Integrated Analytics with Azure Synapse

Project Overview

Azure Synapse Analytics is a cloud data warehouse platform that combines data integration, data exploration, data warehousing, and big data analytics to offer a unified workspace for creating end-to end analytics solutions. In short, Azure Synapse Analytics is a one-stop data analytics solution that enables data sourcing, reporting and analytics. It helps the users to effectively and securely by integrating data from any data source, data warehouse, or big data analytics platform. In this project, I have created a data analytics workspace by using Azure Synapse Analytics to perform data analysis to find solutions for the three problem statements.

Problem statement 1:

Use Azure Synapse Analytics workspace as a resource in Azure Synapse Analytics for integrated data analytics. Ingest dataset from the external source(**https://raw.githubusercontent.com/MicrosoftLearning/DP-900T00A-Azure-Data-Fundamentals/master/Azure-Synapse/products.csv**) and store the data in ADLs Gen2(Azure Data Lake Storage Generation 2) account created on the Azure portal while configuring the workspace.

Problem statement 2:

Use the Data hub to analyze and query the data that has been ingested and create an SQL query to load top 100 rows from the dataset. Modify the SQL query to find the number of products by each category and visualize the resulting data.

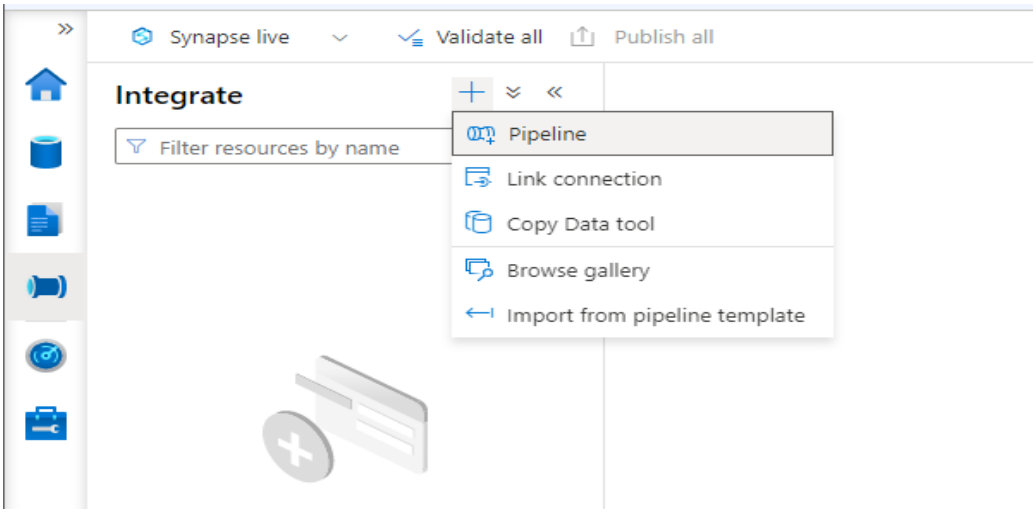
Problem statement 3:

Create an Apache spark pool from Manage Hub and load the data into a data frame to analyze the data in a spark environment.

