**CS5525-S21 Assignment-5**

**(Dataproc and BigQuery on GCP)**

**Name: Vandana Maddi** **ID: 16307527**

**Task-01: Install Cloud SDK**

**Step-01:** Installing GCloud SDK

Graphical user interface, text, application, email

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

Graphical user interface, text, application

Description automatically generated

Graphical user interface, text

Description automatically generated

Graphical user interface, text, application

Description automatically generated

Graphical user interface, text, application

Description automatically generated

**Step-02:** Setup GCloud SDK and using **“gcloud info”** , **“gcloud init”** commands

Text

Description automatically generated

**Step-03:** Configuring account and project using gcloud init command

Text

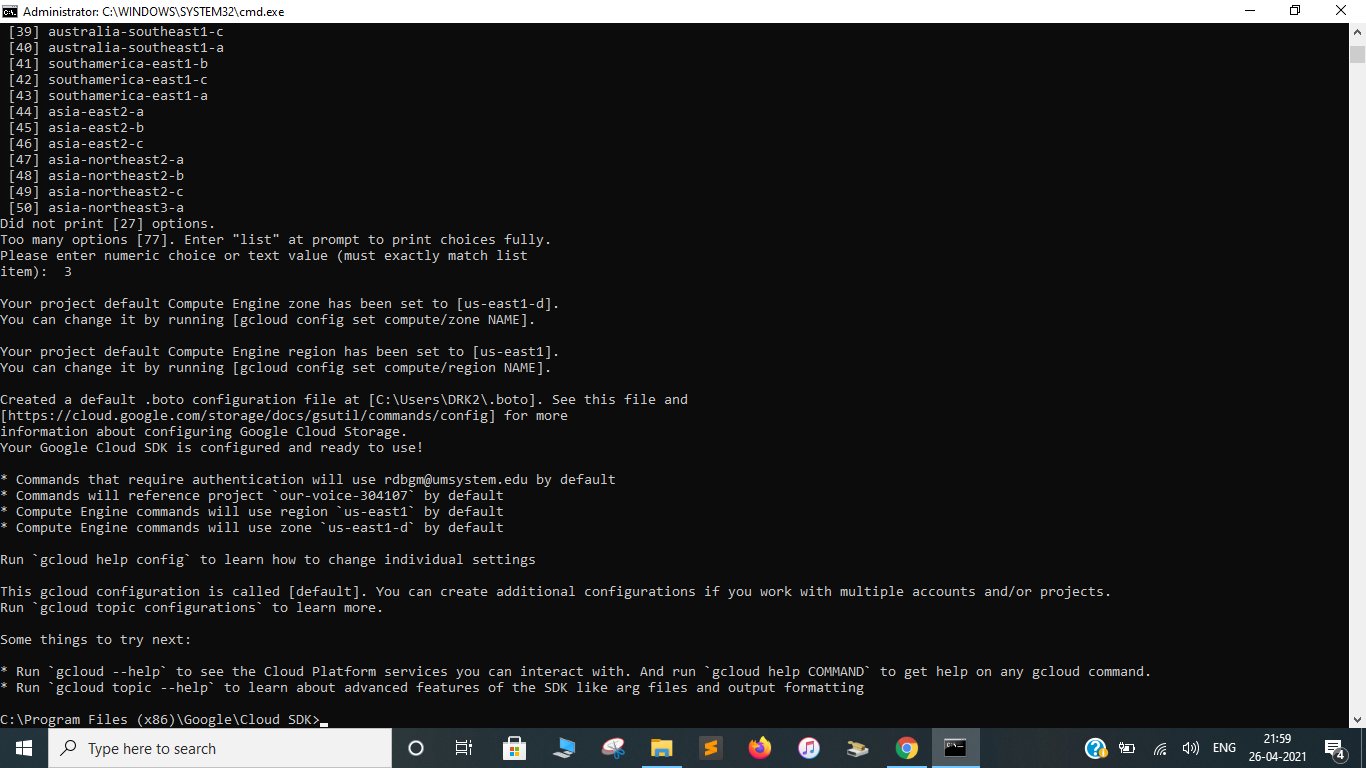
Description automatically generated

Text

Description automatically generated

Text

Description automatically generated



Text

Description automatically generated

Text

Description automatically generated

**Task-02: Dataproc**

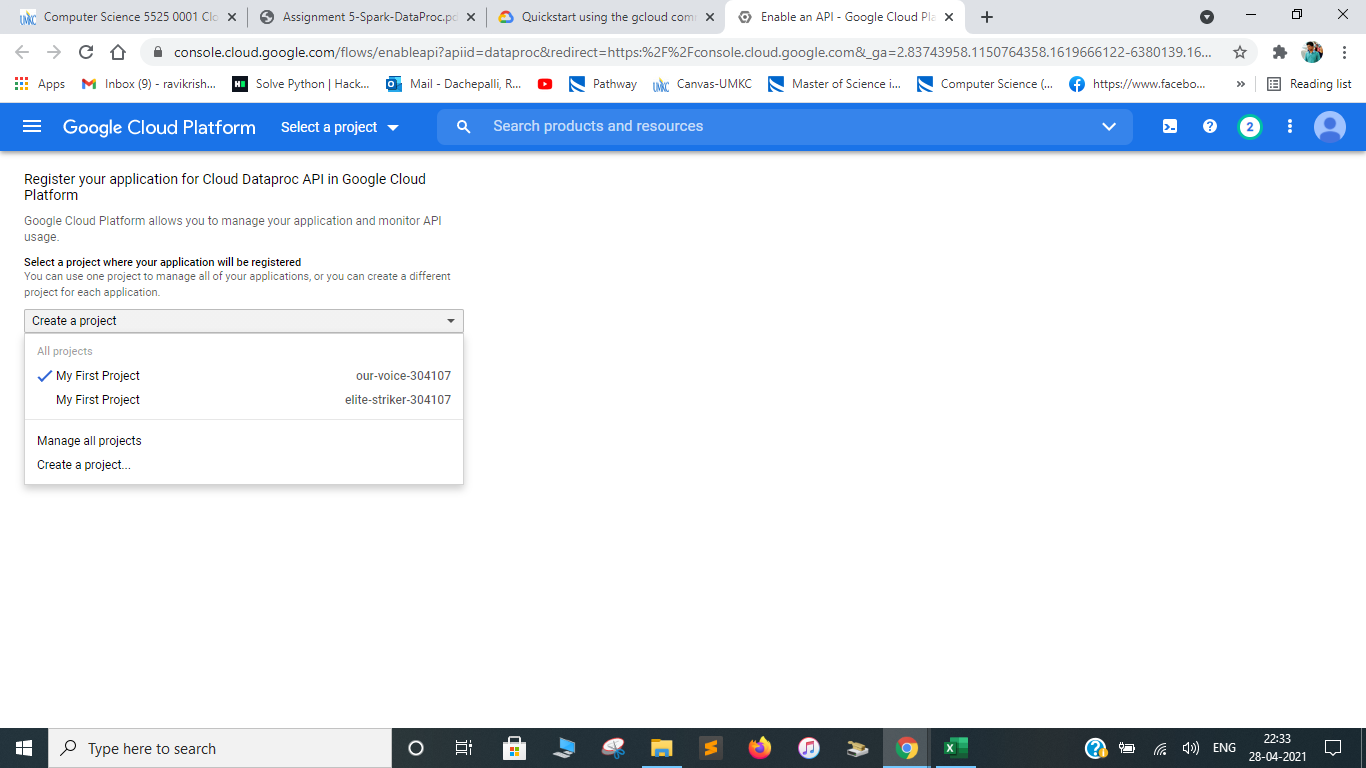
**Step-01:** Going through the GCP documentation, First select the project through GCP Console and Enable API as per the instructions.

Graphical user interface, text, application, email

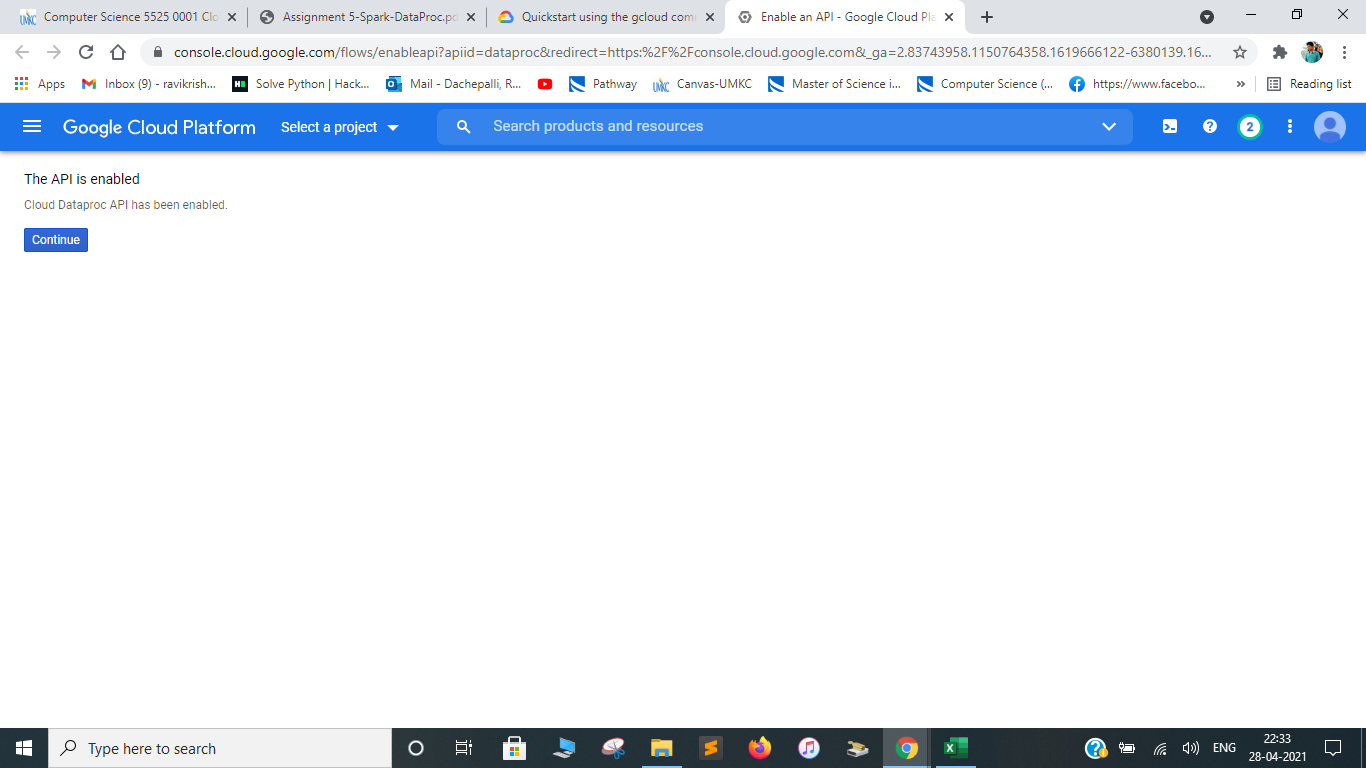
Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

