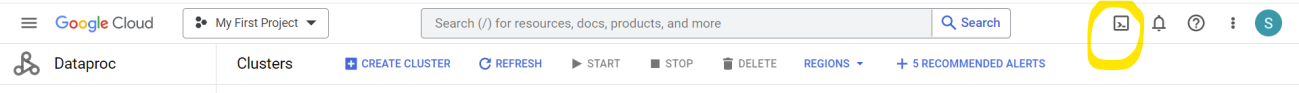
Assignment-5

Google Dataproc (Spark) and BigQuery

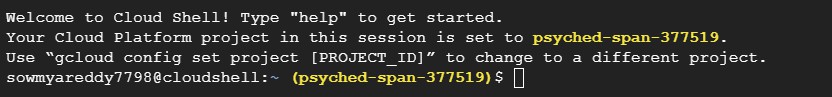
**Name: Sowmya Patlolla**

# Spark exercise:

Step1: Activate the cloud shell by clicking on the following icon.

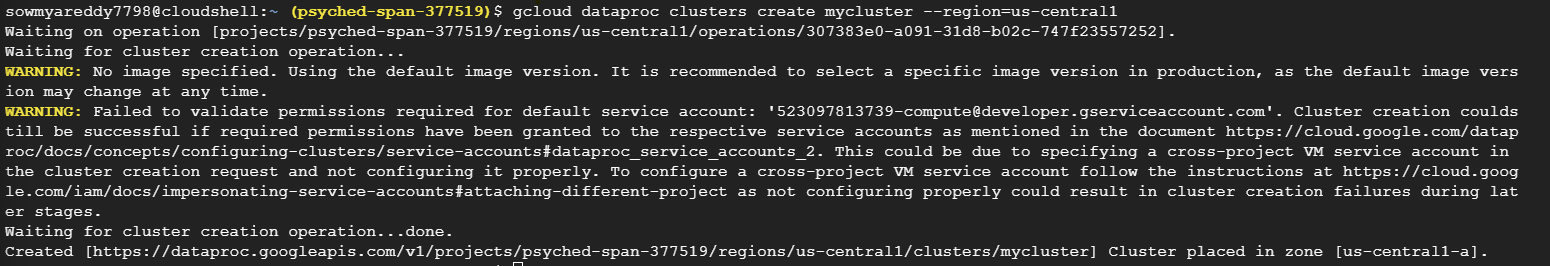


Cloud shell shows the following message after activating it.

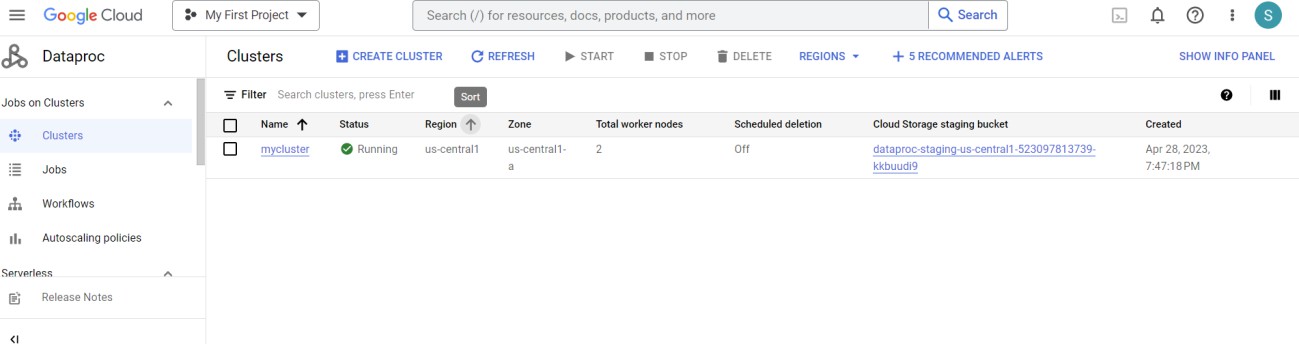


Step2: Created the Dataproc cluster named “mycluster” in the “us-central1” region by running the following command.

gcloud dataproc clusters create mycluster --region=us-central1



The following image shows the created cluster named “mycluster” in the cloud console.



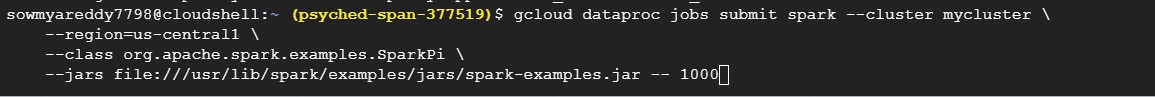
Step3: Submitting the job which calculates the pi values on the created Dataproc cluster by executing the following command.

