MS- DP203T00: DATA ENGINEERING ON MICROSOFT AZURE

DURATION	LEVEL	TECHNOLOGY	DELIVERY METHOD	TRAINING CREDITS
4 Days	Intermediate	Azure	Instructor-led	NA

INTRODUCTION

In this course, the student will learn how to implement and manage data engineering workloads on Microsoft Azure, using Azure services such as Azure Synapse Analytics, Azure Data Lake Storage Gen2, Azure Stream Analytics, Azure Databricks, and others. The course focuses on common data engineering tasks such as orchestrating data transfer and transformation pipelines, working with data files in a data lake, creating and loading relational data warehouses, capturing and aggregating streams of real-time data, and tracking data assets and lineage.

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 includes 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 OBJECTIVES

After completing this course, students will be able to:

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

COURSE CONTENT

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.

Lessons

- What is data engineering.
- Important data engineering concepts.
- Data engineering in Microsoft Azure.

After completing this module, students will be able 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

Microsoft Azure provides a comprehensive platform for data engineering; but what is data engineering? Complete this module to find out.

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.

Lessons

- Understand Azure Data Lake Storage Gen2.
- Enable Azure Data Lake Storage Gen2 in Azure Storage.
- Compare Azure Data Lake Store to Azure Blob storage.
- Understand the stages for processing big data.
- Use Azure Data Lake Storage Gen2 in data analytics workloads.

After completing this module, students will be able 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.

Lessons

- How Azure Synapse Analytics works
- When to use Azure Synapse Analytics
- Exercise - Explore Azure Synapse Analytics

After completing this module, students will be able to:

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

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 analyze data in files, without the need to load the data into a relational database.

Lessons

- When to use Azure Synapse Analytics
- Understand Azure Synapse serverless SQL pool capabilities and use cases
- Query files using a serverless SQL pool
- Create external database objects
- Exercise - Query files using a serverless SQL pool

After completing this module, students 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.

Lessons

- Transform data files with the CREATE EXTERNAL TABLE AS SELECT statement
- Encapsulate data transformations in a stored procedure
- Include a data transformation stored procedure in a pipeline
- Exercise - Transform files using a serverless SQL pool

After completing this module, students will be able to:

- Familiarity with Azure Synapse Analytics.
- Experience using Transact-SQL to query and manipulate data.

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.

Lessons

- Understand lake database concepts
- Explore database templates
- Create a lake database
- Use a lake database
- Exercise - Analyze data in a lake database

After completing this module, students will be able to:

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

Module 7: Secure data and manage users in Azure Synapse serverless SQL pools

Learn how you can set up security when using Azure Synapse serverless SQL pools

Lessons

- Introduction
- Choose an authentication method in Azure Synapse serverless SQL pools
- Manage users in Azure Synapse serverless SQL pools
- Manage user permissions in Azure Synapse serverless SQL pools
- Module assessment
- Summary

After completing this module, students will be able to:

- Choose an authentication method in Azure Synapse serverless SQL pools
- Manage users in Azure Synapse serverless SQL pools
- Manage user permissions in Azure Synapse serverless SQL pools

Module 8: Analyze 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 analyze and visualize data in a data lake.

Lessons

- Introduction
- Get to know Apache Spark
- Use Spark in Azure Synapse Analytics
- Analyze data with Spark
- Visualize data with Spark
- Exercise - Analyze data with Spark
- Module assessment
- Summary

After completing this module, students will be able to:

- 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.

Module 9: 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.

Lessons

- Introduction
- Modify and save dataframes
- Partition data files
- Transform data with SQL
- Exercise: Transform data with Spark in Azure Synapse Analytics
- Module assessment
- Summary

After completing this module, students will be able to:

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

Module 10: 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.

Lessons

- Introduction
- Understand Delta Lake
- Create Delta Lake tables
- Create catalog tables
- Use Delta Lake with streaming data
- Use Delta Lake in a SQL pool
- Exercise - Use Delta Lake in Azure Synapse Analytics
- Module assessment
- Summary

After completing this module, students will be able 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 11: 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.

Lessons

- Understand pipelines in Azure Synapse Analytics
- Create a pipeline in Azure Synapse Studio
- Define data flows
- Run a pipeline
- Exercise - Build a data pipeline in Azure Synapse Analytics
- Module assessment
- Summary

After completing this module, students will be able 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 12: 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.

Lessons

- Introduction
- Understand Synapse Notebooks and Pipelines
- Use a Synapse notebook activity in a pipeline
- Use parameters in a notebook
- Exercise - Use an Apache Spark notebook in a pipeline
- Module assessment
- Summary

After completing this module, students will be able to:

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

Module 13: Analyze 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.

Lessons

- Introduction
- Design a data warehouse schema
- Create data warehouse tables
- Load data warehouse tables
- Query a data warehouse
- Exercise - Explore a data warehouse
- Module assessment
- Summary

After completing this module, students will be able 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 14: 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.

Lessons

- Introduction
- Load staging tables
- Load dimension tables
- Load time dimension tables
- Load slowly changing dimensions
- Load fact tables
- Perform post load optimization
- Exercise - load data into a relational data warehouse
- Module assessment
- Summary

After completing this module, students will be able 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

Module 15: Manage and monitor data warehouse activities in Azure Synapse Analytics

- Learn how to manage and monitor Azure Synapse Analytics.
- Lessons
- Introduction
- Scale compute resources in Azure Synapse Analytics
- Pause compute in Azure Synapse Analytics
- Manage workloads in Azure Synapse Analytics
- Use Azure Advisor to review recommendations
- Use dynamic management views to identify and troubleshoot query performance
- Module assessment
- Summary

After completing this module, students will be able to:

- Scale compute resources in Azure Synapse Analytics
- Pause compute in Azure Synapse Analytics
- Manage workloads in Azure Synapse Analytics
- Use Azure Advisor to review recommendations
- Use Dynamic Management Views to identify and troubleshoot query performance

Module 16: Secure a data warehouse in Azure Synapse Analytics

Learn how to approach and implement security to protect your data with Azure Synapse Analytics.

