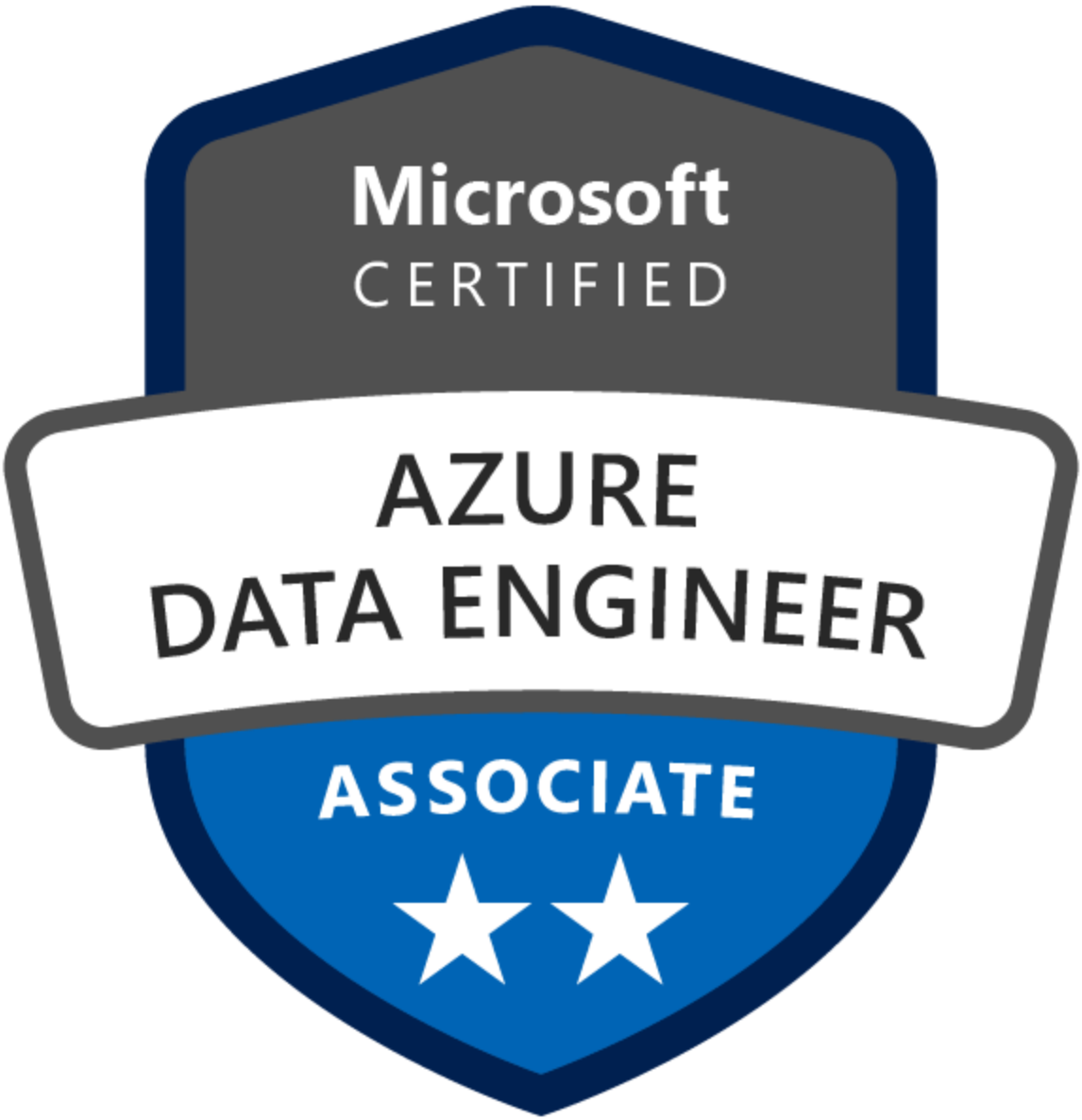

Design and Implement Data Storage (40-45%)

