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

1. Exploring data analytics workspace by using Azure synapse analytics.
2. Creating a ADLS Gen 2 account in the Azure synapse.
3. 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 ADlS gen storage.

Problem statement 1

1. 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.

Problem statement 2

1. Creating a Apache spark pool with given inputs.
2. Loading the data into dataframe by using built-in code.
3. 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.

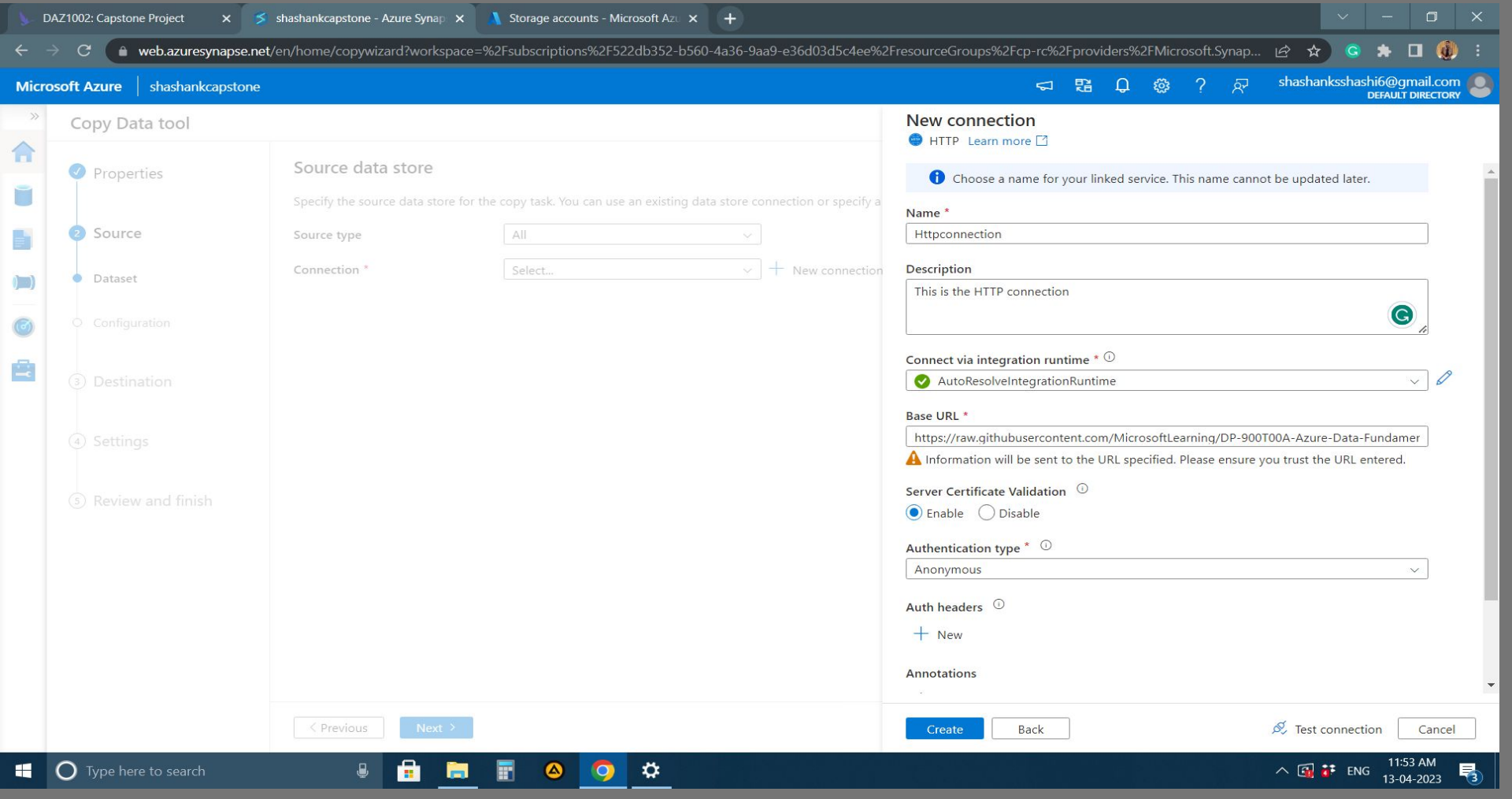
Problem statement 3

Project Outcomes

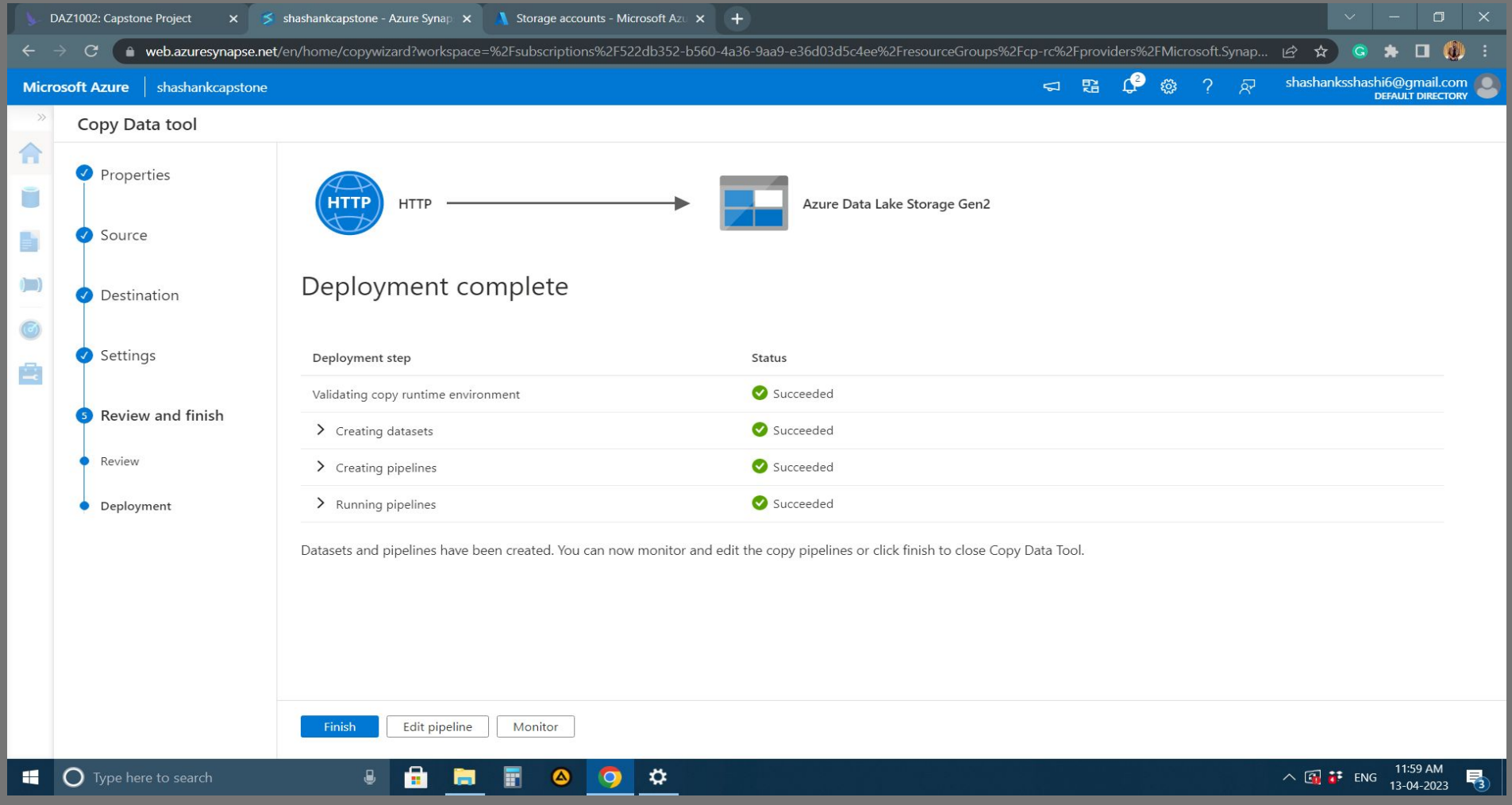
1. Created tha Azure synapse workspace, Ingested the data choosing HTTP and Transformed the data as per the requirements and stored it in ADLS gen2 storage.
2. SQL and SPARK pool has been used to analyse the data.
3. With the help of chart view getting visual information and proper insights about the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

