DP-203: Data Engineering on Microsoft Azure

Audience

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

In addition to their professional experience, students who take this training should have technical knowledge equivalent to the Azure fundamentals course.

Duration

4 days. Hands on.

Course Objectives

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. Students 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 layers and focus on data engineering considerations for working with source files. 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. The student will spend time on the course learning how to monitor and analyze the performance of 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. The student will then show how the data in an analytical system can be used to create dashboards, or build predictive models in Azure Synapse Analytics.

Delegates will appreciate how to:

- Explore compute and storage options for data engineering workloads in Azure
- Design and Implement the serving layer
- Understand data engineering considerations
- Run interactive queries using serverless SQL pools
- Explore, transform, and load data into the Data Warehouse using Apache Spark
- Perform data Exploration and Transformation in Azure Databricks
- 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
- Optimize Query Performance with Dedicated SQL Pools in Azure Synapse
- Analyze and Optimize Data Warehouse Storage
- 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 Content

Module 1: Get started with data engineering on Azure

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 data engineering on Azure
- Introduction to Azure Data Lake Storage Gen2
- Introduction to Azure Synapse Analytics

Lab : Explore Azure Synapse Analytics

Azure Synapse Analytics provides a single, consolidated data analytics platform for end-to end data analytics. In this exercise, you'll explore various ways to ingest and explore data. This exercise is designed as a high-level overview of the various core capabilities of Azure Synapse

Module 2: Build data analytics solutions using Azure Synapse 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

- Use a serverless SQL to query files in a data lake
- Use a serverless SQL pool to transform data
- Create a lake database

Lab : Query files using a serverless SQL pool

SQL is probably the most used language for working with data in the world. Most data analysts are proficient in using SQL queries to retrieve, filter, and aggregate data - most commonly in relational databases. As organizations increasingly take advantage of scalable file storage to create data lakes, SQL is often still the preferred choice for querying the data. Azure Synapse Analytics provides serverless SQL pools that enable you to decouple the SQL query engine from the data storage and run queries against data files in common file formats such as delimited text and Parquet.

Lab : Transform files using a serverless pool

Data analysts often use SQL to query data for analysis and reporting. Data engineers can also make use of SQL to manipulate and transform data; often as part of a data ingestion pipeline or extract, transform, and load (ETL) process.

Lab : Analyze data in a lake database

Azure Synapse Analytics enables you to combine the flexibility of file storage in a data lake with the structured schema and SQL querying capabilities of a relational database through the ability to create a lake database. A lake database is a relational database schema defined on a data lake file store that enables data storage to be separated from the compute used to query it. Lake databases combine the benefits of a structured schema that includes support for data types, relationships, and other features typically only found in relational database systems, with the flexibility of storing data in files that can be used independently of a relational database store. Essentially, the lake database 'overlays' a relational schema onto files in folders in the data lake.

Module 3: Perform data engineering with Azure Synapse Apache Spark Pools

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

- Analyze data with Apache Spark in Azure Synapse Analytics
- Transform data with Apache Spark in Azure Synapse Analytics
- Use Delta lake in Azure Synapse Analytics

Lab : Analyze data with Spark

Apache Spark is an open source engine for distributed data processing, and is widely used to explore, process, and analyze huge volumes of data in data lake storage. Spark is available as a processing option in many data platform products, including Azure HDInsight, Azure Databricks, and Azure Synapse Analytics on the Microsoft Azure cloud platform. One of the benefits of Spark is support for a wide range of programming languages, including Java, Scala, Python, and SQL; making Spark a very flexible solution for data processing workloads including data cleansing and manipulation, statistical analysis and machine learning, and data analytics and visualization.

Lab : Transform data with Spark in Synapse Analytics

Data engineers often use Spark notebooks as one of their preferred tools to perform extract, transform, and load (ETL) or extract, load, and transform (ELT) activities that transform data from one format or structure to another.

Lab : Use Delta Lake in Azure Synapse Analytics

Delta Lake is an open source project to build a transactional data storage layer on top of a data lake. Delta Lake adds support for relational semantics for both batch and streaming data operations, and enables the creation of a Lakehouse architecture in which Apache Spark can be used to process and query data in tables that are based on underlying files in the data lake.

Module 4: 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

- Analyse data in a relational data warehouse
- Load data into a relational data warehouse

Lab : Explore a data warehouse

Azure Synapse Analytics is built on a scalable set capabilities to support enterprise data warehousing; including file-based data analytics in a data lake as well as large-scale relational data warehouses and the data transfer and transformation pipelines used to load them. In this lab, you'll explore how to use a dedicated SQL pool in Azure Synapse Analytics to store and query data in a relational data warehouse.

Lab : Load data into a data warehouse

In this exercise, you're going to load data into a dedicated SQL Pool using COPY INTO, external tables and CTAS, Insert, Update and merge

Module 5: Transfer and transform data with Azure Synapse Analytics Pipelines.

This module teaches students how to build data integration pipelines to ingest from multiple data sources, transform data using mapping data flows, and perform data movement into one or more data sinks.

