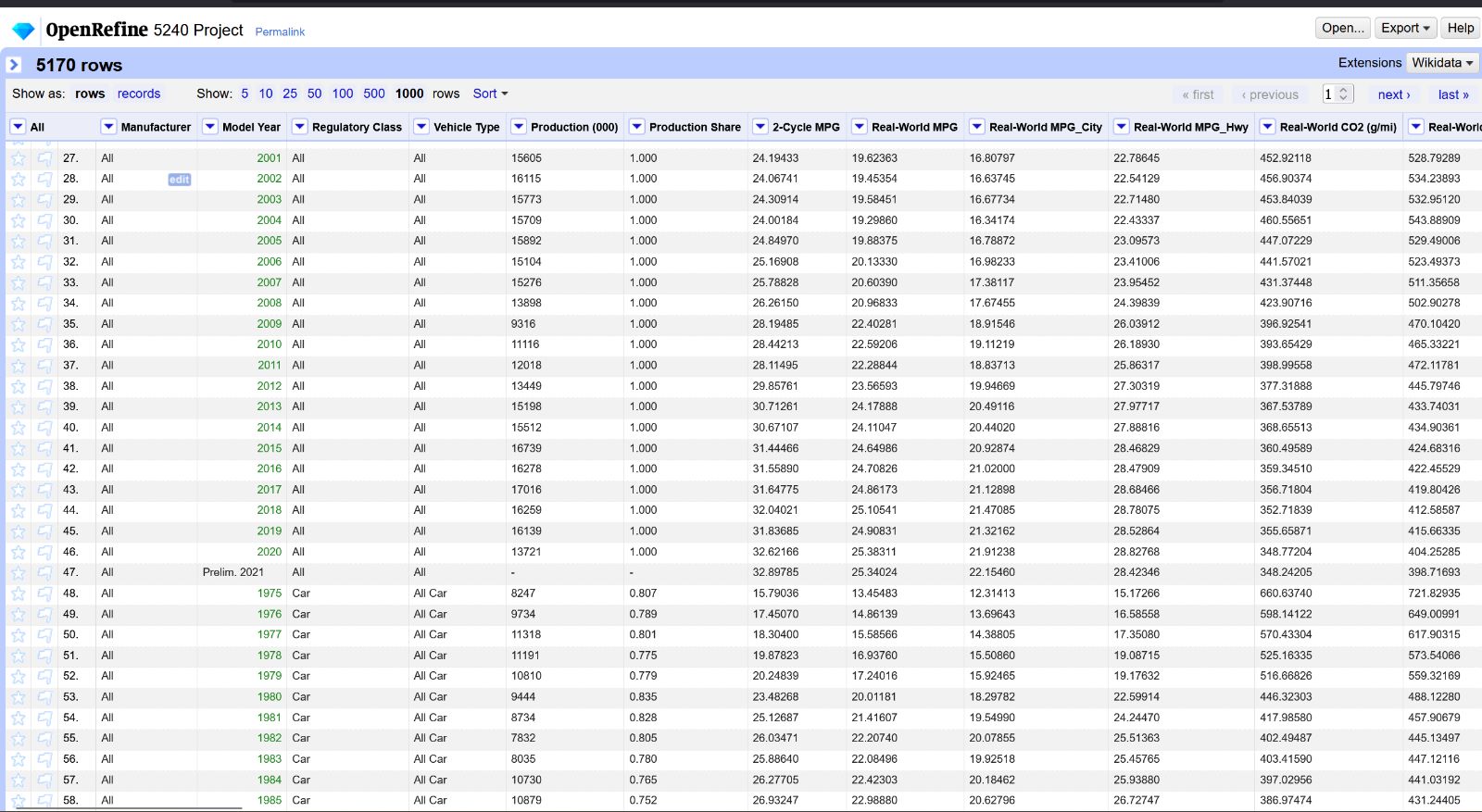
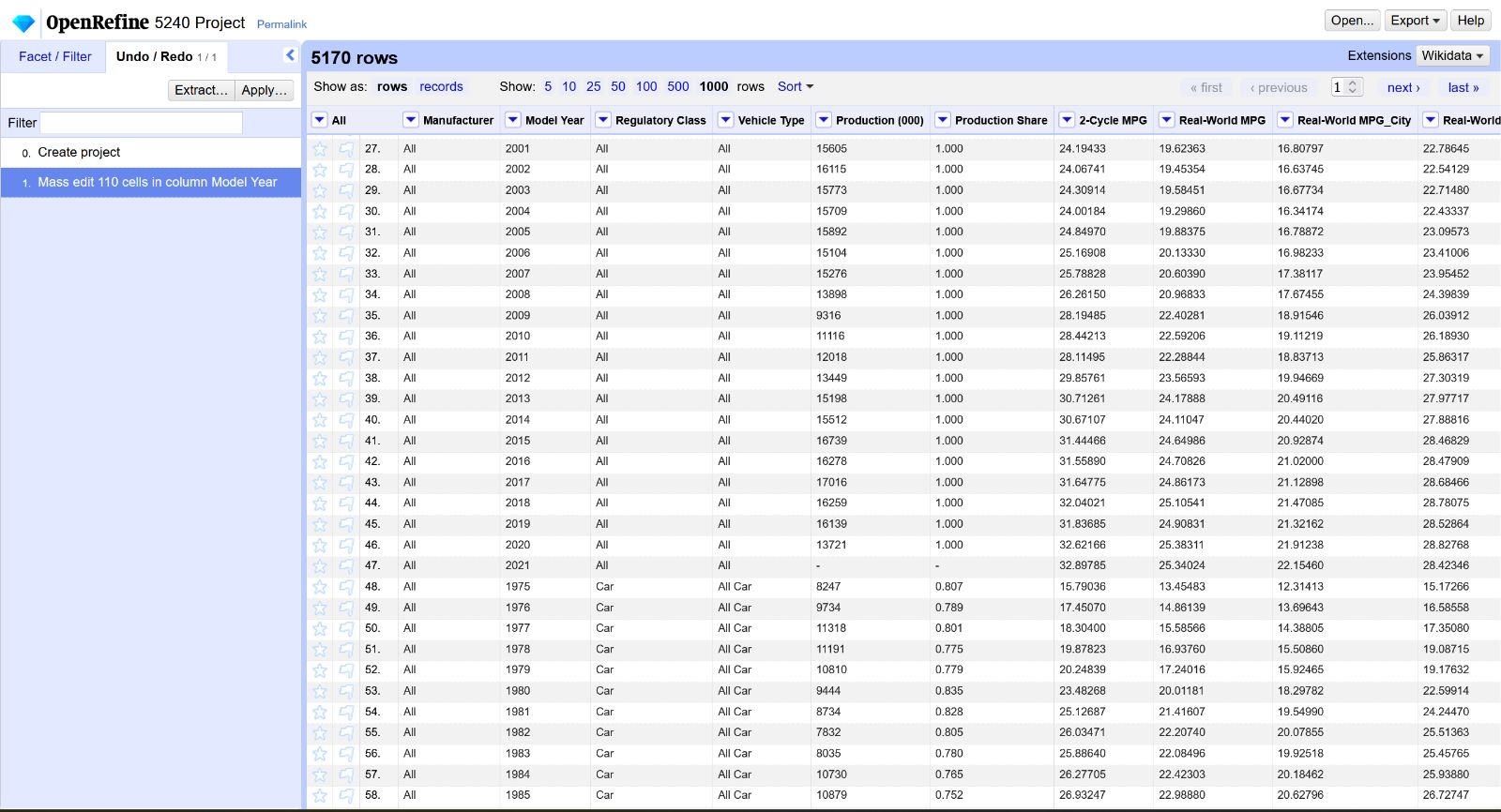
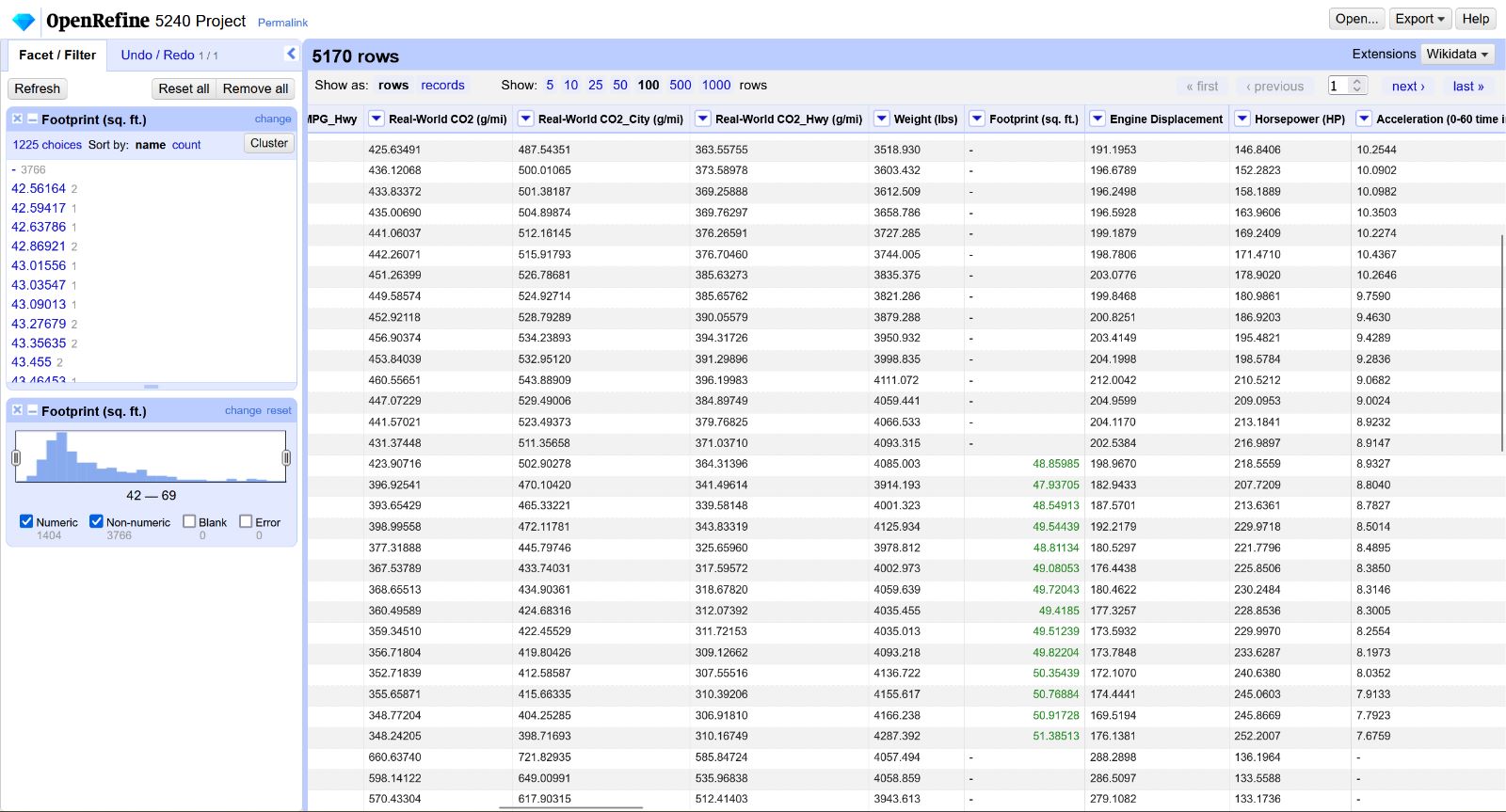
We have taken a static dataset from <https://www.epa.gov/automotive-trends/explore-automotive-trends-data> to work on the requirement.