Select the project

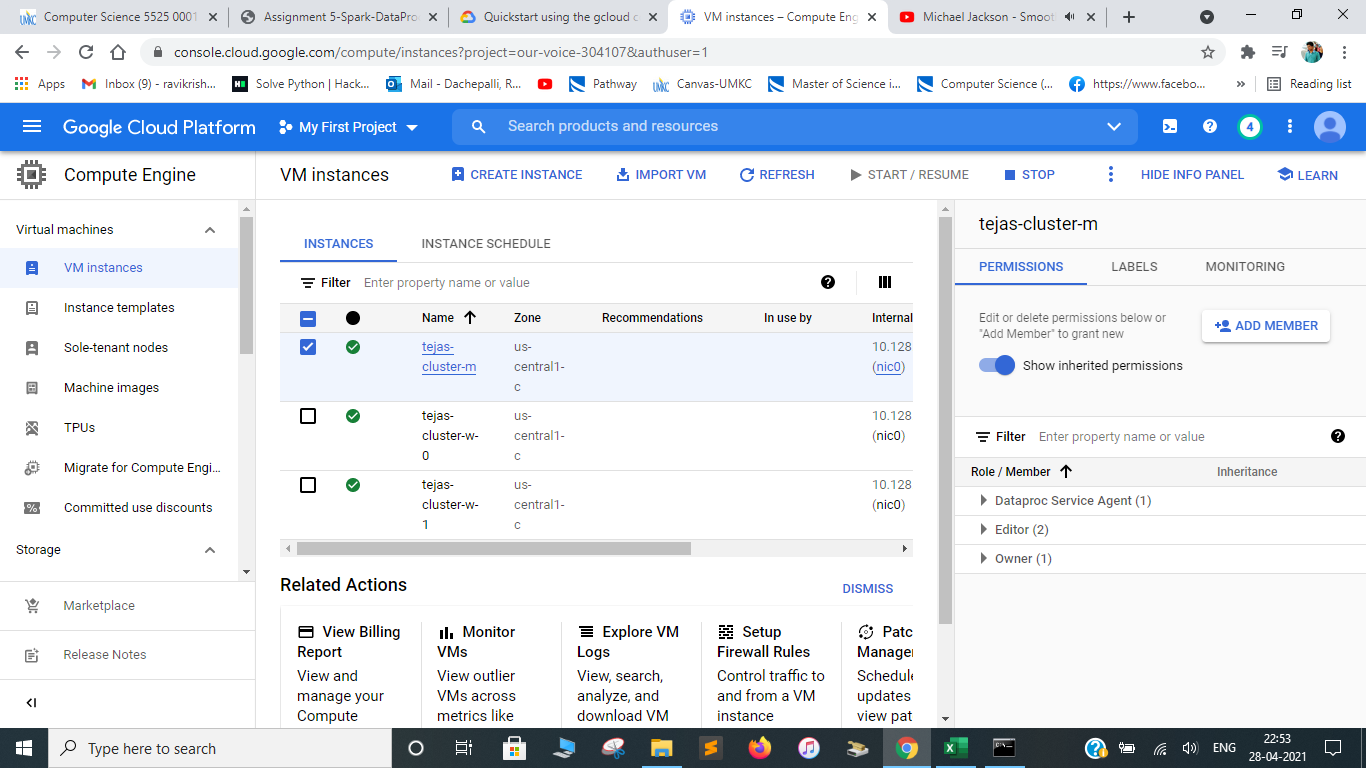
Enable API after selecting the project

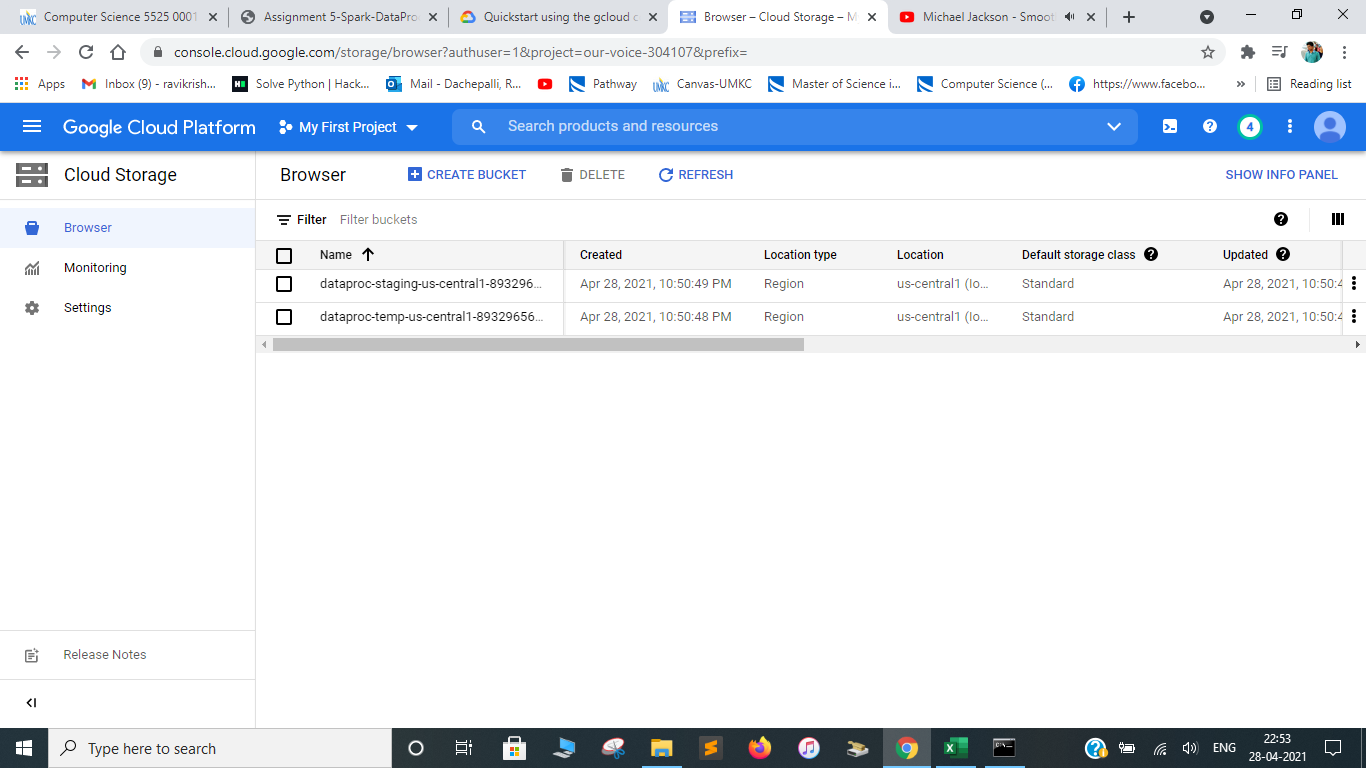


**Step-02:** Using SDK, create a cluster using the following command.

