Complete guide to Azure Databricks with PySpark

Duration: 40 hours

Course Overview

This comprehensive course provides an in-depth understanding of Azure Databricks and its applications within the big data and analytics landscape. Starting with an introduction to Azure Databricks and its role alongside other Azure tools like Synapse and Data Lake, followed by a detailed exploration of its architecture, components, and integration points. Participants will gain hands-on experience with PySpark for data transformations and advanced optimizations, delve into Delta Lake for efficient data management, and learn how to build modular ETL pipelines. The course also covers real-world use cases, orchestration with Azure services, monitoring and tuning for cost and performance, and essential security and deployment strategies. Each module is reinforced with practical labs to solidify learning, enabling participants to confidently apply their skills in data engineering and analytics tasks.

Course contents

Module 1: Introduction to Azure Databricks & Big Data Landscape

- What is Azure Databricks? Why use it?
- Overview of Azure Synapse, Azure Data Lake, and Azure Databricks
- Apache Spark on Azure Databricks: Evolution & Benefits
- Azure Use Cases in Data Engineering

Hands-on Lab:

- Creating an Azure Databricks Workspace
- Launching your first cluster and notebook

Module 2: Azure Databricks Architecture Deep Dive

- Components: Control Plane vs Data Plane
- Azure Integration Points (Key Vault, AAD, Data Lake Gen2, Event Hub)
- Workspace, Repos, Jobs, Pools, Tokens
- Cluster Modes: Standard, High Concurrency, Job Clusters

Hands-on Lab:

- Explore workspace UI
- Create and configure different cluster types

Module 3: PySpark on Azure Databricks

- PySpark Basics: SparkSession, DataFrames, RDDs
- DataFrame Transformations, Actions, and UDFs
- Working with Struct, Arrays, Maps
- Integrating with Azure Data Lake Gen2 using PySpark

Hands-on Lab:

- Load, transform, and write data to ADLS Gen2
- Use UDFs for business rules

Module 4: Advanced PySpark & Optimizations

- Window Functions and Complex Aggregations
- Joins, Broadcast Optimization
- Partitioning, Bucketing, and Caching
- Performance Tips using Spark UI

Hands-on Lab:

- Optimize a slow transformation job
- Use cache/persist effectively

Module 5: Delta Lake on Azure Databricks

- Introduction to Delta Lake in Azure
- Medallion Architecture: Bronze, Silver, Gold
- ACID Transactions and Schema Evolution
- Time Travel and VACUUM

Hands-on Lab:

- Create Delta tables on ADLS Gen2
- Implement schema evolution and time travel

Module 6: ETL Pipeline Development on Azure Databricks

- Building Modular ETL Pipelines with PySpark
- Auto Loader and COPY INTO for Inges on
- Data Transformation with Delta Lake
- Data Quality using Expectations (Delta Live Tables preview)

Hands-on Lab:

- Develop an end-to-end ETL pipeline from raw to gold layer
- Schedule and parameterize the job

Module 7: Data Use Cases in Azure Databricks

- Cleaning and preparing customer data
- Generating daily reports from transaction logs
- Validating and filtering data from CSV files
- Simple trend analysis on historical sales data

Hands-on Labs:

- Build simple use cases with sample datasets
- Implement transformations, filters, and basic aggregations

Module 8: Simple Orchestra on with Azure Services

- Run Databricks notebooks using Azure Data Factory
- Schedule jobs and monitor runs
- Use REST APIs for job automation

Hands-on Lab:

- Create a pipeline in ADF to run a notebook
- Monitor job status and logs from Azure

Module 9: Monitoring, Logging & Tuning in Azure Databricks

- Debugging with Spark UI & Azure Log Analytics
- Cost Optimization: Auto-Termination, Spot Instances
- Adaptive Query Execution and AQE Plans
- Monitoring with Azure Monitor, Alerts & Diagnostics

Hands-on Lab:

- Use Spark UI to troubleshoot job slowness
- Log metrics to Azure Monitor

Module 10: Security and Deployment in Azure Databricks

- Set up access control with Azure AD

- Store and use secrets with Azure Key Vault
- Understand Unity Catalog basics
- Deploy notebooks using Git integration

Hands-on Lab:

- Manage notebook access and permissions
- Use secrets in a Databricks job
- Link a Git repo to Databricks for version control