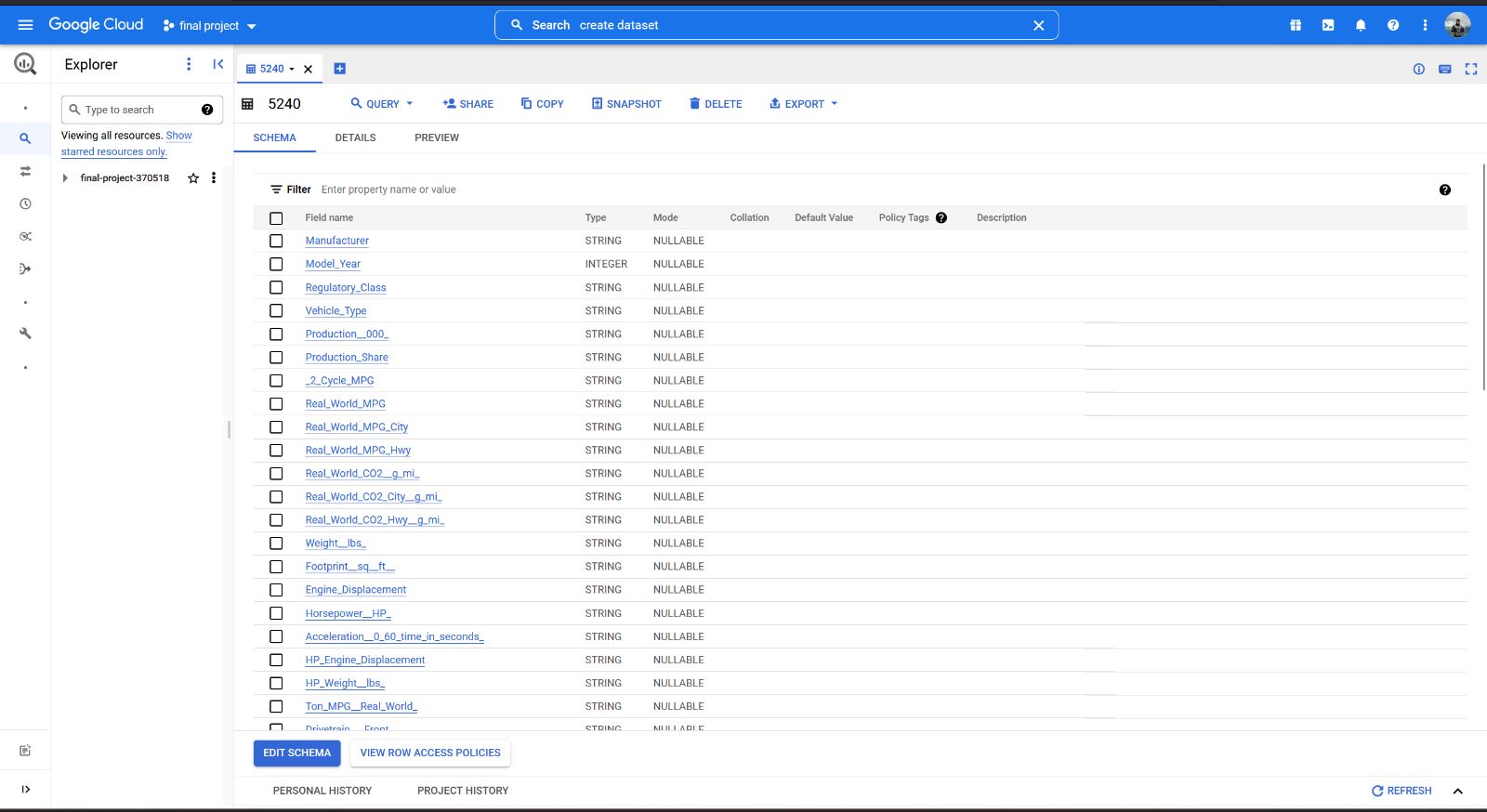
Our dataset seems quite clean. However, moderate cleaning is required, like Model Year contains "Prelim. 2021" rather than 2021. We have used OpenRefine to clean data. Below are the snapshots.

