March 2023

Azure Data Engineer Learning Pathway

www.aka.ms/pathways

Getting started

Azure Data Engineers integrate, transform, and consolidate data from various structured and unstructured data systems into structures that are suitable for building analytics solutions.

Microsoft Learn

- Build your Tech resilience
- 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
- Intro to data protection and privacy regulations

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

Additional Study

- 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
- Lifecycle Management
- Exercise: Flatten nested structures and explode arrays with Apache Spark in synapse
- Preserve metadata and ACLs using copy activity in Azure Data Factory

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
- 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
- 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

Role based certification

Azure Data Engineer

DP-203: Data Engineering on Microsoft Azure

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

Microsoft Learn content:

- Get started with data engineering on Azure
- Build data analytics solutions using Azure Synapse serverless SQL pools
- Perform data engineering with Azure Synapse Apache Spark Pools
- Work with Data Warehouses using Synapse Analytics
- Transfer and transform data with Synapse Analytics pipelines
- Work with Hybrid Transactional and Analytical Processing Solutions using Azure Synapse Analytics
- Implement a Data Streaming Solution with Azure Stream Analytics
- Govern data across an enterprise
- Data engineering with Azure Databricks

Exam Study Guide | Course Page

30 Day Challenge | Exam Page

Azure Data Architecture Guide | Practice Assessment