Implement data engineering solutions using Azure Databricks

Course code: MOC DP-750

This course is designed for data engineers and data professionals who want to learn how to design, implement, and operate complete data engineering solutions using the Azure Databricks platform and the Unity Catalog service. In this course you will understand the key concepts of the Azure Databricks platform, learn how to select and configure the right compute, and practice organizing data objects in Unity Catalog with a focus on security, governance, and data lineage. You will learn how to design data models including dimensional modeling and Slowly Changing Dimensions, ingest data using a variety of approaches (Lakeflow Connect, Auto Loader, Spark Structured Streaming, Lakeflow Spark Declarative Pipelines), cleanse and transform data, and enforce data quality using pipeline expectations. You will also learn how to design and implement data pipelines using the medallion architecture, automate them with Lakeflow Jobs, apply development lifecycle best practices (Git, testing, Declarative Automation Bundles, Databricks CLI), and monitor and optimize workloads including troubleshooting. The course is also a comprehensive preparation for the DP-750 exam, Microsoft Certified: Azure Databricks Data Engineer Associate.

What You Will Learn

- Get started with the Azure Databricks platform and its key concepts
- Learn how to select and configure the right compute for different scenarios
- Practice creating and organizing objects in Unity Catalog, including schemas, tables, views, and volumes
- Understand how to secure data using fine-grained access control, row filtering, column masking, and Azure Key Vault
- Find out how to apply data governance through attribute-based access control, retention policies, data lineage, audit logging, and Delta Sharing
- Learn how to design data models including partitioning, clustering, and Slowly Changing Dimensions (SCD Type 2)
- Practice extracting and ingesting data using Lakeflow Connect, Auto Loader, Spark Structured Streaming, and Lakeflow Spark Declarative Pipelines
- Learn how to cleanse and transform data using PySpark and SQL operations (joins, aggregations, pivots, merge)
- Understand how to enforce data quality using pipeline expectations and manage schema drift
- Learn how to design and implement the medallion architecture (Bronze › Silver › Gold)
- Practice automating data pipelines with Lakeflow Jobs using triggers, scheduling, alerts, and retry policies
- Get familiar with the development lifecycle in Azure Databricks: Git, testing with pytest, Declarative Automation Bundles, and Databricks CLI
- Learn how to monitor and optimize workloads and troubleshoot issues with caching, data skew, memory spill, and shuffle

Who This Course Is For

- Data engineers who want to design and implement data engineering solutions on the Azure Databricks platform using Unity Catalog.
- Data and BI architects who want to understand the architecture of a modern lakehouse solution built on Azure Databricks and Delta Lake.
- Data professionals who want to prepare for the Microsoft DP-750 certification exam.

Prerequisites

- Basic knowledge of SQL and relational databases
- Basic knowledge of Python and the Apache Spark framework (especially PySpark)
- Basic knowledge of data warehouse design principles and ETL/ELT process implementation
- Basic knowledge of Microsoft Azure data services at the level of the MOC DP-900 course is recommended
- Basic familiarity with the Azure Databricks platform and the Delta Lake format is recommended

Course Outline

1 Explore Azure Databricks

GOPAS Praha
Na Strži 2097/63
140 00 Praha 4 - Krč
Tel.: +420 226 201 390
info@gopas.cz

GOPAS Brno
Nové sady 996/25
602 00 Brno
Tel.: +420 530 513 590
info@gopas.cz

GOPAS Bratislava
Dr. Vladimíra Clementisa 10
Bratislava, 821 02
Tel.: +421 902 903 132
info@gopas.sk

- Get started with the Azure Databricks platform and navigate the workspace UI
- Identify the typical workloads Azure Databricks is designed for
- Understand the key concepts of the platform
- Get familiar with data governance through Unity Catalog and Microsoft Purview
- Lab: Upload a dataset to a Unity Catalog volume, work in a notebook, and use Databricks Assistant in the CityMoves Transit scenario

2 Select and Configure Compute

- Learn how to select the appropriate compute type for a given task
- Find out how to configure compute performance and the runtime for running different types of workloads
- Learn how to install libraries at the cluster and notebook level
- Learn how to configure access to compute resources
- Lab: Create a cluster, install libraries, and generate synthetic data using PySpark and the faker library

3 Create and Organize Objects in Unity Catalog

- Get familiar with object naming conventions in Unity Catalog
- Practice creating catalogs, schemas, tables, views, and volumes
- Understand how to perform DDL operations and implement foreign catalogs to connect to external data sources
- Find out how to configure AI/BI Genie instructions
- Lab: Build a complete namespace for a university data platform — medallion schemas, managed tables with PK/FK, views, a volume, and SQL functions

4 Secure Unity Catalog Objects

- Understand the query lifecycle and access control strategies
- Learn how to implement fine-grained access control, row filtering, and column masking
- Find out how to work with secrets using Azure Key Vault
- Learn how to authenticate access to data using service principals and to resources using managed identities
- Lab: Configure permissions, row filters to restrict data access by region, and email masking, and protect sensitive credentials using Azure Key Vault