Design a Data Storage Structure

Design an Azure Data Lake solution

Introduction to Azure Data Lake Storage Gen2

Building your Data Lake on Azure Data Lake Storage gen2

Recommend file types for storage

Example scenarios for core Azure Storage services

Recommend file types for analytical queries

Query data in Azure Data Lake using Azure Data Explorer

Query Azure Storage analytics logs in Azure Log Analytics

Design for efficient querying

Design Azure Table storage for queries

Guidelines for table design

Design for data pruning

Dynamic file pruning

Design a folder structure that represents the levels of data transformation

Copy & transform data in Data Lake Storage using Azure Data Factory

Design a distribution strategy

How to choose the right data distribution strategy for Azure Synapse?

Guidance for designing distributed tables in Azure Synapse

Design a data archiving solution

Designing a data archiving strategy on Microsoft Azure

Solution architecture: Archive on-premises data to the cloud

Design a Partition Strategy

Design a partitioning strategy for files

File Partition using Azure Data Factory

Incrementally copy new files by using the Copy Data tool

Design a partitioning strategy for analytical workloads

Best practices for Azure Databricks

Partitions in tabular models

Automated Partition Management with Azure Analysis Services

Design a partitioning strategy for efficiency/performance

Designing partitions for query performance

Design a partitioning strategy for Azure Synapse Analytics

Partitioning tables in Azure Synapse Analytics

Identify when partitioning is needed in Azure Data Lake Storage Gen2

Partitioning in ADLS Gen2

Design the Serving Layer

Design star schemas

Star schema overview

Designing Star Schema

Design slowly changing dimensions

Design a Slowly Changing Dimension (SCD) in Azure Data Factory

Design a dimensional hierarchy

Simple hierarchical dimensions

Hierarchies in tabular models

Design a solution for temporal data

What is temporal data?

Getting started with temporal tables in Azure SQL Database

Design for incremental loading

Incrementally load data from a source to a destination datastore

Incrementally load data from Azure SQL Database to Blob storage

Design analytical stores

Choosing an analytical data store in Azure

Azure Cosmos DB analytical store

Design meta stores in Azure Synapse Analytics and Azure Databricks

Azure Synapse Analytics shared metadata tables

Manage Apache Hive metastore for Databricks

Implement Physical Data Storage Structures

Implement compression

Data compression in Azure SQL Database

Forum discussion on compression in Azure SQL DB

Implement partitioning

Data partitioning strategies

How to partition your data in Azure Cosmos DB?

Implement sharding

Sharding patterns and strategies

Adding a shard using Elastic Database tools

Implement different table geometries with Azure Synapse Analytics pools

Spatial Types – geometry

Table data types for dedicated SQL pool

Implement data redundancy

Azure Storage redundancy

Change how a storage account is replicated

Implement distributions

Distributions in Azure Synapse Analytics

Examples for table distribution

Implement data archiving

Archive on-premises data to the cloud

Rehydrate blob data from the archive tier

Implement Logical Data Structures

Build a temporal data solution

Azure SQL Temporal Tables

Creating a system-versioned temporal table

Build a slowly changing dimension

Azure Data Factory Data Flow: Building Slowly Changing Dimensions

How to implement Slowly Changing Dimension Type 1?

Slowly Changing Dim Type 2 with ADF Mapping Data Flows

Build a logical folder structure

Creating an Azure Blob Hierarchy

Modeling a directory structure on Azure Blob Storage

Build external tables

Use external tables with Synapse SQL

Create external tables in Azure Storage / Azure Data Lake

Implement file and folder structures for efficient querying and data pruning

Query multiple files or folders

Query folders and multiple files

Implement the Serving Layer

Deliver data in a relational star schema

Data models within Azure Analysis Services

Deliver data in Parquet files

What is a Parquet file?

