DP-203T00: Data Engineering on Microsoft Azure

Course outline

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.

Lessons

- 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

Lab : Explore compute and storage options for data engineering workloads

- Combine streaming and batch processing with a single pipeline
- Organize the data lake into levels of file transformation
- Index data lake storage for query and workload acceleration

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: Design and implement the serving layer

This module teaches how to design and implement data stores in a modern data warehouse to optimize analytical workloads. The student will learn how to design a

multidimensional schema to store fact and dimension data. Then the student will learn how to populate slowly changing dimensions through incremental data loading from Azure Data Factory.

Lessons

- Design a multidimensional schema to optimize analytical workloads
- Code-free transformation at scale with Azure Data Factory
- Populate slowly changing dimensions in Azure Synapse Analytics pipelines

Lab : Designing and Implementing the Serving Layer

- Design a star schema for analytical workloads
- Populate slowly changing dimensions with Azure Data Factory and mapping data flows

After completing this module, students will be able to:

- Design a star schema for analytical workloads
- Populate a slowly changing dimensions with Azure Data Factory and mapping data flows

Module 3: Data engineering considerations for source files

This module explores data engineering considerations that are common when loading data into a modern data warehouse analytical from files stored in an Azure Data Lake, and understanding the security consideration associated with storing files stored in the data lake.

Lessons

- Design a Modern Data Warehouse using Azure Synapse Analytics
- Secure a data warehouse in Azure Synapse Analytics

Lab : Data engineering considerations

- Managing files in an Azure data lake
- Securing files stored in an Azure data lake

After completing this module, students will be able to:

- Design a Modern Data Warehouse using Azure Synapse Analytics
- Secure a data warehouse in Azure Synapse Analytics

Module 4: 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).

Lessons

- 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

Lab : Run interactive queries using serverless SQL pools

- Query Parquet data with serverless SQL pools
- Create external tables for Parquet and CSV files
- Create views with serverless SQL pools
- Secure access to data in a data lake when using serverless SQL pools
- Configure data lake security using Role-Based Access Control (RBAC) and Access Control List

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 5: 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.

Lessons

- 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

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

- Perform Data Exploration in Synapse Studio
- Ingest data with Spark notebooks in Azure Synapse Analytics
- Transform data with DataFrames in Spark pools in Azure Synapse Analytics
- Integrate SQL and 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 6: 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.

Lessons

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

Lab : Data Exploration and Transformation in Azure Databricks

- Use DataFrames in Azure Databricks to explore and filter data
- Cache a DataFrame for faster subsequent queries
- Remove duplicate data
- Manipulate date/time values
- Remove and rename DataFrame columns
- Aggregate data stored in a DataFrame

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 7: 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.

Lessons

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

Lab : Ingest and load Data into the Data Warehouse

- Perform petabyte-scale ingestion with Azure Synapse Pipelines
- Import data with PolyBase and COPY using T-SQL
- Use data loading best practices in Azure Synapse Analytics

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 8: 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.

Lessons

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

Lab : Transform Data with Azure Data Factory or Azure Synapse Pipelines

- Execute code-free transformations at scale with Azure Synapse Pipelines
- Create data pipeline to import poorly formatted CSV files
- Create Mapping Data Flows

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 9: 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.

Lessons

- Orchestrate data movement and transformation in Azure Data Factory

Lab : Orchestrate data movement and transformation in Azure Synapse Pipelines

- Integrate Data from Notebooks with Azure Data Factory or Azure Synapse Pipelines

After completing this module, students will be able to:

- Orchestrate data movement and transformation in Azure Synapse Pipelines

Module 10: Optimize query performance with dedicated SQL pools in Azure Synapse

In this module, students will learn strategies to optimize data storage and processing when using dedicated SQL pools in Azure Synapse Analytics. The student will know how to use developer features, such as windowing and HyperLogLog functions, use data loading best practices, and optimize and improve query performance.