5 Govern Unity Catalog Objects

- Learn how to create and store table definitions and configure attribute-based access control (ABAC) using tags and policies
- Find out how to apply data retention policies (including VACUUM and predictive optimization)
- Learn how to set up and manage data lineage and audit logging
- Find out how to design a secure data sharing strategy using the Delta Sharing protocol
- Lab: Implement governance for a connected vehicle platform — PII tags, retention policies, querying system tables for lineage, and audit log analysis

6 Design and Implement Data Modeling

- Learn how to design the data ingestion logic, select the right tools, and choose the appropriate table format
- Understand how to design and implement partitioning and clustering strategies
- Find out how to select and implement a Slowly Changing Dimension type (especially SCD Type 2) and temporal (history) tables
- Learn how to decide between managed and unmanaged tables and choose the right data aggregation granularity
- Lab: Design a Delta Lake model for retail banking — a customer dimension with SCD Type 2, a fact table with liquid clustering, Change Data Feed, and try out Delta time travel

7 Ingest Data into Unity Catalog

- Learn how to extract and ingest data through Lakeflow Connect, notebooks, and SQL methods
- Find out how to work with a CDC feed and Spark Structured Streaming
- Learn how to use Auto Loader for automatic processing of files from cloud storage

GOPAS Praha
Na Strži 2097/63
140 00 Praha 4 - Krč
Tel.: +420 226 201 390
info@gopas.cz

GOPAS Brno
Nové sady 996/25
602 00 Brno
Tel.: +420 530 513 590
info@gopas.cz

GOPAS Bratislava
Dr. Vladimíra Clementisa 10
Bratislava, 821 02
Tel.: +421 902 903 132
info@gopas.sk

- Practice using Lakeflow Spark Declarative Pipelines to declaratively describe data ingestion
- Lab: Load CSV files from a Unity Catalog volume into Delta tables using PySpark, COPY INTO, and CTAS, and configure Auto Loader to process new files

8 Cleanse, Transform, and Load Data into Unity Catalog

- Learn how to profile data and select the correct column data types
- Find out how to handle duplicate data and NULL values
- Practice transforming data using filters, aggregations, joins, set operators, denormalization, and pivots
- Learn how to load data using merge, insert, and append operations
- Lab: Cleanse and restructure real estate data — choose the correct data types, remove duplicate data, and combine data from different tables for trend analysis

9 Implement and Manage Data Quality Constraints

- Learn how to implement validation checks and data type checks
- Find out how to detect and manage schema drift
- Learn how to manage data quality using pipeline expectations
- Lab: Build a Lakeflow Spark Declarative Pipeline for the insurer ClearCover that enforces the required quality of input data, and try out monitoring data quality metrics

10 Design and Implement Data Pipelines

- Learn how to design the order of operations within a pipeline and decide between notebooks and Lakeflow Pipelines
- Understand how to design Lakeflow job logic and handle error handling
- Practice building pipelines using both notebooks and Lakeflow Spark Declarative Pipelines
- Lab: Build a medallion architecture (Bronze › Silver › Gold) for GlobStay hotel data — deduplication, validation, data aggregation, notebook parameterization, and configuration of a Lakeflow Job with sequential dependencies and retry policies

11 Implement Lakeflow Jobs

- Learn how to configure Lakeflow Jobs
- Find out how to configure triggers (both time-based and event-based) and task scheduling
- Learn how to set up alerts for success/failure and automatic restarts
- Lab: Automate a data pipeline for TelConnect — a parameterized notebook processing call data records through the bronze/silver/gold layers, configure task dependencies, time-based and event-based triggers, notifications, and retry policies

12 Implement Development Lifecycle Processes

- Learn how to apply Git version control and manage branches and pull requests
- Find out how to implement a testing strategy for data pipelines
- Learn how to configure and package Declarative Automation Bundles
- Practice deploying bundles using the Databricks CLI
- Lab: Implement a testing strategy using the pytest library, then package and deploy a transformation pipeline as a Declarative Automation Bundle via the Databricks CLI

13 Monitor, Troubleshoot, and Optimize Workloads

- Learn how to monitor and manage cluster compute consumption
- Find out how to troubleshoot and fix Lakeflow Jobs, Spark jobs, and notebooks
- Learn how to diagnose issues with caching, data skew, memory spill, and shuffle using the Spark UI
- Learn how to implement log streaming to Azure Log Analytics
- Lab: Generate synthetic workloads with intentional data skew and excessive shuffle, diagnose them in the Spark UI, and apply targeted fixes using broadcast joins, Adaptive Query Execution, and shuffle reduction techniques

GOPAS Praha
Na Strži 2097/63
140 00 Praha 4 - Krč
Tel.: +420 226 201 390
info@gopas.cz

GOPAS Brno
Nové sady 996/25
602 00 Brno
Tel.: +420 530 513 590
info@gopas.cz

GOPAS Bratislava
Dr. Vladimíra Clementisa 10
Bratislava, 821 02
Tel.: +421 902 903 132
info@gopas.sk