Lab 5 – Azure Synapse Spark

1. Provision an Azure Synapse Analytics workspace

A screenshot of a computer

AI-generated content may be incorrect.

1. View files in the data lake

A screenshot of a computer

AI-generated content may be incorrect.

1. Use Spark to explore data

A computer screen with a white screen

AI-generated content may be incorrect.

1. Analyze data in a data frame

A computer screen with a computer screen

AI-generated content may be incorrect.

1. Aggregate and group data in a data frame

A computer screen with a message

AI-generated content may be incorrect.

1. Use Spark SQL in PySpark code

A screenshot of a computer

AI-generated content may be incorrect.

1. Run SQL code in a cell

A computer screen with a computer screen

AI-generated content may be incorrect.

1. Visualize data with Spark

A computer screen with a message

AI-generated content may be incorrect.

1. Get Started with Matplotlib

A computer screen with a graph and pie chart

AI-generated content may be incorrect.

1. Use the Seaborn library

A computer screen shot of a graph

AI-generated content may be incorrect.

1. Delete Azure resources

A computer screen with a computer screen

AI-generated content may be incorrect.

Summary:

Apache Spark is a powerful, open-source tool used for processing large amounts of data. It works with different programming languages like Java, Scala, Python, and SQL, making it useful for both data engineering and data science tasks. In a Synapse workspace, we use Spark to analyze orders stored in CSV files. By creating a PySpark notebook and attaching it to a Spark pool runtime, we can read these files into a DataFrame and even define a schema to organize the data. If we only need specific information, we can select certain columns or filter the data based on conditions. Spark also lets us group and summarize the data as needed. For convenience, we can write SQL queries using the spark.sql library to analyze the data directly, with results displayed as tables or charts. To create visualizations, we can convert the Spark DataFrame into a Pandas DataFrame and use Python libraries like Matplotlib or Seaborn. Matplotlib allows us to make detailed visualizations such as bar charts or line charts, while Seaborn makes it easier to create attractive plots with simpler code. This workflow—reading data, analyzing it with PySpark or SQL, and visualizing it with Python—makes it easy to work with large datasets and gain insights effectively.