DP-203 : Data Engineering on Microsoft Azure Syllabus

Section 1: What is Partition strategy

- Implement a partition strategy for files
- Implement a partition strategy for analytical workloads, streaming workloads
- Implement a partition strategy for Azure Synapse Analytics
- Azure Data Lake Storage Gen2

Section 2: Understanding the data exploration layer

- How to create and execute queries
- Compute solution that leverages SQL server less and Spark cluster
- Recommend and implement Azure Synapse Analytics database templates
- Data lineage to Microsoft Purview
- Browse and search metadata in Microsoft Purview Data Catalog

Section 3: Overview of Ingest and transform data

- Design and implement incremental loads
- Transform data by using Apache Spark, Transact-SQL (T-SQL)
- Ingest and transform data by using Azure Synapse Pipelines or Azure Data Factory
- Transform data by using Azure Stream Analytics
- What is Cleanse data
- How to handle duplicate data
- How to handle missing data
- How to handle late-arriving data
- What is Split data
- Concept of Shred JSON
- Encode and decode data
- Configure error handling for a transformation

- Normalize and denormalize data
- How to perform data exploratory analysis

Section 4: Batch processing solution

- Develop batch processing solutions by using
  - Azure Data Lake Storage
  - Azure Data bricks
  - Azure Synapse Analytics
  - Azure Data Factory
- Use Poly Base to load data to a SQL pool
- Implement Azure Synapse Link and query the replicated data
- To create data pipelines
- What is Scale resources
- Configure the batch size
- To create tests for data pipelines
- Integrate Jupyter or Python notebooks into a data pipeline
- Upsert data
- Revert data to a previous state
- Configure exception handling, batch retention
- Read from and write to a delta lake

Section 5: How to develop a stream processing solution

- Create a stream processing solution by using Stream Analytics and Azure Event Hubs
- Process data by using Spark structured streaming
- Create windowed aggregates
- Handle schema drift
- Process time series data
- Process data across partitions
- Process within one partition
- Configure checkpoints and watermarking during processing
- Scale resources
- Create tests for data pipelines
- Optimize pipelines for analytical or transactional purposes

- Handle interruptions
- Configure exception handling
- Upsert data
- Replay archived stream data

Section 6: How to manage batches and pipelines

- Trigger batches
- Handle failed batch loads
- Validate batch loads
- Manage data pipelines in Azure Data Factory or Azure Synapse Pipelines
- Schedule data pipelines in Data Factory or Azure Synapse Pipelines
- Implement version control for pipeline artifacts
- Manage Spark jobs in a pipeline

Section 7: Implement data security

- Implement data masking
- Encrypt data at rest and in motion
- Implement row-level and column-level security
- Implement Azure role-based access control (RBAC)
- Implement POSIX-like access control lists (ACLs) for Data Lake Storage Gen2
- Implement a data retention policy
- Implement secure endpoints (private and public)
- Implement resource tokens in Azure Databricks
- Load a DataFrame with sensitive information
- Write encrypted data to tables or Parquet files
- Manage sensitive information

Section 8: Monitoring data storage and data processing

- Implement logging used by Azure Monitor
- Configure monitoring services
- Monitor stream processing
- Measure performance of data movement

- Monitor and update statistics about data across a system
- Monitor data pipeline performance
- To measure query performance
- How to schedule and monitor pipeline tests
- Interpret Azure Monitor metrics and logs
- How to implement a pipeline alert strategy

Section 9: Optimize and troubleshoot data storage and data processing

- Compact small files
- How to handle skew in data
- Handle data spill
- Optimize resource management
- Tune queries by using indexers, cache
- Troubleshoot a failed Spark job
- Troubleshoot a failed pipeline run, including activities executed in external services