DATA ENGINEERING AND ANALYTICS

Data Engineering on Microsoft Azure (DP-203)

Level: Practitioner • 2 days (expandable to 3) • Virtual, In-person

Overview

Data engineering on Azure is not one skill but a chain of them: designing storage that will still make sense in two years, building pipelines that move and transform data reliably, processing it in batch and in real time, and securing and monitoring the whole system. The DP-203 certification covers this chain, and the challenge for most learners is that the pieces are usually taught as disconnected services rather than as one coherent architecture.

This is a hands-on, practitioner course. It is grounded in the DP-203 domains (data storage, data processing, and securing, monitoring, and optimizing) but it deliberately does not walk the exam objectives as a checklist. Instead it builds one end-to-end architecture in the order a real project would: storage and lake design first, then batch pipelines, then streaming, then security and operations, each layer resting on the one before. In keeping with a less-but-deeper philosophy, we go deep on the patterns that carry most of the exam and most real projects, and point to the documentation for the long tail. Every module ends with a lab, and each module builds on the one before.

Who Should Attend

- Data engineers building or inheriting data platforms on Azure
- Database developers and ETL developers moving into cloud data engineering
- Architects and analytics engineers preparing for the DP-203 certification

Learners new to data concepts should start with *Microsoft Azure Data Fundamentals (DP-900)*.

Prerequisites

- Solid SQL, including joins, aggregation, and window functions
- Basic Python or Scala familiarity for Spark exercises (reading code is enough)
- Working Azure experience: portal navigation, resource groups, and storage accounts

What You Will Learn

- Design a data lake on Azure Data Lake Storage Gen2: zones, folder structure, file formats, and partitioning
- Choose and design serving layers, including dedicated SQL pools and lakehouse tables
- Build and orchestrate batch pipelines with Azure Data Factory and Synapse pipelines
- Process data at scale with Apache Spark on Azure
- Build a streaming solution with Event Hubs and Stream Analytics
- Secure, monitor, and optimize a data platform, and connect it all to DP-203 exam readiness

Course Outline

Day one: storage, the lake, and batch processing

- Designing Data Storage

  - The end-to-end Azure data architecture: where each service fits and why
  - Azure Data Lake Storage Gen2: zones, hierarchy, and file formats (Parquet and Delta)
  - Partitioning strategies and the query patterns that should drive them
  - Lab: design and build a data lake structure and load raw data into it
- The Serving Layer
  - Dedicated SQL pools: distribution strategies, and when a warehouse still earns its cost
  - Serverless SQL: querying the lake directly and when that is enough
  - Star schemas and slowly changing dimensions in an Azure context
  - Lab: create a serving layer and query lake data through both serverless and dedicated options
- Batch Pipelines
  - Orchestrating ingestion and transformation with Data Factory and Synapse pipelines
  - Incremental loading patterns and designing for safe reruns
  - Lab: build a pipeline that ingests, transforms, and loads data into the serving layer

Day two: Spark, streaming, and running the platform

- Processing with Apache Spark
  - Spark on Azure: Synapse Spark pools and Azure Databricks, and how to choose
  - DataFrames, transformations, and writing results back to the lake
  - Lab: transform lake data with a Spark notebook and land it as Delta tables
- Streaming Data
  - Batch versus streaming: latency, windows, and what real time actually requires
  - Event Hubs for ingestion and Stream Analytics for windowed processing
  - Handling late and out-of-order data
  - Lab: build a streaming job that aggregates events into the serving layer
- Security, Monitoring, and Optimization
  - Securing the platform: managed identities, RBAC and ACLs, Key Vault, and data masking
  - Monitoring pipelines and queries; finding and fixing the slow parts
  - Mapping what you have built to the DP-203 exam domains and planning your preparation
  - Lab: lock down and monitor the pipeline built across the course, then tune one slow query

Extended Version

The three-day version keeps the same gradient and adds depth and dedicated exam preparation:

- Deeper Spark work: performance tuning, partitioning, and handling skew
- Advanced streaming patterns, including Spark Structured Streaming
- Structured DP-203 exam preparation with practice questions and objective-by-objective review
- A capstone that designs and builds a complete batch-plus-streaming data platform from a business brief