gcloud dataproc jobs submit spark --cluster mycluster \

--region=us-central1 \

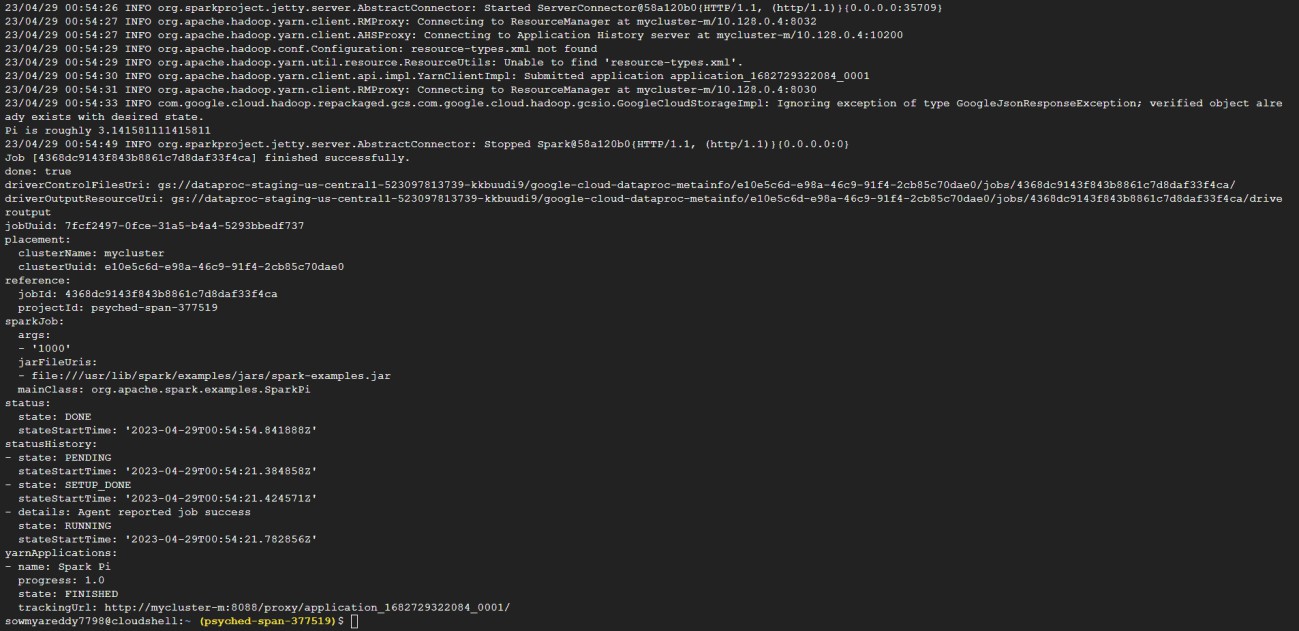
--class org.apache.spark.examples.SparkPi \

--jars file:///usr/lib/spark/examples/jars/spark-examples.jar -- 1000

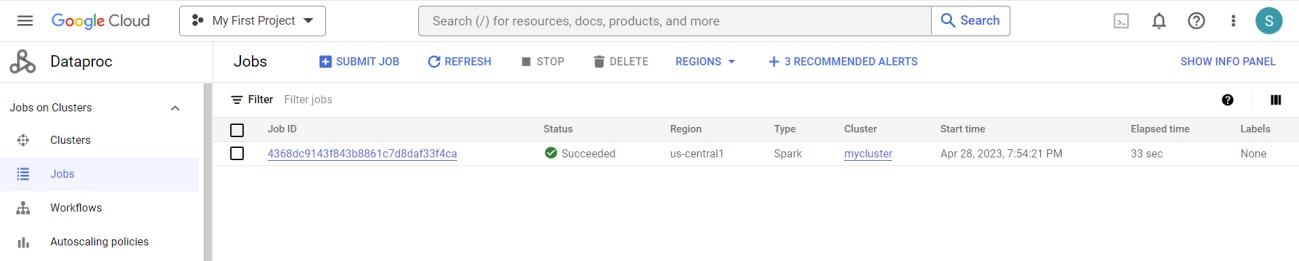


The following image shows the output of running the Dataproc job which results in the value of the pi.

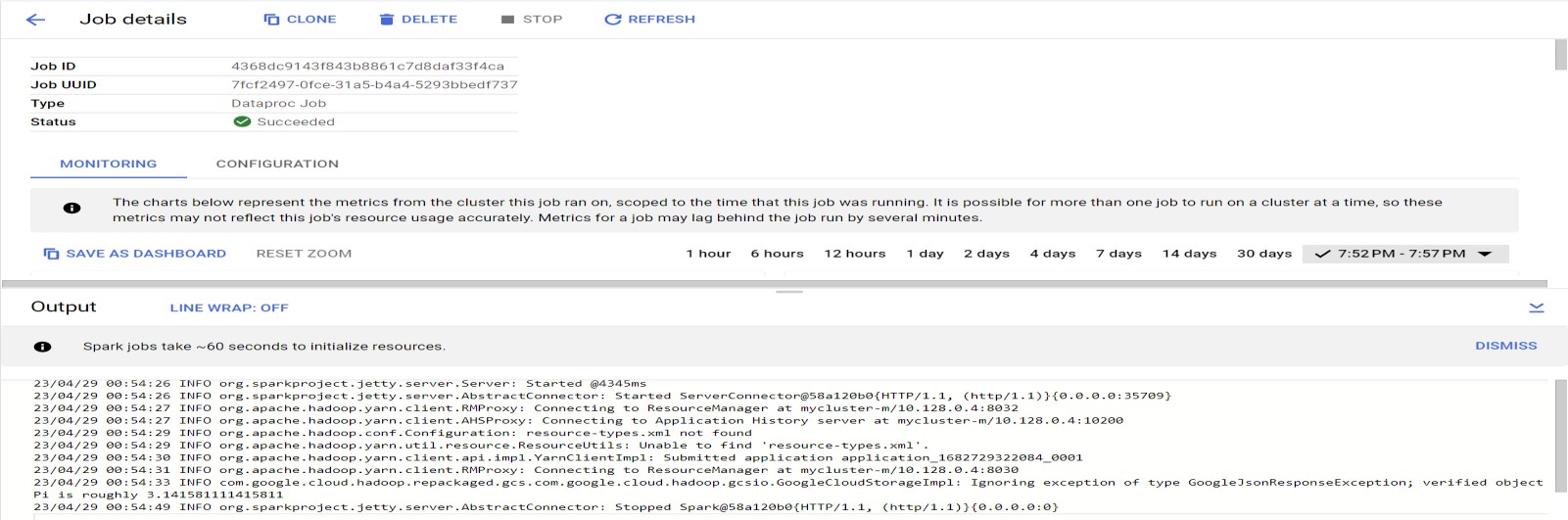




The following image shows the created job and its complete information on the cloud console.