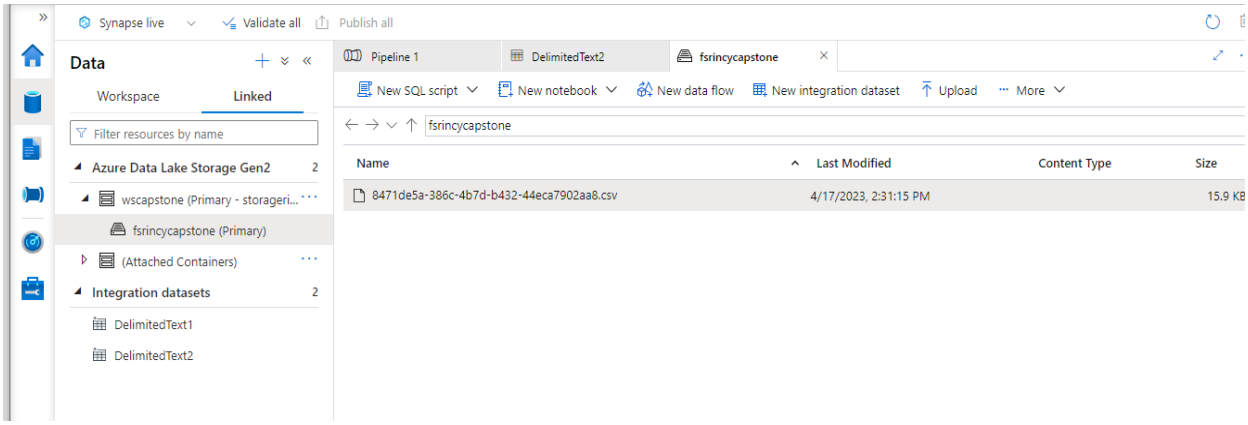
Project Procedure

Step1: Creation of Azure Synapse Workspace and loading data

Here, I have created Azure Synapse Workspace and a data storage account using Azure Synapse Analytics and to load the data from the external source(https://raw.githubusercontent.com/MicrosoftLearning/DP-900T00A-Azure-Data-Fundamentals/master/Azure-Synapse/products.csv), first I have created a pipeline as shown below



To copy the data from source to the storage account, I've used the copy data option from the move and transform section and updated the source and sink parameters. Once the data is copied to the storage account, we can view the same in the **data hub** as shown below**.**

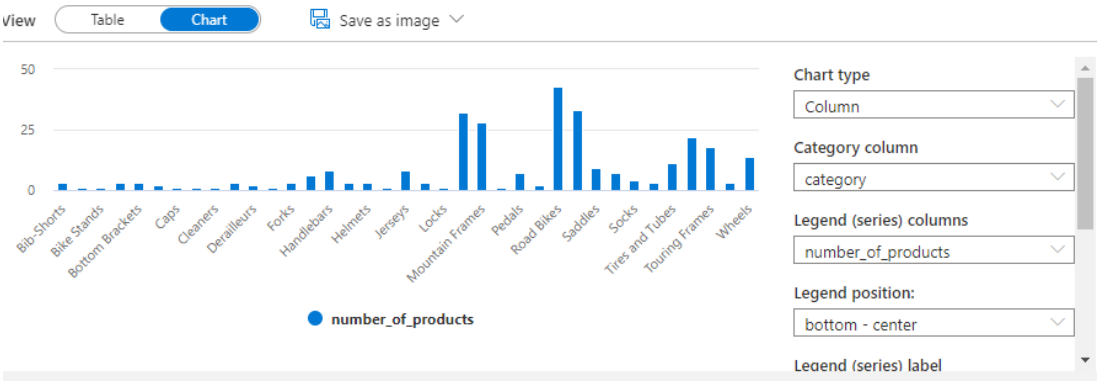


Step 2: Analyzing and querying the data.

To analyze and query the data that has been loaded into the Azure Synapse Workspace, I've used SQL queries and charts for visualization. First, I ran the SQL query to find the top 100 rows, then ran the query to find the number of products by each category and finally used charts to visualize the resulting data.

```
-- This is auto-generated code
SELECT
    TOP 100 *
FROM
    OPENROWSET(
        BULK 'https://storagerincy.dfs.core.windows.net/fsrincycapstone/8471de5a
        FORMAT = 'CSV',
        PARSER_VERSION = '2.0',
        HEADER_ROW = TRUE
    ) AS [result]
```

```
-- This is auto-generated code
SELECT
    category, count(*) as number_of_products
FROM
    OPENROWSET(
        BULK 'https://storagerincy.dfs.core.windows.net/fsrincycapstone/8471de5a-
        FORMAT = 'CSV',
        PARSER_VERSION = '2.0',
        HEADER_ROW = TRUE
    ) AS [result]
    GROUP BY category
```

