Data Engineering on Microsoft Azure (DP-203)

Course 22920 – 32 Hours

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.

On Completion, Delegates will be able to

- Explore compute and storage options for data engineering workloads in Azure
- Run interactive queries using serverless SQL pools
- Perform data Exploration and Transformation in Azure Databricks
- Explore, transform, and load data into the Data Warehouse using Apache Spark
- Ingest and load Data into the Data Warehouse
- Transform Data with Azure Data Factory or Azure Synapse Pipelines
- Integrate Data from Notebooks with Azure Data Factory or Azure Synapse Pipelines
- Support Hybrid Transactional Analytical Processing (HTAP) with Azure Synapse Link
- Perform end-to-end security with Azure Synapse Analytics
- Perform real-time Stream Processing with Stream Analytics
- Create a Stream Processing Solution with Event Hubs and Azure Databricks

Who Should Attend

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

Course Contents

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

This module provides an overview of the Azure compute and storage technology options that are available to data engineers building analytical workloads. This module teaches ways to structure the data lake, and to optimize the files for exploration, streaming, and batch workloads. The student will learn how to organize the data lake into levels of data refinement as they transform files through batch and stream processing. Then they will learn how to create indexes on their datasets, such as CSV, JSON, and Parquet files, and use them for potential query and workload acceleration.

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

After completing this module, students will be able to:

- Describe Azure Synapse Analytics
- Describe Azure Databricks
- Describe Azure Data Lake storage
- Describe Delta Lake architecture
- Describe Azure Stream Analytics

Module 2: Run interactive queries using Azure Synapse Analytics serverless SQL pools

In this module, students will learn how to work with files stored in the data lake and external file sources, through T-SQL statements executed by a serverless SQL pool in Azure Synapse Analytics. Students will query Parquet files stored in a data lake, as well as CSV files stored in an external data store. Next, they will create Azure Active Directory security groups and enforce access to files in the data lake through Role-Based Access Control (RBAC) and Access Control Lists (ACLs).

- 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

After completing this module, students will be able to:

- Understand 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

This module teaches how to use various Apache Spark DataFrame methods to explore and transform data in Azure Databricks. The student will learn how to perform standard DataFrame methods to explore and transform data. They will also learn how to perform more advanced tasks, such as removing duplicate data, manipulate date/time values, rename columns, and aggregate data.

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

After completing this module, students will be able to:

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

Module 4: Explore, transform, and load data into the Data Warehouse using Apache Spark
This module teaches how to explore data stored in a data lake, transform the data, and load data into a relational data store. The student will explore Parquet and JSON files and use techniques to query and transform JSON files with hierarchical structures. Then the student will use Apache Spark to load data into the data warehouse and join Parquet data in the data lake with data in the dedicated SQL pool.

- Understand big data engineering with Apache Spark in Azure Synapse Analytics
- Ingest data with Apache Spark notebooks in Azure Synapse Analytics
- Transform data with DataFrames in Apache Spark Pools in Azure Synapse Analytics
- Integrate SQL and Apache Spark pools in Azure Synapse Analytics

After completing this module, students will be able to:

- Describe big data engineering with Apache Spark in Azure Synapse Analytics
- Ingest data with Apache Spark notebooks in Azure Synapse Analytics
- Transform data with DataFrames in Apache Spark Pools in Azure Synapse Analytics
- Integrate SQL and Apache Spark pools in Azure Synapse Analytics

Module 5: Ingest and load data into the data warehouse
This module teaches students how to ingest data into the data warehouse through T-SQL scripts and Synapse Analytics integration pipelines. The student will learn how to load data into Synapse dedicated SQL pools with PolyBase and COPY using T-SQL. The student will also learn how to use workload management along with a Copy activity in a Azure Synapse pipeline for petabyte-scale data ingestion.

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

After completing this module, students will be able to:

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

Module 6: Transform data with Azure Data Factory or Azure Synapse Pipelines
This module teaches students how to build data integration pipelines to ingest from multiple data sources, transform data using mapping data flowss, and perform data movement into one or more data sinks.

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

After completing this module, students will be able to:

- Perform data integration with Azure Data Factory
- Perform code-free transformation at scale with Azure Data Factory

Module 7: Orchestrate data movement and transformation in Azure Synapse Pipelines

In this module, you will learn how to create linked services, and orchestrate data movement and transformation using notebooks in Azure Synapse Pipelines.

- Orchestrate data movement and transformation in Azure Data Factory

After completing this module, students will be able to:

- Orchestrate data movement and transformation in Azure Synapse Pipelines

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

In this module, students will learn how to secure a Synapse Analytics workspace and its supporting infrastructure. The student will observe the SQL Active Directory Admin, manage IP firewall rules, manage secrets with Azure Key Vault and access those secrets through a Key Vault linked service and pipeline activities. The student will understand how to implement column-level security, row-level security, and dynamic data masking when using dedicated SQL pools.

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

After completing this module, students will be able to:

- Secure a data warehouse in Azure Synapse Analytics
- 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

In this module, students will learn how Azure Synapse Link enables seamless connectivity of an Azure Cosmos DB account to a Synapse workspace. The student will understand how to enable and configure Synapse link, then how to query the Azure Cosmos DB analytical store using Apache Spark and SQL serverless.

- 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 pools
- Query Azure Cosmos DB with serverless SQL pools

After completing this module, 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
In this module, students will learn how to process streaming data with Azure Stream Analytics. The student will ingest vehicle telemetry data into Event Hubs, then process that data in real time, using various windowing functions in Azure Stream Analytics. They will output the data to Azure Synapse Analytics. Finally, the student will learn how to scale the Stream Analytics job to increase throughput.

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

After completing this module, students will be able to:

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

Module 11: Create a Stream Processing Solution with Event Hubs and Azure Databricks
In this module, students will learn how to ingest and process streaming data at scale with Event Hubs and Spark Structured Streaming in Azure Databricks. The student will learn the key features and uses of Structured Streaming. The student will implement sliding windows to aggregate over chunks of data and apply watermarking to remove stale data. Finally, the student will connect to Event Hubs to read and write streams.

- Process streaming data with Azure Databricks structured streaming

After completing this module, students will be able to:

- Process streaming data with Azure Databricks structured streaming