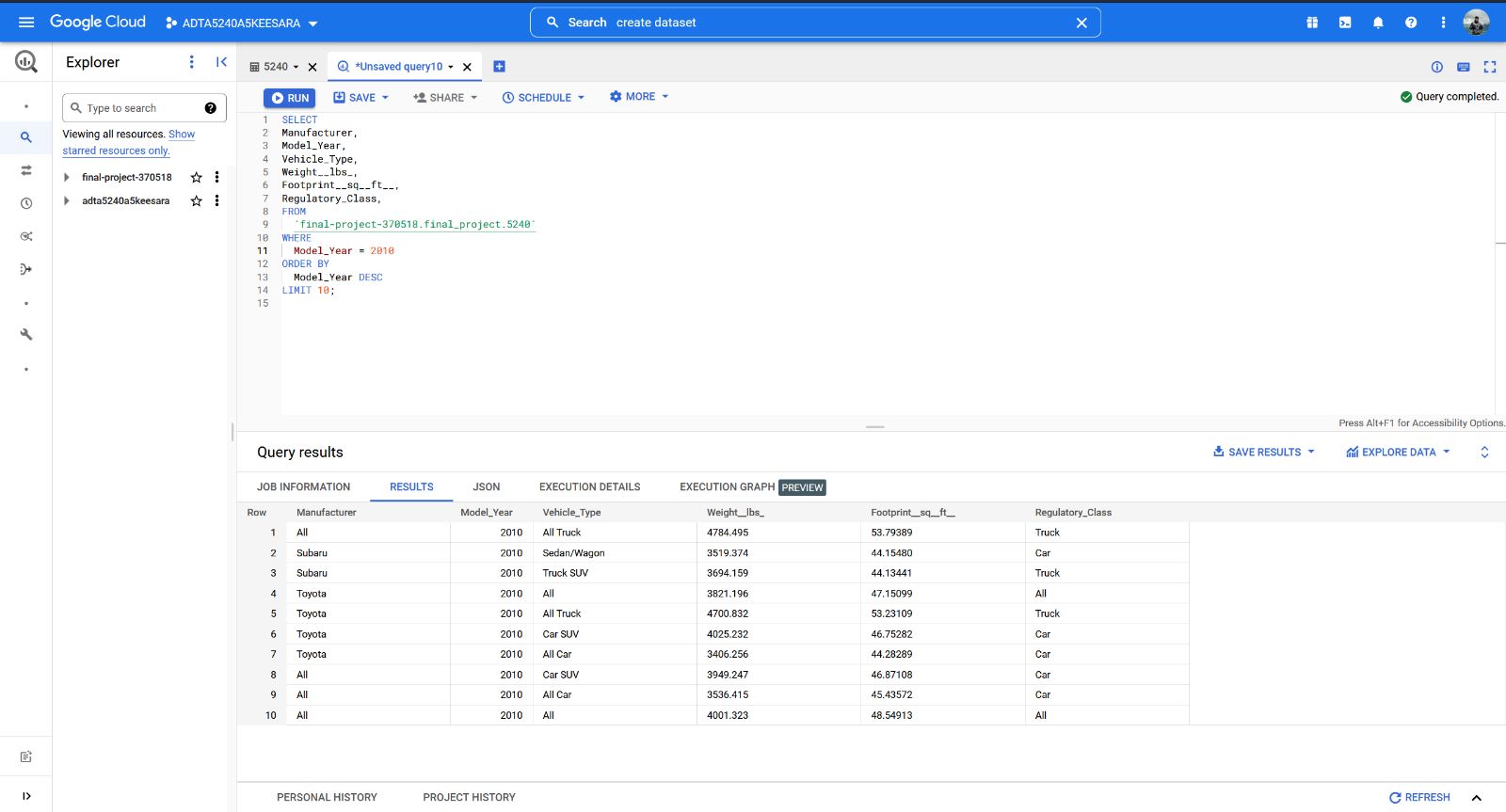


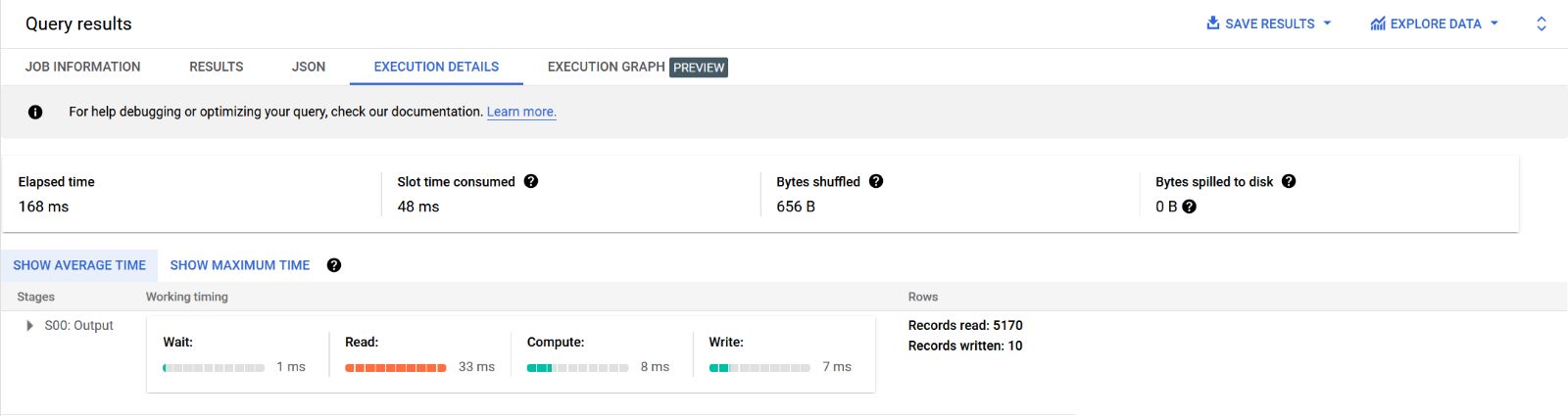




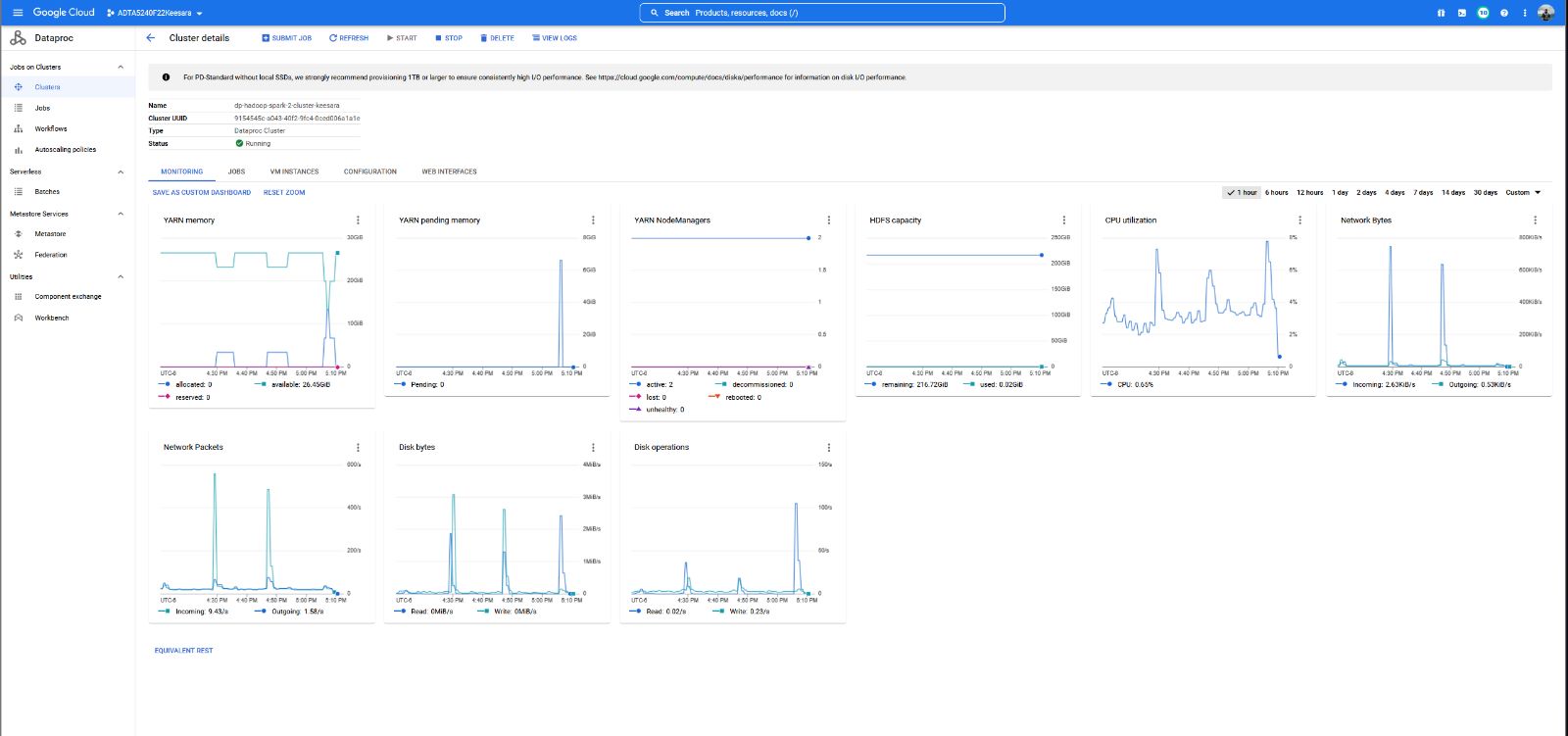
Imported data to BigQuery to store data and make access of data for analysts in our company, performed query, and below are the snapshots.