**Command: gcloud dataproc clusters create example-cluster --region=region**

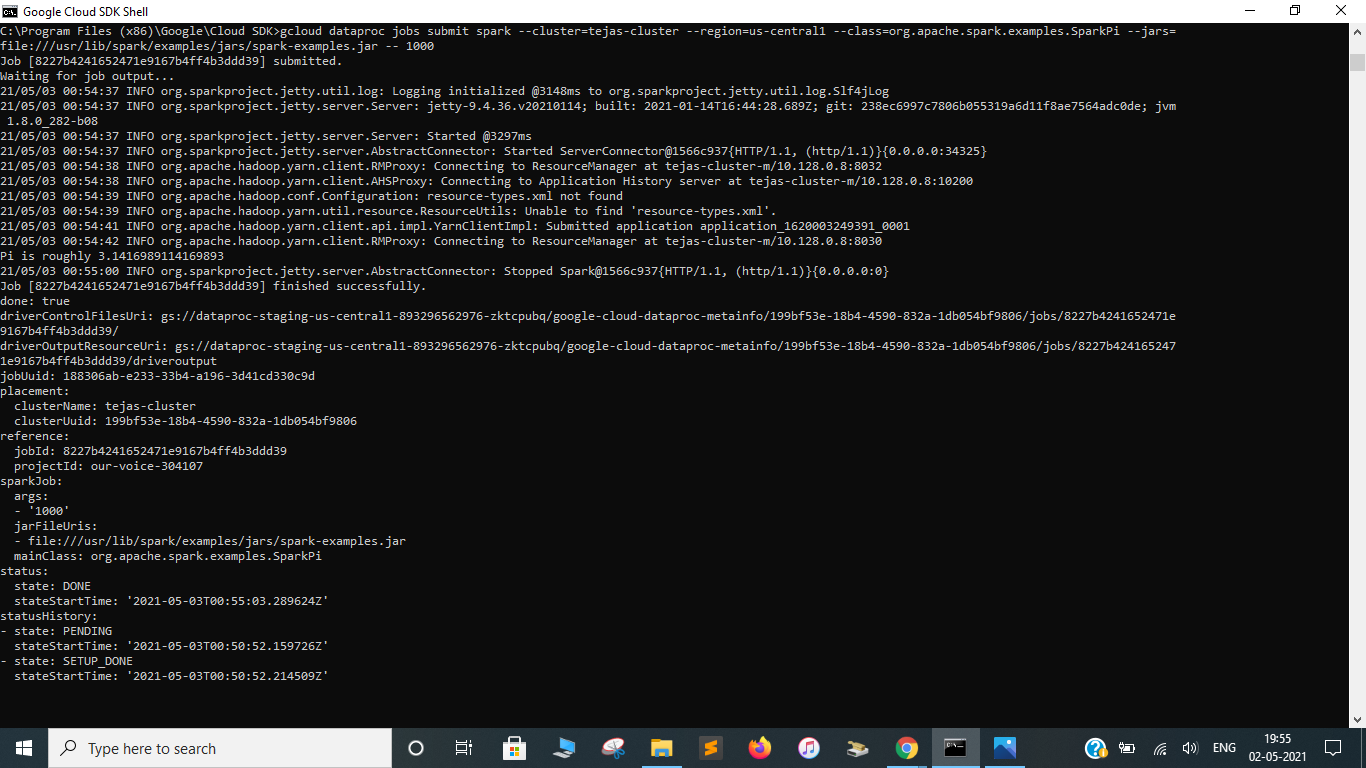
A cluster will be created with one master node and two worker nodes as in the following picture.