Job info along with output in the cloud console.

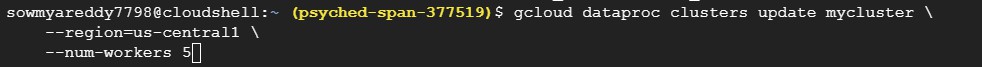


Step4: Tried to increase the number of workers nodes in the created cluster by updating it using the following command.

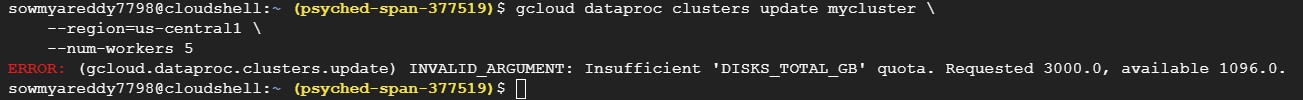
gcloud dataproc clusters update mycluster \

--region=us-central1 \

--num-workers 5



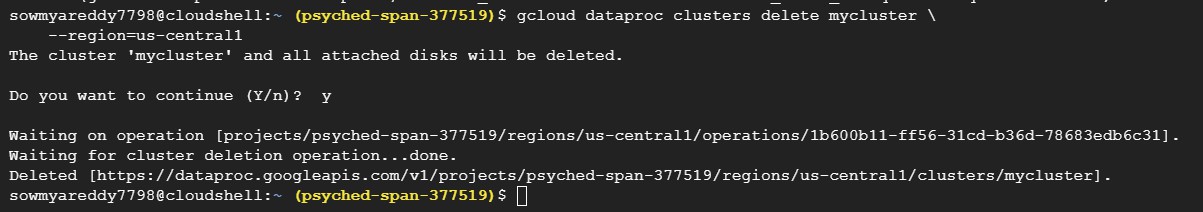
Following is the output of the above command. But the cluster didn’t update due to the insufficient disk space in my cloud account.



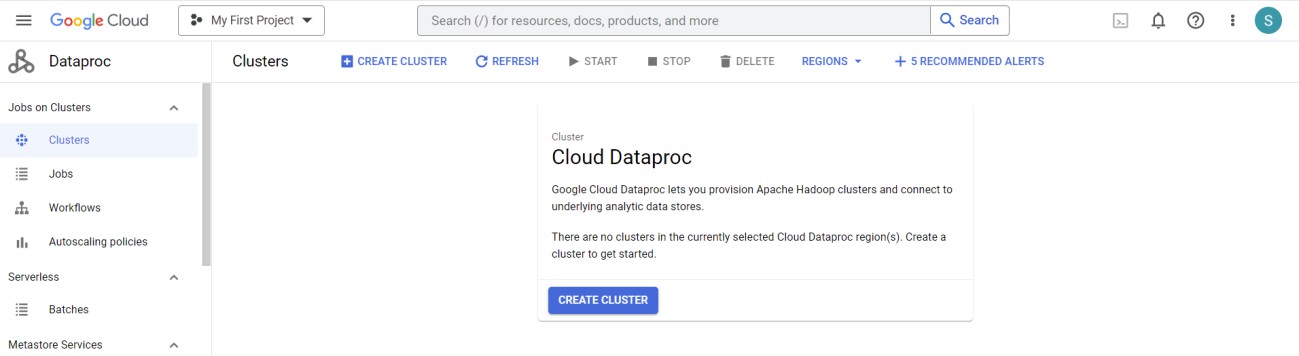
Step5: After finishing the job delete the created Dataproc cluster by running the following command to reduce the costs of your cloud project.