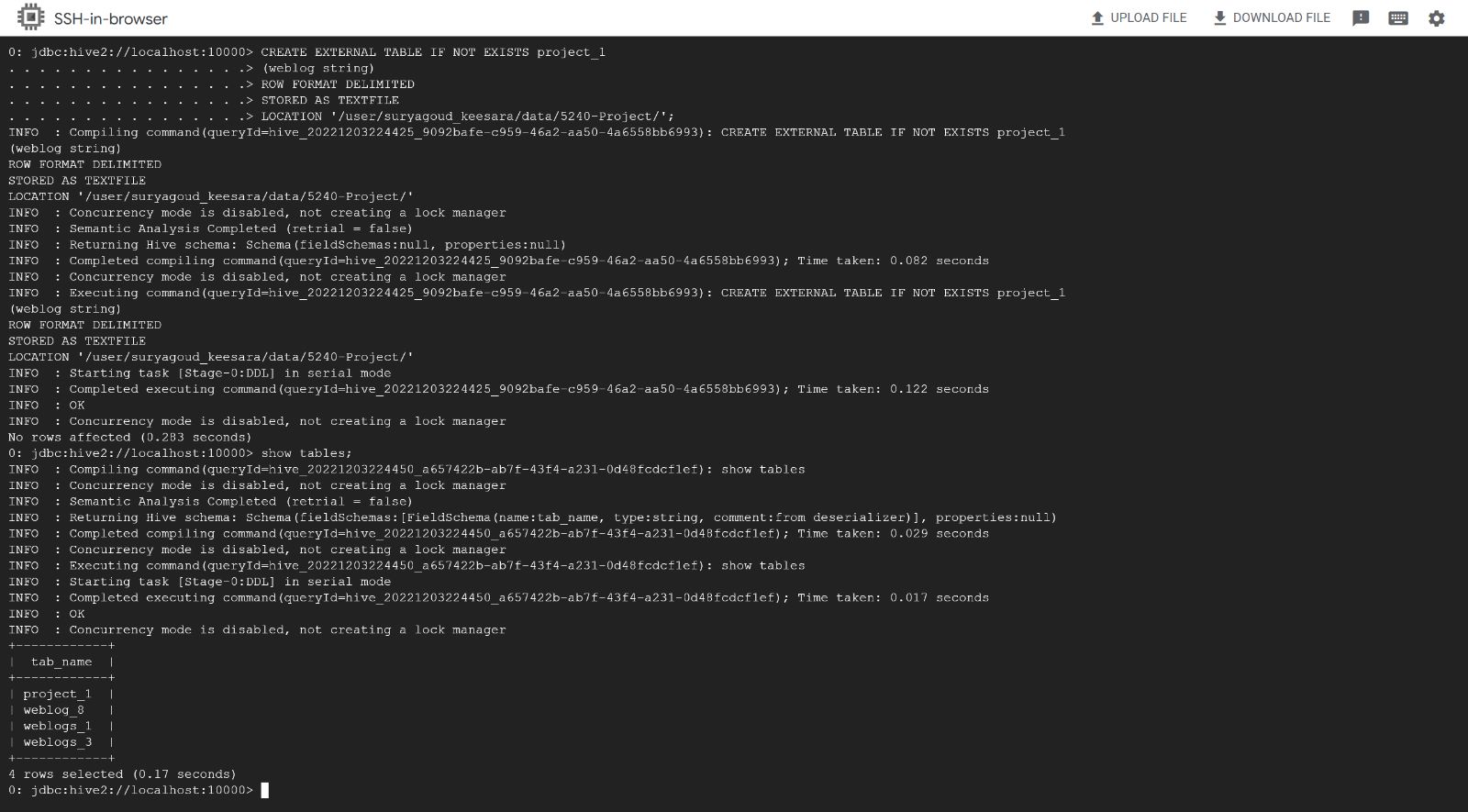


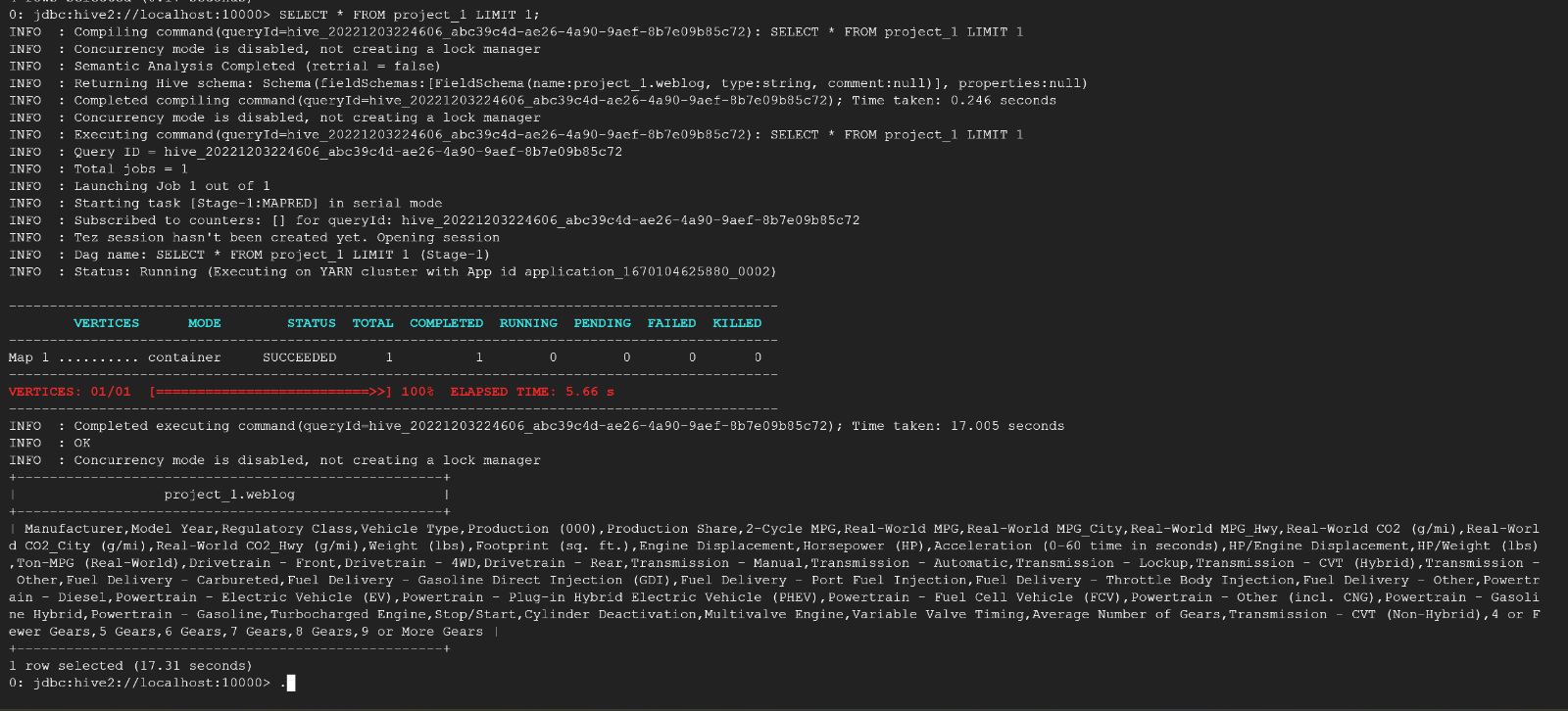


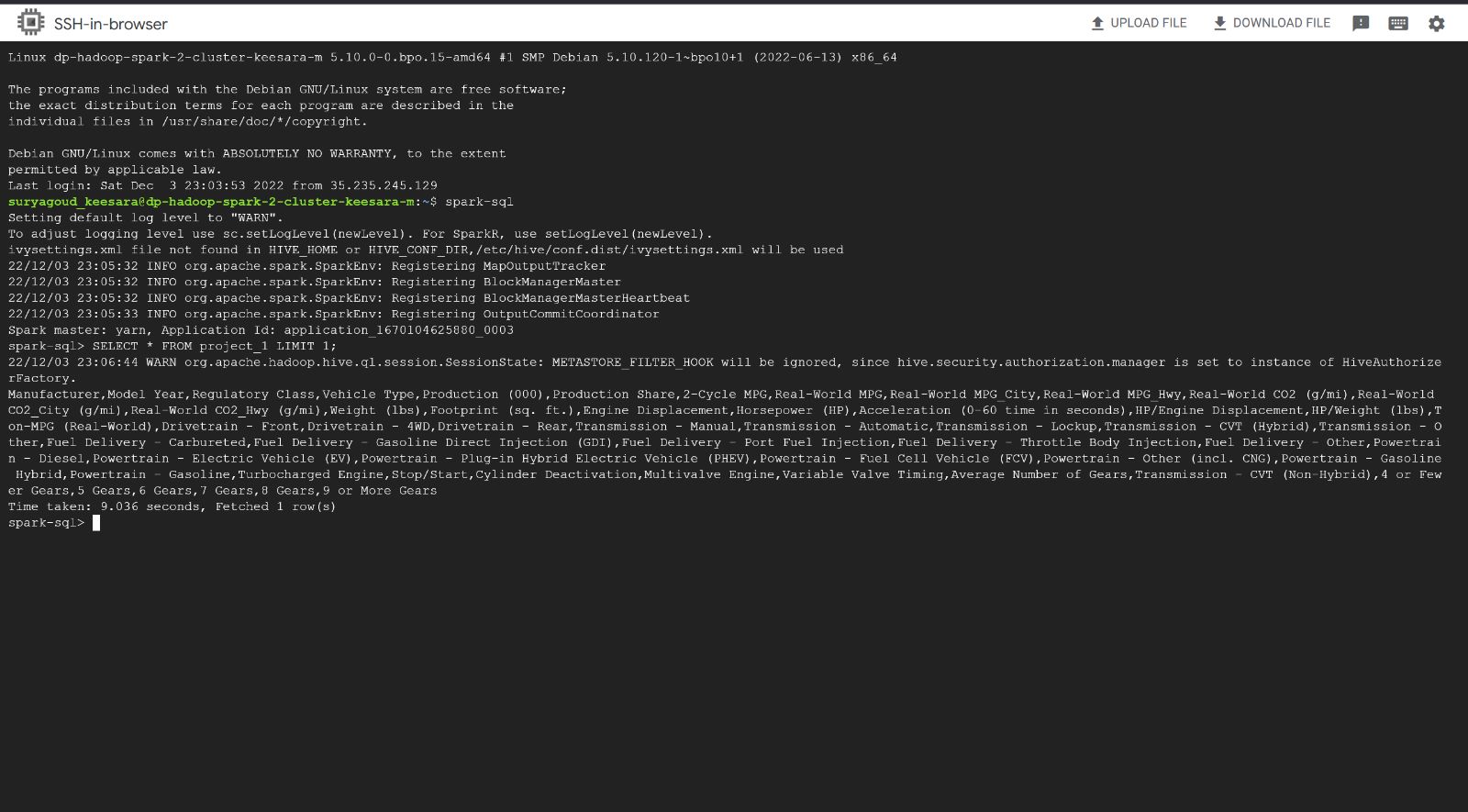


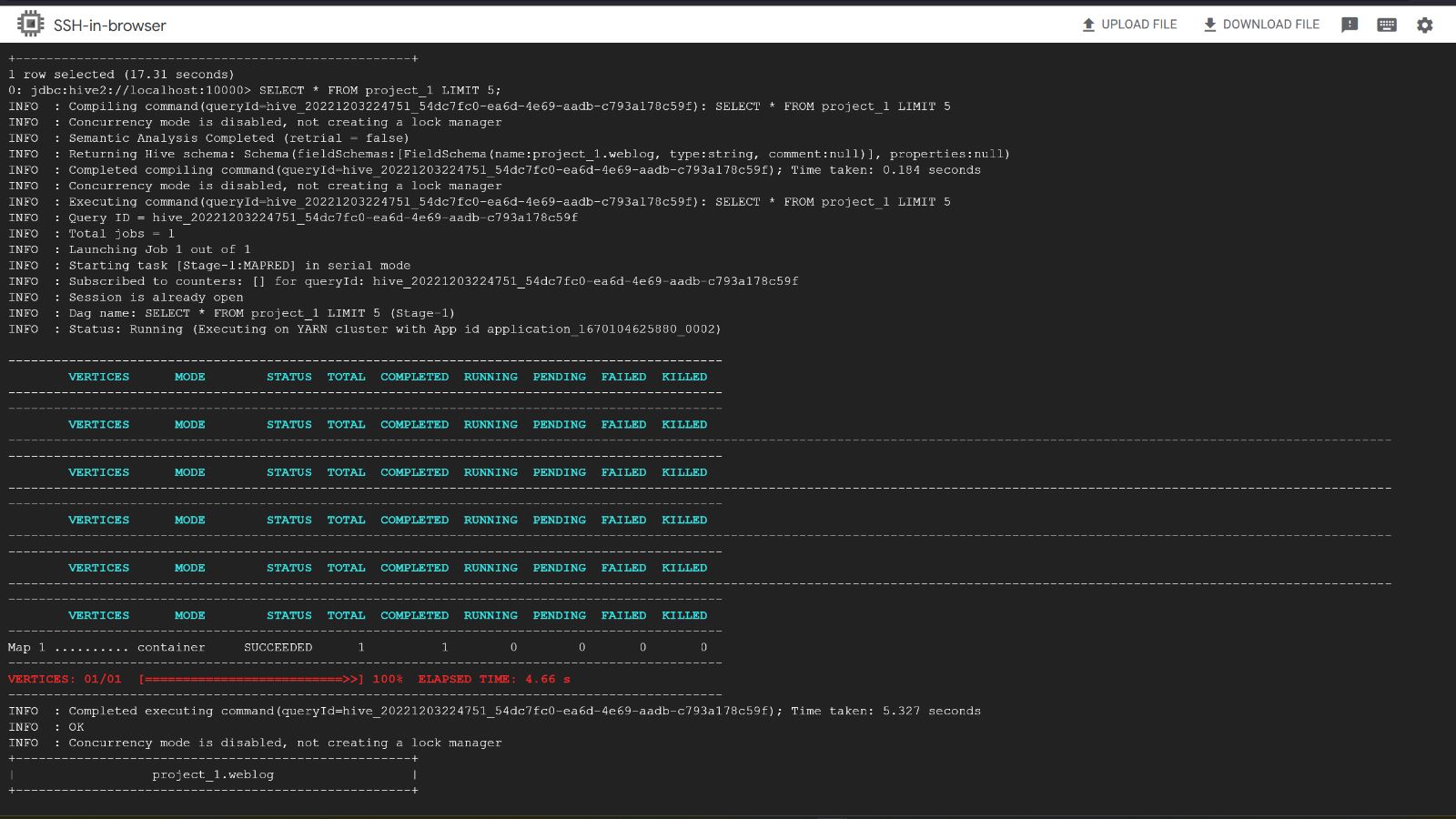
Below are the cluster monitoring details. By monitoring these details, we can check CPU utilization, Hadoop Distributed File System capacity, VM Instances, etc. It would be helpful