Learnomate Technologies is the Information technology company which provide training on different IT Technologies.
Out of that **GCP Data Engineer** is the one of the technology.
Course structure design in such a way that student will learn from Basic concepts to advance.

Welcome to the world of Google Cloud Platform (GCP) Data Engineering! In this training course, you'll dive into the exciting realm of managing and analyzing data on one of the world's most powerful cloud platforms.
This course is designed to equip you with the knowledge and skills necessary to excel in the role of a GCP Data Engineer. Whether you're new to the field or looking to enhance your existing expertise, this training program will provide you with a comprehensive understanding of GCP's data services, tools, and best practices.

Course Overview

Module 1: GCP Introduction

- Why we need Cloud.
- Overview of Google Cloud Platform (GCP)
- Key GCP Services and Products
- How to create Free Tier Account in GCP

Module 2: GCP Interfaces

- **Cloud Console**
  - ▶Navigating the GCP Console
  - ▶Configuring the GCP Console for Efficiency
  - ▶ Using the GCP Console for Service Management
- **Cloud Shell**
  - ▶ Introduction to GCP Shell
  - ▶ Command-line Interface (CLI) Basics
  - ▶ GCP Shell Commands for Service Deployment and Management
- **Cloud SDK**
  - ▶ Overview of GCP Software Development Kits (SDKs)
  - ▶ Installing and Configuring SDKs
  - ▶ Writing and Executing GCP SDK Commands

Module 3: GCP Locations

- **Regions**
  - ▶ Understanding GCP Regions
  - ▶ Selecting Regions for Service Deployment
  - ▶ Impact of Region on Service Performance

- **Zones**
  - ▶ Exploring GCP Zones
  - ▶ Distributing Resources Across Zones
  - ▶ High Availability and Disaster Recovery Considerations

- **Importance**
  - ▶ Significance of Choosing the Right Location
  - ▶ Global vs. Regional Resources
  - ▶ Factors Influencing Location Decisions

Module 4: GCP IAM & Admin

- **Identities**
  - ▶ Introduction to Identity and Access Management (IAM)
  - ▶ Users, Groups, and Service Accounts
  - ▶ Best Practices for Identity Management

- **Roles**
  - ▶ GCP IAM Roles Overview
  - ▶ Defining Custom Roles
  - ▶ Role-Based Access Control (RBAC) Implementation

- **Policy**
  - ▶ Resource-based Policies
  - ▶ Understanding and Implementing Organization Policies
  - ▶ Auditing and Monitoring Policies

- **Resource Hierarchy**
  - ▶ GCP Resource Hierarchy Structure
  - ▶ Managing Resources in a Hierarchy
  - ▶ Organizational Structure Best Practices

Module 5 :Linux Basics

- Overview of Linux
- Basic Command Line Interface (CLI)
  - Navigation: ls, cd, pwd
  - File operations: cp, mv, rm, mkdir, rmdir
  - Viewing file contents: cat, less, more, head, tail
- GCP service-related commands

Module 6: Python for Data Engineer

CHAPTER 1 - Python Basics

- Strings
- Operators
- Numbers (Int, Float)
- Booleans

CHAPTER 2 - Data Types & Data Structures

- Lists
- Tuple
- Dictionary
- Sets

CHAPTER 3 - Python Programming Constructs

- if, elif, else statements
- for loops, while loops
- Exception Handling
- File I/O operations

CHAPTER 4- Modular Programming in Python

- Functions
- Lambda Functions and Classes

Module 7 : Google Cloud Storage

- **Introduction to Cloud Storage**

▶Overview of Cloud Storage as a scalable and durable object storage service.

▶Understanding buckets and objects in Cloud Storage.

▶Use cases for Cloud Storage, such as data backup, multimedia storage, and website content delivery.

- **Cloud Storage Operations**

▶Creating and managing Cloud Storage buckets.

▶Uploading and downloading objects to and from Cloud Storage.

▶Setting access controls and permissions for buckets and objects.

- **Data Transfer and Lifecycle Management**

▶Strategies for efficient data transfer to and from Cloud Storage.

