DP-601T00: IMPLEMENTING A LAKEHOUSE WITH MICROSOFT FABRIC

About This Course

This course is designed to build your foundational skills in data engineering on Microsoft Fabric, focusing on the Lakehouse concept. This course will explore the powerful capabilities of Apache Spark for distributed data processing and the essential techniques for efficient data management, versioning, and reliability by working with Delta Lake tables. This course will also explore data ingestion and orchestration using Dataflows Gen2 and Data Factory pipelines. This course includes a combination of lectures and hands-on exercises that will prepare you to work with lakehouses in Microsoft Fabric.

Audience Profile

The primary audience for this course is data professionals who are familiar with data modeling, extraction, and analytics. It is designed for professionals who are interested in gaining knowledge about Lakehouse architecture, the Microsoft Fabric platform, and how to enable end-to-end analytics using these technologies.

Prerequisites

You should be familiar with basic data concepts and terminology.

Duration

8 hours (1 day)

Course Content

Module 1: Introduction to end-to-end analytics using Microsoft Fabric

Discover how Microsoft Fabric can meet your enterprise's analytics needs in one platform. Learn about Microsoft Fabric, how it works, and identify how you can use it for your analytics needs.

- Introduction to Microsoft Fabric
- Data teams and Fabric
- Enable and use Microsoft Fabric

Module 2: Get started with lakehouses in Microsoft Fabric

Lakehouses merge data lake storage flexibility with data warehouse analytics. Microsoft Fabric offers a lakehouse solution for comprehensive analytics on a single SaaS platform.

- What is a Lakehouse
- Work with a Fabric Lakehouse
- Explore, transform and visualize data in the Lakehouse

Module 3: Use Apache Spark in Microsoft Fabric

Apache Spark is a core technology for large-scale data analytics. Microsoft Fabric provides support for Spark clusters, enabling you to analyze and process data in a Lakehouse at scale.

- Prepare to use Apache Spark
- Run Spark in Fabric
- Load data in a Spark DataFrame
- Transform data in a Spark DataFrame
- Partition the output file
- Work with data using Spark SQL
- Query Data using Spark SQL API
- Visualize Data

Module 4: Work with Delta Lake tables in Microsoft Fabric

Tables in a Microsoft Fabric lakehouse are based on the Delta Lake storage format commonly used in Apache Spark. By using the enhanced capabilities of delta tables, you can create advanced analytics solutions.

- Understand Delta Lake
- Create delta tables using code in Spark
- Managed vs External Tables
- Work with delta tables in Spark
- Data versioning and Time Travel
- Use delta tables with Streaming data

Module 5: Ingest Data with Dataflows Gen2 in Microsoft Fabric

Data ingestion is crucial in analytics. Microsoft Fabric's Data Factory offers Dataflows (Gen2) for visually creating multi-step data ingestion and transformation using Power Query Online.

- Understand Dataflows (Gen2)
- Dataflow (Gen2) benefits and limitations
- Explore Dataflows (Gen2) in Microsoft Fabric
- Integrate Dataflows (Gen2) and Pipelines in Microsoft Fabric

Module 6: Use Data Factory pipelines in Microsoft Fabric

Microsoft Fabric includes Data Factory capabilities, including the ability to create pipelines that orchestrate data ingestion and transformation tasks.

- Pipelines in Microsoft Fabric
- Common Activities – Copy Data
- Common Activities – pipeline templates
- Run and monitor pipelines