gcloud dataproc clusters delete mycluster \

--region=us-central1

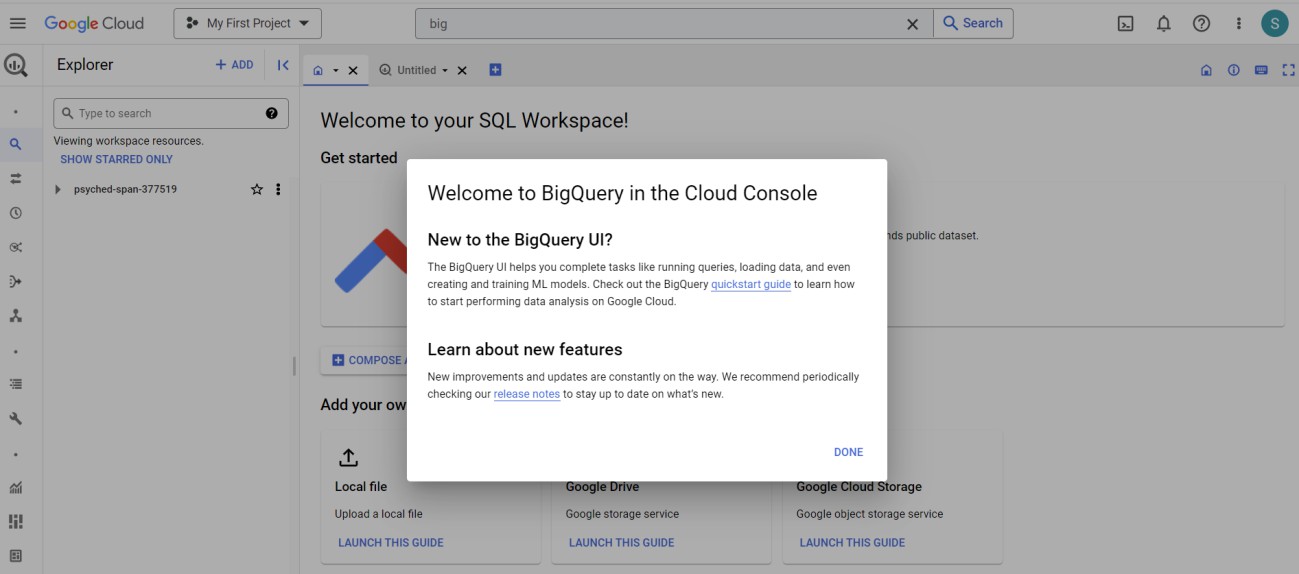


After deleting the cluster, the console appears as follows.

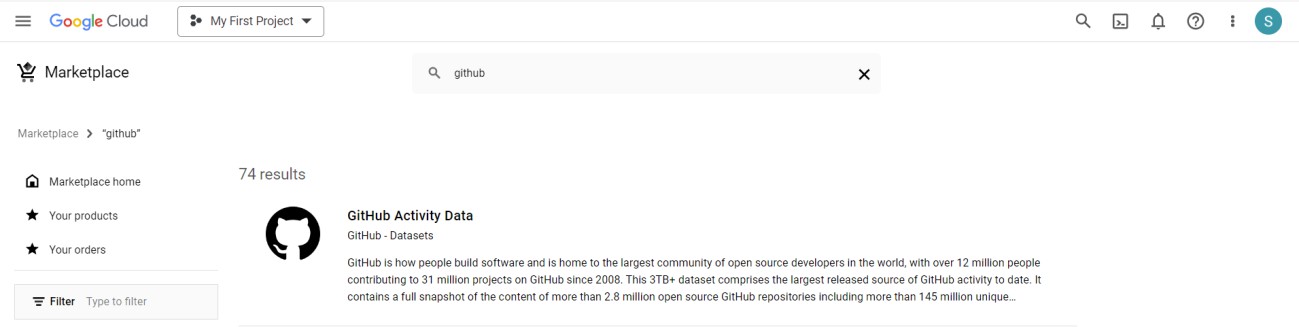
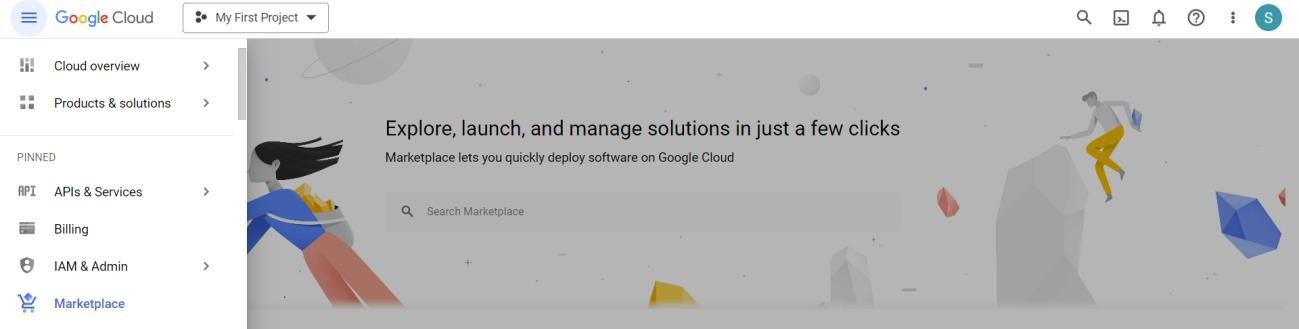


# Bigquery exercise:

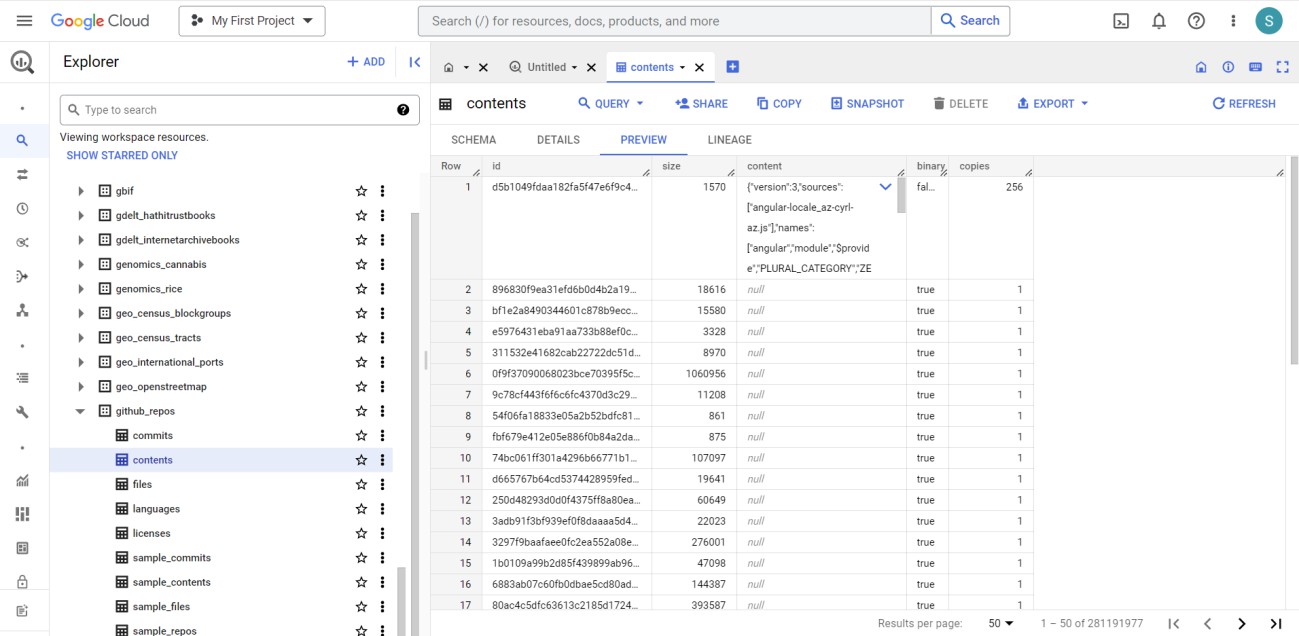
Step1: Enabling the bigquery in the cloud console.



