# Course Outline for DP-203: Microsoft Azure Data Engineer

### Course Objectives and Exam Mapping

After completing this course, students will be able to:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Mod**  **#** | **Module Title** | **Module Goal** | **Learning Objective** | **Associated Learn materials that act as a source or inspires content (no content)** | **Exam OD Mapping (From CourseOD doc)** |
| 0 | About the course | This module sets the goals and expectations of the course. | In this module the student will be able to:   * Data Engineering Curriculum and Exams * About this course * Course Syllabus | N/A | N/A |
| 1 | Designing and Implementing Data Storage | 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. | In this module the student will be able to:   * 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 | N/A | * 1. Design a data storage structure   2. Design a partition strategy   1.4 Implement physical data storage structures  1.5 Implement logical data structures  2.2 Design and develop a batch processing solution  2.3 Design and develop a stream processing solution |
| 2 | Designing and Implementing 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. | In this module, the student will be able to:   * Design a star schema for analytical workloads (OLAP) * Populate slowly changing dimensions with Azure Data Factory and mapping data flows |  | * 1. Design the serving layer   1.6 Implementing the serving layer  2.2 Design and develop a batch processing solution |
| 3 | High level data engineering design |  | Some kind of module to whiteboard out a design/solution. The first two modules shock and awe the students, the module here pauses to think about it using the **Whiteboard asset** |  |  |
| 4 | Determining functional requirements |  |  |  |  |
| 5. | Running interactive queries using serverless SQL pool with Azure Synapse Analytics | 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). | In this module the student will be able to:   * 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 Lists (ACLs) | N/A | 2.1 Ingest and transform data  3.1 Design security for data policies and standards  3.2 Implement data security |
| 6 | Exploring, Transforming, and Loading 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. | In this module the student will be able to:   * 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 | N/A | 2.1 Ingest and transform data |
| 7 | 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. | In this module the student will be able to:   * 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 |  | 2.1 Ingest and transform data  2.2 Design and develop a batch processing solution (remove duplicate data) |
| 8 | Ingesting and Loading 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. | In this module the student will be able to:   * 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 | N/A | 2.4 Manage batches and pipelines  2.1 Ingest and transform data  1.6 Implement the serving layer |
| 9 | Transform Data with Synapse Pipelines | This module teaches students how to build data integration pipelines to ingest from multiple data sources, transform data using mapping data flows and notebooks, and perform data movement into one or more data sinks. | In this module the student will be able to:   * Execute code-free transformations at scale with Azure Synapse Pipelines * Create data pipeline to import poorly formatted CSV files * Create Mapping Data Flows | N/A | 2.4 Manage batches and pipelines  2.1 Ingest and transform data  1.6 Implement the serving layer  4.1 Monitor data storage and data processing |
| 10 | Integrate Data with Synapse Pipelines | The student will learn how to create linked services, and orchestrate data movement and transformation in Azure Synapse Pipelines. | * Orchestrate data movement and transformation in Azure Synapse Pipelines |  | 2.4 Manage batches and pipelines  2.1 Ingest and transform data |
| 11 | Analyze Data and 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. | In this module the student will be able to:   * Understand developer features of Azure Synapse Analytics * Optimize data warehouse query performance in Azure Synapse Analytics * Improve query performance | N/A | 1.2 Design a partition strategy  1.3 Design the serving layer  4.1 Monitor data storage and data processing  4.2 Optimize and troubleshoot data storage and data processing |
| 12 | Analyzing and Optimizing 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. | In this module the student will be able to:   * Check for skewed data and space usage * Understand column store storage details * Study the impact of materialized views * Explore rules for minimally logged operations | N/A | 1.2 Design a partition strategy  1.3 Design the serving layer  1.4 Implement physical data storage structures  4.1 Monitor data storage and data processing  4.2 Optimize and troubleshoot data storage and data processing |
| 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. | In this module the student will be able to:   * Secure Azure Synapse Analytics supporting infrastructure * Secure the Azure Synapse Analytics workspace and managed services * Secure Azure Synapse Analytics workspace data | N/A | 3.1 Design security for data policies and standards  3.2 Implement data security |
| 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. | In this module the student will be able to:   * 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 | N/A | 2.1 Ingest and transform data  2.3 Design and develop a stream processing solution |
| 15 | Creating 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. | In this module the student will be able to:   * Know the 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 |  | 2.1 Ingest and transform data  2.3 Design and develop a stream processing solution |
| 16 | Optimize and Visualize Data in the Data Warehouse | 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 datasource 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. | In this module the student will be able to:   * Integrate a 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 | N/A | 1.3 Design the serving layer  1.5 Implement logical data structures |
| 17 | Delivering the modern data warehouse with Azure Synapse Analytics, Azure Databricks, and Azure Data Factory | In this module, students will learn how Azure Data Factory (ADF), Azure Databricks, and Azure Synapse Analytics can be used together to build a modern data warehouse. The student will start by using Azure Data Factory (ADF) to automate the movement of data in various formats gathered from various sources, including Cosmos DB, into a centralized Azure Data Lake Storage Gen2 (ADLS Gen2) repository. The student will then use Azure Databricks to prepare and analyze those data, and finally write the aggregations to Azure Synapse Analytics. As part of the process, the student will also use Databricks to connect to the Cosmos DB Change Feed to stream near-real-time vehicle telemetry data directly into the Data Warehouse using Spark Structured Streaming. | In this module the student will be able to:   * Use Azure Data Factory to automate data movement in various formats from multiple sources into a data lake * Use Azure Databricks to prepare and analyze data, then write aggregations to Azure Synapse Analytics * Perform real-time processing of telemetry data using Spark Structured Streaming | N/A | 1.3 Design the serving layer  1.4 Implement physical data storage structures  2.1 Ingest and transform data  2.2 Design and develop a batch processing solution  2.3 Design and develop a stream processing solution  2.4 Manage batches and pipelines |