Implement a Data Analytics Solution with Azure Databricks

Courses Code: DP-3011
Duration: 1 Day
Delivery Mode: Instructor-led Training (ILT) | Virtual Instructor led Training (VILT)

Overview

This course explores how to use Databricks and Apache Spark on Azure to take data projects from exploration to production. You'll learn how to ingest, transform, and analyze large-scale datasets with Spark Data Frames, Spark SQL, and PySpark, while also building confidence in managing distributed data processing. Along the way, you'll get hands-on with the Databricks workspace—navigating clusters and creating and optimizing Delta tables. You'll also dive into data engineering practices, including designing ETL pipelines, handling schema evolution, and enforcing data quality. The course then moves into orchestration, showing you how to automate and manage workloads with Lake flow Jobs and pipelines. To round things out, you'll explore governance and security capabilities such as Unity Catalog and Purview integration, ensuring you can work with data in a secure, well-managed, and production-ready environment.

Skills Covered

- Use the Azure Databricks workspace to ingest, transform, and analyse large-scale datasets with Spark DataFrames, Spark SQL and PySpark.
- Navigate and configure Spark clusters and Delta tables (including schema enforcement, versioning and time-travel).
- Design and build production-ready ETL/ELT data pipelines using declarative constructs (e.g., Delta Live Tables) and orchestration mechanisms (e.g., Lakeflow Jobs & pipelines).
- Implement governance, security and management features (such as Unity Catalog and integration with Microsoft Purview) to ensure your data-lakehouse solution is production-ready.

Who Should Attend

Before taking this course, learners should already be comfortable with the fundamentals of Python and SQL. This includes being able to write simple Python scripts and work with common data structures, as well as writing SQL queries to filter, join, and aggregate data. A basic understanding of common file formats such as CSV, JSON, or Parquet will also help when working with datasets. In addition, familiarity with the Azure portal and core services like Azure Storage is important, along with a general awareness of data concepts such as batch versus streaming processing and structured versus unstructured data. While not mandatory, prior exposure to big data frameworks like Spark, and experience working with Jupyter notebooks, can make the transition to Databricks smoother.

Prerequisites

- Before starting this, you should already be comfortable with the fundamentals of Python and SQL. This includes being able to write simple Python scripts and work with common data structures, as well as writing SQL queries to filter, join, and aggregate data. A basic understanding of common file formats such as CSV, JSON, or Parquet will also help when working with datasets.
- In addition, familiarity with the Azure portal and core services like Azure Storage is important, along with a general awareness of data concepts such as batch versus streaming processing and structured versus unstructured data. While not mandatory, prior exposure to big data frameworks like Spark, and experience working with Jupyter notebooks, can make the transition to Databricks smoother.

Course Modules

1. Explore Azure Databricks

Azure Databricks is a cloud service that provides a scalable platform for data analytics using Apache Spark.

2. Perform data analysis with Azure Databricks

Learn how to perform data analysis using Azure Databricks. Explore various data ingestion methods and how to integrate data from sources like Azure Data Lake and Azure SQL Database. This module guides you through using collaborative notebooks to perform exploratory data analysis (EDA), so you can visualize, manipulate, and examine data to uncover patterns, anomalies, and correlations.

3. Use Apache Spark in Azure Databricks

Azure Databricks is built on Apache Spark and enables data engineers and analysts to run Spark jobs to transform, analyze and visualize data at scale.

4. Manage data with Delta Lake

Delta Lake is a data management solution in Azure Databricks providing features including ACID transactions, schema enforcement, and time travel ensuring data consistency, integrity, and versioning capabilities.

5. Build data pipelines with Delta Live Tables

Building data pipelines with Delta Live Tables enables real-time, scalable, and reliable data processing using Delta Lake's advanced features in Azure Databricks

6. Deploy workloads with Azure Databricks Workflows

Deploying workloads with Azure Databricks Workflows involves orchestrating and automating complex data processing pipelines, machine learning workflows, and analytics tasks. In this module, you learn how to deploy workloads with Databricks Workflows.

Ref: https://trainocate.com/courses/microsoft/dp-3011-implement-a-data-analytics-solution-with-azure-databricks-training

END OF THE PAGE