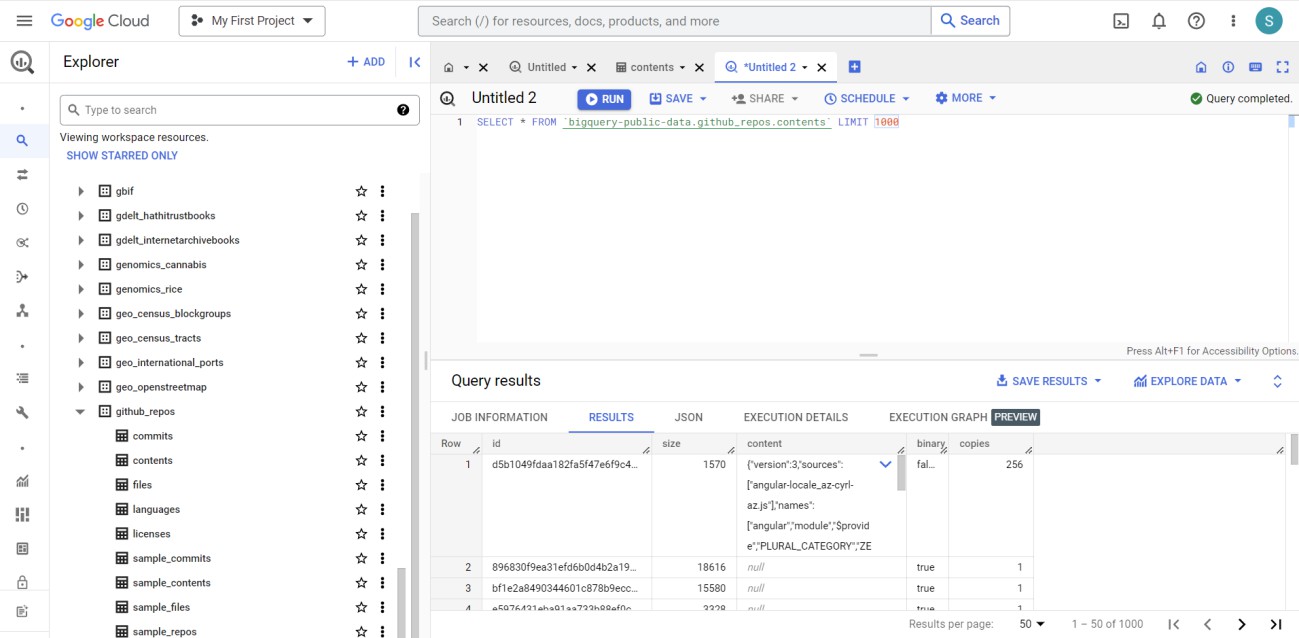
Step2: Go to the cloud marketplace and load the public dataset (GitHub activity data) to bigquery.



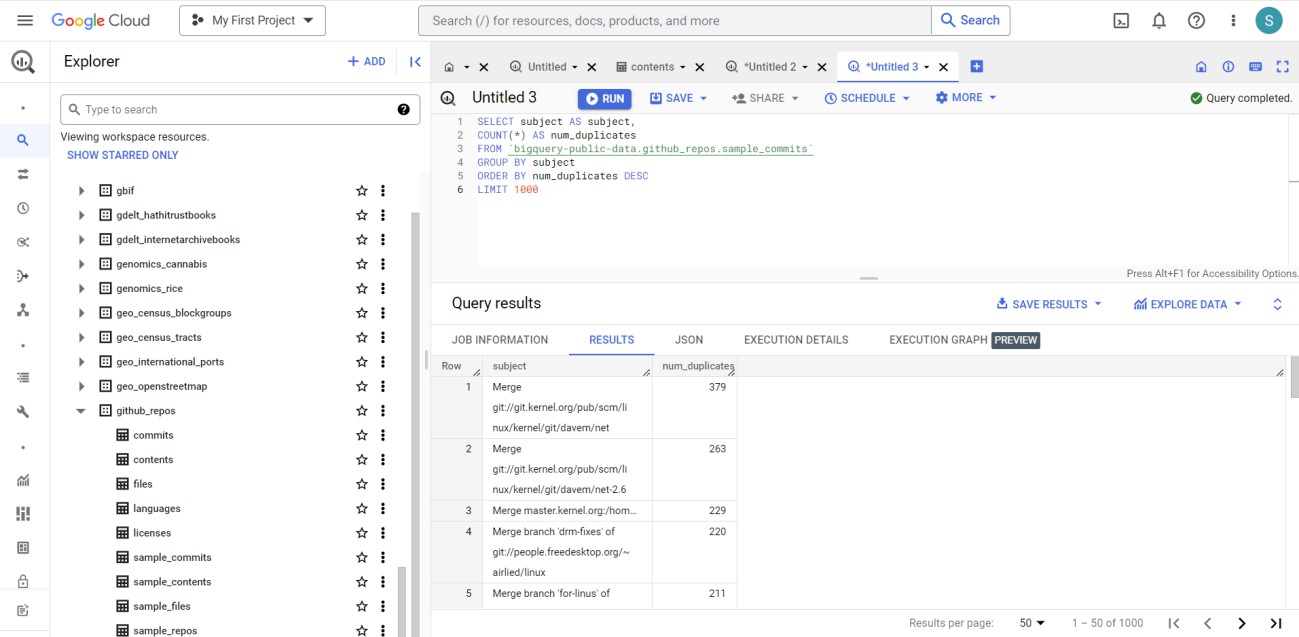
Step3: Let’s make a quick preview on GitHub Public Activity Data.