▶Implementing data lifecycle policies for automatic object deletion or archival.

▶Utilizing Transfer Service for large-scale data transfers.

- **Versioning and Object Versioning**

▶Enabling and managing versioning for Cloud Storage buckets.

▶Understanding how object versioning works.

▶Use cases for object versioning in data resilience and recovery.

- **Integration with Other GCP Services**

▶Integrating Cloud Storage with BigQuery for data analytics.

▶Using Cloud Storage as a data source for Dataflow and Dataproc.

▶Exploring options for serving static content on websites.

Module 8 : Cloud SQL

Introduction to Cloud SQL

- Overview of Cloud SQL as a fully managed relational database service.
- Supported database engines and use cases for Cloud SQL.

Creating and Managing Cloud SQL Instances

- Creating MySQL or PostgreSQL instances.
- Configuring database settings, users, and access controls.
- Importing and exporting data in Cloud SQL.

Backups and High Availability

- Configuring automated backups and performing manual backups.
- Implementing high availability with failover replicas.
- Strategies for restoring data from backups.

Scaling and Performance Optimization

- Vertical and horizontal scaling options in Cloud SQL.
- Performance optimization tips for database queries.
- Monitoring and troubleshooting database performance.

Integration with Other GCP Services

- Connecting Cloud SQL with App Engine, Compute Engine, and Kubernetes Engine.
- Using Cloud SQL as a backend database for applications.
- Data synchronization with Cloud Storage and BigQuery.

End to End Database migration Project

- Offline: Export and Import method
- Online: DMS method

Module 9 : BigQuery (SQL development)

Introduction to BigQuery

- Overview of BigQuery as a fully managed, serverless data warehouse.
- Use cases for BigQuery in business intelligence and analytics.
- Various method of creating table in BigQuery
- BigQuery Data Sources and File Formats
- Native table and External Tables

SQL Queries and Performance Optimization

- Writing and optimizing SQL queries in BigQuery.
- Understanding query execution plans and best practices.
- Partitioning and clustering tables for performance.

Data Integration and Export

- Loading data into BigQuery from Cloud Storage, Cloud SQL, and other sources.
- Exporting data from BigQuery to various formats.
- Real-time data streaming into BigQuery.

Configuring access controls and permissions in BigQuery.

- Implementing encryption for data in BigQuery.
- Auditing and monitoring for security compliance.

BigQuery Views

- Views, Materialized Views, Authorized Views

Integration with Other GCP Services

- Integrating BigQuery with Dataflow for ETL processes.
- Using BigQuery in conjunction with Data Studio for visualization.
- Building data pipelines with BigQuery and Composer.

Case Study-1: Spotify

Case Study-2: Social Media

Module 10 : DataProc (Pyspark Development)

- Introduction to DataProc
- Introduction to Hadoop and Apache Spark
- Understanding the difference between Spark and MapReduce
- What is Spark and Pyspark.
- Understanding Spark framework and its functionalities
- Overview of DataProc as a fully managed Apache Spark and Hadoop service.
- Use cases for DataProc in data processing and analytics.

Cluster Creation and Configuration

- Creating and managing DataProc clusters.
- Configuring cluster properties for performance and scalability.
- Preemptible instances and cost optimization.

Running Jobs on DataProc

- Submitting and monitoring Spark and Hadoop jobs on DataProc.
- Use of initialization actions and custom scripts.
- Job debugging and troubleshooting.

Integration with Storage and BigQuery

- Reading and writing data from/to Cloud Storage and BigQuery.
- Integrating DataProc with other storage solutions.
- Performance optimization for data access.

Scaling and Automation

- Autoscaling DataProc clusters based on workload.
- Using Dataprep or other tools for data preparation before processing.
- Automation and scheduling of recurring jobs.

