Lab 6 –Transform Spark

1. Provision an Azure Synapse Analytics workspace

A screenshot of a computer

AI-generated content may be incorrect.

1. Preview of the source files

A screenshot of a computer

AI-generated content may be incorrect.

1. Import a Spark Notebook to transform data

A screenshot of a computer

AI-generated content may be incorrect.

1. Load Source data

A screenshot of a computer

AI-generated content may be incorrect.

1. Transform the data structure

A computer screen with a computer screen

AI-generated content may be incorrect.

1. Save the transformed data

A computer screen with a message box

AI-generated content may be incorrect.

1. Partition data

A screenshot of a computer

AI-generated content may be incorrect.

1. Convert dataframe to external tables

A computer screen with a computer screen

AI-generated content may be incorrect.

1. Create partition data using SQL

A screenshot of a computer

AI-generated content may be incorrect.

1. Query the table meta store

A screenshot of a computer

AI-generated content may be incorrect.

Summary:

Data engineers use Spark notebooks as an ETL tool to transform data from one format or structure to another. The process begins by provisioning a Synapse Analytics workspace and previewing the sales data. An existing Spark notebook that performs ETL is imported, and all CSV files in the data folder are loaded using spark.read.csv with the specified path and parameters. The initial cell execution takes longer due to the Spark pool startup time, but subsequent cells execute faster. Once the source data is loaded, a new DataFrame is created, and the CustomerName column is split into FirstName and LastName, after which the CustomerName column is deleted. The Spark SQL library is used for filtering, deriving, naming, and applying other data modifications. The result is saved as a Parquet file in a new folder named "transformed data." To enhance performance and data structuring, the data is partitioned by Year and Month fields and stored in a new folder with subfolders for each year (2019, 2020, 2021). For analysis, the Spark DataFrame is converted to a table, and similar partitions are created using SQL. Finally, the data files abstracted by the table in the metastore are queried, concluding the lab.