for the Data Engineering team to monitor the performance and connectivity.

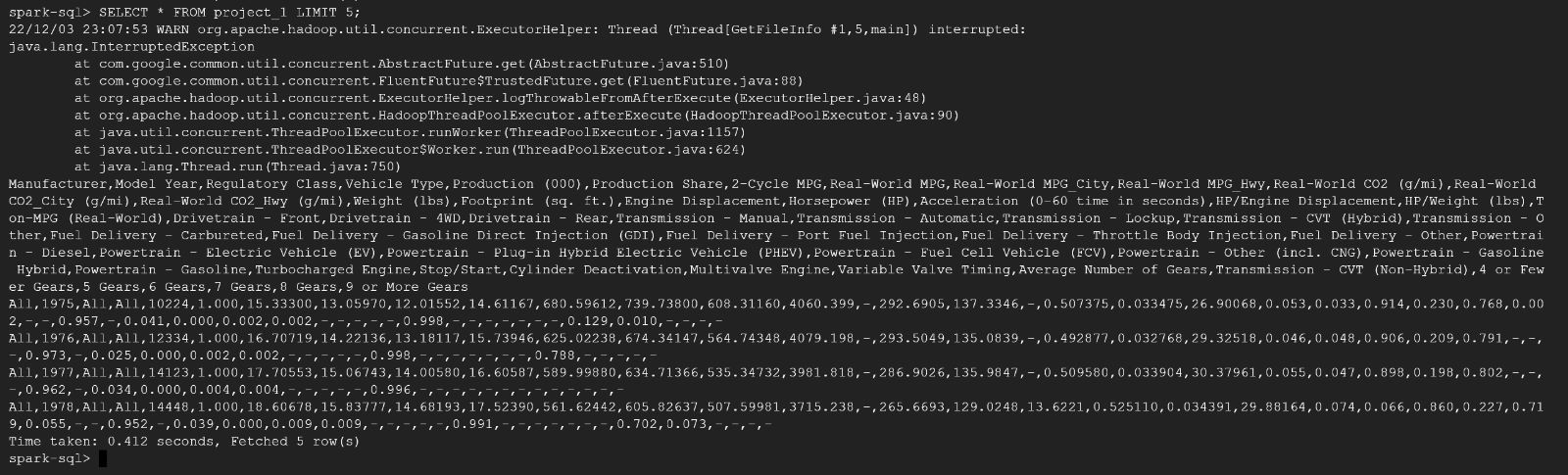
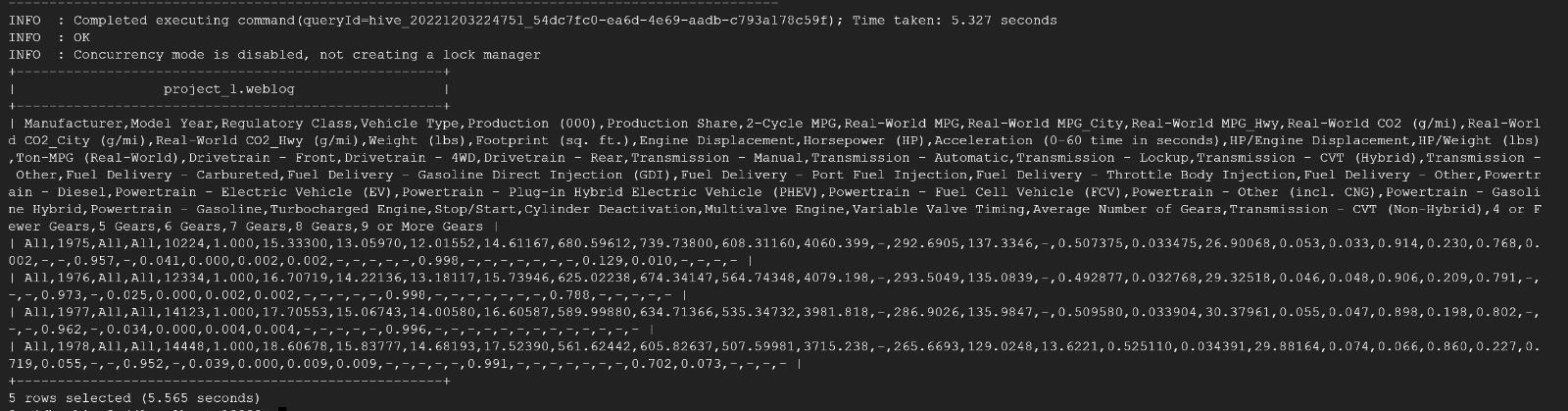


Both Hive and Spark are Relational Database Management Systems and are helpful to perform complex data analytics on big data. Below are the query results related to Hive and Spark.