Case Study-1: Data Cleaning of Employee Travel Records

End to End Batch Pyspark pipeline using Dataproc, BigQuery, GCS

Module 11 : DataFlow (Apache Beam development)

- Introduction to DataFlow
- Use cases for DataFlow in real-time analytics and ETL.
- Understanding the difference between Apache Spark and Apache Beam
- How Dataflow is different from Dataproc

Building Data Pipelines with Apache Beam

- Writing Apache Beam pipelines for batch and stream processing.
- Custom Pipelines and Pre-defined pipelines
- Transformations and windowing concepts.

Integration with Other GCP Services

- Integrating DataFlow with BigQuery, Pub/Sub, and other GCP services.
- Real-time analytics and visualization using DataFlow and BigQuery.
- Workflow orchestration with Composer.

End to End Streaming Pipeline using Apache beam with Dataflow, Python app, PubSub, BigQuery, GCS

Template method of creating pipelines

Module 12 : Cloud Pub/Sub

Introduction to Pub/Sub

- Understanding the role of Pub/Sub in event-driven architectures.
- Key Pub/Sub concepts: topics, subscriptions, messages, and acknowledgments.

Creating and Managing Topics and Subscriptions

- Using the GCP Console to create Pub/Sub topics and subscriptions.
- Configuring message retention policies and acknowledgment settings.

Publishing and Consuming Messages

- Writing and deploying code to publish messages to a topic.
- Implementing subscribers to consume and process messages from subscriptions.

Integration with Other GCP Services

- Connecting Pub/Sub with Cloud Functions for serverless event-driven computing.
- Integrating Pub/Sub with Dataflow for real-time stream processing.

Streaming use-case using Dataflow

Module 13 : Cloud Composer (DAG Creations)

Introduction to Composer

- Overview of Airflow Architecture
- Use cases for Composer in managing and scheduling workflows.

Creating and Managing Workflows

- Creating and configuring Composer environments.
- Defining and scheduling workflows using Apache Airflow.
- Monitoring and managing workflow executions.

Integration with Data Engineering Services

- Orchestrating workflows involving BigQuery, DataFlow, and other services.
- Coordinating ETL processes with Composer.
- Integrating with external systems and APIs.

Error Handling and Troubleshooting

- Handling errors and retries in Composer workflows.
- Debugging and troubleshooting failed workflow executions.
- Logging and monitoring for Composer workflows

- **Level-1-DAG: Orchestrating the BigQuery pipelines**
- **Level-2-DAG: Orchestrating the DataProc pipelines**
- **Level-3-DAG: Orchestrating the Dataflow pipelines**
- **Implementing CI/CD in Composer Using Cloud Build and GitHub**

Module 14: Cloud Functions

- **Cloud Functions Introduction**
- **Setting up Cloud Functions in GCP**
- **Event-driven architecture and use cases**
- **Writing and deploying Cloud Functions**
- **Triggering Cloud Functions:**
  HTTP triggers
  Pub/Sub triggers
  Cloud Storage triggers

Monitoring and logging Cloud Functions

Usecase-1: Loading the files from GCS to BigQuery as soon as it is uploaded.

Azur Data Engineering Tools

Module 15 : Azur Introduction

Introduction Azure Platform and Overview

Multiple Interaction Methods

- Console
- Shell
- SDK

Azure Account Creation

Service Overview, subscription and Resource Container

Azure DataLake Storage (ADLS)

Introduction to Azure Data Lake Storage (ADLS)

- Core Components of ADLS Gen2
- Blobs, Containers, and Storage Accounts
- Hierarchical Namespace o Why Use Azure Data Lake Storage?
- Scalability, Cost-Effectiveness, Security
- Integration with Analytics Services

Access and Management of ADLS

- Data Management via Console, SDK, CLI, REST API
- Role-Based Access Control (RBAC) and Access Control Lists (ACLs)

Pricing Overview

- Pricing Tiers (Hot, Cool, Archive)
- Cost Factors

Best Practices for ADLS

- Data Partitioning
- Tiering Strategy for Cost Optimization
- Monitoring and Usage Tracking

Interactive Access Methods

- Managing ADLS via Console
- CLI/Shell Interaction
- Programmatic Access via SDKs

Azure SQL DB

- **Introduction to Azure SQL Database**
- **Key Features**
- **Creating a SQL db for practice**

Module 16 : AzurData Factory

What is ADF and Use cases

Core components of ADF and understanding & configuring

- Integration runtime
- Linked Services
- Datasets
- Pipelines
- Activities
- Triggers

Key Features of Azure Data Factory

Pipelines in ADF

- Creating and Configuring Pipelines
- Activities and Triggers
- Monitoring and Managing Pipeline Runs
- Orchestration

Use case: End to End Incremental pipeline with ADF, ADLS, SQL DB.

