Advanced Azure Databricks Course Content

Duration: 30 hours

Module 1: Python and PySpark Fundamentals

- Introduction to Python programming
- Jupyter Notebooks
- Setting up PySpark environment in Databricks
- Basic PySpark operations
- Data Frames

Module 2: Introduction to Azure Databricks

- Overview of Azure Databricks
- Features and advantages of Databricks
- Understanding Apache Spark
- Databricks architecture
- Creating Databricks workspaces
- Managing Databricks workspaces

Module 3: Spark SQL

- Introduction to Spark SQL
- Data Frame creation
- Working with joins
- Aggregations
- Window functions
- Userdefined functions (UDFs)

Module 4: Databricks Interface and Setup

- User Interface (UI), CLI, and REST API
- Creating an Azure free account
- Databricks workspace
- Exploring the Azure portal
- Databricks community workspace

Module 5: Clusters and Notebooks

- Types of clusters: Job and Interactive
- Creating and configuring clusters
- Creating, importing, exporting, and managing notebooks

Module 6: Data Ingestion Techniques

- Loading data from various sources (CSV, JSON, Parquet)
- Working with Azure Blob Storage
- Working with Azure Data Lake Storage (ADLS)
- Creating and managing mount points for data access

Module 7: Data Management in Databricks

- Understanding DBFS
- Databases and tables
- Creating and managing databases and tables
- Working with Hive tables in Databricks
- Unity Catalog: Managing Data Access and Governance

Module 8: Data Processing with Apache Spark

- Introduction to Spark SQL
- DataFrame API
- Data wrangling
- Transformations, aggregations, and joins
- Handling missing data and data quality issues

Module 9: Delta Lake Fundamentals

- Understanding Delta Lake and its advantages

- ACID transactions, time travel, and data versioning
- Managing and optimizing data storage with Delta Lake
- Delta Live Tables: Simplifying ETL with Declarative Pipelines

Module 10: Databricks Utilities and Widgets

- Using dbutils for file and library management
- Creating and managing widgets for dynamic parameters
- Notebook parameters and interactivity

Module 11: CLI and Automation

- Installing and using the Databricks CLI
- Automating cluster management and notebook operations
- Creating and managing jobs with the CLI

Module 12: Integrating with Azure Services

- Integrating Databricks with Azure Data Factory
- Accessing data from Azure SQL Database
- Working with Azure Synapse Analytics

Module 13: RealTime Stream Processing

- Introduction to Spark Streaming
- Processing realtime data with Databricks
- Integrating with Apache Kafka for event streaming
- Structured Streaming
- Building Scalable Streaming Applications

Module 14: Advanced Databricks Features

- Job scheduling and automation
- Model Management with MLflow
- Tracking and Versioning
- Interactive dashboards and visualization tools

Module 15: Machine Learning with MLFlow in Databricks

- Overview of MLflow
- Setting up MLflow in Databricks
- Logging parameters, metrics, and artifacts
- Viewing experiment runs in the MLflow UI
- Registering models in the MLflow Model Registry
- Preprocessing Data for Machine Learning
- Data cleaning and transformation techniques
- Feature engineering and selection

Module 16: Deploying Machine Learning Models in Databricks

- Deploying Machine Learning Models in Databricks
- Understanding model deployment
- Configuring deployment settings and endpoints
- Setting up model serving with MLflow
- Monitoring model performance post deployment
- Implementing A/B testing for model comparisons