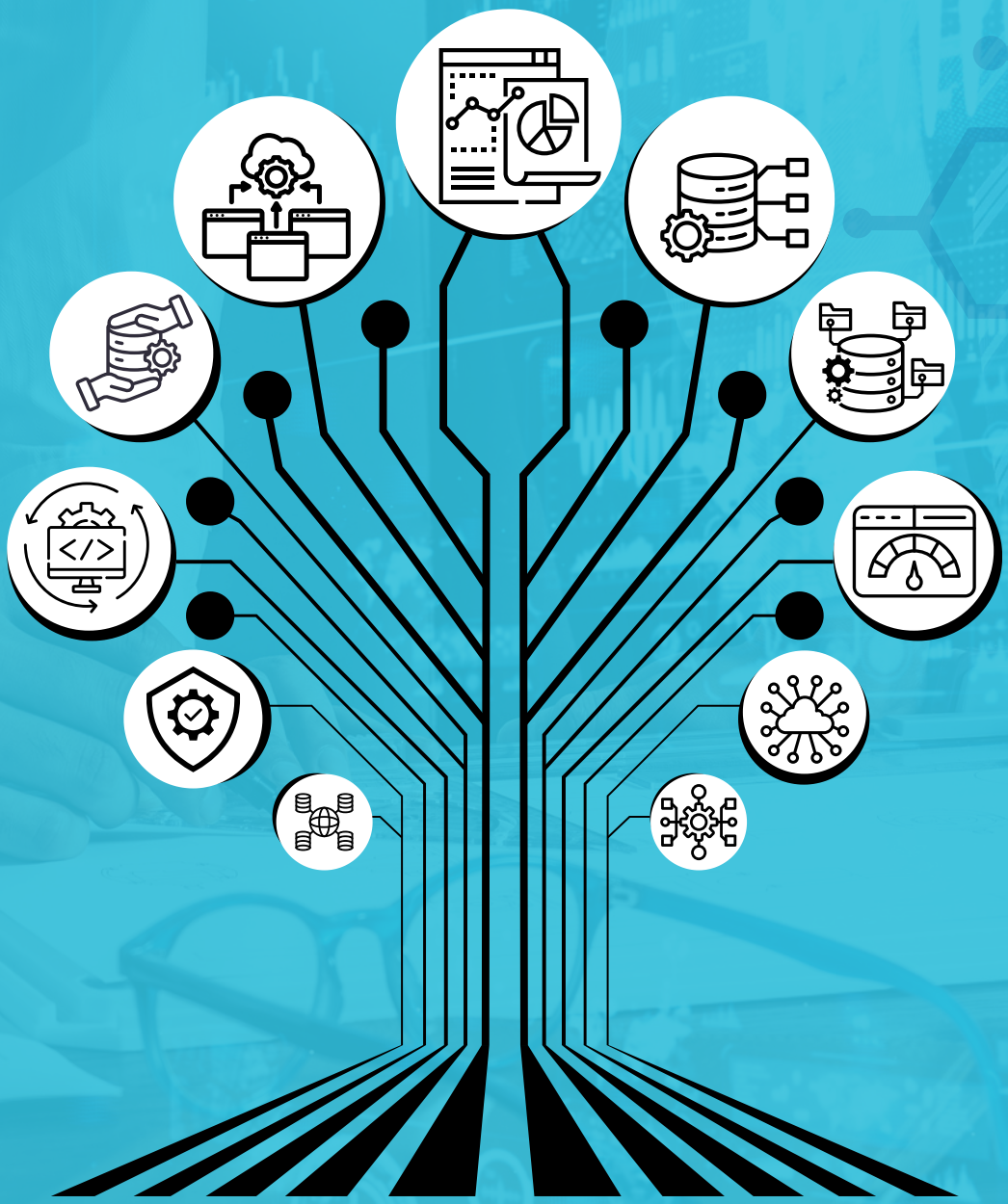
LEARNOMATE TECHNOLOGIES

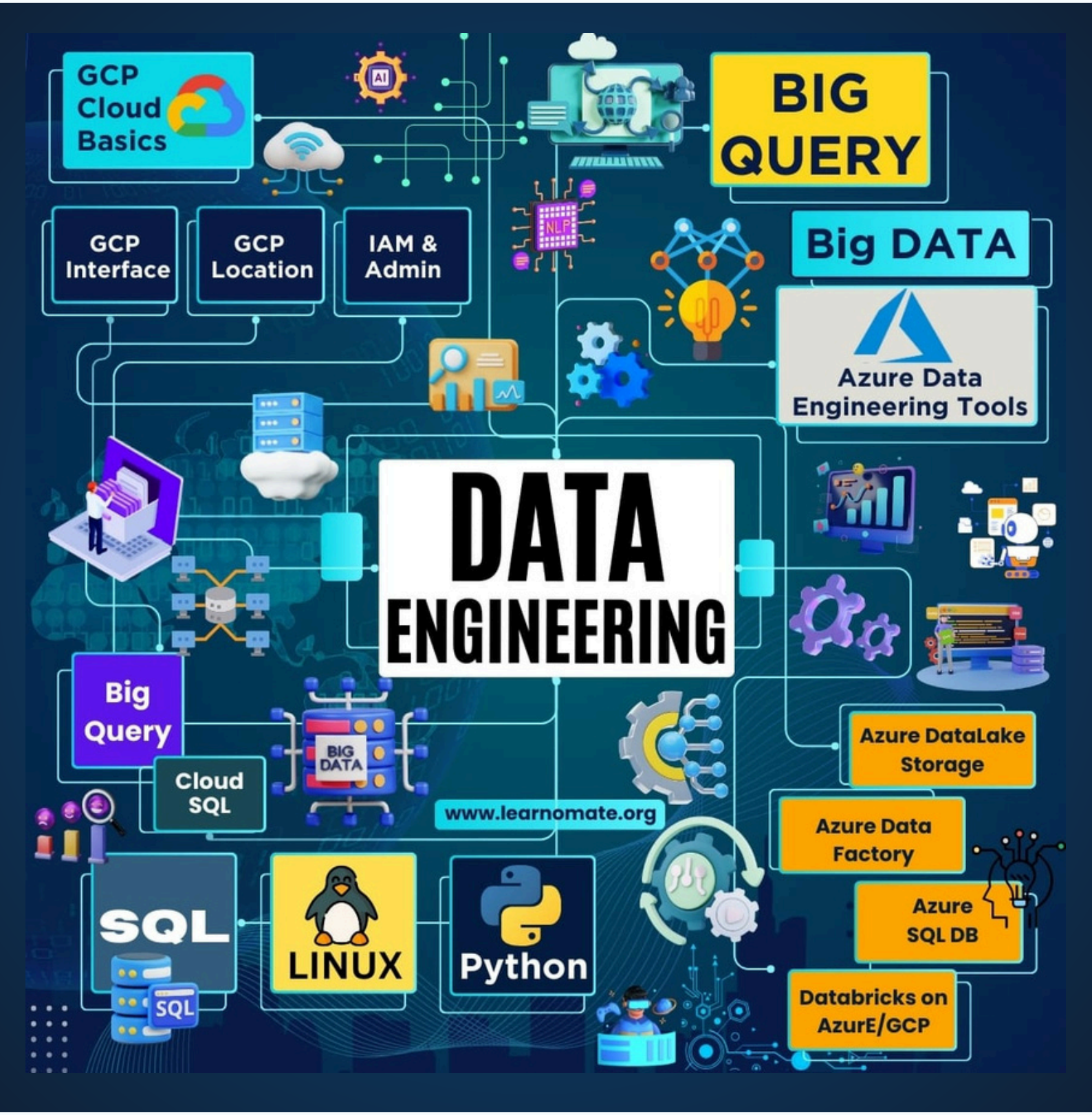


DATA ENGINEER SYLLABUS

+91 9325408926,+91 9225093995

info@learnomate.org

DATA ENGINEER ROADMAP



+91 9325408926,+91 9225093995 info@learnomate.org

About Learnomate

At Learnomate Technologies, we take pride in being one of the best software training institutes, offering a wide range of IT training programs and services that cater to learners across the globe. Our mission is simple: to empower aspiring IT professionals with the skills they need to excel in today's competitive tech world. With our global reputation, we are known for delivering high-quality training that transforms careers.

Our CEO, Ankush Thavali, is a renowned Oracle DBA expert, and his training programs are widely recognized for their depth, practicality, and success rates. Backed by a team of highly skilled trainers, we provide the best training material and an exceptionally supportive team to ensure a seamless learning journey for our students.

At Learnomate, we believe that financial constraints should never stop you from achieving your IT dreams. That's why we offer flexible payment options, including EMI and installment plans, making quality IT education accessible to everyone. Whether you're an aspiring professional or looking to upskill, our programs are designed to suit your needs and help you succeed.

Join us today and experience why thousands of students worldwide trust Learnomate Technologies to be their partner in IT growth. Let's build your future together!

What we offer in this training?

Data drives decision-making in today's digital era, and Data Engineers play a pivotal role in ensuring seamless data workflows. Our Data Engineering Training equips you with the skills to design, build, and maintain data pipelines while working with tools like Azure Data Factory, Google Cloud Platform (GCP), Hadoop, Spark, and more. From handling massive datasets to creating optimized pipelines, this course prepares you for the growing demand for data engineers.

Salient Features:

- Comprehensive training on cloud platforms like Azure and GCP.
- Practical exposure to big data tools like Hadoop and Spark.
- Real-time projects to simulate industry work environments.
- Insights into data modeling, ETL processes, and scalability.
- Guidance on how to ace data engineering interviews.

Duration: 3 Months

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.

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.

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.

Access Controls and Security

Configuring access controls and permissions in BigQuery.