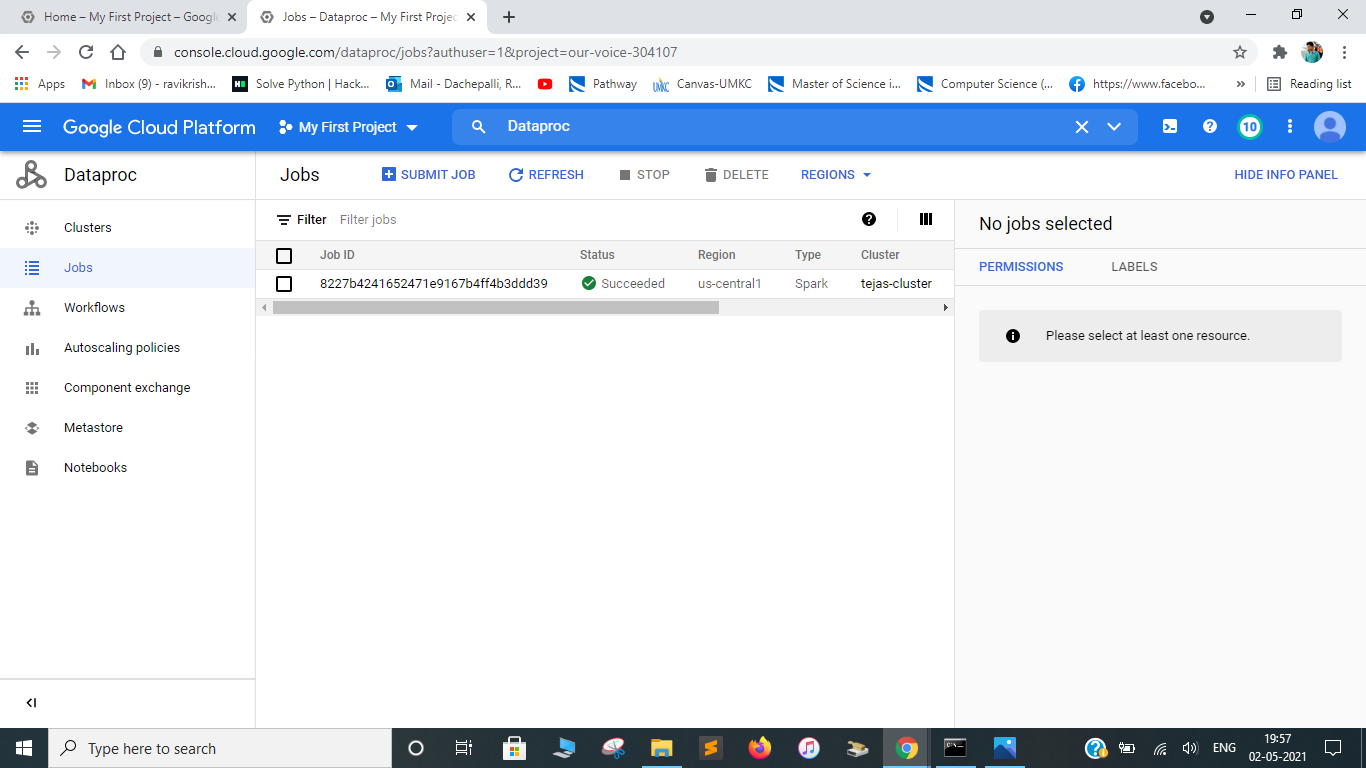


**Step-03:** To submit a Job, use the following command.

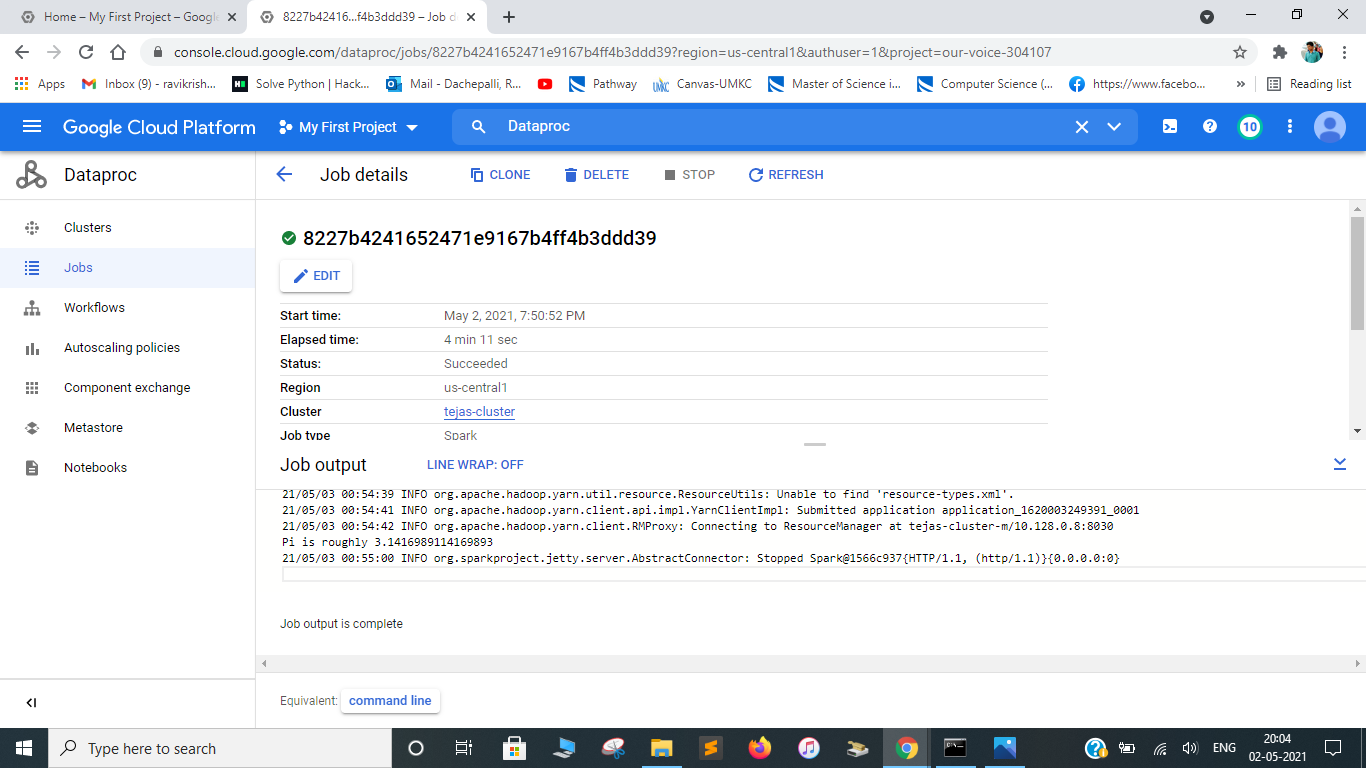
**Command:** gcloud dataproc jobs submit spark --cluster tejas-cluster \ --region=*us-central1* \ --class=org.apache.spark.examples.SparkPi \ --jars=file:///usr/lib/spark/examples/jars/spark-examples.jar – 1000