Parquet format in Azure Data Factory

Parquet format in Azure Data Lake Analytics

Maintain metadata

Preserve metadata using copy activity in Azure Data Factory

Implement a dimensional hierarchy

Create and manage hierarchies

Design and Develop Data Processing (25-30%)

Ingest and Transform Data

Transform data by using Apache Spark

Transform data in the cloud by using a Spark activity in ADF

Transform data using Spark activity in Azure Data Factory

Transform data by using Transact-SQL

Apply SQL Transformation in AML

Transform data by using Data Factory

Transform data in Azure Data Factory

Transform data using mapping data flows

Transform data by using Azure Synapse Pipelines

Use Azure Synapse Analytics to create a pipeline for data transformation

Transform data by using Stream Analytics

Transform data by using Azure Stream Analytics

Cleanse data

Data Cleansing

Clean Missing Data module

Split data

Split data

Split Data module

Shred JSON

JSON in your Azure SQL Database? Let’s benchmark some options!

Encode and decode data

Azure Data Factory copy activity with Base64 encoded string

Handling data encoding issues while loading data to SQL Data Warehouse

Configure error handling for the transformation

Handle SQL truncation error rows in Data Factory mapping data flows

Troubleshoot mapping data flows in Azure Data Factory

Error row handling

Normalize and denormalize values

Normalize data in AML

Normalize Data module

How do I denormalize data in Azure Machine Learning Studio?

Transform data by using Scala

ETL by using Azure Databricks & Scala

Perform data exploratory analysis

Exploratory Data Analysis with Azure Synapse Analytics

Perform EDA in Azure Data Explorer with Web UI

Design and Develop a Batch Processing Solution

Develop batch processing solutions by using Data Factory, Data Lake, Spark, Azure Synapse Pipelines, PolyBase, and Azure Databricks

Batch processing in Azure

Choosing a batch processing technology in Azure

Building batch data processing solutions in Microsoft Azure

Process large-scale datasets by using Data Factory & Batch

Run Spark Jobs using Azure Container Registry & Blob storage

Batch Processing with Databricks and Data Factory in Azure

Create data pipelines

Create a pipeline in Azure Data Factory

Build a data pipeline by using ADF, DevOps, & Machine Learning

Design and implement incremental data loads

Load data incrementally from Azure SQL Database to Blob storage

Implement incremental data loading with ADF

Incremental data loading using Azure Data Factory

Design and develop slowly changing dimensions

Processing Slowly Changing Dimensions with ADF Data Flows

Handle security and compliance requirements

Azure security baseline for Batch

Policy Regulatory Compliance controls for Azure Batch

Scale resources

Automatically scale compute nodes in an Azure Batch pool

Configure the batch size

Choose a VM size & image for compute nodes

Design and create tests for data pipelines

Unit testing Azure Data Factory pipelines

Integrate Jupyter/IPython notebooks into a data pipeline

Set up a Python development environment for AML

Explore Azure Machine Learning with Jupyter Notebooks

Handle duplicate data

Handle duplicate data in Azure Data Explorer

Dedupe rows by using data flow snippets

Remove duplicate rows module

Handle missing data

Clean missing data module

Methods for handling missing values

Handle late-arriving data

Late arriving events

Late arrival tolerance

Upsert data

Optimize Azure SQL Upsert scenarios

Implement Upsert using Dataflow

Regress to a previous state

Monitor Batch solutions by counting tasks & nodes by state

Design and configure exception handling

Error handling and detection in Azure Batch

Configure batch retention

Manage task lifetime

Design a batch processing solution

Batch processing

Debug Spark jobs by using the Spark UI

Debug Apache Spark jobs with the Spark UI

Design and Develop a Stream Processing Solution

Develop a stream processing solution by using Stream Analytics, Azure Databricks, and Azure Event Hubs

Implement a data streaming solution with Azure Streaming Analytics

Stream processing with Azure Databricks

Stream data into Azure Databricks using Event Hubs

Process data by using Spark structured streaming

Structured Streaming

Overview of Apache Spark Structured Streaming

Structured Streaming tutorial

Monitor for performance and functional regressions

Understand Stream Analytics job monitoring

Design and create windowed aggregates

Introduction to Stream Analytics windowing functions

Windowing functions (Azure Stream Analytics)

Handle schema drift

Schema drift in the mapping data flow

Process time-series data

Time series solutions

Understand time handling in Azure Stream Analytics

Process across partitions

Stream processing with Azure Stream Analytics

Use repartitioning to optimize processing with Stream Analytics

Process within one partition

Maximize throughput with repartitioning

Configure checkpoints/watermarking during processing

Checkpoints in Azure Stream Analytics jobs

Watermarks

Illustrated example of watermarks

How to calculate watermark for Streaming Analytics?