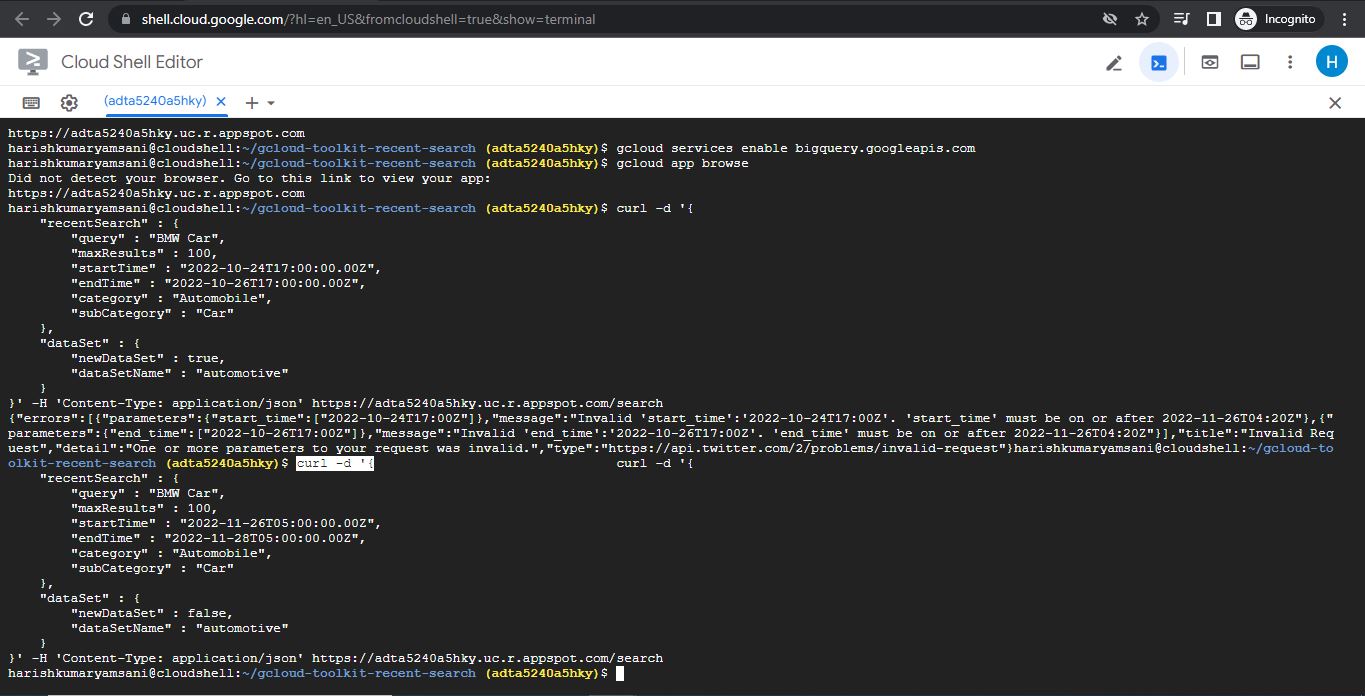






We have created an API that will fetch data from Twitter based on the criteria. We can cleanse this data by querying it and exporting it to either the GCP bucket or google sheets for further processing or storing it for further use.

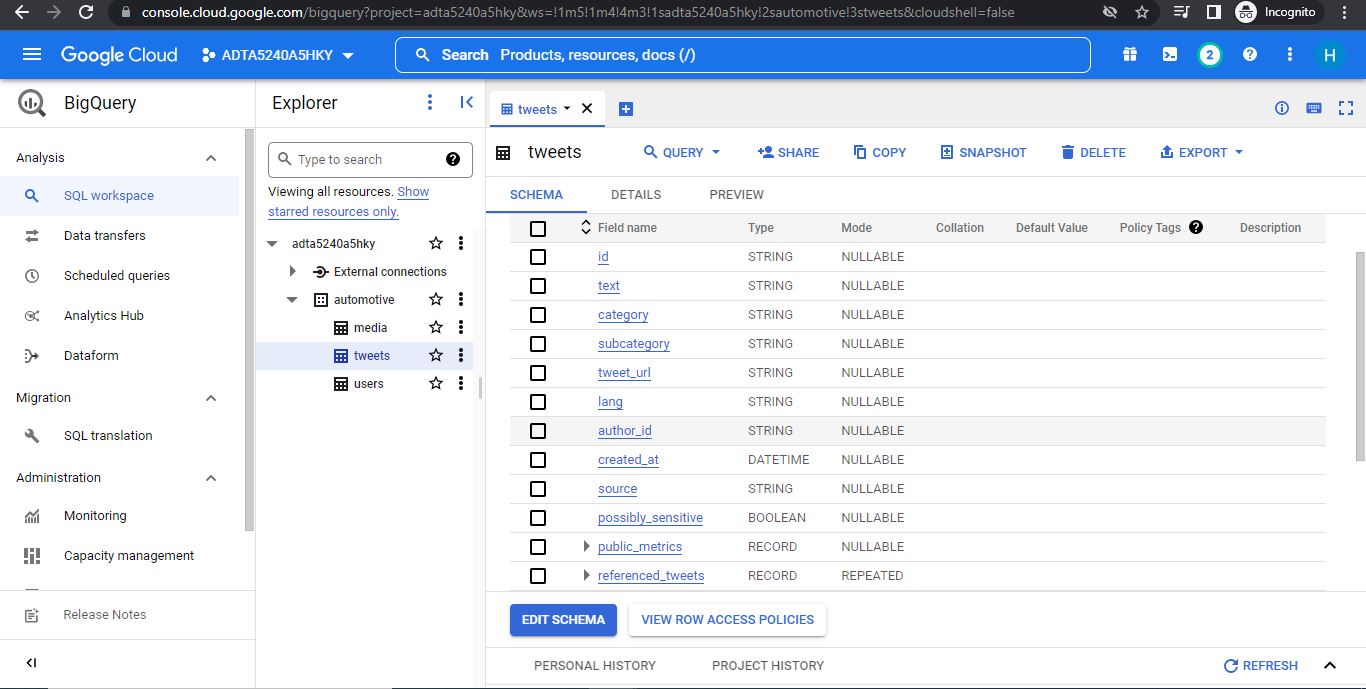
Written an API to fetch data related to Automobiles and the query results are particular to BMW cars between the given start and end times (Source: [URL](https://developer.twitter.com/en/docs/tutorials/developer-guide--twitter-api-toolkit-for-google-cloud%23:~:text=The%20Twitter%20API%20toolkit%20for,Twitter%20data%2C%20easily%20and%20quickly.))