Job created

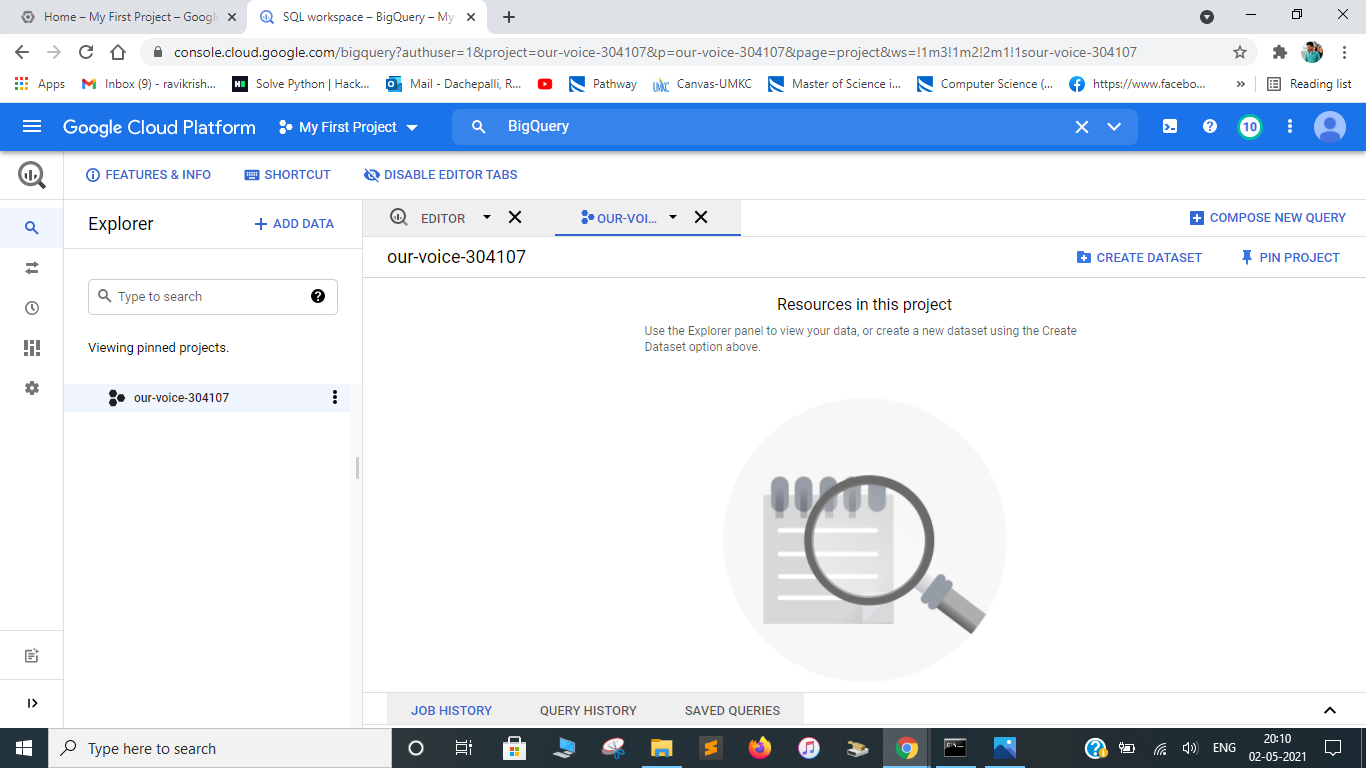


**Step-04:** Calculate the pi-value. Here, the pi value is 3.1416989114169893

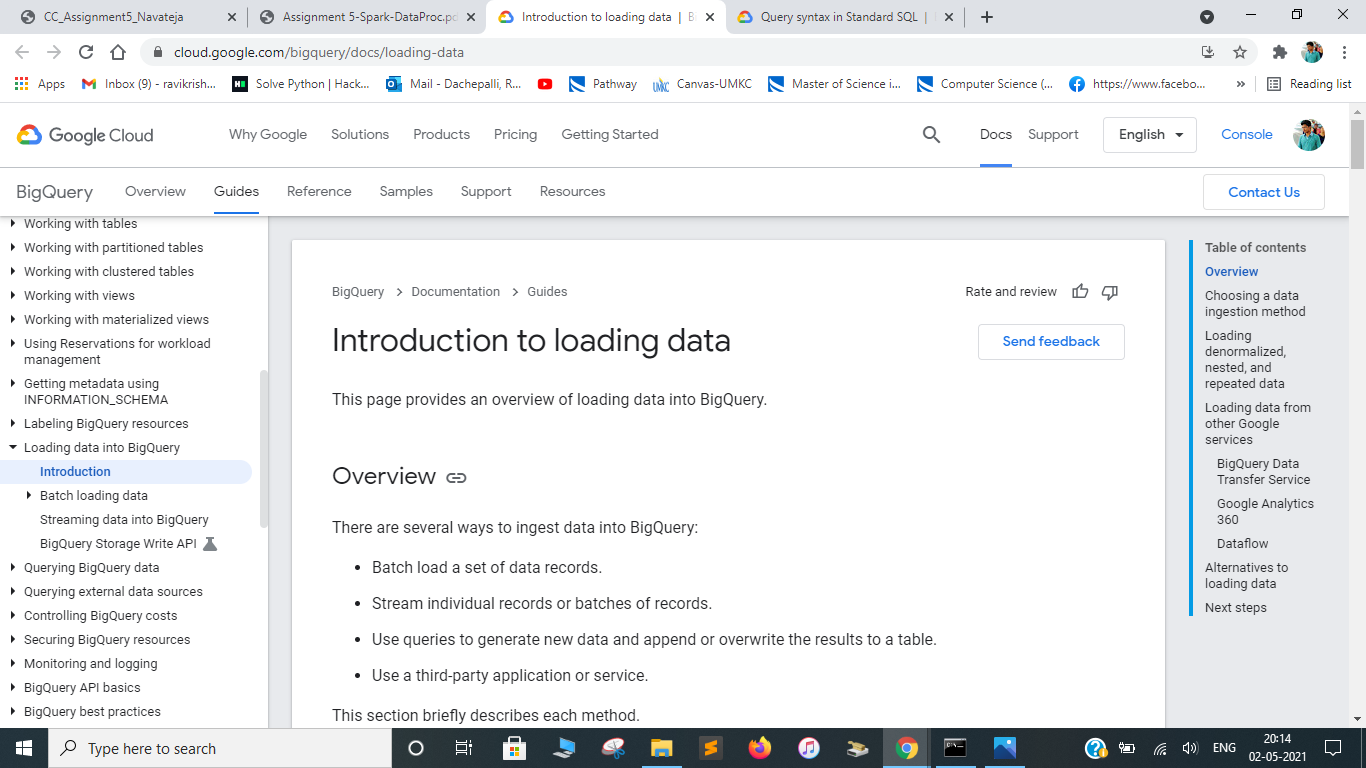


