring Streaming Pipelines — Lag, Throughput, Errors

MODULE 12

10 Hours

Vertex AI & GenAI on GCP

Machine Learning & Generative AI — Vertex AI, Gemini APIs, RAG pipelines, and MLOps on GCP.

Vertex AI Platform

- Vertex AI Overview — Training, Prediction, Pipelines
- AutoML — Tabular, Vision, Text & Video Models
- Custom Training with TensorFlow & PyTorch
- Vertex AI Pipelines — Kubeflow-Based MLOps
- Feature Store for ML Feature Management
- Model Registry & Model Monitoring

Generative AI on GCP

- Gemini API — Text, Multimodal & Code Generation
- Vertex AI Studio — Prompt Design & Tuning
- Embeddings API & Vector Search (Matching Engine)
- RAG Pipeline — Cloud Storage → Embeddings → BigQuery Vector
- Grounding with Google Search & Enterprise Data
- Responsible AI — Safety Filters & Guardrails
- Gen AI Pricing — Gemini Pro, Flash & Nano Comparison

MODULE 13

Cloud Functions, Cloud Run & Serverless

6 Hours

Serverless compute on GCP — Cloud Functions, Cloud Run, App Engine & event-driven processing.

Serverless Compute

- Cloud Functions (2nd Gen) — Event & HTTP Triggers
- Cloud Functions → BigQuery, GCS, Pub/Sub Integration
- Cloud Run — Containerized Serverless Applications
- Cloud Run Jobs — Batch Processing Without Servers
- Cloud Scheduler — Cron Jobs for Serverless Workflows
- Cloud Build — CI/CD for Cloud Functions & Cloud Run

Serverless Data Processing Patterns

- GCS Upload → Cloud Function → BigQuery Load
- Pub/Sub → Cloud Function → Data Transformation
- Scheduled Cloud Function for API Data Ingestion
- Cloud Run as a Lightweight Spark Alternative
- Serverless vs Dataflow — When to Use Which
- Cost Optimization for Serverless Workloads

MODULE 14

Terraform & Infrastructure as Code on GCP

6 Hours

Infrastructure automation — Terraform for GCP, Deployment Manager, and DevOps best practices.

Terraform for GCP

- Terraform Basics — Providers, Resources, Variables
- GCP Provider — BigQuery, GCS, Dataproc, Composer Resources
- Terraform State Management with GCS Backend
- Terraform Modules for Reusable GCP Infrastructure
- Terraform Import — Bring Existing Resources Under IaC
- CI/CD with Terraform — Cloud Build + GitHub Actions

GCP DevOps Essentials

- Cloud Deployment Manager — GCP-Native IaC
- Cloud Monitoring & Cloud Logging for Data Pipelines
- Error Reporting & Alerting Policies
- Cloud Trace & Cloud Profiler for Performance
- SRE Practices for Data Engineering Teams
- Cost Management — Budgets, Exports & Recommendations

MODULE 15

GCP Data Engineering Capstone Projects

16 Hours

End-to-end projects combining multiple GCP services — batch, streaming, lakehouse & ML pipelines.

Project 1: Batch Data Lake Pipeline

- GCS (Raw) → Dataflow (Transform) → BigQuery (Analytics)
- Partitioned & Clustered Tables for Performance
- Cloud Composer DAG for End-to-End Orchestration
- Data Quality Checks with Great Expectations
- Terraform for Full Infrastructure Provisioning

Project 2: Real-Time Streaming Analytics

- Pub/Sub (Ingest) → Dataflow (Window/Aggregate) → BigQuery
- Real-Time Dashboard with Looker Studio
- Dead-Letter Handling & Alert Policies
- Streaming Pipeline Monitoring & Autoscaling

Project 3: Lakehouse with Databricks on GCP

- GCS Data Lake → Auto Loader → Delta Lake Bronze/Silver/Gold
- Unity Catalog for Governance & Access Control
- Delta Live Tables for Declarative ETL
- Lakehouse Federation — Query BigQuery from Databricks
- Workflows for Scheduling & Alerting

Project 4: GenAI-Powered Data Pipeline

- Document Ingestion → Gemini API → Structured Data Extraction
- Embeddings → Vertex AI Vector Search → RAG Application
- BigQuery ML for Predictive Analytics
- End-to-End ML Pipeline with Vertex AI Pipelines

MODULE 16

GCP Data Engineer Certification Prep

8 Hours

Focused preparation for the Google Cloud Professional Data Engineer certification exam.

Exam Strategy & Key Topics

- Exam Format — 50 Questions, 2 Hours, Case Studies
- Storage & Database Selection Decision Tree
- ETL/ELT Pipeline Architecture Patterns
- Security & Compliance — IAM, Encryption, DLP
- Data Processing — Batch vs Streaming Decision Framework
- ML & AI on GCP — BigQuery ML, Vertex AI, AutoML
- Cost Optimization Strategies Across All Services
- Practice Exams & Case Study Walkthroughs

Why Choose Sreyobhilashi IT?

1. **Industry Expert Trainers**
   Learn from professionals with 10+ years in data engineering and cloud platforms.
2. **Hands-On Lab Environment**
   Every topic includes live coding exercises on real GCP infrastructure.
3. **Real-World Capstone Projects**
   Build production-grade pipelines that showcase on your portfolio.
4. **Placement Assistance**
   Resume building, mock interviews, and direct company referrals.
5. **Multi-Cloud Coverage**
   AWS, Azure & GCP — become a truly versatile data engineer.
6. **Certification Focused**
   Aligned with Google Cloud Professional Data Engineer certification exam.
7. **Lifetime Access**
   Access to course materials, recordings, and community support.
8. **Flexible Batches**
   Weekday, weekend, and fast-track batches available.

Enroll Now — Start Your GCP Data Engineering Journey!

www.databrickstraining.in | WhatsApp: +91-8500002025