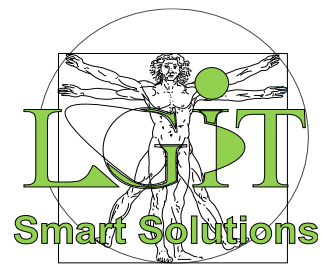# Data Engineering on Microsoft Azure: DP-203T00

Course Overview

In this course, the student will learn about the data engineering as it pertains to working with batch and real-time analytical solutions using Azure data platform technologies. Students will begin by understanding the core compute and storage technologies that are used to build an analytical solution. The students will learn how to interactively explore data stored in files in a data lake. They will learn the various ingestion techniques that can be used to load data using the Apache Spark capability found in Azure Synapse Analytics or Azure Databricks, or how to ingest using Azure Data Factory or Azure Synapse pipelines. The students will also learn the various ways they can transform the data using the same technologies that is used to ingest data. They will understand the importance of implementing security to ensure that the data is protected at rest or in transit. The student will then show how to create a real-time analytical system to create real-time analytical solutions.

Audience Profile

The primary audience for this course is data professionals, data architects, and business intelligence professionals who want to learn about data engineering and building analytical solutions using data platform technologies that exist on Microsoft Azure. The secondary audience for this course data analysts and data scientists who work with analytical solutions built on Microsoft Azure.

Prerequisites

Successful students start this course with knowledge of cloud computing and core data concepts and professional experience with data solutions.
Specifically completing:
- AZ-900 - Azure Fundamentals
- DP-900 - Microsoft Azure Data Fundamentals

Duration: 24 Hours (8 x 3hrs)

Course Outline

Module 1: Explore compute and storage options for data engineering workloads

- Understand Azure Synapse Analytics
- Describe Azure Databricks
- Describe Azure Databricks Delta Lake architecture
- Understand Azure Data Lake storage
- Work with data streams by using Azure Stream Analytics

Module 2: Run interactive queries using serverless SQL pools

- Explore Azure Synapse serverless SQL pools capabilities
- Query data in the lake using Azure Synapse serverless SQL pools
- Create metadata objects in Azure Synapse serverless SQL pools
- Secure data and manage users in Azure Synapse serverless SQL pools

Module 3: Data Exploration and Transformation in Azure Databricks

- Describe Azure Databricks
- Read and write data in Azure Databricks
- Work with Data Frames in Azure Databricks
- Work with Data Frames advanced methods in Azure Databricks

Module 4: Explore, transform, and load data into the Data Warehouse using Apache Spark

- Understand big data engineering with Apache Spark pools in Azure Synapse Analytics
- Ingest data with Apache Spark notebooks in Azure Synapse Analytics
- Integrate SQL and Apache Spark pools in Azure Synapse Analytics
- Transform data with Data Frames in Apache Spark pools in Azure Synapse Analytics
- Monitor and manage data engineering workloads with Apache Spark pools in Azure Synapse Analytics

Module 5: Ingest and load data into the Data Warehouse

- Use data loading best practices in Azure Synapse Analytics
- Perform Petabyte-scale ingestion with Azure Data Factory or Azure Synapse Pipelines

Module 6: Transform Data with Azure Data Factory or Azure Synapse Pipelines

- Perform Data integration with Azure Data Factory or Azure Synapse Pipelines
- Perform Code-free transformation at scale with Azure Data Factory or Azure Synapse Pipelines

Module 7: Integrate Data from Notebooks with Azure Data Factory or Azure Synapse Pipelines

- Understand how to add an activity to the control flow to orchestrate data from other technologies
- Understand how to use parameters in Azure Data Factory/Synapse pipelines

Module 8: End-to-end security with Azure Synapse Analytics

- Secure a data warehouse
- Configure and manage secrets in Azure Key Vault
- Implement compliance controls for sensitive data

Module 9: Support Hybrid Transactional Analytical Processing (HTAP) with Azure Synapse Link

- At the end of this module, the students will be able to
- Design hybrid transactional and analytical processing using Azure Synapse Analytics
- Configure Azure Synapse Link with Azure Cosmos DB
- Query Azure Cosmos DB with Apache Spark for Azure Synapse Analytics
- Query Azure Cosmos DB with SQL Serverless for Azure Synapse Analytics

Module 10: Real-time Stream Processing with Stream Analytics

- Work with data streams by using Azure Stream Analytics
- Enable reliable messaging for Big Data applications using Azure Event Hubs
- Ingest data streams with Azure Stream Analytics

Module 11: Create a Stream Processing Solution with Event Hubs and Azure Databricks

- Learn the key features and uses of Structured Streaming
- Stream data from a file and write it out to a distributed file system.
- Use sliding windows to aggregate over chunks of data rather than all data
- Apply watermarking to throw away stale old data that you do not have space to keep.
- Connect to Event Hubs read and write streams