SCREENSHOTS

Creating a NEW CONNECTION



PIPELINE creation



View of the UPDATED QUERY

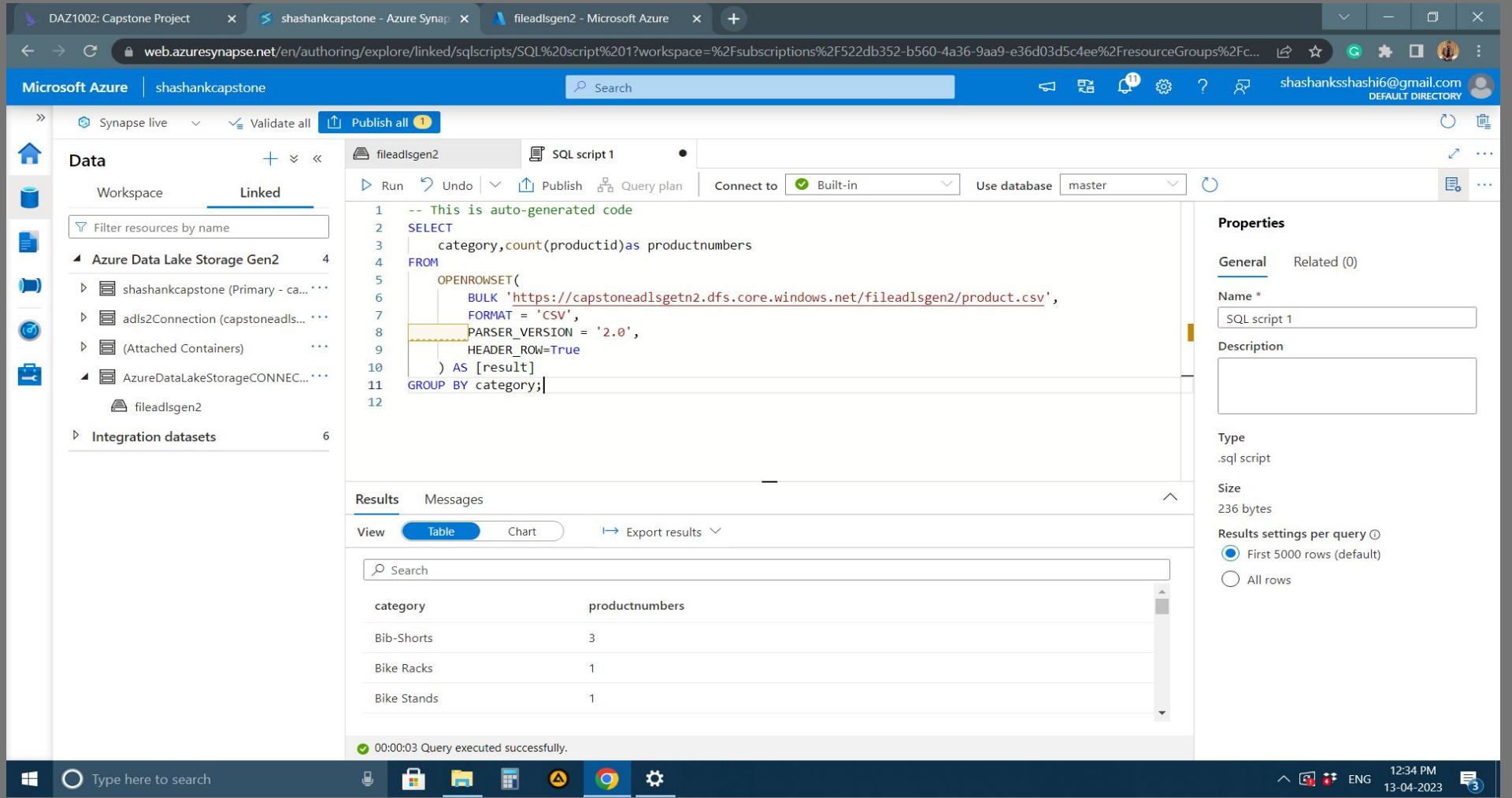