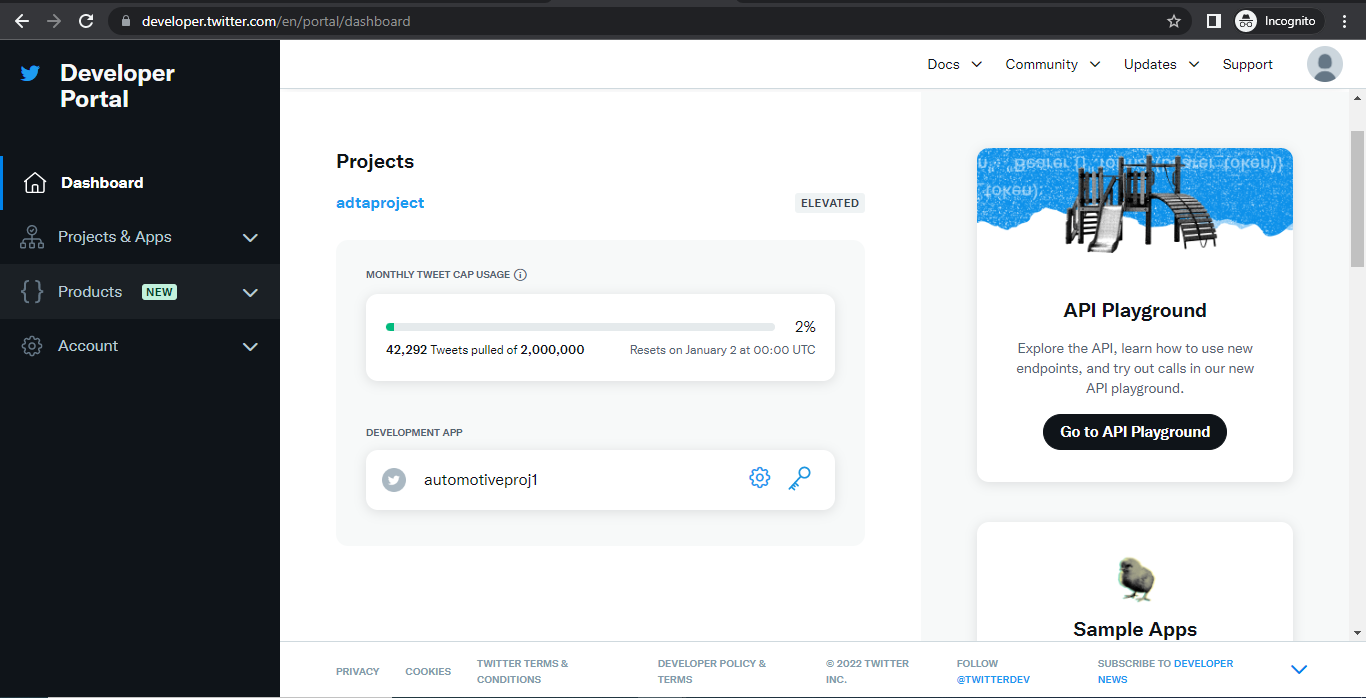
Upon the successful run of API, we can see the below three folders in the dataset.

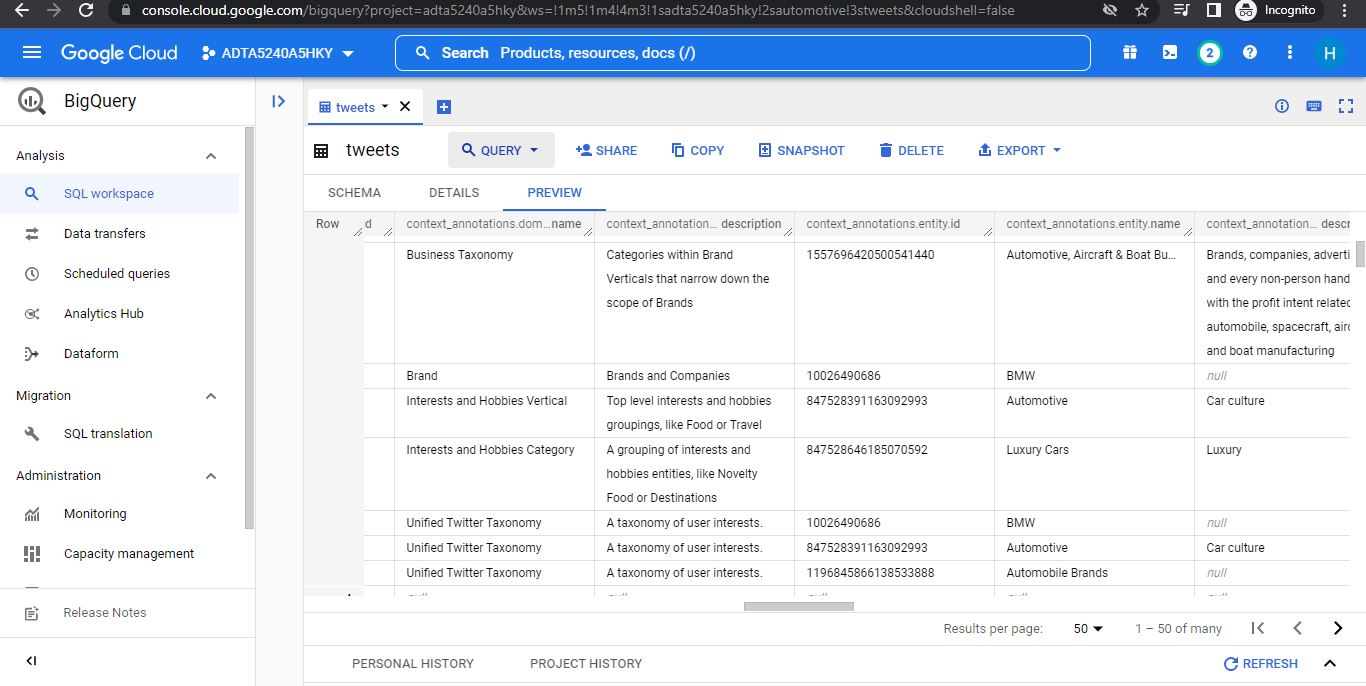
media, tweets and users

tweets data contains the information pulled from Twitter.

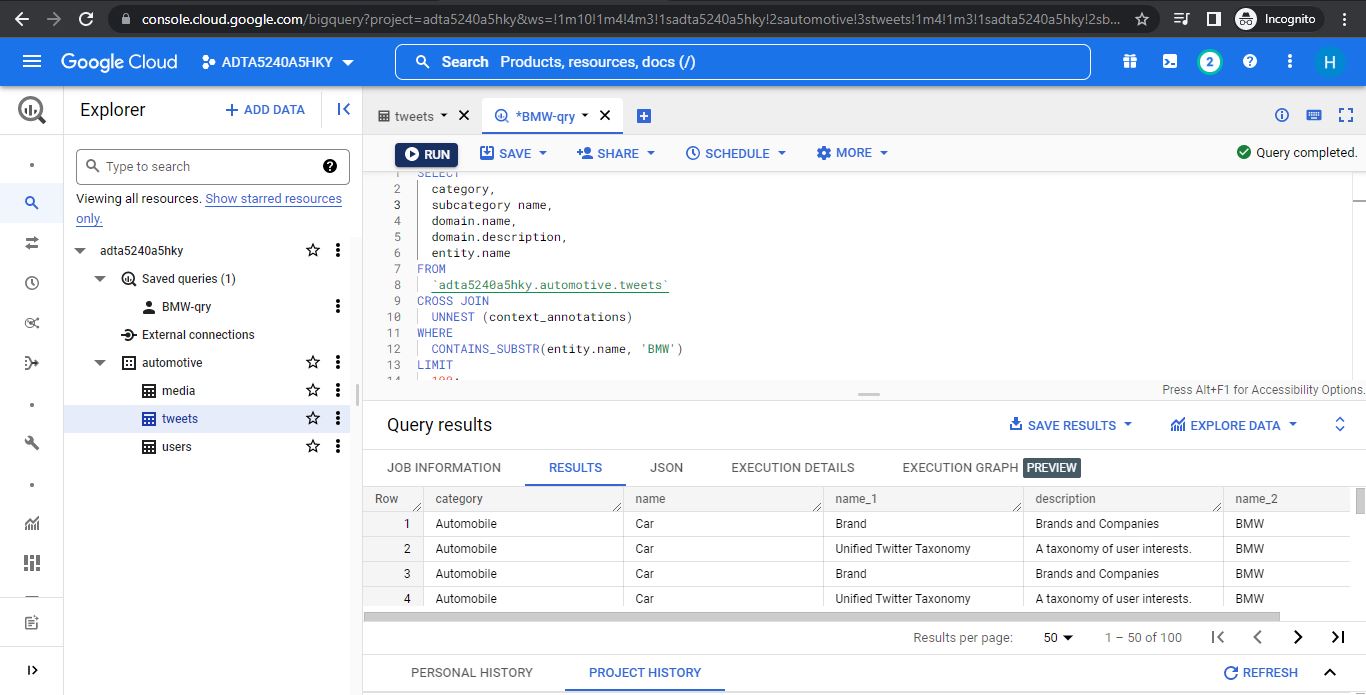


When we log in to the Twitter Developer account, it will show the number of tweets pulled using API, and show the remaining limit.

