Implement Data Engineering Solutions Using Azure Databricks

Course DP-750 4 days Instructor-led

Introduction

Master end-to-end data engineering with Azure Databricks and Unity Catalog. This course moves from foundational setup to production deployment, covering environment configuration and enterprise-grade governance. Learn to build robust ingestion pipelines, implement security with Unity Catalog, and deploy optimized workloads. By the end, you will have the practical skills to implement, secure, and maintain scalable lakehouse solutions that meet rigorous enterprise requirements.

Who Should Attend?

The target audience is data engineers who have fundamental knowledge of data analytics concepts, a basic understanding of cloud storage, and familiarity with data organization principles. They should be comfortable working with SQL and have experience using Python, including notebooks, for data engineering tasks.

At Course Completion

After completing this course, students will be able to:

- Explore Azure Databricks
- Understand Azure Databricks architecture
- Understand Azure Databricks Integrations
- Select and Configure Compute in Azure Databricks
- Create and organize objects in Unity Catalog
- Secure Unity Catalog objects
- Govern Unity Catalog objects
- Design and implement data modeling with Azure Databricks
- Ingest data into Unity Catalog
- Cleanse, transform, and load data into Unity Catalog
- Implement and manage data quality constraints with Azure Databricks
- Design and implement data pipelines with Azure Databricks
- Implement Lakeflow Jobs with Azure Databricks
- Implement development lifecycle processes in Azure Databricks
- Monitor, troubleshoot and optimize workloads in Azure Databricks

Prerequisites

Prior to attending this course, learners are expected to have a good understanding of Azure Databricks workspaces and Unity Catalog, along with familiarity with data access patterns and core data engineering and data warehouse concepts. In addition, they should have foundational knowledge of Azure security, including Microsoft Entra ID, and be familiar with Git version control fundamentals.

Course Outline

Module 1: Explore Azure Databricks

- Get started with Azure Databricks
- Identify Azure Databricks workloads
- Understand key concepts
- Data governance using Unity Catalog and Microsoft Purview
- Module assessment

Module 2: Understand Azure Databricks architecture

- Understand Azure Databricks architecture
- Understand Unity Catalog managed storage
- Understand external storage
- Understand default storage
- Module assessment

Module 3: Understand Azure Databricks Integrations

- Understand integration with Microsoft Fabric
- Understand integration with Power BI
- Understand integration with VS Code
- Understand integration with Power Platform
- Understand integration with Copilot Studio
- Understand integration with Microsoft Purview
- Understand integration with Microsoft Foundry
- Module assessment

Module 4: Select and Configure Compute in Azure Databricks

- Choose an appropriate compute type
- Configure compute performance
- Configure compute features
- Install libraries for compute
- Configure compute access
- Module assessment

Module 5: Create and organize objects in Unity Catalog

- Apply naming conventions
- Create catalog
- Create schema
- Create tables and views
- Create volumes
- Implement DDL operations
- Implement foreign catalog
- Configure AI/BI Genie instructions

Module 6: Secure Unity Catalog objects

- Understand query lifecycle
- Implement access control strategies
- Understand fine-grained access control
- Implement row filtering and column masking
- Access Azure Key Vault secrets
- Authenticate data access with service principals
- Authenticate resource access with managed identities
- Module assessment

Module 7: Govern Unity Catalog objects

- Create and preserve table definitions
- Configure ABAC with tags and policies
- Apply data retention policies
- Set up and manage data lineage
- Configure audit logging
- Design secure Delta Sharing strategy
- Module assessment

Module 8: Design and implement data modeling with Azure Databricks

- Design ingestion logic and data source configuration
- Choose a data ingestion tool
- Choose a data table format
- Design and implement a data partitioning scheme
- Choose a slowly changing dimension (SCD) type
- Implement a slowly changing dimension (SCD) type 2
- Design and implement a temporal (history) table to record changes over time
- Choose granularity on a column or table based on requirements
- Choose managed vs external tables
- Design and implement a clustering strategy

Module 9: Ingest data into Unity Catalog

- Ingest data with Lakeflow Connect
- Ingest data with notebooks
- Ingest data with SQL methods
- Ingest data with CDC feed
- Ingest data with Spark Structured Streaming
- Ingest data with Auto Loader
- Ingest data with Lakeflow Spark Declarative Pipelines
- Module assessment

Module 10: Cleanse, transform, and load data into Unity Catalog

- Profile data
- Choose column data types
- Resolve duplicates and nulls
- Transform data with filters and aggregations
- Transform data with joins and set operators
- Transform data with denormalization and pivots
- Load data with merge, insert, and append
- Module assessment

Module 11: Implement and manage data quality constraints with Azure Databricks

- Implement validation checks
- Implement data type checks
- Detect and manage schema drift
- Manage data quality with pipeline expectations
- Module assessment

Module 12: Design and implement data pipelines with Azure Databricks

- Design order of operations for a pipeline
- Choose notebook vs Lakeflow Pipelines
- Design Lakeflow job logic
- Design error handling in pipelines and jobs
- Create pipeline with notebook
- Create pipeline with Lakeflow Spark Declarative Pipelines
- Module assessment

Module 13: Implement Lakeflow Jobs with Azure Databricks

- Create job setup and configuration

- Configure job triggers
- Schedule a job
- Configure job alerts
- Configure automatic restarts
- Module assessment

Module 14: Implement development lifecycle processes in Azure Databricks

- Apply Git version control best practices
- Manage branching and pull requests
- Implement testing strategy
- Configure and package Declarative Automation Bundles
- Deploy bundle with Databricks CLI
- Module assessment

Module 15: Monitor, troubleshoot and optimize workloads in Azure Databricks

- Monitor and manage cluster consumption
- Troubleshoot and repair Lakeflow Jobs
- Troubleshoot Spark jobs and notebooks
- Investigate caching, skewing, spilling, shuffle
- Implement log streaming with Azure Log Analytics
- Module assessment