KPMG Learning Academy

Microsoft Data Engineering on Microsoft Azure Training (DP-203)

Module 1: Introduction to data engineering on Azure

Microsoft Azure provides a comprehensive platform for data engineering; but what is data engineering? Complete this module to find out. In this module you will learn how to:

- Identify common data engineering tasks
- Describe common data engineering concepts
- Identify Azure services for data engineering

Module 2: Introduction to Azure Data Lake Storage Gen2

Data lakes are a core element of data analytics architectures. Azure Data Lake Storage Gen2 provides a scalable, secure, cloud-based solution for data lake storage. In this module you will learn how to:

- 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

Module 3: Introduction to Azure Synapse Analytics

Learn about the features and capabilities of Azure Synapse Analytics - a cloud-based platform for big data processing and analysis. In this module, you'll learn how to:

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

KPMG Learning Academy

Module 4: Use Azure Synapse serverless SQL pool to query files in a data lake

With Azure Synapse serverless SQL pool, you can leverage your SQL skills to explore and analyse data in files, without the need to load the data into a relational database. After the completion of this module, you will be able to:

- 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

Module 5: Use Azure Synapse serverless SQL pools to transform data in a data lake

By using a serverless SQL pool in Azure Synapse Analytics, you can use the ubiquitous SQL language to transform data in files in a data lake. After completing this module, you'll be able to:

- 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

Module 6: Create a lake database in Azure Synapse Analytics

Why choose between working with files in a data lake or a relational database schema? With lake databases in Azure Synapse Analytics, you can combine the benefits of both. After completing this module, you will be able to:

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

Module 7: Analyse data with Apache Spark in Azure Synapse Analytics

Apache Spark is a core technology for large-scale data analytics. Learn how to use Spark in Azure Synapse Analytics to analyse and visualise data in a data lake. After completing this module, you will be able to:

- Identify core features and capabilities of Apache Spark
- Configure a Spark pool in Azure Synapse Analytics
- Run code to load, analyse, and visualise data in a Spark notebook

KPMG Learning Academy

Module 8: Transform data with Spark in Azure Synapse Analytics

Data engineers commonly need to transform large volumes of data. Apache Spark pools in Azure Synapse Analytics provide a distributed processing platform that they can use to accomplish this goal. In this module, you will learn how to:

- Use Apache Spark to modify and save dataframes
- Partition data files for improved performance and scalability
- Transform data with SQL

Module 9: Use Delta Lake in Azure Synapse Analytics

Delta Lake is an open source relational storage area for Spark that you can use to implement a data lakehouse architecture in Azure Synapse Analytics. In this module, you'll learn how to:

- 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

Module 10: Analyse data in a relational data warehouse

Relational data warehouses are a core element of most enterprise Business Intelligence (BI) solutions, and are used as the basis for data models, reports, and analysis. In this module, you'll learn how to:

- 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

Module 11: Load data into a relational data warehouse

A core responsibility for a data engineer is to implement a data ingestion solution that loads new data into a relational data warehouse. In this module, you'll learn how to:

- 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

KPMG Learning Academy

Module 12: Build a data pipeline in Azure Synapse Analytics

Pipelines are the lifeblood of a data analytics solution. Learn how to use Azure Synapse Analytics pipelines to build integrated data solutions that extract, transform, and load data across diverse systems. In this module, you will learn how to:

- 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

Module 13: Use Spark Notebooks in an Azure Synapse Pipeline

Apache Spark provides data engineers with a scalable, distributed data processing platform, which can be integrated into an Azure Synapse Analytics pipeline. In this module, you will learn how to:

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

Module 14: Plan hybrid transactional and analytical processing using Azure Synapse Analytics

Learn how hybrid transactional / analytical processing (HTAP) can help you perform operational analytics with Azure Synapse Analytics. After completing this module, you'll be able to:

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

Module 15: Implement Azure Synapse Link for SQL

Azure Synapse Link for SQL enables low-latency synchronization of operational data in a relational database to Azure Synapse Analytics. In this module, you'll learn how to:

- 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

KPMG Learning Academy

Module 16: Get started with Azure Stream Analytics

Azure Stream Analytics enables you to process real-time data streams and integrate the data they contain into applications and analytical solutions. In this module, you'll learn how to:

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

Module 17: Ingest streaming data using Azure Stream Analytics and Azure Synapse Analytics

Azure Stream Analytics provides a real-time data processing engine that you can use to ingest streaming event data into Azure Synapse Analytics for further analysis and reporting. After completing this module, you'll be able to:

- 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

Module 18: Visualise real-time data with Azure Stream Analytics and Power BI

By combining the stream processing capabilities of Azure Stream Analytics and the data visualization capabilities of Microsoft Power BI, you can create real-time data dashboards. In this module, you'll learn how to:

- 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

Module 19: Introduction to Microsoft Purview

In this module, you'll evaluate whether Microsoft Purview is the right choice for your data discovery and governance needs. By the end of this module, you'll be able to:

- Evaluate whether Microsoft Purview is appropriate for your data discovery and governance needs
- Describe how the features of Microsoft Purview work to provide data discovery and governance

KPMG Learning Academy

Module 20: Integrate Microsoft Purview and Azure Synapse Analytics

Learn how to integrate Microsoft Purview with Azure Synapse Analytics to improve data discoverability and lineage tracking. After completing this module, you'll be able to:

- 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

Module 21: Explore Azure Databricks

Azure Databricks is a cloud service that provides a scalable platform for data analytics using Apache Spark. In this module, you'll learn how to:

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

Module 22: Use Apache Spark in Azure Databricks

Azure Databricks is built on Apache Spark and enables data engineers and analysts to run Spark jobs to transform, analyse and visualise data at scale. In this module, you'll learn how to:

- Describe key elements of the Apache Spark architecture
- Create and configure a Spark cluster
- Describe use cases for Spark
- Use Spark to process and analyse data stored in files
- Use Spark to visualise data

Module 23: Run Azure Databricks Notebooks with Azure Data Factory

Using pipelines in Azure Data Factory to run notebooks in Azure Databricks enables you to automate data engineering processes at cloud scale. In this module, you'll learn how to:

- 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