Scale resources

Understand and adjust Streaming Units

Scale an Azure Stream Analytics job to increase throughput

Design and create tests for data pipelines

Test live data locally using Azure Stream Analytics tools

Test an Azure Stream Analytics job in the portal

Optimize pipelines for analytical or transactional purposes

Use repartitioning to optimize processing

Leverage query parallelization

Handle interruptions

Avoid service interruptions in Azure Stream Analytics jobs

Design and configure exception handling

Azure Stream Analytics output error policy

Exception handling in Azure Stream Analytics

Upsert data

Upserts from Stream Analytics

Azure Stream Processing upsert to DocumentDB

Replay archived stream data

Estimate replay catch-up time

Design a stream processing solution

Stream processing with Azure Stream Analytics

Manage Batches and Pipelines

Trigger batches

Trigger a Batch job using Azure Functions

Handle failed batch loads

Check for pool and node errors

Validate batch loads

Job and task error checking

Manage data pipelines in Data Factory/Synapse Pipelines

Monitor and manage Azure Data Factory pipelines

Managing the mapping data flow graph

Schedule data pipelines in Data Factory/Synapse Pipelines

Create a trigger that runs a pipeline on a schedule

Implement version control for pipeline artifacts

Source control in Azure Data Factory

Manage Spark jobs in a pipeline

Monitor a pipeline with Spark activity

Design and Implement Data Security (10-15%)

Design Security for Data Policies and Standards

Design data encryption for data at rest and in transit

Azure Data Encryption at rest

Azure Storage Encryption for data at rest

Protect data in transit

Design a data auditing strategy

Auditing for Azure SQL Database & Synapse Analytics

Design a data masking strategy

Dynamic data masking

Static Data Masking for Azure SQL Database

Design for data privacy

Data privacy in the trusted cloud

Design a data retention policy

Understand data retention in Azure Time Series Insights

Design to purge data based on business requirements

Data purge

Enable data purge on your Azure Data Explorer cluster

Design Azure role-based access control (Azure RBAC) and POSIX-like Access Control List (ACL) for Data Lake Storage Gen2

Role-based access control (Azure RBAC)

Access control lists in Azure Data Lake Storage Gen2

Design row-level and column-level security

Row-level security in Azure SQL Database

Column-level security

Implement Data Security

Implement data masking

Get started with SQL Database dynamic data masking

Encrypt data at rest and in motion

Transparent data encryption for SQL Database

Implement row-level and column-level security

Row-level security in Azure SQL Database

Column-level security

Implement Azure RBAC

Use the portal to assign a role for access to blob & queue data

Implement POSIX-like ACLs for Data Lake Storage Gen2

Use PowerShell to manage ACLs in Data Lake Storage Gen2

Implement a data retention policy

Configuring retention in Azure Time Series Insights

Implement a data auditing strategy

Set up auditing for your server

Manage identities, keys, and secrets across different data platform technologies

Manage keys, secrets, for secure data with Key Vault

Implement secure endpoints (private and public)

Use private endpoints for Azure Storage

Use Azure SQL MI securely with public endpoints

Configure public endpoint in Managed Instance

Implement resource tokens in Azure Databricks

Authentication using Databricks personal access tokens

Load a DataFrame with sensitive information

DataFrames tutorial

Write encrypted data to tables or Parquet files

Use Parquet with Azure Data Lake Analytics

Manage sensitive information

Security Control: Data protection

Monitor and Optimize Data Storage and Data Processing (10-15%)

Monitor Data Storage and Data Processing

Implement logging used by Azure Monitor

Azure Monitor Logs overview

Collect custom logs with Log Analytics agent in Azure Monitor

Configure monitoring services

Monitoring Azure resources with Azure Monitor

Enable Azure Monitor for VMs overview

Measure the performance of data movement

Copy activity performance and scalability guide

Monitor and update statistics about data across a system

Update statistics in Synapse SQL

Update Statistics (Transact-SQL)

Monitor data pipeline performance

Monitor and alert Data Factory by using Azure Monitor

Measure query performance

Query Performance Insight for Azure SQL Database

How to measure the performance of the Azure SQL DB?