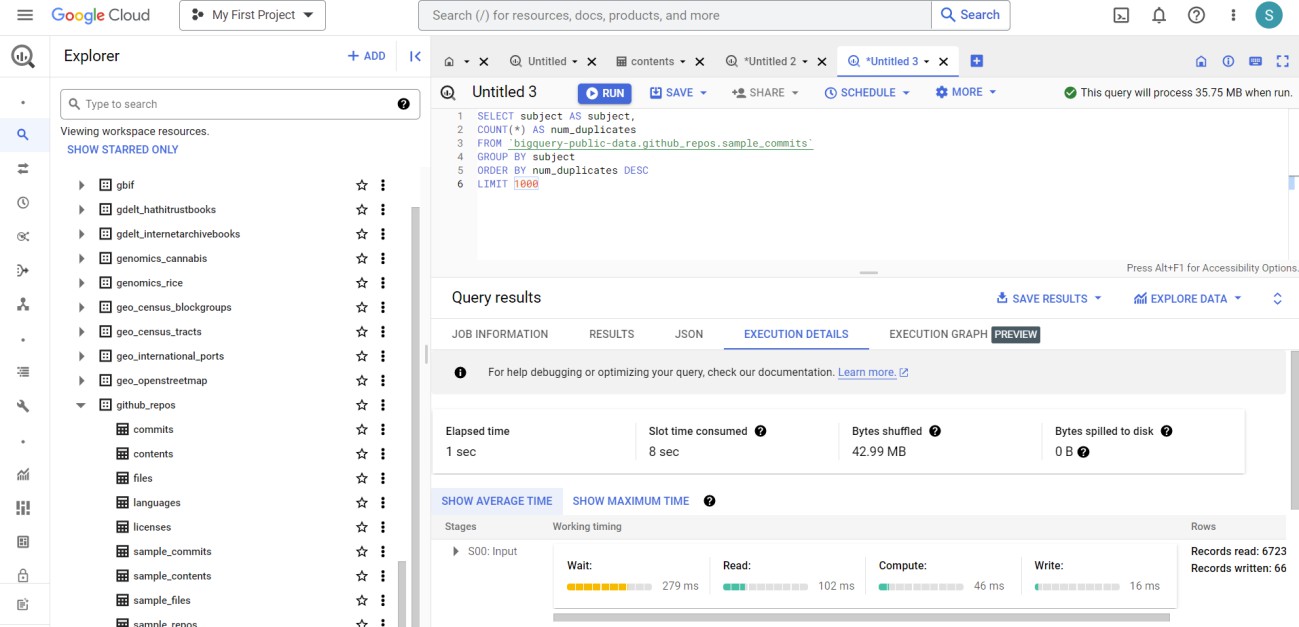


Step4: Let’s make a quick preview of how data looks by running the following command.