Lessons

- Introduction
- Understand network security options for Azure Synapse Analytics
- Configure Conditional Access
- Configure authentication
- Manage authorization through column and row level security
- Exercise - Manage authorization through column and row level security

- Manage sensitive data with Dynamic Data Masking
- Implement encryption in Azure Synapse Analytics
- Module assessment
- Summary

After completing this module, students will be able to:

- Understand network security options for Azure Synapse Analytics
- Configure Conditional Access
- Configure Authentication
- Manage authorization through column and row level security
- Manage sensitive data with Dynamic Data masking
- Implement encryption in Azure Synapse Analytics

Module 17: 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.

Lessons

- Introduction
- Understand hybrid transactional and analytical processing patterns
- Describe Azure Synapse Link
- Module assessment
- Summary

After completing this module, students will be able to:

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

Module 18: Implement Azure Synapse Link with Azure Cosmos DB

Azure Synapse Link for Azure Cosmos DB enables HTAP integration between operational data in Azure Cosmos DB and Azure Synapse Analytics runtimes for Spark and SQL.

Lessons

- Introduction
- Enable Cosmos DB account to use Azure Synapse Link
- Create an analytical store enabled container
- Create a linked service for Cosmos DB
- Query Cosmos DB data with Spark
- Query Cosmos DB with Synapse SQL
- Exercise - Implement Azure Synapse Link for Cosmos DB
- Module assessment
- Summary

After completing this module, students will be able to:

- 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.

Module 19: 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.

Lessons

- Introduction
- What is Azure Synapse Link for SQL?
- Configure Azure Synapse Link for Azure SQL Database
- Configure Azure Synapse Link for SQL Server 2022
- Exercise - Implement Azure Synapse Link for SQL
- Module assessment
- Summary

After completing this module, students will be able 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.

Module 20: 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.

Lessons

- Introduction
- Understand data streams
- Understand event processing
- Understand window functions
- Exercise - Get started with Azure Stream Analytics
- Module assessment
- Summary

After completing this module, students will be able to:

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

Module 21: 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.

Lessons

- Introduction
- Stream ingestion scenarios
- Configure inputs and outputs
- Define a query to select, filter, and aggregate data
- Run a job to ingest data
- Exercise - Ingest streaming data into Azure Synapse Analytics
- Module assessment
- Summary

After completing this module, students will 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 22: Visualize 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.

Lessons

- Introduction
- Use a Power BI output in Azure Stream Analytics
- Create a query for real-time visualization
- Create real-time data visualizations in Power BI
- Exercise - Create a real-time data visualization
- Module assessment
- Summary

After completing this module, students will be able 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 23: Explore Azure Databricks

Azure Databricks is a cloud service that provides a scalable platform for data analytics using Apache Spark.

Lessons

- Introduction
- Get started with Azure Databricks
- Identify Azure Databricks workloads
- Understand key concepts

- Data governance using Unity Catalog and Microsoft Purview
- Exercise - Explore Azure Databricks
- Module assessment
- Summary

After completing this module, students will be able to:

- Provision an Azure Databricks workspace
- Identify core workloads for Azure Databricks
- Use Data Governance tools Unity Catalog and Microsoft Purview
- Describe key concepts of an Azure Databricks solution

Module 24: Perform data analysis with Azure Databricks

Learn how to perform data analysis using Azure Databricks. Explore various data ingestion methods and how to integrate data from sources like Azure Data Lake and Azure SQL Database. This module guides you through using collaborative notebooks to perform exploratory data analysis (EDA), so you can visualize, manipulate, and examine data to uncover patterns, anomalies, and correlations.

Lessons

- Introduction
- Ingest data with Azure Databricks
- Data exploration tools in Azure Databricks
- Data analysis using DataFrame APIs
- Exercise - Explore data with Azure Databricks
- Module assessment
- Summary

After completing this module, students will be able to:

- Ingest data using Azure Databricks.
- Using the different data exploration tools in Azure Databricks.
- Analyze data with DataFrame APIs.

Module 25: Manage data with Delta Lake

Delta Lake is a data management solution in Azure Databricks providing features including ACID transactions, schema enforcement, and time travel ensuring data consistency, integrity, and versioning capabilities.

Lessons

- Introduction
- Get started with Delta Lake
- Manage ACID transactions
- Implement schema enforcement
- Data versioning and time travel in Delta Lake
- Data integrity with Delta Lake
- Exercise - Use Delta Lake in Azure Databricks
- Module assessment
- Summary

After completing this module, students will be able to:

- What Delta Lake is
- How to manage ACID transactions using Delta Lake
- How to use schema versioning and time travel in Delta Lake
- How to maintain data integrity with Delta Lake

Module 26: Build data pipelines with Delta Live Tables

Building data pipelines with Delta Live Tables enables real-time, scalable, and reliable data processing using Delta Lake's advanced features in Azure Databricks

Lessons

- Introduction
- Explore Delta Live Tables
- Data ingestion and integration
- Real-time processing
- Exercise - Create a data pipeline with Delta Live Tables
- Module assessment
- Summary

After completing this module, students will be able to:

- Describe Delta Live Tables
- Ingest data into Delta Live Tables
- Use Data Pipelines for real time data processing

ASSOCIATED CERTIFICATIONS & EXAM

This course will prepare delegates to write the Microsoft DP-203: Data Engineering on Microsoft Azure exam.