



Training Course DP-203: Data Engineering on Microsoft Azure

Overview: In this course, the student will learn about the data engineering patterns and practices as it pertains to working with batch and real-time analytical solutions using Azure data platform technologies. Participants will begin by understanding the core compute and storage technologies that are used to build an analytical solution. They will then explore how to design an analytical serving layer and focus on data engineering considerations for working with source files. 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. Attendees will also learn the various ways they can transform the data using the same technologies that are used to ingest data. Students will spend time on the course learning how to monitor and analyze the performance of an analytical system so that they can optimize the performance of data loads, or queries that are issued against the systems. They will understand the importance of implementing security to ensure that the data is protected at rest or in transit. Participants will then show how the data in an analytical system can be used to create dashboards or build predictive models in Azure Synapse Analytics.

Duration: 4 Days.

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.

Certification: This course prepares you for the DP-203: Data Engineering on Microsoft Azure.

Course Objectives: After completing this course, students 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.

Course Outline

- 1- Introduction to data engineering on Azure.
 - Identify common data engineering tasks.
 - Describe common data engineering concepts.
 - Identify Azure services for data engineering.

 - 2- Introduction to Azure Data Lake Storage Gen2.
 - Describe the key features and benefits of Azure Data Lake Storage Gen2.
 - Enable Azure Data Lake Storage Gen2 in an Azure Storage account.
 - Compare Azure Data Lake Storage Gen2 and Azure Blob storage.
 - Describe where Azure Data Lake Storage Gen2 fits in the stages of analytical processing.
 - Describe how Azure data Lake Storage Gen2 is used in common analytical workloads.

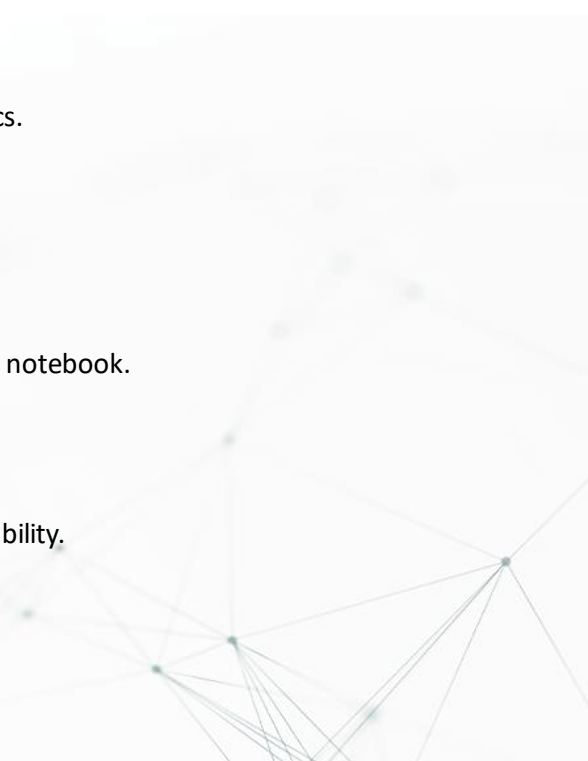
 - 3- Introduction to Azure Synapse Analytics.
 - Identify the business problems that Azure Synapse Analytics addresses.
 - Describe core capabilities of Azure Synapse Analytics.
 - Determine when to use Azure Synapse Analytics.

 - 4- Use Azure Synapse serverless SQL pool to query files in a data lake.
 - Identify capabilities and use cases for serverless SQL pools in Azure Synapse Analytics.
 - Query CSV, JSON, and Parquet files using a serverless SQL pool.
 - Create external database objects in a serverless SQL pool.

 - 5- Use Azure Synapse serverless SQL pools to transform data in a data lake.
 - Use a CREATE EXTERNAL TABLE AS SELECT (CETAS) statement to transform data.
 - Encapsulate a CETAS statement in a stored procedure.
 - Include a data transformation stored procedure in a pipeline.

 - 6- Create a lake database in Azure Synapse Analytics.
 - Understand lake database concepts and components.
 - Describe database templates in Azure Synapse Analytics.
 - Create a lake database.

 - 7- Analyze data with Apache Spark in Azure Synapse Analytics.
 - Identify core features and capabilities of Apache Spark.
 - Configure a Spark pool in Azure Synapse Analytics.
 - Run code to load, analyze, and visualize data in a Spark notebook.

 - 8- Transform data with Spark in Azure Synapse Analytics.
 - Use Apache Spark to modify and save dataframes.
 - Partition data files for improved performance and scalability.
- 

- Transform data with SQL.

9- Use Delta Lake in Azure Synapse Analytics.

- Describe core features and capabilities of Delta Lake.
- Create and use Delta Lake tables in a Synapse Analytics Spark pool.
- Create Spark catalog tables for Delta Lake data.
- Use Delta Lake tables for streaming data.
- Query Delta Lake tables from a Synapse Analytics SQL pool.

10- Analyze data in a relational data warehouse.

- Design a schema for a relational data warehouse.
- Create fact, dimension, and staging tables.
- Use SQL to load data into data warehouse tables.
- Use SQL to query relational data warehouse tables.

11- Load data into a relational data warehouse.

- Load staging tables in a data warehouse.
- Load dimension tables in a data warehouse.
- Load time dimensions in a data warehouse.
- Load slowly changing dimensions in a data warehouse.
- Load fact tables in a data warehouse.
- Perform post-load optimizations in a data warehouse.

12- Build a data pipeline in Azure Synapse Analytics.

- Describe core concepts for Azure Synapse Analytics pipelines.
- Create a pipeline in Azure Synapse Studio.
- Implement a data flow activity in a pipeline.
- Initiate and monitor pipeline runs.

13- Use Spark Notebooks in an Azure Synapse Pipeline.

