Azure Databricks with Pyspark

Duration: 60 hours

Overview

This course provides a practical deep dive into PySpark and Azure Databricks, equipping learners with essential skills for big data engineering. Starting with PySpark fundamentals, participants explore Spark's architecture, transformations, and actions, before diving into advanced data processing with SQL, MLlib, and ETL pipelines. The course then covers Azure Databricks, its Lakehouse and Medallion architectures, and Delta Lake optimizations, leading to hands-on experience in building scalable data workflows using PySpark in Databricks notebooks.

Pre-requisites

- Basic knowledge of Python & SQL (prior experience with PySpark is beneficial but not mandatory).
- Fundamentals of data engineering & cloud platforms (Azure, AWS, or GCP experience is a plus).

Course Syllabus

Module 1: Introduction to PySpark

- What is PySpark?
- Components of Spark: Driver, Executors, Cluster Manager
- Spark RDDs vs. DataFrames vs. Datasets
- Spark Transformations and Actions
- Setting Up Development Environment (Colab Free Tier, Local Machine optional)
- Hands-on: Running PySpark code on local and cluster mode

Module 2: PySpark in Detail

- Working with DataFrames in PySpark
- PySpark SQL and DataFrame API
- Data Manipulation in PySpark (Filtering, Joins, Aggregations)
- PySpark MLlib (Introduction to Machine Learning in Spark)
- Hands-on: Writing PySpark scripts for data transformation

Module 3: Azure Databricks Architecture & Lakehouse

- Introduction to Azure Databricks
- Databricks Components: Clusters, Notebooks, Jobs, Workflows Databricks File System (DBFS) and Data Management
- Introduction to Lakehouse Architecture
- Benefits of Lakehouse vs. Traditional Data Warehousing
- Hands-on: Creating Databricks clusters and working with DBFS

Module 4: Azure Databricks Medallion Architecture

- Introduction to Medallion Architecture (Bronze, Silver, Gold Layers)
- Data Ingestion Strategies in Databricks
- Implementing ETL Pipelines in Databricks
- Delta Lake Overview and Optimizations
- Hands-on: Building a Medallion Architecture-based Data Pipeline

Module 5: PySpark in Databricks Notebooks

- Using PySpark in Databricks Notebooks
- Writing and Executing PySpark Code in Databricks
- Databricks Widgets and Parameterization
- Visualizing Data in Databricks (Matplotlib, Seaborn, Plotly)
- Automating Workflows with Databricks Jobs
- Hands-on: Implementing a complete Data Pipeline with PySpark in Databricks