Databricks Basics

Duration – 16 Hours

Program Description

This Databricks training program focuses on building expertise in data analytics and processing within the Databricks environment. It begins with mastering SQL operations for data transformations and analytics, followed by foundational Python concepts using libraries like Pandas, NumPy, Matplotlib, and Seaborn for data handling and visualization.

Participants will learn to set up and navigate a Databricks workspace on Azure, covering initial data ingestion and environment setup.

Practical skills in file handling, including using Auto Loader for file streaming and working with various file formats, are also included. The capstone project consolidates these skills by building an end-to-end data analytics and machine learning pipeline for related data

Learning Goals

- Perform advanced SQL operations within Databricks, including Joins, Subqueries, Common Table Expressions, and Window Functions, for data analytics and transformation tasks.
- Leverage Python libraries like Pandas, NumPy, Matplotlib, and Seaborn for data manipulation and visualization within the Databricks environment.
- Set up and manage Databricks environments on Azure, understand the Intelligent Lakehouse architecture, and perform efficient data ingestion.
- Utilize Databricks Auto Loader to handle and process various file formats such as CSV, JSON, AVRO, and Parquet.
- Integrate Databricks features to design and deploy a complete end-to-end pipeline tailored for respective related data analytics

Course Topics

- SQL Essentials for Databricks - Joins, Subqueries, Common Table Expressions, Window Functions
- Minimal Python for Databricks - Pandas, NumPy, Matplotlib, Seaborn
- Introduction to Databricks Environment - Setting up Databricks Account on Azure, Databricks Intelligent Lakehouse Environment, Data Ingestion
- File Handling in Databricks - Auto Loader, Handling CSV, JSON, AVRO, Parquet Files
- Capstone Project - End-to-End Databricks Pipeline for related Data Analytics

Databricks Intermediate

Duration – 16 Hours

Program Description

This Databricks training program focuses on mastering SQL and Delta Table functionalities for dynamic dataset management, including schema evolution and time travel.

Participants will learn to design streaming pipelines and automate workflows effectively.

The program also covers creating reliable ETL pipelines with Delta Live Tables and ensuring data quality through incremental updates. Governance and security are addressed with Unity Catalog, enabling centralized control, lineage tracking, and collaboration.

The capstone project consolidates these concepts by building scalable real-time pipelines for related data, ensuring compliance, and delivering actionable insights.

Learning Goals

- Equip participants with advanced SQL and Delta Table skills to manage dynamic datasets, including schema evolution, schema drift, and time travel.
- Enable participants to design real-time streaming pipelines and automate workflows using Databricks tools.
- Develop expertise in building automated, reliable ETL pipelines with Delta Live Tables and ensuring data quality.
- Provide knowledge of centralized data governance and compliance using Unity Catalog for secure and collaborative data management

Course Topics

- Databricks SQL and Delta Tables - Advanced SQL Queries, Delta Tables, Time Travel, Schema Evolution, Schema Drift
- Streaming and Workflows - Streaming Data Ingestion, Structured Streaming, Job Scheduling, Workflows, Task Dependencies
- Delta Live Tables - Creating Delta Live Tables, Incremental Updates, Data Quality Checks
- Unity Catalog - Data Governance, Lineage Tracking, Access Controls, Multi-Workspace Collaboration
- Capstone Project - End-to-End Databricks Pipeline for related Data Analytics

Databricks Advanced

Duration – 24 Hours

Program Description

This advanced Databricks training program emphasizes optimizing large-scale data processing through performance tuning and query optimization.

Participants will master real-time data streaming using Kafka/Event Hubs, integrate streaming data into Delta Lake, and perform advanced analytics.

This course also covers scalable machine learning, focusing on distributed training and hyperparameter tuning. Security and monitoring are reinforced with Unity Catalog policies and workflow tracking.

The capstone project consolidates these skills by building an optimized, scalable, and secure data pipeline for relevant datasets, providing real-time insights trials, monitoring, and regulatory compliance.

Learning Goals

- Equip with advanced skills in performance tuning, partition pruning, and query optimization for large-scale data processing on Databricks.
- Develop expertise in integrating and analysing real-time streaming data from Kafka/Event Hubs into Delta Lake.
- Learn to train and optimize machine learning models at scale using Databricks, including distributed training and hyperparameter tuning.
- Gain proficiency in designing secure, scalable, and optimized data pipelines with real-time insights, governance, and workflow monitoring, tailored to specific industries

Course Topics

- Advanced Databricks Optimization - Performance Tuning, Partition Pruning, Query Optimization
- Advanced Streaming in Databricks - Event Hubs/Kafka Integration, Streaming Analytics
- Databricks ML at Scale - Distributed Training, Hyperparameter Tuning
- Security and Monitoring in Databricks - Unity Catalog Policies, Monitoring Workflows
- Capstone Project - Optimized Data Pipeline with Real-Time Insights