- Describe notebook and pipeline integration.
- Use a Synapse notebook activity in a pipeline.
- Use parameters with a notebook activity.

14- Plan hybrid transactional and analytical processing using Azure Synapse Analytics.

- Describe Hybrid Transactional / Analytical Processing patterns.
- Identify Azure Synapse Link services for HTAP.

15- Implement Azure Synapse Link with Azure Cosmos DB.

- Configure an Azure Cosmos DB Account to use Azure Synapse Link.
- Create an analytical store enabled container.
- Create a linked service for Azure Cosmos DB.

- Analyze linked data using Spark.
- Analyze linked data using Synapse SQL.

16- Implement Azure Synapse Link for SQL.

- Understand key concepts and capabilities of Azure Synapse Link for SQL.
- Configure Azure Synapse Link for Azure SQL Database.
- Configure Azure Synapse Link for Microsoft SQL Server.

17- Get started with Azure Stream Analytics.

- Understand data streams.
- Understand event processing.
- Understand window functions.
- Get started with Azure Stream Analytics.

18- Ingest streaming data using Azure Stream Analytics and Azure Synapse Analytics.

- Describe common stream ingestion scenarios for Azure Synapse Analytics.
- Configure inputs and outputs for an Azure Stream Analytics job.
- Define a query to ingest real-time data into Azure Synapse Analytics.
- Run a job to ingest real-time data, and consume that data in Azure Synapse Analytics.

19- Visualize real-time data with Azure Stream Analytics and Power BI.

- Configure a Stream Analytics output for Power BI.
- Use a Stream Analytics query to write data to Power BI.
- Create a real-time data visualization in Power BI.

20- Introduction to Microsoft Purview.

- Knowledge of Azure accounts and services
- Knowledge of various data sources such as SQL Server and Azure Cosmos DB
- Knowledge of the concepts around data governance.

21- Integrate Microsoft Purview and Azure Synapse Analytics.

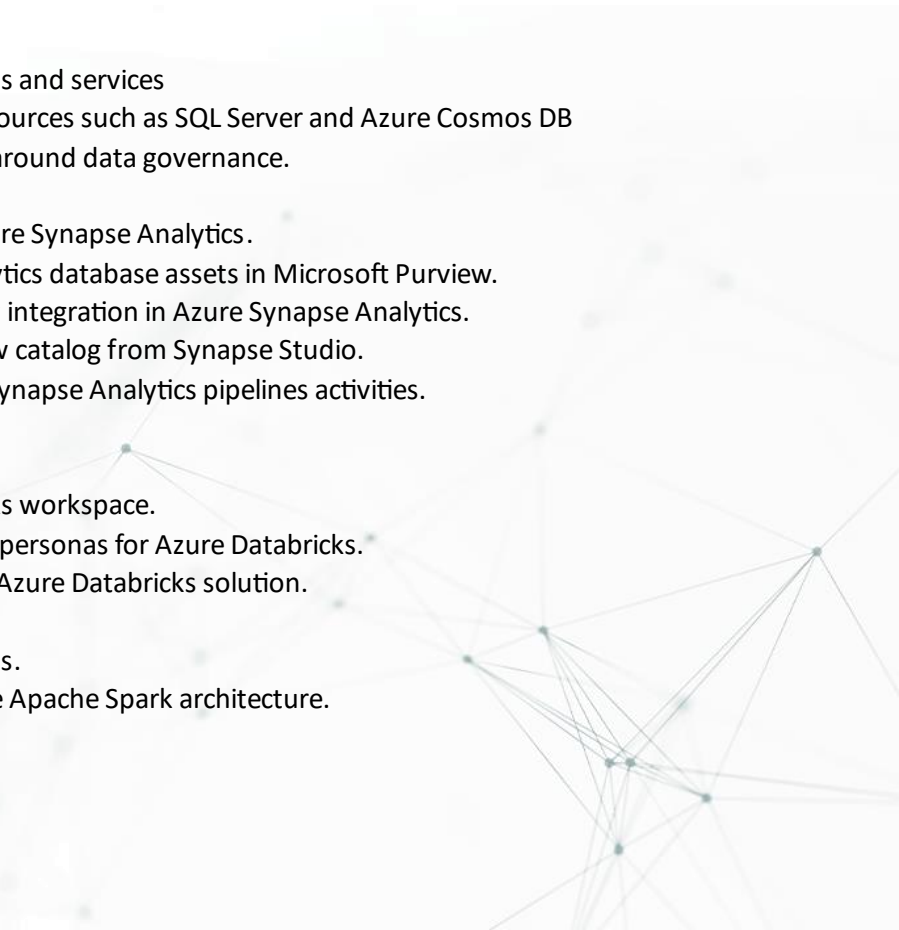
- Catalog Azure Synapse Analytics database assets in Microsoft Purview.
- Configure Microsoft Purview integration in Azure Synapse Analytics.
- Search the Microsoft Purview catalog from Synapse Studio.
- Track data lineage in Azure Synapse Analytics pipelines activities.

22- Explore Azure Databricks.

- Provision an Azure Databricks workspace.
- Identify core workloads and personas for Azure Databricks.
- Describe key concepts of an Azure Databricks solution.

23- Use Apache Spark in Azure Databricks.

- Describe key elements of the Apache Spark architecture.



- Create and configure a Spark cluster.
- Describe use cases for Spark.
- Use Spark to process and analyze data stored in files.
- Use Spark to visualize data.

24- Run Azure Databricks Notebooks with Azure Data Factory.

- Describe how Azure Databricks notebooks can be run in a pipeline.
- Create an Azure Data Factory linked service for Azure Databricks.
- Use a Notebook activity in a pipeline.
- Pass parameters to a notebook.