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

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.

Module 10 : DataProc (Pyspark Development)

Introduction to DataProc

- 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.

Module 11 : DataFlow (Apache Beam development)

Introduction to DataFlow

- Overview of DataFlow as a fully managed stream and batch processing service.
- Use cases for DataFlow in real-time analytics and ETL.

Building Data Pipelin es with Apache Beam

- Writing Apache Beam pipelines for batch and stream processing.
- Transformations and windowing concepts.
- Error handling and testing of DataFlow pipelines.

Monitoring and Optimization

- Monitoring and troubleshooting DataFlow pipelines.
- Optimizing pipeline performance and resource utilization.
- Utilizing DataFlow templates for reusable pipelines.

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.

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.

Error Handling and Retry Policies

- Configuring error handling mechanisms.
- Implementing retry policies for fault-tolerant message processing.

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.

Module 13 : Cloud Composer (DAG Creations

Introduction to Composer

- Overview of Composer as a fully managed workflow orchestration service.
- 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.

Extending and Customizing Composer

- Extending Apache Airflow with custom operators and sensors.
- Creating and managing Composer plugins.
- Versioning and managing workflow code.

Error Handling and Troubleshooting

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

Module 14 : Terraform

Terraform Basics

- Installing and configuring Terraform.
- Writing Terraform configurations using HashiCorp Configuration Language (HCL).
- Initializing and applying Terraform configurations.

Infrastructure Provisioning

- Creating and managing infrastructure resources using Terraform.
- Terraform state and remote backends.
- Importing existing infrastructure into Terraform.

Module and Provider Usage

- Organizing Terraform configurations using modules.
- Utilizing different providers for various cloud services.
- Best practices for reusable and modular Terraform code.

Variables, Outputs, and Functions

- Defining and using variables in Terraform.
- Outputting values from Terraform configurations.

Terraform Workflow and Best Practices

- Terraform workflows: plan, apply, and destroy.
- Managing Terraform environments and workspaces.

Module 15 : Azur Data Factory

Overview of Azure Data Factory

- What is ADF?
- Key components and architecture
- Common use cases

Basic Concepts

- Pipelines, activities, and triggers
- Data integration and orchestration

Working with Azure Data Factory

▶Linked Services and Datasets

- Configuring linked services
- Creating and managing datasets

▶Pipelines and Activities

- Building and running pipelines
- Exploring different types of activities (data movement, data transformation etc.)

▶Triggers and Scheduling

- Types of triggers (scheduled, tumbling window, event-based)
- Configuring and managing triggers

▶ Data Transformation and Mapping Data Flows

- Data Transformation Basics
- Understanding data transformation activities
- Using the Copy Data activity

▶Mapping Data Flows

- Introduction to mapping data flows
- Designing and implementing data flows
- Data flow transformations (e.g., join, aggregate, filter, etc.

Module 16 : Azure Databricks

Overview of Azure Databricks

- What is Databricks?
- Key features and architecture
- Common use cases

Workspace, Notebooks and Clusters

- Workspace creation
- Creating and managing notebooks
- Understanding clusters and cluster management

Working with Data in Databricks

- **Data Ingestion and Exploration**
  - Connecting to data sources (Azure Blob Storage, Azure Data Lake, etc.)
  - Loading and exploring data in Databricks

Data Frames and Spark SQL

- Introduction to Data Frames
- Performing operations with Data Frames
- Using Spark SQL for querying data

Basic ETL and workflows in Databricks

- Writing ETL scripts in Databricks
- Transforming and cleaning data

Advanced Databricks Features

- Delta Lake and ACID transactions
- Delta tables
- Optimizing and tuning Spark jobs
- Unity Catalog
- Auto loader

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.

LEARNOMATE TECHNOLOGIES

www.learnomate.org

A Steps towards the bright future

Training Highlights

- Online Training available
- All sessions are practical based
- Recording Access shared to students
- Job Assistance
- Resume Preparation
- Dedicated Support Team to solve issues

Course Details

- Trainer: Saidul Sir
- Duration: 3 Months

+91 9325408926,+91 9225093995 info@learnomate.org

LEARNOMATE TECHNOLOGIES

www.learnomate.org

A Steps towards the bright future

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)

Call/WhatsApp: +91 9325408926, +91 9225093995

- **Kalewadi Branch.**

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

Call/WhatsApp: +91 8983069523

- **Kharadi Branch**

City Vista Buisness park. A Wing, 3rd floor, Office No. 12A. Fountain Road, Ashoka Nagar, Kharadi, Pune, Maharashtra 411014

Call/WhatsApp: +91 9325408926,+91 9225093995

LEARNOMATE TECHNOLOGIES

www.learnomate.org

A Steps towards the bright future

THANK YOU

FOLLOW US

Join Learnomate Technologies Channel

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

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

https://www.facebook.com/learnomate

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

+91 9325408926,+91 9225093995 info@learnomate.org