September 2023

Azure Data Engineer Learning Pathway (1/2)

www.aka.ms/pathways

Role based certification

Azure Data Engineer

DP-203: Azure Data Engineer

Skills measured:

- Design and implement data storage
- Design and develop data processing
- Design and implement data security
- Monitor and optimize data storage and data processing

Exam Study Guide | Course Page | Azure Data Architecture Guide

30 Day Challenge | Exam Page | Practice Assessment

Official Curriculum

- New to the Cloud or Azure? Start with Azure Fundamentals
- New to data solutions on Azure? Build your knowledge with Data Fundamentals
- Intro to data classification and protection

Get started with data engineering on Azure

- Introduction to data engineering on Azure
- Introduction to Azure Data Lake Storage Gen2
- Introduction to Azure Synapse Analytics

Build data analytics solutions using Azure Synapse serverless SQL pools

- Use Azure Synapse serverless SQL pool to query files in a data lake
- Use Azure Synapse serverless SQL pools to transform data in a data lake
- Create a lake database in Azure Synapse Analytics
- Secure data and manage users in Azure Synapse serverless SQL pools

Perform data engineering with Azure Synapse Apache Spark Pools

- Analyze data with Apache Spark in Azure Synapse Analytic
- Transform data with Spark in Azure Synapse Analytics
- Use Delta Lake in Azure Synapse Analytics

Work with Data Warehouses using Azure Synapse Analytics

- Analyze data in a relational data warehouse
- Load data into a relational data warehouse
- Manage and monitor data warehouse activities in Azure Synapse Analytics
- Analyze and optimize data warehouse storage in Azure Synapse Analytics
- Secure a data warehouse in Azure Synapse Analytics

Transfer and transform data with Azure Synapse Analytics pipeline

- Build a data pipeline in Azure Synapse Analytics
- Use Spark Notebooks in an Azure Synapse Pipeline

Work with Hybrid Transactional and Analytical Processing Solutions using Azure Synapse Analytics

- Plan hybrid transactional and analytical processing using Azure Synapse Analytics
- Implement Azure Synapse Link with Azure Cosmos DB
- Implement Azure Synapse Link for SQL

Implement a Data Streaming Solution with Azure Stream Analytics

- Get started with Azure Stream Analytics
- Ingest streaming data using Azure Stream Analytics and Azure Synapse Analytics
- Visualize real-time data with Azure Stream Analytics and Power BI

Govern data across an enterprise

- Introduction to Microsoft Purview
- Discover trusted data using Microsoft Purview
- Catalog data artifacts by using Microsoft Purview
- Manage Power BI assets by using Microsoft Purview
- Integrate Microsoft Purview and Azure Synapse Analytics

Data engineering with Azure Databricks

- Explore Azure Databricks
- Use Apache Spark in Azure Databricks
- Use Delta Lake in Azure Databricks
- Use SQL Warehouses in Azure Databricks
- Run Azure Databricks Notebooks with Azure Data Factory

Additional Study

Design and implement data storage:

- Understand Azure Data Lake Storage Gen2
- Access tiers for Azure Blob Storage
- Storage considerations when using Azure Synapse serverless SQL pools
- Query a Parquet file using Azure Synapse serverless SQL pools
- Dynamic file pruning
- Understand table distribution design
- Partitioning tables in dedicated SQL pool
- Understand table distribution design
- Best practices for dedicated SQL pools in Azure Synapse Analytics
- Star Schema
- Multidimensional Schemas and Data
- Manage retention of historical data in system-versioned temporal tables
- Getting started with temporal tables
- Create and configure a self-hosted integration runtime
- Manage self-hosted integration runtime
- Choosing an analytical data store in Azure
- Synapse Analytics shared metadata tables
- When do you use Apache Spark pools?
- Data Compression
- Exercise - Use table distribution and indexes to improve performance
- Change storage account is replication
- Slowly Changing Dimension Transformation
- Populate slowly changing dimensions
- Create external tables in Azure Synapse serverless SQL pools
- Views in Synapse serverless SQL pools
- Tutorial: Load data to Azure Synapse Analytics SQL pool
- Create, develop, and maintain Synapse notebooks in Azure Synapse Analytics
- Quickstart: Create a serverless Apache Spark pool in Synapse Analytics using web tools

Design and develop data processing:

