Advanced Data Engineering in Microsoft Fabric

Course code: GOC681

An advanced training program for data professionals who want to master modern data engineering in Microsoft Fabric, with a strong emphasis on practical development in Python and PySpark. Most of the course is hands-on coding in Notebooks — you will implement data transformations using Python (Polars, DuckDB) or PySpark, automate ETL processes, and work with advanced data-processing techniques in a distributed environment.You will learn how to design and implement a medallion architecture within a Lakehouse environment. The course explores multiple data ingestion approaches - from Dataflows Gen2 and orchestration Pipelines to fully custom ingestion logic in Notebooks. You will gain proficiency in data storage, understand the differences between data warehouses and Lakehouses, learn how to query them effectively, and work with advanced components such as stored procedures, functions, and data masking.Through automation and orchestration of data workflows using Pipelines, you will learn to coordinate complex processes and integrate the individual layers of the medallion architecture. The course focuses on performance optimization - partitioning, data compression, and Spark job tuning. You will learn how to monitor Fabric capacities and measure processing efficiency. Practical exercises will also guide you through code versioning and deploying changes using Git integration and deployment pipelines.Together with the course Advanced Techniques for Data Analysis and Reporting in Microsoft Fabric [GOC682], this training forms a complete preparation path for the certification exam DP-600: Fabric Analytics Engineer Associate.

What You Will Learn

- Design and implement a medallion architecture in Microsoft Fabric within a Lakehouse environment
- Implement data logic and transformations using Python (Polars, DuckDB) and PySpark in Notebooks
- Work with various data ingestion methods – Dataflows Gen2, Pipelines, and custom code
- Copy and reuse data across OneLake
- Profile, clean, and transform data using code in a variety of practical scenarios
- Work with both Lakehouse and Data Warehouse, including data security features
- Automate and orchestrate data workflows using Pipelines
- Optimize performance (partitioning, compression, Spark job optimization)
- Version code and deploy changes using Git integration and deployment pipelines

Who the Course Is For

This course is intended primarily for data engineers and developers who want to work with Microsoft Fabric at the code level and design, implement, and operate data solutions in a production environment. It is also suitable for advanced analysts and data architects with Python experience who want to move toward data engineering and distributed data processing.

Prerequisites

- Basic knowledge of Microsoft Fabric, at least at the level of course GOC680
- Knowledge of Python (pandas, list comprehensions, functions, error handling) and PySpark, at least at the level of course GOC685
- Basic understanding of relational databases and SQL
- Basic experience with data warehouses or data lakes
- Understanding of data extraction, ingestion, profiling, and transformation concepts
- Experience with data analysis and data integration tools (ETL processes, data pipelines)
- Knowledge of version control and Git integration is an advantage

Course Outline

1. Environment Setup and Core Principles
   - Medallion architecture – principles and components
   - Lakehouse, Data Warehouse, analytical engines, semantic layers
   - Tenant configuration, capacity selection, performance and cost implications

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

2. Data Ingestion and Data Copying
   - Data ingestion methods
   - Dataflows Gen2
   - Pipelines
   - Custom ingestion using Python / PySpark in Notebooks
   - Copying and reusing data in OneLake
   - Shortcuts
   - Decision methodology and architectural implications
   - Practical implementation
3. Data Profiling, Cleaning, and Transformation
   - Data profiling
   - Principles and methods
   - Implementation in Python / PySpark (Notebooks)
   - Data cleaning and transformation
   - Designing cleaning mechanisms based on profiling results
   - Code-based data transformations
   - Slowly Changing Dimensions and advanced scenarios
4. Data Storage
   - Lakehouse vs. Data Warehouse – differences and use cases
   - Querying data
   - SQL queries
   - Querying Lakehouse and Warehouse
   - Advanced components
   - Stored procedures, functions, roles, schemas
   - RLS, CLS, data masking
5. Automation
   - Orchestration Pipelines
   - Coordination and dependencies
   - Integration of notebooks, dataflows, and SQL objects
   - Notebook orchestration
   - Managing dependent steps in Python / PySpark
   - Fail-over and error handling
6. Monitoring and Optimization
   - Performance optimization for Spark workloads
   - Partitioning, compression, V-order, vacuuming
   - Monitoring Fabric capacity and processing efficiency
7. Versioning and Deployment
   - Git integration
   - Deployment pipelines

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