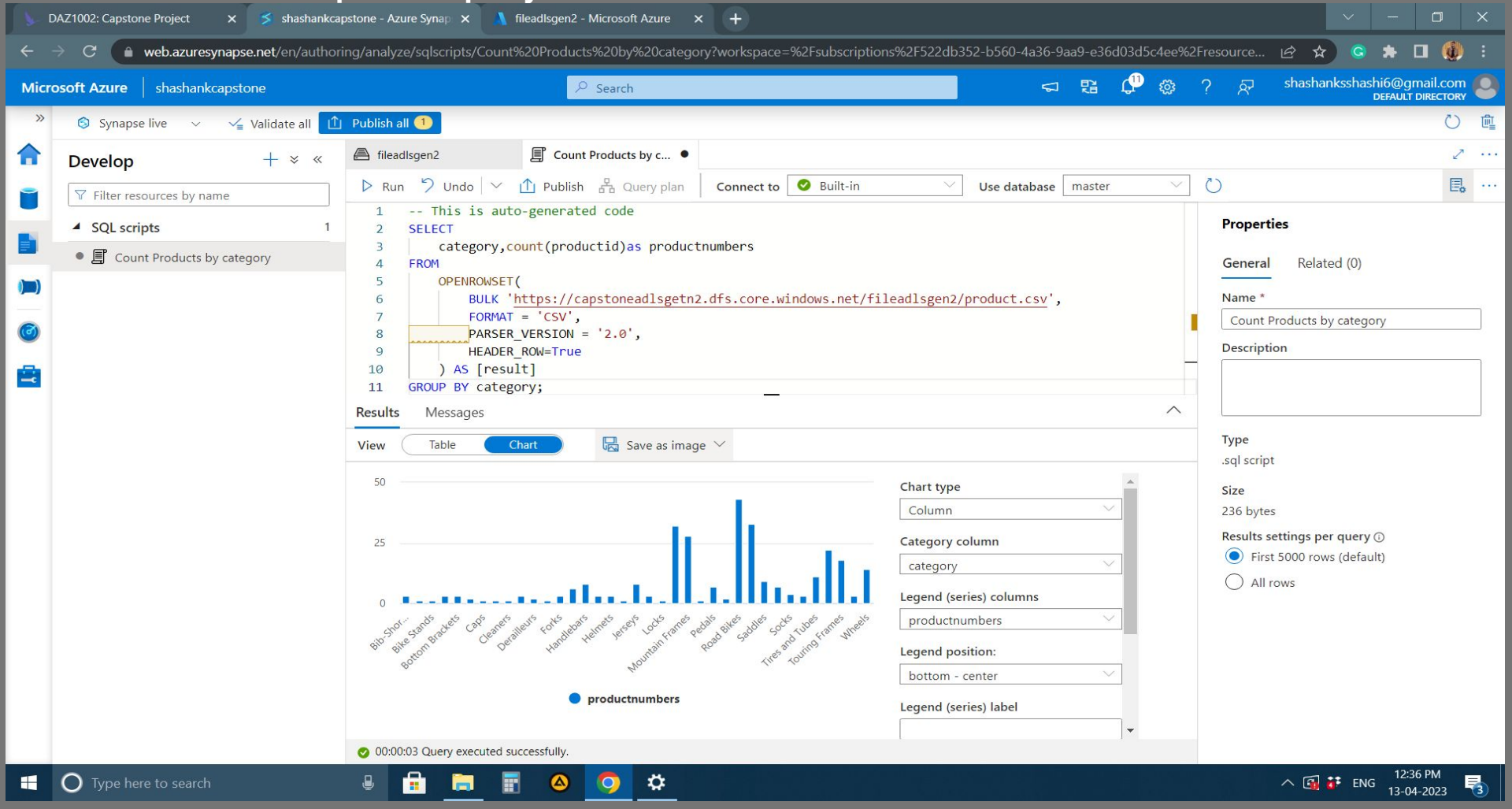
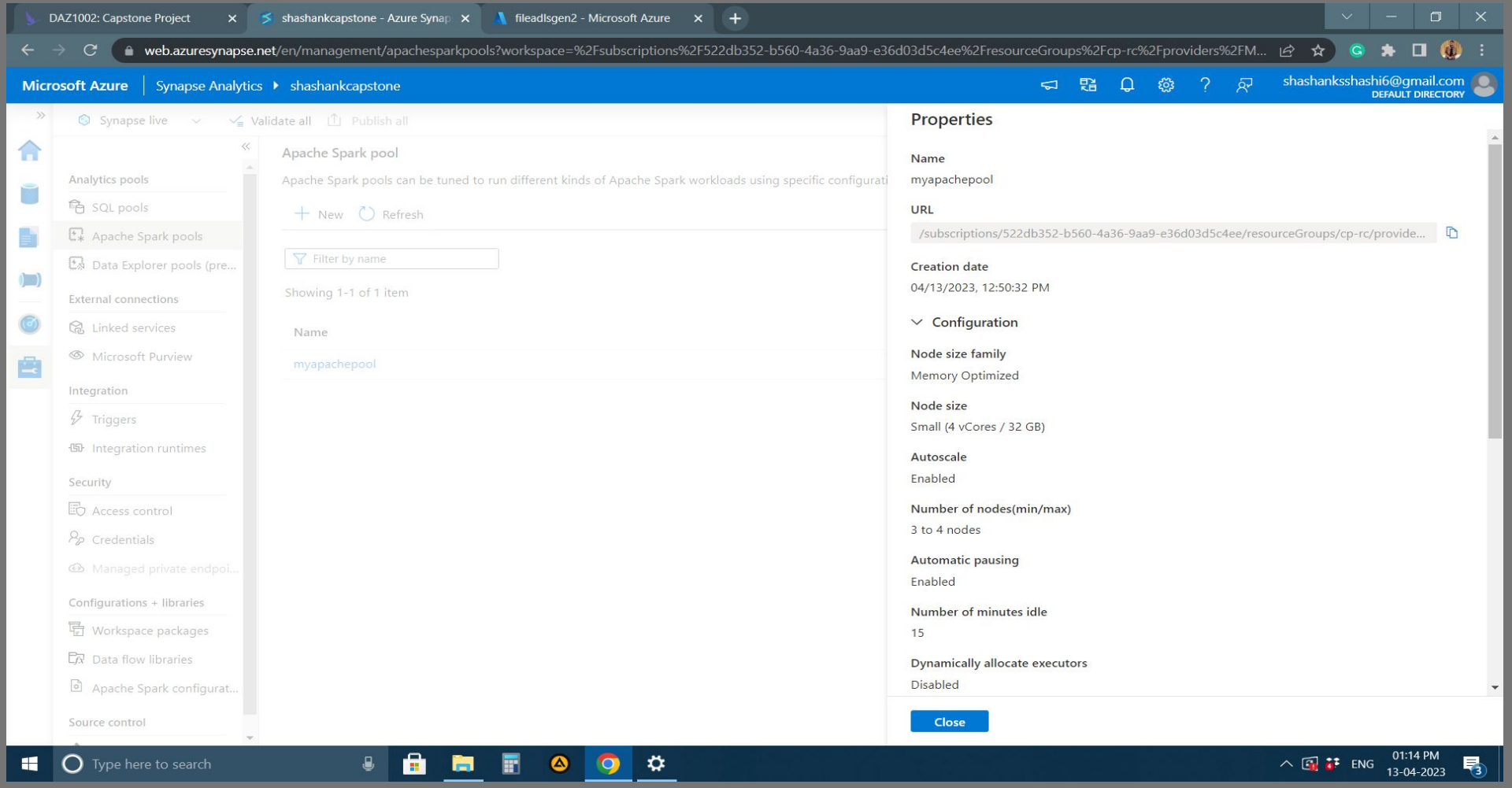


