Data Engineering on Microsoft Azure
DP-203T00

Course Name	**Data Engineering on Microsoft Azure**
Course Code	**DP-203T00**
Course Duration	4 Days
Course Structure	Instructor-Led
Course Overview	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.
Course 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 Outcome	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
Assessment/Evaluation	This course will prepare delegates to take the DP-203: Data Engineering on Microsoft Azure exam. Successfully passing this exam will result in the attainment of the Data Engineering on Microsoft Azure Certification and Certificate of Attendance issued by IT-IQ Botswana.

Course Details	
Topic	**TOPIC 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. **Learning objectives** In this module you will learn how to: • Identify common data engineering tasks. • Describe common data engineering concepts. • Identify Azure services for data engineering. **Topic 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. **Learning objectives** 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. **Topic 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. **Learning objectives** 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. **Topic 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. **Learning objectives** 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.

	Topic 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. **Learning objectives** 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. **Topic 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. **Learning objectives** 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. **Topic 7: 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. **Learning objectives** 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, analyze, and visualize data in a Spark notebook. **Topic 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. **Learning objectives** In this module, you will learn how to: • Use Apache Spark to modify and save data frames. • Partition data files for improved performance and scalability. • Transform data with SQL. **Topic 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. **Learning objectives** 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. **Topic 10: 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.

	Learning objectives 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. **Topic 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. **Learning objectives** 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. **Topic 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. **Learning objectives** 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. **Topic 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. **Learning objectives** 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. **Topic 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. **Learning objectives** After completing this module, you'll be able to: • Describe Hybrid Transactional / Analytical Processing patterns. • Identify Azure Synapse Link services for HTAP.

	Topic 15: 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. **Learning objectives** After completing this module, you'll 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. **Topic 16: 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. **Learning objectives** 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. **Topic 17: 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. **Learning objectives** In this module, you'll learn how to: • Understand data streams.

	• Understand event processing. • Understand window functions. • Get started with Azure Stream Analytics. **Topic 18: 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. **Learning objectives** 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. **Topic 19: 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. **Learning objectives** 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.

	Topic 20: Introduction to Microsoft Purview In this module, you'll evaluate whether Microsoft Purview is the right choice for your data discovery and governance needs. **Learning objectives** By the end of this module, you'll be able to: • Evaluate whether Microsoft Purview is appropriate for data discovery and governance needs. • Describe how the features of Microsoft Purview work to provide data discovery and governance. **Topic 21: Integrate Microsoft Purview and Azure Synapse Analytics** Learn how to integrate Microsoft Purview with Azure Synapse Analytics to improve data discoverability and lineage tracking. **Learning objectives** 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.

	Topic 22: Explore Azure Databricks Azure Databricks is a cloud service that provides a scalable platform for data analytics using Apache Spark. **Learning objectives** 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. **Topic 23: 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, analyze and visualize data at scale. **Learning objectives** 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 analyze data stored in files. • Use Spark to visualize data.

	Topic 24: 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. **Learning objectives** 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.