Lessons

- Build a data pipeline in Azure Synapse Analytics
- Use a spark notebook in an Azure Synapse Pipeline

Lab : Build a pipeline in Azure Synapse Pipelines

In this exercise, you'll load data into a dedicated SQL Pool using a pipeline in Azure Synapse Analytics Explorer. The pipeline will encapsulate a data flow that loads product data into a table in a data warehouse.

Lab : Use an Apache Spark notebook in a Pipeline.

In this exercise, we're going to create an Azure Synapse Analytics pipeline that includes an activity to run an Apache Spark notebook.

Module 6: Work with hybrid transactional and analytical (HTAP) solutions using Azure Synapse Analytics

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

- Plan hybrid transactional and analytical processing
- Implement Azure Synapse Link with Azure Cosmos DB
- Implement Azure Synapse link for SQL

Lab : Use Azure Synapse Link for Azure Cosmos DB

Azure Synapse Link for Azure Cosmos DB is a cloud-native hybrid transactional analytical processing (HTAP) technology that enables you to run near-real-time analytics over operational data stored in Azure Cosmos DB from Azure Synapse Analytics.

Lab : Use Azure Synapse Link for SQL

Azure Synapse Link for SQL enables you to automatically synchronize a transactional database in SQL Server or Azure SQL Database with a dedicated SQL pool in Azure Synapse Analytics. This synchronization enables you to perform low-latency analytical workloads in Synapse Analytics without incurring query overhead in the source operational database.

Module 7: Implement a data streaming solution with Azure 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

- Get started with Azure Stream Analytics
- Ingest streaming data using Stream Analytics and Azure Synapse Analytics
- Visualize real-time data with Azure Stream Analytics and Power BI

Lab : Get started with Azure Stream Analytics

In this exercise you'll provision an Azure Stream Analytics job in your Azure subscription, and use it to query and summarize a stream of real-time event data and store the results in Azure Storage.

Lab : Ingest streaming data into Azure Synapse Analytics

In this exercise, you'll use Azure Stream Analytics to process a stream of sales order data, such as might be generated from an online retail application. The order data will be sent to Azure Event Hubs, from where your Azure Stream Analytics jobs will read the data and ingest it into Azure Synapse Analytics.

Lab : Create a real-time data visualisation

In this exercise, you'll use Azure Stream Analytics to process a stream of sales order data, such as might be generated from an online retail application. The order data will be sent to Azure Event Hubs, from where your Azure Stream Analytics job will read and summarize the data before sending it to Power BI, where you will visualize the data in a report.

Module 8: Govern data across an enterprise

Introduction to Purview to track data being passed through the end-to-end Business Intelligence solution.

Lessons

- Introduction to Microsoft Purview
- Integrate Microsoft Purview and Azure Synapse Analytics

Lab : Integrate Azure Synapse Analytics and Microsoft Purview

Microsoft Purview enables you to catalog data assets across your data estate and track the flow of data as it is transferred from one data source to another - a key element of a comprehensive data governance solution.

Module 9: Recap of using Azure Databricks and data lakes rather than using Synapse Analytics.

Covering each of the subjects delivered previously in the course with Synapse Analytics with the Spark cluster with Azure Databricks and Azure Data Factory as standalone services.

Lessons

- Explore Azure Databricks
- Use Apache Spark in Azure Databricks
- Run Databricks notebooks in Azure Data Factory
- Ingest streaming data using Stream Analytics and Azure Synapse Analytics
- Visualize real-time data with Azure Stream Analytics and Power BI

Lab : Explore Azure Databricks

Azure Databricks is a Microsoft Azure-based version of the popular open-source Databricks platform. Similarly to Azure Synapse Analytics, an Azure Databricks workspace provides a central point for managing Databricks clusters, data, and resources on Azure.

Lab : Analyze files in Azure Databricks

Azure Databricks is a Microsoft Azure-based version of the popular open-source Databricks platform. Azure Databricks is built on Apache Spark, and offers a highly scalable solution for data engineering and analysis tasks that involve working with data in files. One of the benefits of Spark is support for a wide range of programming languages, including Java, Scala, Python, and SQL; making Spark a very flexible solution for data processing workloads including data cleansing and manipulation, statistical analysis and machine learning, and data analytics and visualization.

Lab : Delta-Lake in Azure Databricks

Delta Lake is an open source project to build a transactional data storage layer for Spark on top of a data lake. Delta Lake adds support for relational semantics for both batch and streaming data operations, and enables the creation of a Lakehouse architecture in which Apache Spark can be used to process and query data in tables that are based on underlying files in the data lake.

Lab : Azure Databricks SQL

SQL is an industry-standard language for querying and manipulating data. Many data analysts perform data analytics by using SQL to query tables in a relational database. Azure Databricks includes SQL functionality that builds on Spark and Delta Lake technologies to provide a relational database layer over files in a data lake..

Lab : Azure Databricks with Azure Data Factory

You can use notebooks in Azure Databricks to perform data engineering tasks, such as processing data files and loading data into tables. When you need to orchestrate these tasks as part of a data engineering pipeline, you can use Azure Data Factory.