Lessons

- Optimize data warehouse query performance in Azure Synapse Analytics
- Understand data warehouse developer features of Azure Synapse Analytics

Lab : Optimize Query Performance with Dedicated SQL Pools in Azure Synapse

- Understand developer features of Azure Synapse Analytics
- Optimize data warehouse query performance in Azure Synapse Analytics
- Improve query performance

After completing this module, students will be able to:

- Optimize data warehouse query performance in Azure Synapse Analytics
- Understand data warehouse developer features of Azure Synapse Analytics

Module 11: Analyze and Optimize Data Warehouse Storage

In this module, students will learn how to analyze then optimize the data storage of the Azure Synapse dedicated SQL pools. The student will know techniques to understand table space usage and column store storage details. Next the student will know how to compare storage requirements between identical tables that use different data types. Finally, the student will observe the impact materialized views have when executed in place of complex queries and learn how to avoid extensive logging by optimizing delete operations.

Lessons

- Analyze and optimize data warehouse storage in Azure Synapse Analytics

Lab : Analyze and Optimize Data Warehouse Storage

- Check for skewed data and space usage
- Understand column store storage details
- Study the impact of materialized views

- Explore rules for minimally logged operations

After completing this module, students will be able to:

- Analyze and optimize data warehouse storage in Azure Synapse Analytics

Module 12: 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.

Lessons

- 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

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

- Configure Azure Synapse Link with Azure Cosmos DB
- Query Azure Cosmos DB with Apache Spark for Synapse Analytics
- Query Azure Cosmos DB with serverless SQL pool for Azure Synapse Analytics

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 13: 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.

Lessons

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

Lab : End-to-end security with Azure Synapse Analytics

- Secure Azure Synapse Analytics supporting infrastructure
- Secure the Azure Synapse Analytics workspace and managed services
- Secure Azure Synapse Analytics workspace 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 14: 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.

Lessons

- 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

Lab : Real-time Stream Processing with Stream Analytics

- Use Stream Analytics to process real-time data from Event Hubs
- Use Stream Analytics windowing functions to build aggregates and output to Synapse Analytics

- Scale the Azure Stream Analytics job to increase throughput through partitioning
- Repartition the stream input to optimize parallelization

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 15: 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.

Lessons

- Process streaming data with Azure Databricks structured streaming

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

- Explore 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 remove stale data
- Connect to Event Hubs read and write streams

After completing this module, students will be able to:

- Process streaming data with Azure Databricks structured streaming

Module 16: Build reports using Power BI integration with Azure Synapase Analytics

In this module, the student will learn how to integrate Power BI with their Synapse workspace to build reports in Power BI. The student will create a new data source and Power BI report in Synapse Studio. Then the student will learn how to improve query performance with materialized views and result-set caching. Finally, the

student will explore the data lake with serverless SQL pools and create visualizations against that data in Power BI.

Lessons

- Create reports with Power BI using its integration with Azure Synapse Analytics

Lab : Build reports using Power BI integration with Azure Synapase Analytics

- Integrate an Azure Synapse workspace and Power BI
- Optimize integration with Power BI
- Improve query performance with materialized views and result-set caching
- Visualize data with SQL serverless and create a Power BI report

After completing this module, students will be able to:

- Create reports with Power BI using its integration with Azure Synapse Analytics

Module 17: Perform Integrated Machine Learning Processes in Azure Synapse Analytics

This module explores the integrated, end-to-end Azure Machine Learning and Azure Cognitive Services experience in Azure Synapse Analytics. You will learn how to connect an Azure Synapse Analytics workspace to an Azure Machine Learning workspace using a Linked Service and then trigger an Automated ML experiment that uses data from a Spark table. You will also learn how to use trained models from Azure Machine Learning or Azure Cognitive Services to enrich data in a SQL pool table and then serve prediction results using Power BI.

Lessons

- Use the integrated machine learning process in Azure Synapse Analytics

Lab : Perform Integrated Machine Learning Processes in Azure Synapse Analytics

- Create an Azure Machine Learning linked service
- Trigger an Auto ML experiment using data from a Spark table
- Enrich data using trained models
- Serve prediction results using Power BI

After completing this module, students will be able to:

- Use the integrated machine learning process in Azure Synapse Analytics