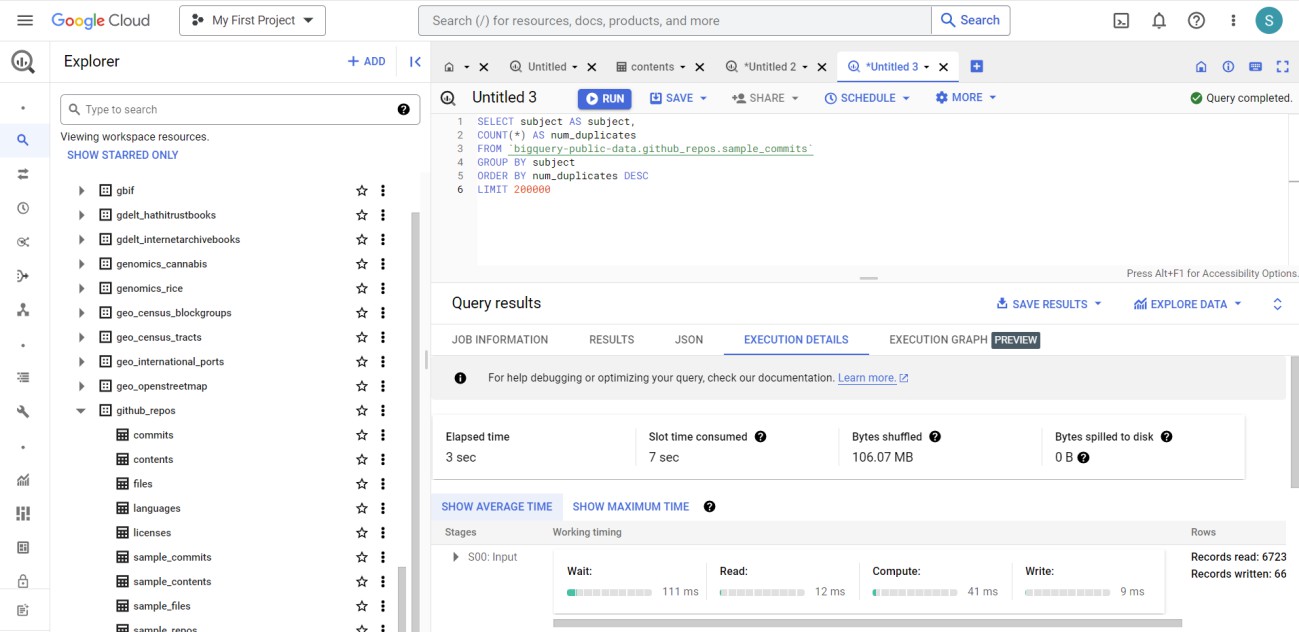


Step5: Executing the simple query on GitHub commits data.

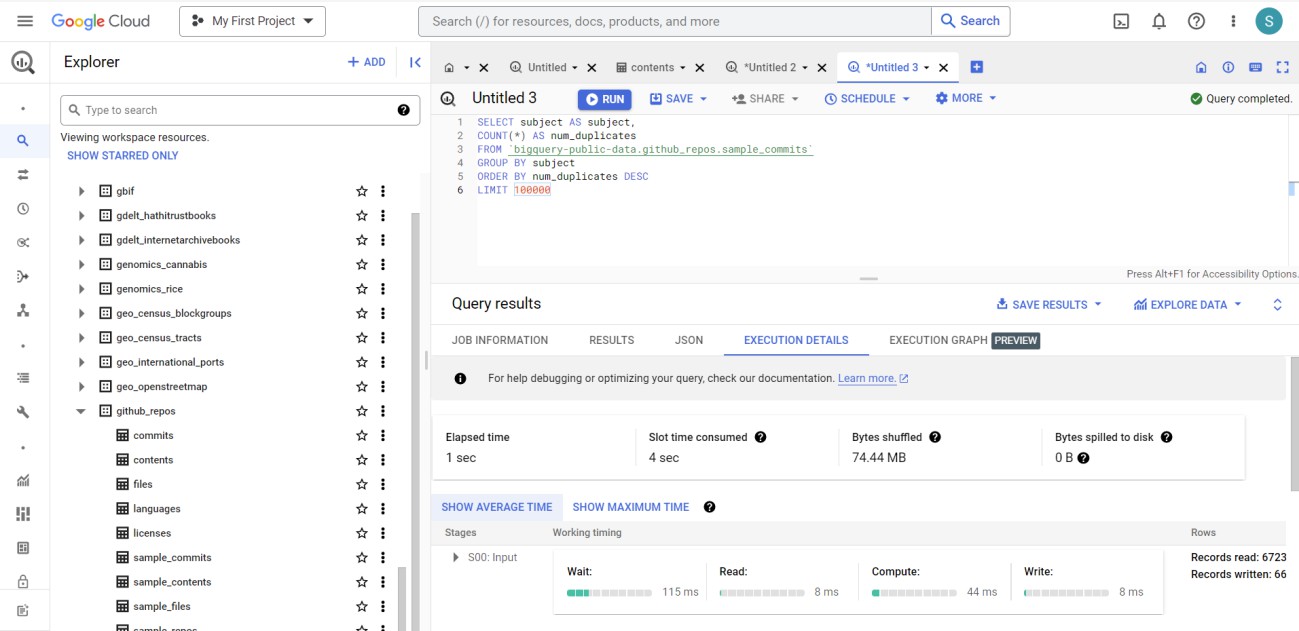




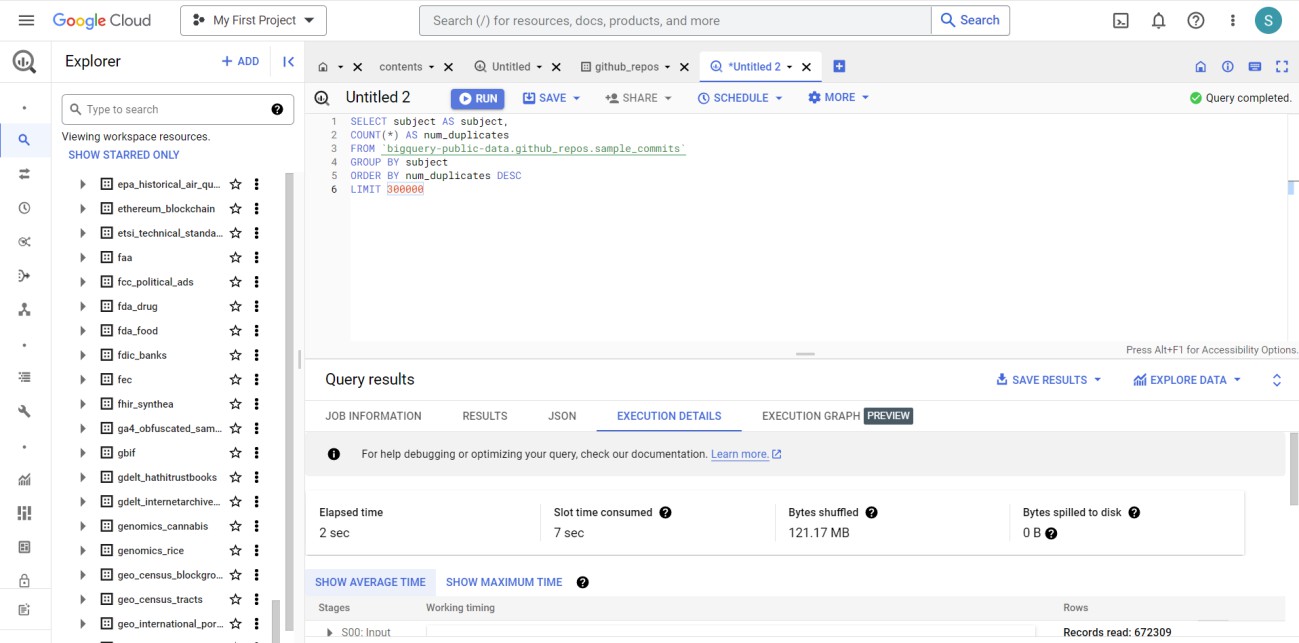
Step6: Executing the query to fetch 200000 rows of data.



