Capstone Project

INTEGRATED ANALYTICS WITH AZURE SYNAPSE

Shashank S, Batch: DS19M

Project abstract

Exploring data analytics workspace by using Azure Synapse Analytics and using Azure Synapse Analytics Workspace as a resource in Azure Synapse Analytics for Integrated Data Analytics.

Finding out solution for Three problem statements which included,

- Ingesting dataset from an external resource, transforming the dataset and storing the transformed data into ADLS Gen-2(Azure Data-lake Storage Generation 2) storage.
- Exploring datahub and performing SQL queries on top of transformed data and visualize the result set.
- Creating a Pyspark environment(Apache Spark Pool) and performing operations on the same set of data.

Project Procedure

- Exploring data analytics workspace by using Azure synapse analytics.
- 2. Creating a ADLS Gen 2 account in the Azure synapse.
- By using Built-in Copy task option Ingesting the data from a source through HTML link.
- 4. After ingestion, creating a connection and transforming the data and load the transformed data into ADIS gen storage.

- Analyse and query the data which has been ingested in azure synapse workspace.
- 2. Through SQL query performing AGGREGATION and GROUPING the data as per the requirement
- 3. Analysing the result set through Chart view.

- Creating a Apache spark pool with given inputs.
- 2. Loading the data into dataframe by using built-in code.
- Analysing the data in spark pool by setting the language as pyspark.
- 4. Performing the same query and getting insights from the help of chart view.

Problem statement 1

Problem statement 2

Problem statement 3

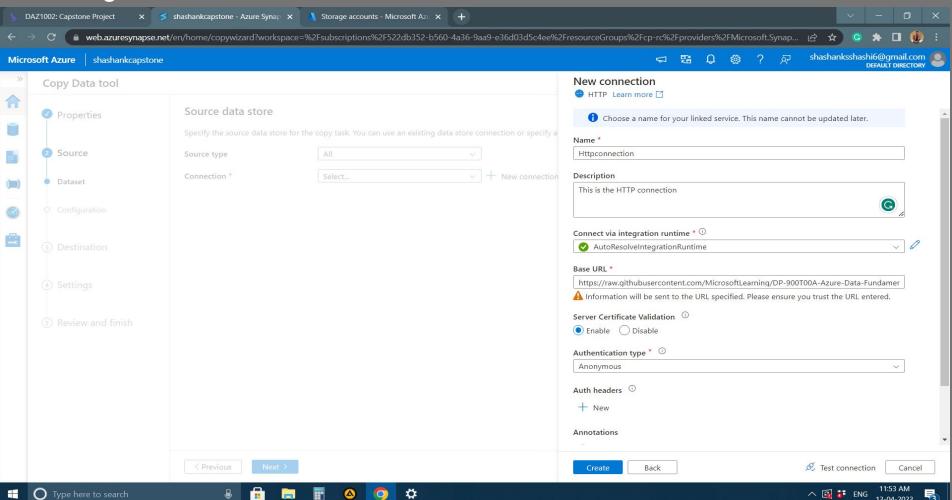
Project Outcomes

- Created tha Azure synapse workspace, Ingested the data choosing HTTP and Transformed the data as per the requirements and stored it in ADLS gen2 storage.
- 2. SQL and SPARK pool has been used to analyse the data.
- 3. With the help of chart view getting visual information and proper insights about the data.