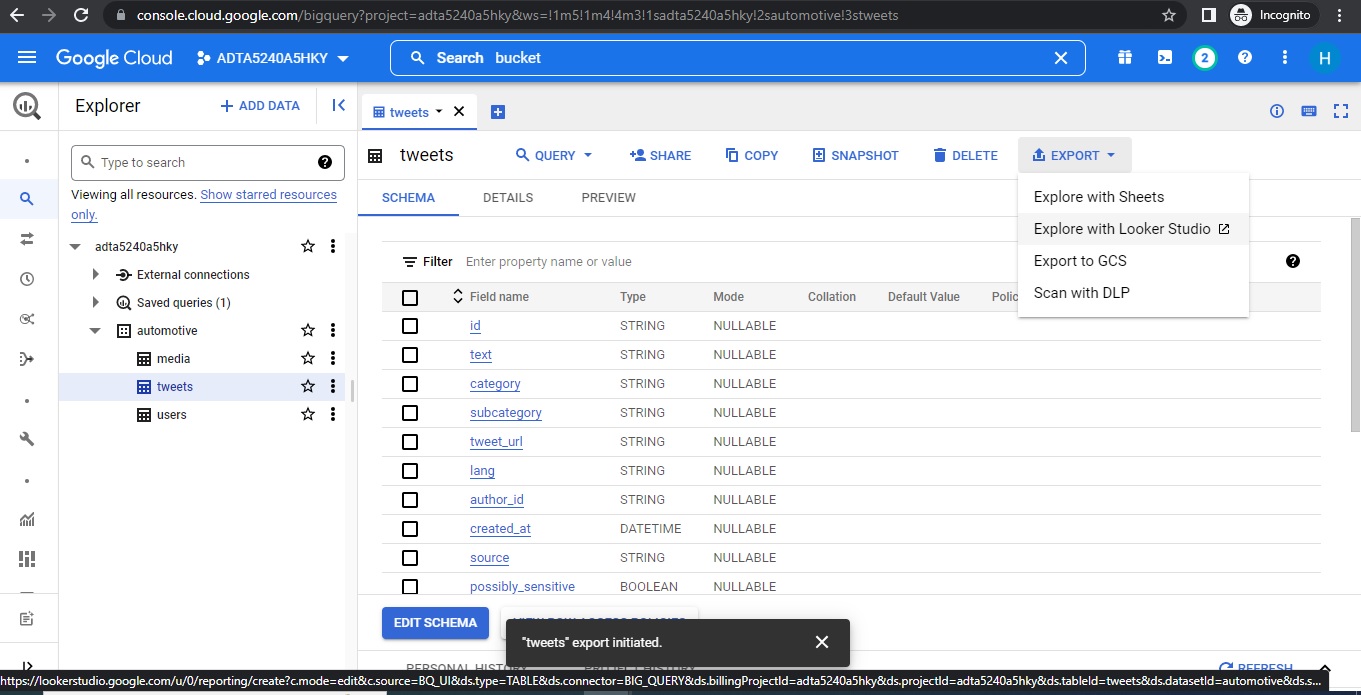


Below is the snapshot of the preview pane. Here we can see the data that we pulled from the Twitter feed. 

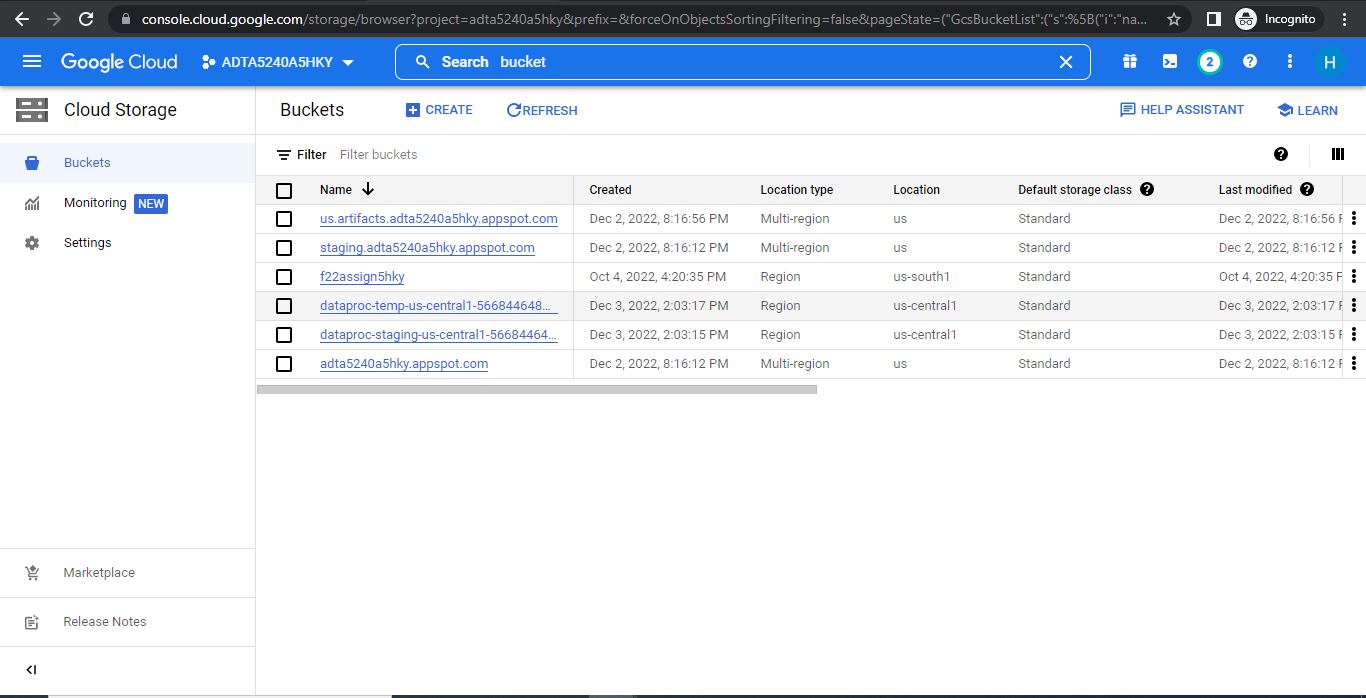
We can query it based on the requirement, and here is the snapshot of a sample query. Unlike traditional queries, we need to use 'CROSS JOIN' to take every individual element of your unnested array and join it back to its parent row.



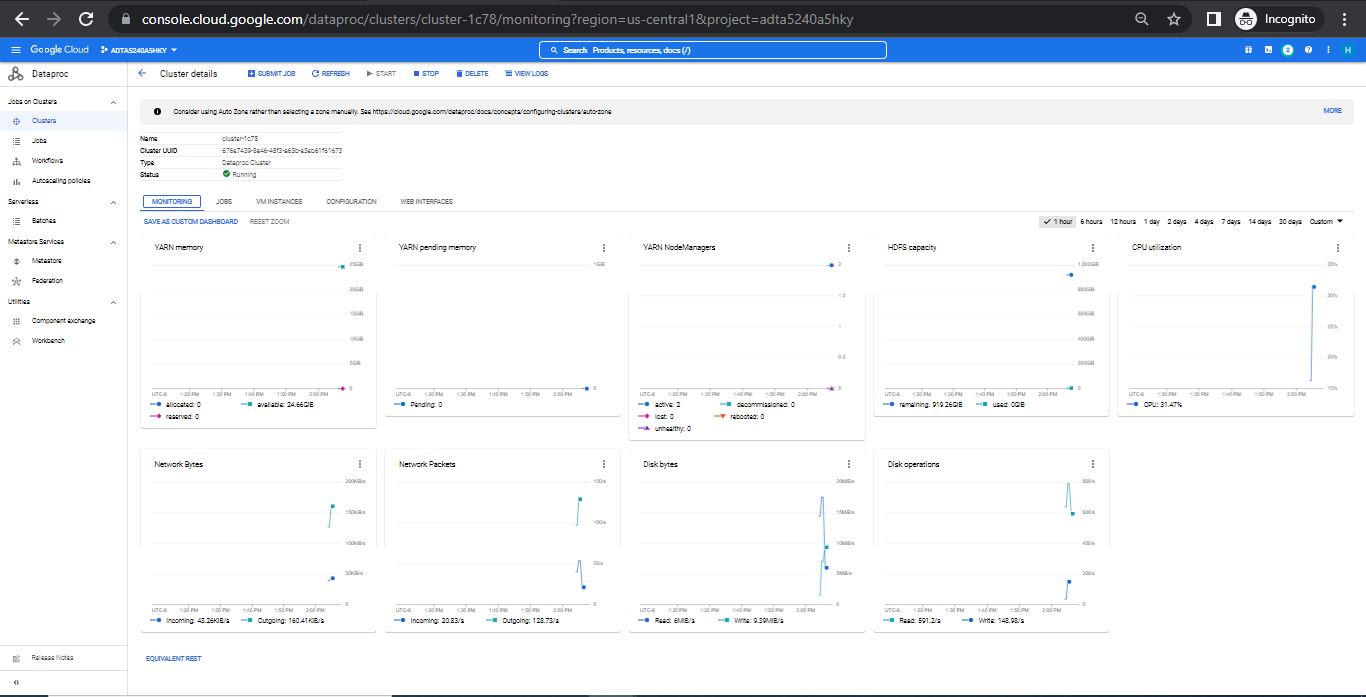
We can export data to a storage bucket, or a Google spreadsheet for further processing or visualization purposes.



Automatically metadata will be stored in the storage buckets. However, if we wish to export data from BigQuery to a storage bucket we can create a new file or can append the old one.



Below are the cluster monitoring details. By monitoring these details, we can check CPU utilization, Hadoop Distributed File System capacity, VM Instances, etc. It would be helpful for the Data Engineering team to monitor the performance and connectivity.



I have completed this entire work using my existing project id. For the Data Analyst team, I have created another project called ‘Automobile Project’ and will move everything to this project.