- Common practices for data loading
- Tutorial: Extract, transform, and load data by using Azure Databricks
- Understand the Streaming Analytics Workflow
- Handling bad records and files
- Prepare and transform data with Azure Synapse Analytics
- Analyse complex data types in Azure Synapse Analytics
- Understand data store models
- Prepare and transform data
- Define a modern data warehouse architecture
- Choosing a batch processing technology
- Manage source data files
- Copy activity in Azure Data Factory
- MERGE (Transact-SQL)
- Continuous integration and delivery for Azure Synapse workspace
- Handle SQL truncation error rows in Data Factory mapping data flows
- Backup and restore in Azure Synapse Dedicated SQL pool
- Implement workload management
- Use extended Apache Spark history server to debug and diagnose Apache Spark applications
- Enterprise Data Warehouse Architecture
- Stream processing with Azure Databricks
- Azure Synapse Analytics
- Monitoring for performance efficiency
- Work with windowing functions
- Schema drift
- Time handling in Stream Analytics
- Checkpoint and replay concepts in Azure Stream Analytics jobs
- Scale an Azure Stream Analytics job to increase throughput

Azure Data Engineer Learning Pathway (2/2)

www.aka.ms/pathways

Additional Study

Design and develop data processing:
- Use repartitioning to optimize processing
- Azure Stream Analytics output error policy
- Stream Analytics output to Cosmos DB
- Stream processing with Stream Analytics
- Data Loading best practices
- Get Started with Synapse Analytics
- Monitor your Synapse Workspace

Design and implement data security :
- Implement encryption
- Data ingestion security considerations
- Configure authentication
- Access control lists (ACLs) in Azure Data Lake Storage Gen2
- Synapse access control
- Column-level security
- Manage authorization through column and row level security
- Manage user permissions
- Auditing for Azure SQL Database and Azure Synapse Analytics
- Retention Policy on storage accounts
- Understand network security options
- Dynamic Data Masking
- Secure a dedicated SQL pool

Monitor and optimize data storage and data processing :
- Monitor and Alert Data Factory by using Azure Monitor
- Exercise - implement workload management
- Monitor your Azure Synapse Analytics dedicated SQL pool workload using DMVs
- Collect custom logs with Log Analytics agent
- Use Synapse Studio to monitor your workspace pipeline runs
- Deploying Apache Airflow in Azure to build and run data pipelines

Monitor and optimize data storage and data processing :
- Auto Optimize in Azure Databricks
- Modify user-defined functions
- Designing distributed tables
- Data spillage scenario - Search and purge
- Quickstart: Create an Azure Synapse workspace using an ARM template
- Indexing dedicated SQL pool tables
- Performance tuning with result set caching
- Optimize Apache Spark jobs
- Troubleshoot library installation errors
- Debug data factory pipelines

Getting started with Microsoft Fabric

Reshape how everyone accesses, manages, and acts on data and insights by connecting every data source and analytics service together—on a single, AI-powered platform.

Discover how Microsoft Fabric can meet your enterprise's analytics needs in one platform. Learn about Microsoft Fabric, how it works, and identify how you can use it for your analytics needs.

- Introduction to end-to-end analytics using Microsoft Fabric
- Get started with lakehouses in Microsoft Fabric
- Use Apache Spark in Microsoft Fabric
- Work with Delta Lake tables in Microsoft Fabric
- Use Data Factory pipelines in Microsoft Fabric
- Ingest Data with Dataflows Gen2 in Microsoft Fabric
- Get started with data warehouses in Microsoft Fabric
- Get started with Real-Time Analytics in Microsoft Fabric

Implement a Lakehouse with Microsoft Fabric

A Lakehouse combined the flexible and scalable storage of a data lake with the analytical querying and modelling capabilities of a data warehouse. Microsoft Fabric provides a lakehouse solution that powers end-to-end data analytics in a single software-as-a-service platform.. This learning path introduces the foundational components of implementing a data lakehouse with Microsoft Fabric.

- Introduction to end-to-end analytics using Microsoft Fabric
- Get started with lakehouses in Microsoft Fabric
- Use Apache Spark in Microsoft Fabric
- Work with Delta Lake tables in Microsoft Fabric
- Ingest Data with Dataflows Gen2 in Microsoft Fabric
- Use Data Factory pipelines in Microsoft Fabric

Other Fabric Resources

Fabric, a unified software-as-a-service (SaaS) offering, reshapes how you use data. Data is now stored in a single open format in OneLake, accessible by all the analytics engines in the platform. Fabric offers scalability, cost-effectiveness, accessibility from anywhere with an internet connection, and continuous updates and maintenance provided by Microsoft.

- Fabric website
- Microsoft Fabric Documentation
- End-to-end tutorials in Microsoft Fabric
- Get the free e-book on getting started with Fabric
- Microsoft Fabric Guided Tour
- Webinar Series: Introduction to Microsoft Fabric
- Join the Fabric Community