CTR-911 DataBricks for Data Engineers

Overview

Course Duration: 50 Hours

About This Course

This course provides a practical introduction to Databricks ,a leading cloud-based platform for big data processing, analytics, and AI. Students will learn how to build, manage, and optimize data pipelines using Apache Spark ,explore data transformation and visualization ,and implement machine learning workflows directly within the Databricks environment. The course combines hands-on labs with real-world examples to prepare learners for roles in data engineering, data science, and analytics .

Audience Profile

Data professionals who wish to expand their knowledge into the realms of Big Data, AI, and MLOps using Databricks tools.

At Course Completion

participants will have full proficiency with the Databricks platform across all key domains, with an emphasis on democratizing data using tools such as Data Lake, including:

- Understanding the architecture of Databricks and the fundamentals of Apache Spark, including concepts like clusters, notebooks, and workspaces
- Working with Data Frames and Spark SQL to process and transform data into various formats (CSV, JSON, Parquet, Delta)
- Gaining in-depth knowledge of Delta Lake, including data versioning (Time Travel) and ACID transactions, and applying performance optimization techniques
- Integrating with Azure services, including Azure Data Lake, Azure Data Factory, and Power BI, with a focus on permissions and security
- Designing, developing, and presenting a comprehensive final project—an ETL/ELT pipeline—including automation, scheduling, monitoring, and error handling

Course Outline

Module 1: Introduction to Databricks and Spark Fundamentals

- Overview of Azure Databricks platform and architecture
- Introduction to Apache Spark and its ecosystem
- Understanding clusters, notebooks, and workspaces
- Setting up the environment and exploring the Databricks interface

A practical foundation to the Databricks environment and the Spark ecosystem, understanding distributed processing and preparing the working environment.

Module 2: Working with Data Frames and Spark SQL

- Understanding Data Frames and Datasets
- Data ingestion from CSV, JSON, Parquet, and Delta formats
- Transformations, actions, and schema management
- Querying data using Spark SQL

Focuses on core Spark APIs, transformations, and data manipulation through SQL-style operations and schema definitions.

Module 3: Advanced Transformations and Delta Lake

- Delta Lake fundamentals and ACID transactions
- Data versioning and time travel
- Building and maintaining reliable data pipelines
- Performance optimization techniques

Deep dive into Delta Lake and its importance in ensuring data reliability, consistency, and performance in large-scale systems.

Module 4: Integration with Azure Services

- Connecting Databricks to Azure Data Lake and Blob Storage
- Automating data ingestion with Azure Data Factory
- Working with Power BI and visualization tools
- Managing permissions and authentication

Hands-on integration with Azure ecosystem tools, connecting Databricks pipelines with real data sources and visualization dashboards.

Module 5: Data Pipeline Project & Automation

- Designing a complete ETL/ELT pipeline
- Job scheduling and orchestration in Databricks
- Monitoring and error handling
- Final project implementation and presentation

Full project day — participants will design, implement, and present a complete end-to-end data pipeline on Databricks.