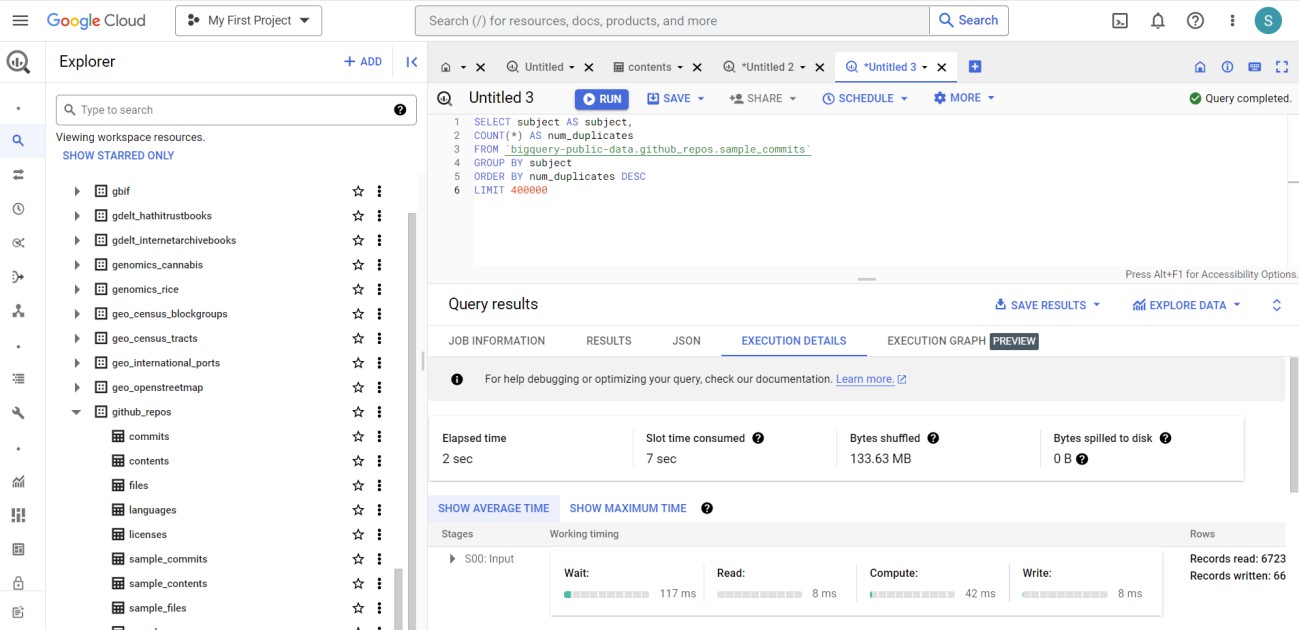
Step7: Executing the query to fetch 100000 rows of data.



Step8: Executing the query to fetch 300000 rows of data.



Step9: Executing the query to fetch 400000 rows of data.



Step10: Analyzing the queries results of 100000,200000,300000 and 400000 rows of data based on Elapsed time, Slot time consumed, and bytes shufled.

|  |  |  |  |
| --- | --- | --- | --- |
| No. of rows  computed | Elapsed time | Slot time  consumed | Bytes shufled |
| 100000 | 1 sec | 4 sec | 74.44 MB |
| 200000 | 3 sec | 7 sec | 106.07 MB |
| 300000 | 2 sec | 7 sec | 121.17 MB |
| 400000 | 2 sec | 7 sec | 133.63 MB |

By observing the above table, even though the number of rows computed were double, there was only minor difference with elapsed and slot time. But the

amount of bytes shufled will depend on the number of rows computed.

# Epilog:

Setting up and configuring a Dataproc cluster can be a complex process, especially while configuring the right number of resources, choosing the right cluster size, and managing the network settings. Maintaining and managing data on the cluster can also be a challenge. We must ensure that the cluster is up to date, that the software stack is compatible with the job requirements or not and the data is

stored in the right format or not. Dataproc clusters can be expensive, and we must keep an eye on the cost of running the cluster. We must ensure that they are not over-provisioning the cluster and that they are not running unnecessary jobs that can drive up the cost. Security is another challenge that we must consider when working with Dataproc clusters. We must ensure that the cluster is secure from unauthorized access.

Query performance can be a challenge with BigQuery, particularly for complex or ad-hoc queries. We must ensure that the queries are optimized to run efficiently, and we may need to partition the data or use clustering to improve performance. BigQuery can be expensive, particularly for users with large amounts of data. must keep an eye on the cost of running queries.