Monitor cluster performance

Monitor cluster performance in Azure HDInsight

Understand custom logging options

Collect custom logs with Log Analytics agent in Azure Monitor

Schedule and monitor pipeline tests

How to monitor & manage big data pipelines with ADF?

Monitor and manage Azure Data Factory pipelines

Interpret Azure Monitor metrics and logs

Azure Monitor Metrics overview

Overview of Azure platform logs

Interpret a Spark directed acyclic graph (DAG)

Directed Acyclic Graph DAG in Apache Spark

Understanding your Apache Spark application through visualization

Optimize and Troubleshoot Data Storage and Data Processing

Compact small files

Auto Optimize

Rewrite user-defined functions (UDFs)

Modify user-defined functions

Handle skew in data

Resolve data-skew problems

Handle data spill

Data security Q&A (See Question 7)

Tune shuffle partitions

Use Unravel to tune Spark data partitioning

Find shuffling in a pipeline

Lightning-fast query performance with Azure SQL Data Warehouse

Optimize resource management

How to optimize your Azure environment?

Azure resource management tips to optimize a cloud deployment

Tune queries by using indexers

Automatic tuning for SQL Database

Tune queries by using cache

Performance tuning with a result set caching

Optimize pipelines for analytical or transactional purposes

Hyperspace: An indexing subsystem for Apache Spark

Optimize pipeline for descriptive versus analytical workloads

Optimize Apache Spark jobs in Azure Synapse Analytics

Troubleshoot a failed spark job

Troubleshoot Apache Spark by using Azure HDInsight

Troubleshoot a slow or failing job on an HDInsight cluster

Troubleshoot a failed pipeline run

Troubleshoot pipeline orchestration in Azure Data Factory

Before taking Microsoft's DP-203 exam, I recommend reading the documentation at the following links because these specific topics were not covered anywhere else in this learning path:

- Designing partitions for query performance
- Incrementally load data from a source data store to a destination data store
- Azure Synapse Analytics shared metadata tables
- Use external tables with Synapse SQL
- Secure a dedicated SQL pool (formerly SQL DW) in Azure Synapse Analytics
- Azure SQL Transparent Data Encryption with customer-managed key
- Using IDENTITY to create surrogate keys using dedicated SQL pool in AzureSynapse Analytics
- External Apache Hive metastore
- Parquet file
- Preserve metadata and ACLs using copy activity in Azure Data Factory
- Shredding JSON (only read “Our test bench” section)
- Handling data encoding issues while loading data to SQL Data Warehouse
- Security considerations for data movement in Azure Data Factory
- Dedupe rows and find nulls by using data flow snippets
- Understanding Pipeline Failures and Error Handling
- Keeping Azure Data Factory metrics and pipeline-run data
- Handle SQL truncation error rows in Data Factory mapping data flows
- Tumbling window trigger
- Read input in any format using .NET custom deserializers (Preview)
- Debug Spark jobs by using the Spark UI
- Spark Structured Streaming tutorial
- Monitoring for performance efficiency
- Checkpoint and replay concepts in Azure Stream Analytics jobs
- Session window (Azure Stream Analytics)
- Exception handling in Azure Stream Analytics
- Source control in Azure Data Factory
- Column-level security
- Use private endpoints for Azure Storage
- Copy activity performance and scalability guide
- Collect custom logs with Log Analytics agent in Azure Monitor
- Managing dependencies in data pipelines
- Autoscaling types
- Troubleshoot performance bottlenecks in Azure Databricks
- Troubleshoot pipeline orchestration and triggers in Azure Data Factory