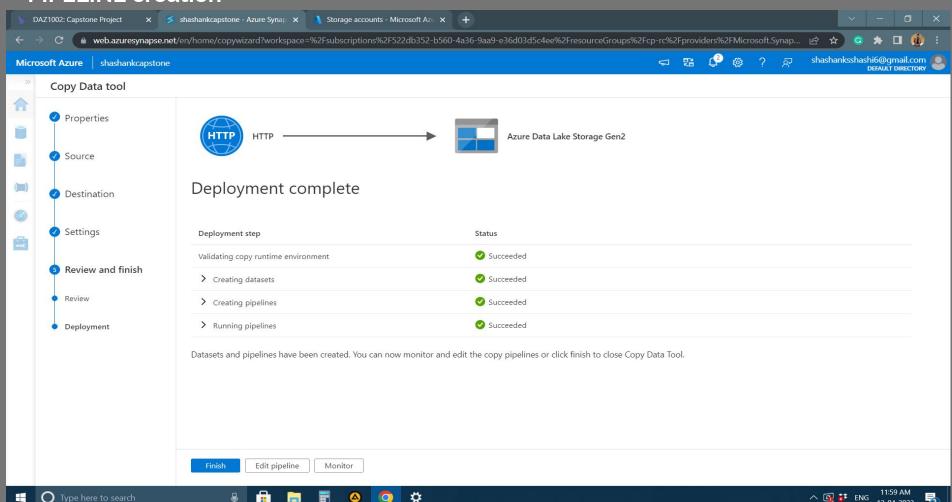
_

SCREENSHOTS

Creating a NEW CONNECTION

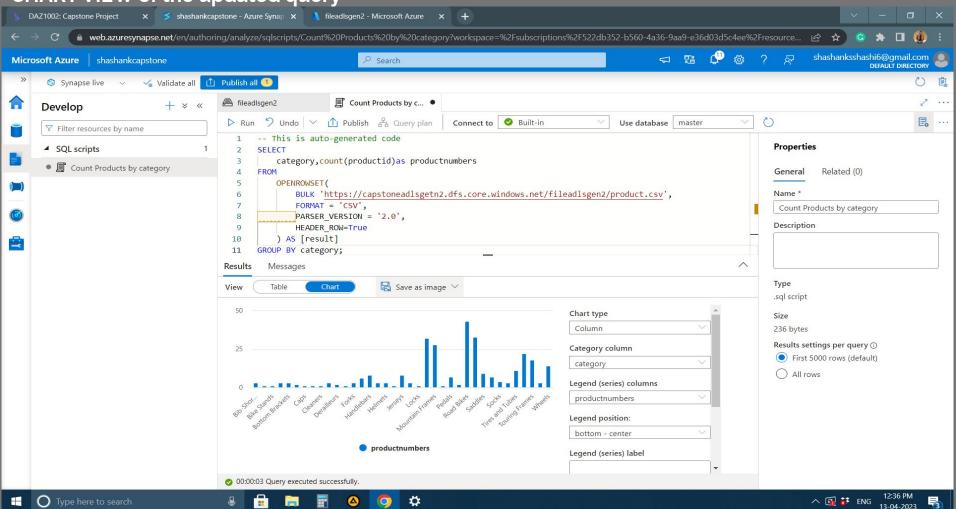


PIPELINE creation

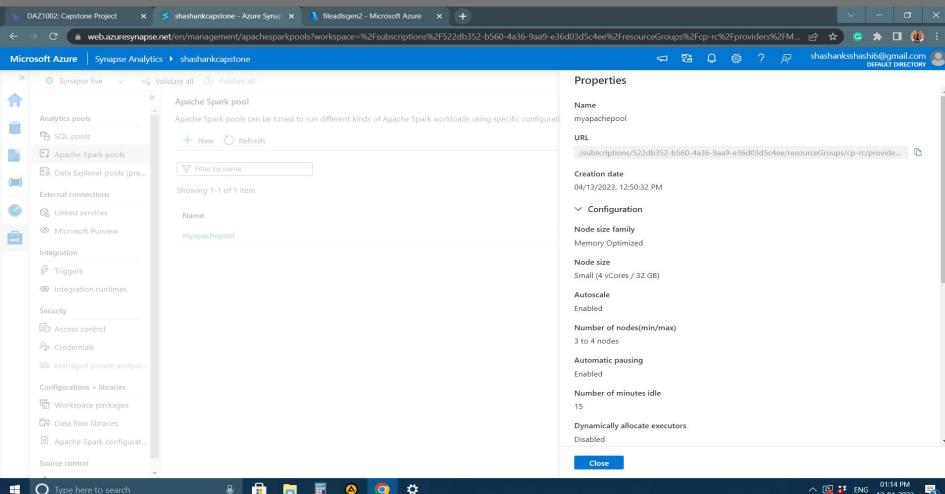


View of the UPDATED QUERY x 🔰 shashankcapstone - Azure Synap x 🔪 fileadlsgen2 - Microsoft Azure x 🛨 🎍 web.azuresynapse.net/en/authoring/explore/linked/sqlscripts/SQL%20script%201?workspace=%2Fsubscriptions%2F522db352-b560-4a36-9aa9-e36d03d5c4ee%2FresourceGroups%2Fc... 🖄 ☆ □ 點 ₽ ※ ? ₽ shashanksshashi6@gmail.com Microsoft Azure shashankcapstone ∠ Search Synapse live ✓ ✓ Validate all Publish all 1 SQL script 1 fileadlsgen2 Data 7 Undo V ↑ Publish ♣ Query plan Connect to Built-in Use database master Workspace Linked -- This is auto-generated code **Properties** ▼ Filter resources by name 2 SELECT 3 category, count(productid) as productnumbers ■ Azure Data Lake Storage Gen2 4 FROM General Related (0) 5 OPENROWSET(BULK 'https://capstoneadlsgetn2.dfs.core.windows.net/fileadlsgen2/product.csv', 6 Name * 7 FORMAT = 'CSV'. adls2Connection (capstoneadls... *** SQL script 1 PARSER VERSION = '2.0', 8 (Attached Containers) Description 9 HEADER ROW=True) AS [result] 10 ▲ AzureDataLakeStorageCONNEC... *** GROUP BY category; 11 12 fileadlsgen2 Integration datasets Type .sql script Size Results Messages 236 bytes View Chart → Export results ∨ Results settings per query ① First 5000 rows (default) O Search () All rows productnumbers category Bib-Shorts 3 Bike Racks Bike Stands 00:00:03 Query executed successfully. Type here to search * △ 🚳 👬 ENG

CHART VIEW of the updated query



Resource creation - MYAPACHE POOL



Query runned in APACHE SPARK POOL

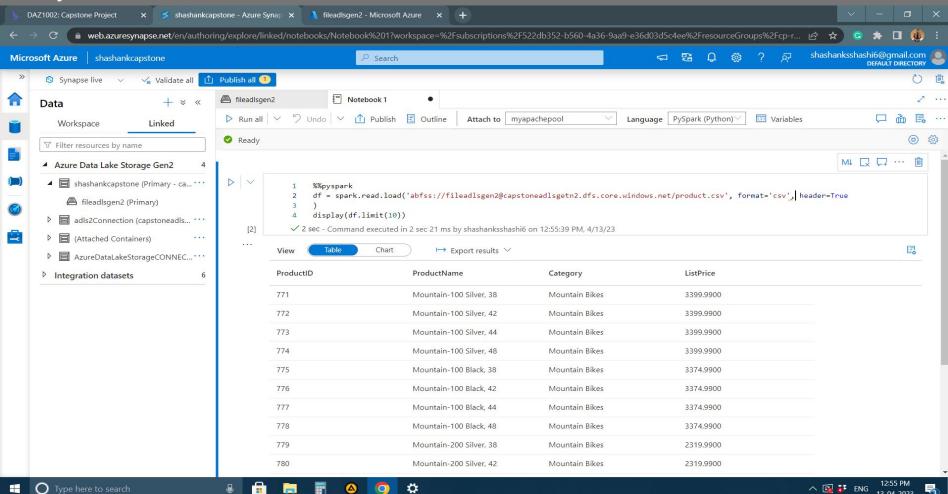
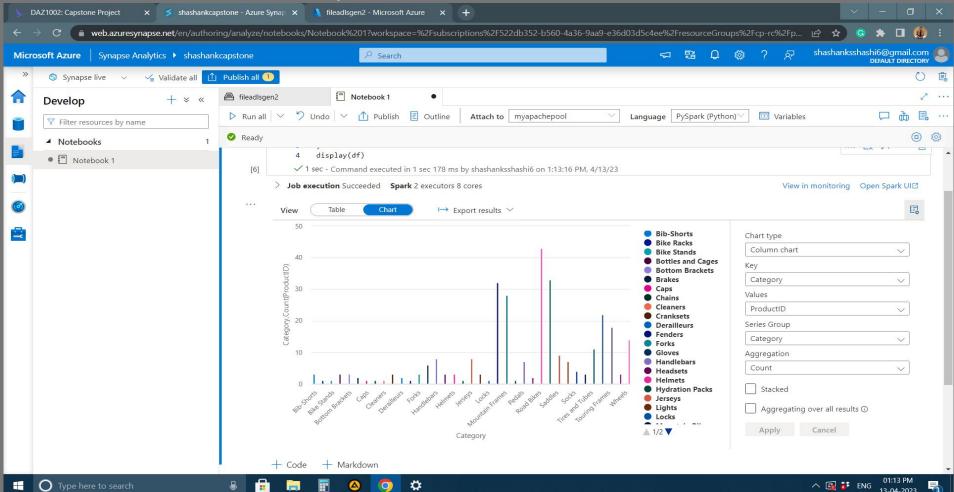


CHART view for the same sql query



THANK YOU