Activity 02: Data Engineering Discussion

Audience Poll 1

Q: You need to perform some data preparation on data stored in ADLS Gen 2. Which option should you use to run the transformations (pick only one)?

A) Synapse dedicated SQL pool / serverless SQL pool

B) Apache Spark in Azure Synapse Analytics

C) Integration Runtime

Audience Poll 2

Q: You are performing initial exploration of the data and experimenting with the necessary transformations. Which option should you use to run the transformations (pick only one)?

A) Synapse dedicated SQL pool / serverless SQL pool

B) Apache Spark in Azure Synapse Analytics

C) Integration Runtime

Audience Poll 3

Q: You want to process a subset of files in folder filled with CSV files, all having the same schema. Which option should you use to run the transformations (pick only one)?

A) Synapse dedicated SQL pool / serverless SQL pool

B) Apache Spark in Azure Synapse Analytics

C) Integration Runtime

Audience Poll 4

Q: You need to transform the data on-premises or within a specific VNET before loading it. Which option should you use to run the transformations (pick only one)?

A) Synapse dedicated SQL pool / serverless SQL pool

B) Apache Spark in Azure Synapse Analytics

C) Integration Runtime

Audience Poll 5

Q: You want to flatten hierarchical fields in JSON to a tabular structure. Which option should you use to run the transformations (pick only one)?

A) Synapse dedicated SQL pool / serverless SQL pool

B) Apache Spark in Azure Synapse Analytics

C) Integration Runtime

Audience Poll 6

Q: You are handling file formats other than delimited (CSV), JSON or Parquet. Which option should you use to run the transformations (pick only one)?

A) Synapse dedicated SQL pool / serverless SQL pool

B) Apache Spark in Azure Synapse Analytics

C) Integration Runtime

Audience Poll 7

Q: The delimited data is badly formatted. Which option should you use to run the transformations (pick only one)?

A) Synapse dedicated SQL pool / serverless SQL pool

B) Apache Spark in Azure Synapse Analytics

C) Integration Runtime

Decision Matrix Summary

Decision Point	dedicated SQL pool / serverless SQL pool	Apache Spark Pool	Integration Runtime	Discussion Comment
Initial exploration of the data and experimenting with the necessary transformation	X			Start with T-SQL, generally
Process a folder filled with CSV files of the same schema	X			Use **T-SQL OPENROWSET** statement
Process a subset of files in folder filled with CSV files of the same schema	X			Use **T-SQL OPENROWSET** statement with wildcards (*) in the path
Transform the data on-premises or within a specific VNET before loading it			X	Use a **Self-Hosted Integration Runtime** on-premises
Transform the data in a code free way			X	Use an **Azure Integration Runtime**
Need to flatten hierarchical fields in JSON to a tabular structure	X			Use Azure Synapse SQL Pools or serverless along with the T-SQL OPENJSON, JSON_VALUE, and JSON_QUERY statements
Need to unpack or flatten deeply nested JSON		X		Use **Spark** to deal with very complex JSON
Handling file formats other than delimited (CSV), JSON or Parquet		X		Use **Spark** to handle the broadest set of file formats (e.g., Orc, Avro, others)
Handling ZIP archived data files		X		Use Spark to unzip the files to storage before processing
Delimited data is badly formatted		X		Use **Spark** to handle particularly poorly formatted files
You want to leverage open source libraries to help you with the data cleansing		X		Use Python or Scala opens source libraries with **Spark**