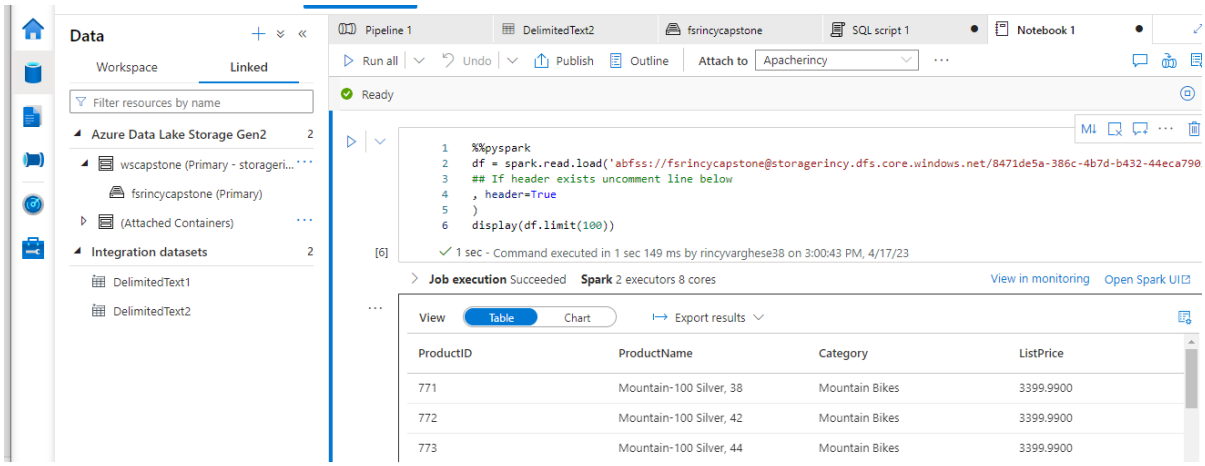


Step3: Analyzing the data using the Spark Pool

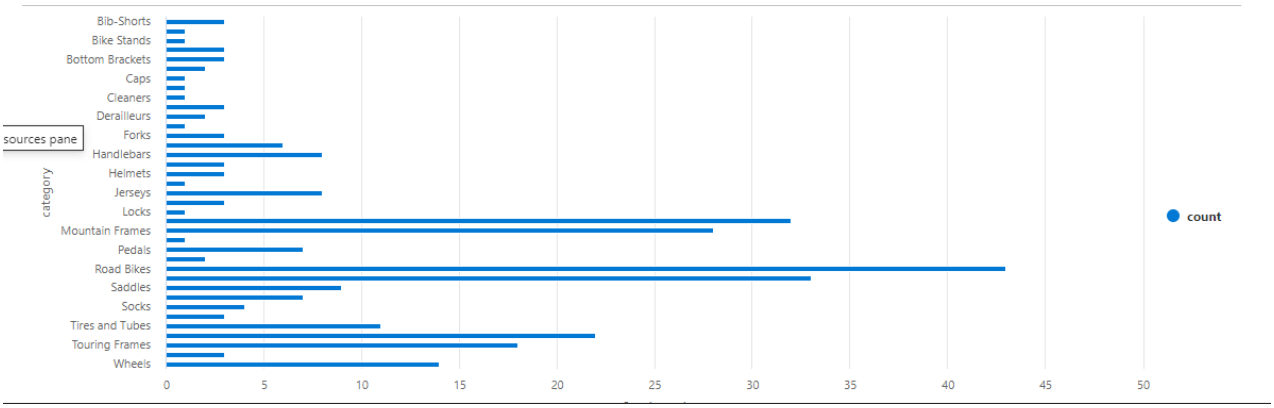
To create a spark pool, I selected the Apache Spark pools from the manage hub and given the required specification as follows:

- Apache spark pool name – Apacherincy
- Node size family – as default
- Node size – default
- Autoscale – Enabled
- Number of nodes – 3 --- 5

Once the spark pool is deployed, I have created a new notebook using the data hub with an auto generated pyspark code to read the data which is shown below.



By modifying the pyspark code, we can perform the same set of operations which we have completed using SQL. The visualization using pyspark code is given below.



Project Outcomes

- Created Azure Synapse Workspace, ingested the data from an external source like HTTP, transformed the data as per the requirements and stored it in an ADLS Gen2 storage.
- SQL pool and Spark pool has been used to analyze the data and visualize the data