CHART VIEW of the updated query



Resource creation - MYAPACHE POOL



Query runned in APACHE SPARK POOL

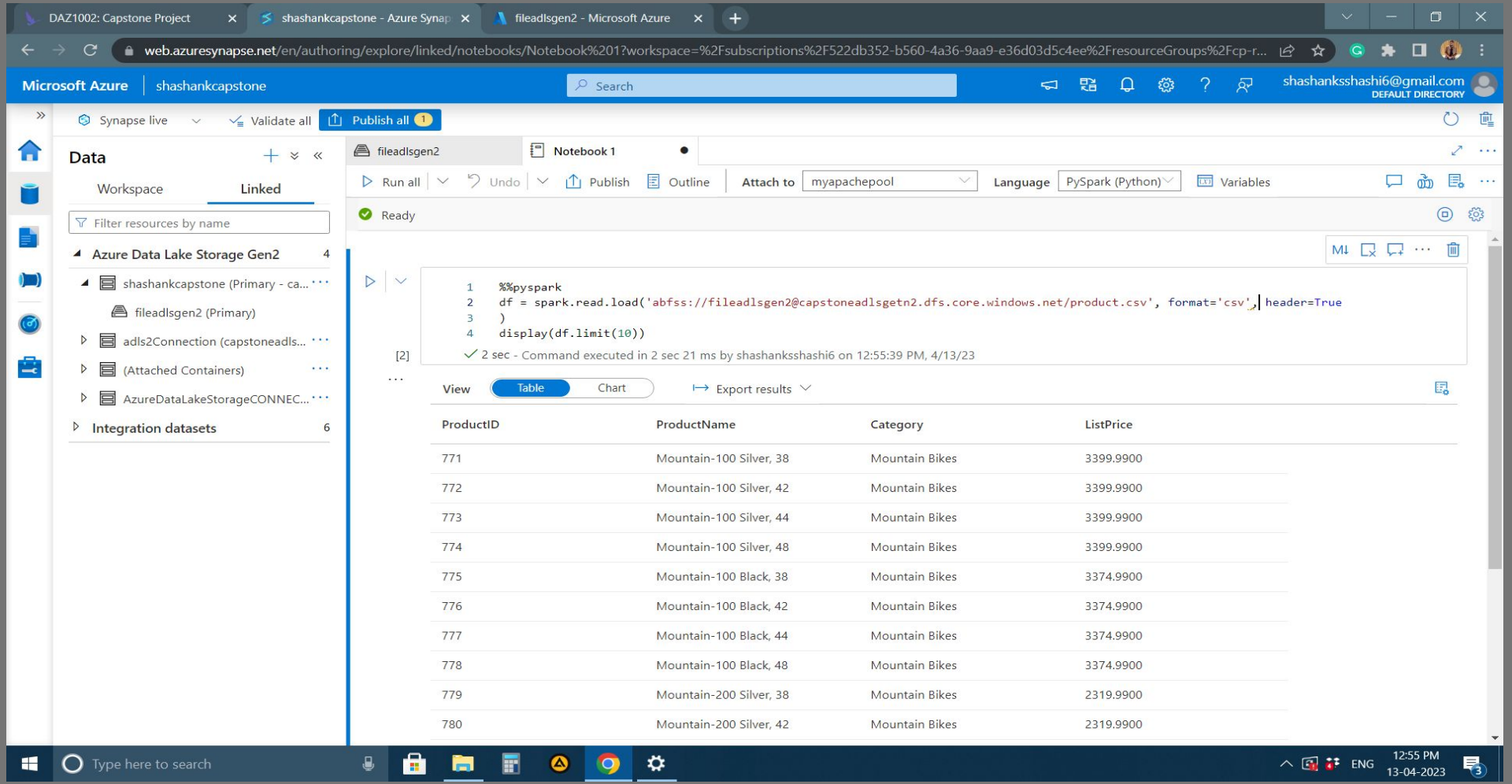
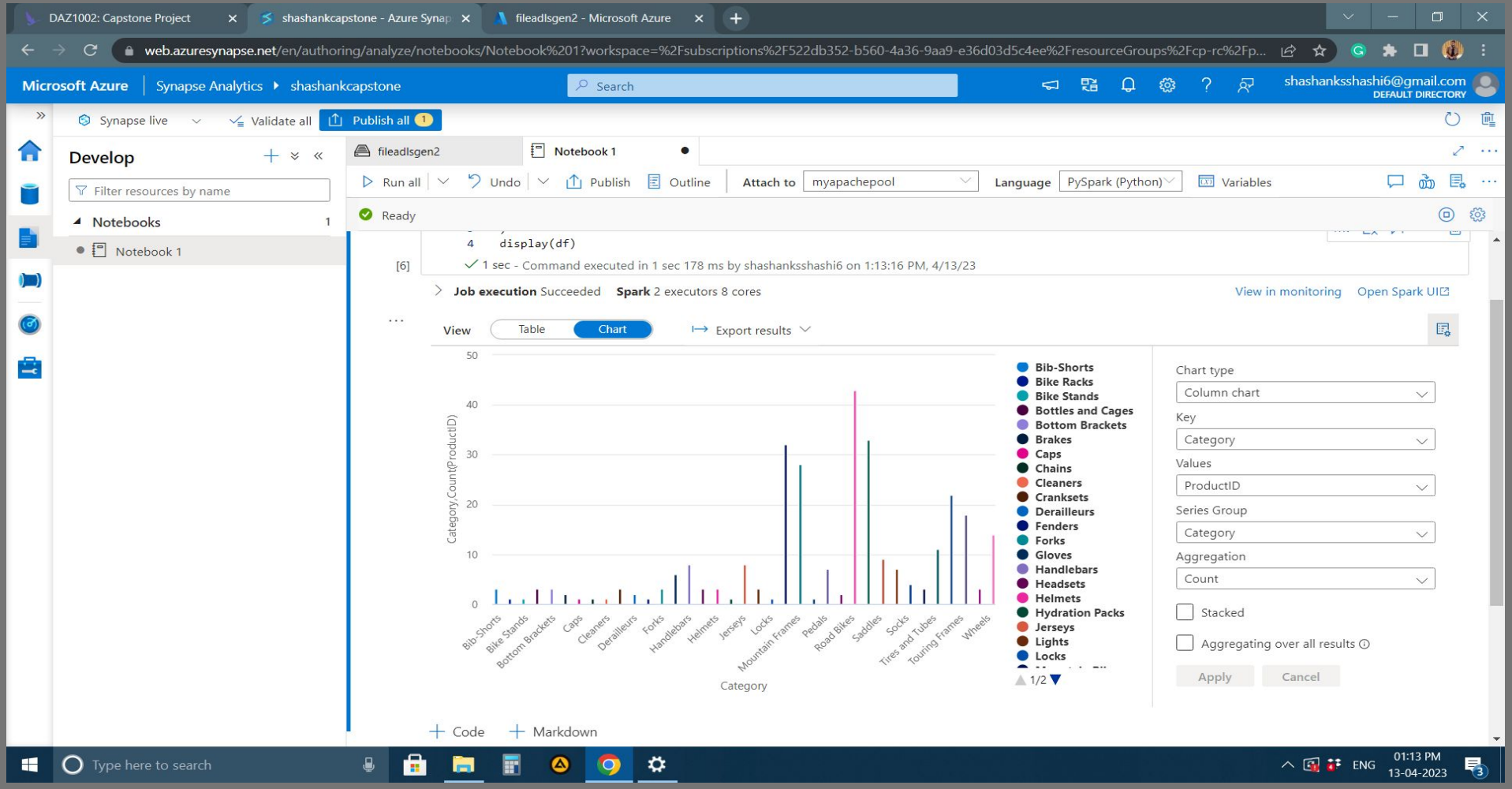


CHART view for the same sql query



THANK YOU