



## **Data Engineering Training Program**

---

**Duration: 8 Months**

# 1. GCP Cloud (Google Cloud Platform)

## **Module 1: Introduction to GCP**

---

- ✓ Overview of Cloud Computing
- ✓ GCP Architecture and Core Services
- ✓ IAM (Identity and Access Management)
- ✓ Billing and Pricing

## **Module 2: Storage and Databases**

---

- ✓ Google Cloud Storage (GCS)
- ✓ BigQuery (Data warehouse)
- ✓ Cloud SQL vs Spanner vs Firestore

## **Module 3: Compute Services**

---

- ✓ Compute Engine
- ✓ Cloud Functions
- ✓ Cloud Run
- ✓ Kubernetes Engine (GKE) – intro only

## **Module 4: Data Engineering on GCP**

---

- ✓ Dataflow (Apache Beam)
- ✓ Pub/Sub
- ✓ Cloud Composer (Airflow on GCP)
- ✓ ETL/ELT Patterns using GCP

# 1. Python (Core)

## **Module 1: Python Basics**

---

- ✓ Variables, Data Types, and Operators
- ✓ Conditionals and Loops
- ✓ Functions and Modules

## **Module 2: Working With Data**

---

- ✓ Lists, Dictionaries, Tuples, Sets
- ✓ File Handling (CSV, JSON)
- ✓ Exception Handling

## **Module 3: Object-Oriented Program**

---

- ✓ Classes and Objects
- ✓ Inheritance and Polymorphism

## **3. Advanced Python**

### **Module 1: Intermediate Concepts**

---

- ✓ List/Dict Comprehensions
- ✓ Lambda, Map, Filter, Reduce
- ✓ Decorators and Generators

### **Module 2: Libraries for Data Engineering**

---

- ✓ pandas, numpy
- ✓ requests, json, os, datetime
- ✓ Logging and Error Handling

### **Module 3: Testing and Packaging**

---

- ✓ Unit Testing with unittest / pytest
- ✓ Virtual Environments
- ✓ Creating Packages

## **4. SQL**

### **Module 1: SQL Basics**

---

- ✓ SELECT, WHERE, ORDER BY, GROUP BY
- ✓ JOINS, Subqueries

## **Module 2: Advanced SQL**

---

- ✓ Window Functions
- ✓ CTEs
- ✓ Query Optimization

## **Module 3: Practice with BigQuery / PostgreSQL**

---

## **5. Databricks**

### **Module 1: Introduction to Databricks**

---

- ✓ Workspace and Notebooks
- ✓ Databricks Runtime & Clusters

### **Module 2: Data Engineering with Databricks**

---

- ✓ Delta Lake Concepts
- ✓ Spark in Databricks (DataFrame API)
- ✓ Notebooks for ETL Pipelines

### **Module 3: MLflow and Integration**

---

- ✓ MLflow for model tracking
- ✓ Databricks CLI and REST APIs

## **6. CI/CD (GitHub Actions)**

### **Module 1: Version Control with Git & GitHub**

---

- ✓ Git Basics (clone, commit, push, pull, merge)
- ✓ Branching Strategy

## **Module 2: GitHub Actions**

---

- ✓ Workflow Syntax (YAML)
- ✓ Build-Test-Deploy Pipelines
- ✓ Secrets & Environment Variables

## **7. Docker**

### **Module 1: Docker Fundamentals**

---

- ✓ Images, Containers, Volumes
- ✓ Dockerfile & Docker Compose

### **Module 2: Docker for Data Pipelines**

---

- ✓ Packaging Python/SQL jobs
- ✓ Docker with Databricks CLI or Spark Jobs

## **8. Apache Spark**

### **Module 1: Spark Basics**

---

- ✓ Spark Architecture (RDD, DAG, Executors)
- ✓ DataFrame API

### **Module 2: Spark SQL & Pyspark**

---

- ✓ Spark SQL, Caching, Partitioning
- ✓ PySpark for ETL/EDA

### **Module 3: Performance Tuning**

---

- ✓ Shuffle, Partitioning, Broadcast Join

## 9. Hive

### **Module 1: Hive Basics**

---

- ✓ Hive Architecture & Metastore
- ✓ HiveQL (DDL, DML)

### **Module 2: Advanced Hive**

---

- ✓ Partitions, Bucketing
- ✓ UDFs, Views, Indexes

## 10. Terraform

### **Module 1: IAC Basics**

---

- ✓ Terraform Core Concepts
- ✓ Providers, Resources, Modules

### **Module 2: GCP Infrastructure Automation**

---

- ✓ GCP-specific Terraform Resources (GCS, BigQuery, IAM)

### **Module 3: CI/CD with Terraform**

---

- ✓ Terraform State Management
- ✓ GitHub Actions + Terraform Workflow

## 11. Jira

### **Module 1: Agile Project Management**

---

- ✓ Boards, Epics, Stories, Tasks
- ✓ Scrum/Kanban Workflows

## **Module 2: Jira for Engineers**

---

- ✓ Creating and Managing Sprints
- ✓ Linking Commits & PRs to Jira Tasks

## **12. DBT (Data Build Tool)**

### **Module 1: DBT Basics**

---

- ✓ DBT Core vs Cloud
- ✓ Model, Seed, Snapshot, Sources

### **Module 2: SQL-based Transformations**

---

- ✓ Writing Modular SQL
- ✓ Jinja Macros and Tests

### **Module 3: Deployment & Orchestration**

---

- ✓ Running dbt via CLI
- ✓ dbt with Airflow or GitHub Actions

# Learning Management System

## Key Features **Learning Management System**



### **Live Sessions with Class Recordings**

Gain access to interactive live training sessions along with recorded classes to reinforce learning at your own pace.



### **Expert-Led Training Sessions**

Learn from industry experts through structured training sessions designed to enhance your skills and practical knowledge.



### **Earn Your Training Certificate**

Receive an official training certification upon successful course completion to validate your learning and expertise.



### **Experience Certificate for Professionals**

Get an experience certificate based on your hands-on project work and practical assessments.

