

# DP-203: Data Engineering on Microsoft Azure [203T00-A] training including Exam Voucher

## Group Training

Training code	CGADP203CE
Spoken Language	English
Language Materials	English
Dayparts	8
Price	€2.100,00 excl. VAT No extra costs.

## What is DP-203: Data Engineering on Microsoft Azure [203T00-A] including Exam Voucher

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.

This course uses MOC (Microsoft Official Courseware) and will be given by an experienced MCT (Microsoft Certified Trainer).

See the below modules for more information:

Module 1: Explore compute and storage options for data engineering workloads

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 Azure Synapse Analytics
- Describe Azure Databricks
- Introduction to Azure Data Lake storage
- Describe Delta Lake architecture
- Work with data streams by using Azure Stream Analytics

#### Module 2: Design and implement 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.

#### Lessons

- Design a multidimensional schema to optimize analytical workloads
- Code-free transformation at scale with Azure Data Factory
- Populate slowly changing dimensions in Azure Synapse Analytics pipelines

#### Module 3: Data engineering considerations for source files

This module explores data engineering considerations that are common when loading data into a modern data warehouse analytical from files stored in an Azure Data Lake, and understanding the security consideration associated with storing files stored in the data lake.

#### Lessons

- Design a Modern Data Warehouse using Azure Synapse Analytics
- Secure a data warehouse in Azure Synapse Analytics

#### Module 4: Run interactive queries using Azure Synapse Analytics 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

- Explore Azure Synapse serverless SQL pools capabilities
- Query data in the lake using Azure Synapse serverless SQL pools
- Create metadata objects in Azure Synapse serverless SQL pools
- Secure data and manage users in Azure Synapse serverless SQL pools

#### Module 5: Explore, transform, and load 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.

Lessons

- Understand big data engineering with Apache Spark in Azure Synapse Analytics
- Ingest data with Apache Spark notebooks in Azure Synapse Analytics
- Transform data with DataFrames in Apache Spark Pools in Azure Synapse Analytics
- Integrate SQL and Apache Spark pools in Azure Synapse Analytics

Module 6: 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.

Lessons

- Describe Azure Databricks
- Read and write data in Azure Databricks
- Work with DataFrames in Azure Databricks
- Work with DataFrames advanced methods in Azure Databricks

Module 7: 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

- Use data loading best practices in Azure Synapse Analytics
- Petabyte-scale ingestion with Azure Data Factory

Module 8: Transform data with Azure Data Factory or Azure 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 perform data movement into one or more data sinks.

Lessons

- Data integration with Azure Data Factory or Azure Synapse Pipelines
- Code-free transformation at scale with Azure Data Factory or Azure Synapse Pipelines

Module 9: Orchestrate data movement and transformation in Azure Synapse Pipelines

In this module, you will learn how to create linked services, and orchestrate data movement and transformation using notebooks in Azure Synapse Pipelines.

Lessons

- Orchestrate data movement and transformation in Azure Data Factory

Module 10: 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.

## Lessons

- Optimize data warehouse query performance in Azure Synapse Analytics
- Understand data warehouse developer features of Azure Synapse Analytics

### Module 11: Analyze and Optimize 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.

## Lessons

- Analyze and optimize data warehouse storage in Azure Synapse Analytics

### Module 12: Support Hybrid Transactional Analytical Processing (HTAP) with Azure Synapse Link

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

- Design hybrid transactional and analytical processing using Azure Synapse Analytics
- Configure Azure Synapse Link with Azure Cosmos DB
- Query Azure Cosmos DB with Apache Spark pools
- Query Azure Cosmos DB with serverless SQL pools

### Module 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.

## Lessons

- Secure a data warehouse in Azure Synapse Analytics
- Configure and manage secrets in Azure Key Vault
- Implement compliance controls for sensitive data

### Module 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.

## Lessons

- Enable reliable messaging for Big Data applications using Azure Event Hubs
- Work with data streams by using Azure Stream Analytics
- Ingest data streams with Azure Stream Analytics

#### Module 15: Create 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.

##### Lessons

- Process streaming data with Azure Databricks structured streaming

#### Module 16: Build reports using Power BI integration with Azure Synapse Analytics

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 data source 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.

##### Lessons

- Create reports with Power BI using its integration with Azure Synapse Analytics

#### Module 17: Perform Integrated Machine Learning Processes in Azure Synapse Analytics

This module explores the integrated, end-to-end Azure Machine Learning and Azure Cognitive Services experience in Azure Synapse Analytics. You will learn how to connect an Azure Synapse Analytics workspace to an Azure Machine Learning workspace using a Linked Service and then trigger an Automated ML experiment that uses data from a Spark table. You will also learn how to use trained models from Azure Machine Learning or Azure Cognitive Services to enrich data in a SQL pool table and then serve prediction results using Power BI.

##### Lessons

- Use the integrated machine learning process in Azure Synapse Analytics

## Who should attend DP-203: Data Engineering on Microsoft Azure [203T00-A] including Exam Voucher

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

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

## Objectives

After completing this course, you will be able 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
- Build reports using Power BI integration with Azure Synapse Analytics
- Perform Integrated Machine Learning Processes in Azure Synapse Analytics

## Exam information

- Exam duration (minutes): 120
- % extra time for non-native speakers: -
- Number of exam questions: 40-60
- Minimum correct questions: Variable
- Exam style: Multiple choice
- Open book: No

If a third-party copyright applies to this course, you will find the copyright on <https://academy.capgemini.nl/en/topic/trademarks/>

Capgemini Academy's general terms and conditions are applied to all products and services mentioned within this document. For the latest version please check <https://academy.capgemini.com/>. The rates of products and services mentioned in this document are subject to change. For the most recent rates, please also visit our website.

---

## About Capgemini Academy

Capgemini Academy's professionals offer what people in IT need. Our professionals have a keen eye for motivation, talent and are aware of specific contexts and circumstances. They move people to move. Programmes and courses that originate from daily experience of our both didactical and substantively strong trainers, light a fire within the individual IT professionals. Real life stories of our professionals' experience that tell how to solve problems and work with the people around it, do the rest.

An organization, like ours, helps people and their organizations day by day to get the best out of themselves and each other. We prepare them to defy tomorrow's challenges. We stimulate learning and curiosity. In order for individual IT professionals and their employers, to build better, longer and more intensive relationships. For mutual benefit.

Capgemini Academy. We transform IT professionals  
[academy.capgemini.nl](https://academy.capgemini.nl)

IN/3A-018.18