Best Practices for ADF

- Designing Efficient Pipelines
- Data Partitioning and Parallelism

Monitoring and Optimization for Performance

Module 17 : Databricks on Azure/GCP

Introduction to Databricks Lakehouse Platform

Databricks Architecture and Core Components

- Explore the architecture, including clusters, jobs, and the workspace environment.
- Key components like Databricks Runtime, Spark, and Delta Lake.

Setting Up and Administering a Databricks Workspace

- Step-by-step guide on workspace creation and management.
- Admin tasks including user management, permissions, and billing.

Databricks Unity Catalog

- Unified governance solution for managing data and metadata across the lakehouse. Organizing data assets and enforcing security policies.

Managing Data with Delta Lake

- Understanding Delta Lake's ACID transactions and versioning capabilities.
- Best practices for efficient data management.
- Creating Delta Lake Tables
- Querying and managing data across multiple file formats.

Working with Notebooks and Clusters

- Collaborative data science using Databricks notebooks.
- Cluster setup, management, and autoscaling.

Building ELT Pipelines with Spark SQL and Python

Performance Optimization in Databricks

Incremental Data Processing with Delta Lake

Real-time data processing with Spark Structured streaming

Databricks Autoloader for file-based ingestion utility

Case Study: Creating an End-to-End Incremental Workflow

By the End of the course What Students can Expect

Proficient in SQL Development:

- Mastering SQL for querying and manipulating data within Google BigQuery and Cloud SQL.
- Writing complex queries and optimizing performance for large-scale datasets.
- Understanding schema design and best practices for efficient data storage.

Pyspark Development Skills:

- Proficiency in using PySpark for large-scale data processing on Google Cloud.
- Developing and optimizing Spark jobs for distributed data processing.
- Understanding Spark's RDDs, Data Frames, and transformations for data manipulation.

Apache Beam Development Mastery:

- Creating data processing pipelines using Apache Beam.
- Understanding the concepts of parallel processing and data parallelism.
- Implementing transformations and integrating with other GCP services.

DAG Creations with Cloud Composer:

- Designing and implementing Directed Acyclic Graphs (DAGs) for orchestrating workflows.
- Using Cloud Composer for workflow automation and managing dependencies.
- Developing DAGs that integrate various GCP services for end-to-end data processing.

Architecture Planning:

- Proficient in architecting end-to-end data solutions on GCP and Azure.
- Understanding the principles of designing scalable, reliable, and cost-effective data architectures.

Certification Readiness

- Prepare for the Google Cloud Professional Data Engineer (PDE) and
- Associate Cloud Engineer (ACE) certifications through a combination of theoretical knowledge and hands-on experience.

Contact Details

If you required any further information, please fill free to contact us.

Learnomate Technologies Pvt. Ltd

- **Main Branch:**

(Sai Luxuria, Office No 15, 3rd Floor,Bhumkar Chowk,
Wakad, Pune, Maharashtra, 411057 India)

Contact Details:

Call/WhatsApp: +91 7757062955
+91 7822917585
Email: info@learnomate.org

- **Kalewadi Branch.**

Office no.216, Solitaire business hub, 2nd floor, Kaspate Wasti, Wakad, Pune,
Maharashtra 411057

Contact Details:

Call/WhatsApp: +91 8983069523

Email: info@learnomate.org

THANK YOU

FOLLOW US

https://www.youtube.com/@learnomate

https://www.linkedin.com/company/learnomatetechnologies/

https://www.facebook.com/learnomate

https://www.instagram.com/learnomate/