**Task-03: BigQuery**

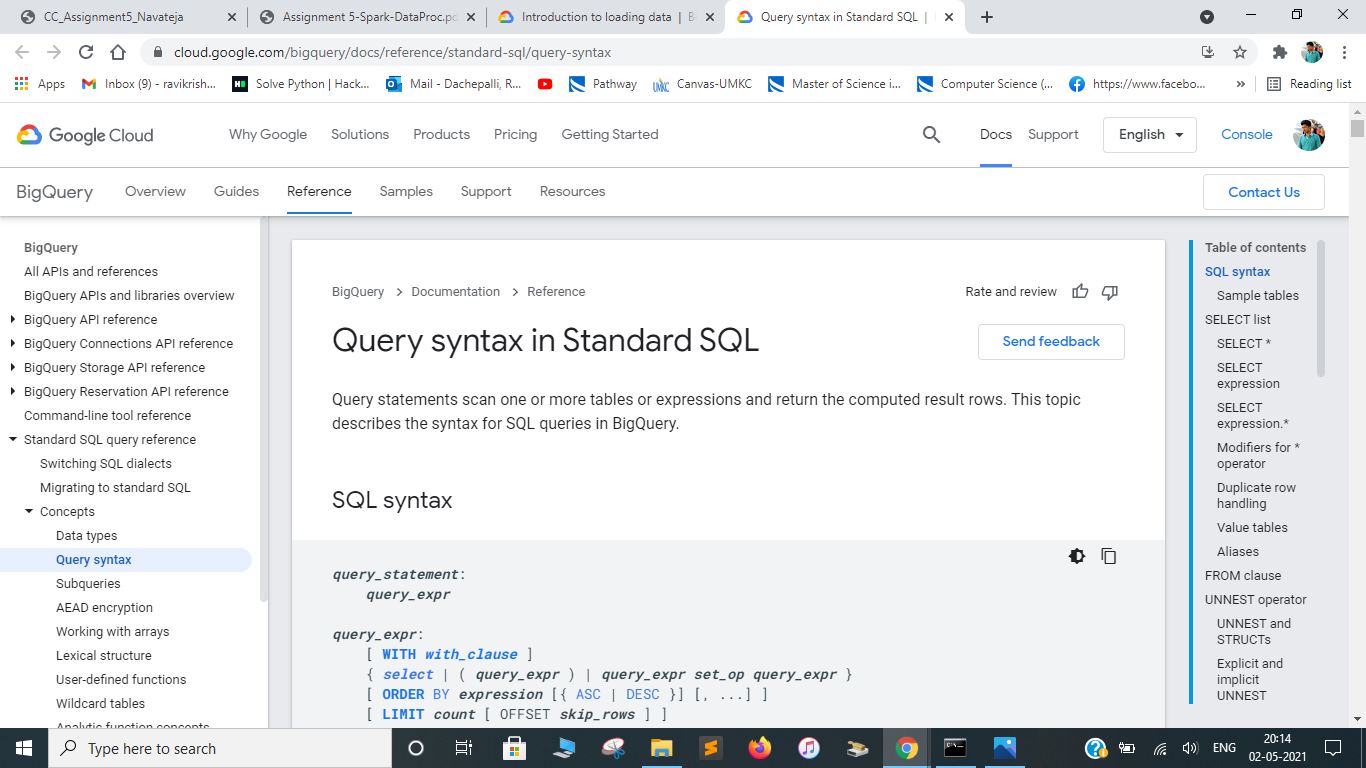
**Step-01:** Exploring BigQuery and learning how to use front-end UI to perform query.



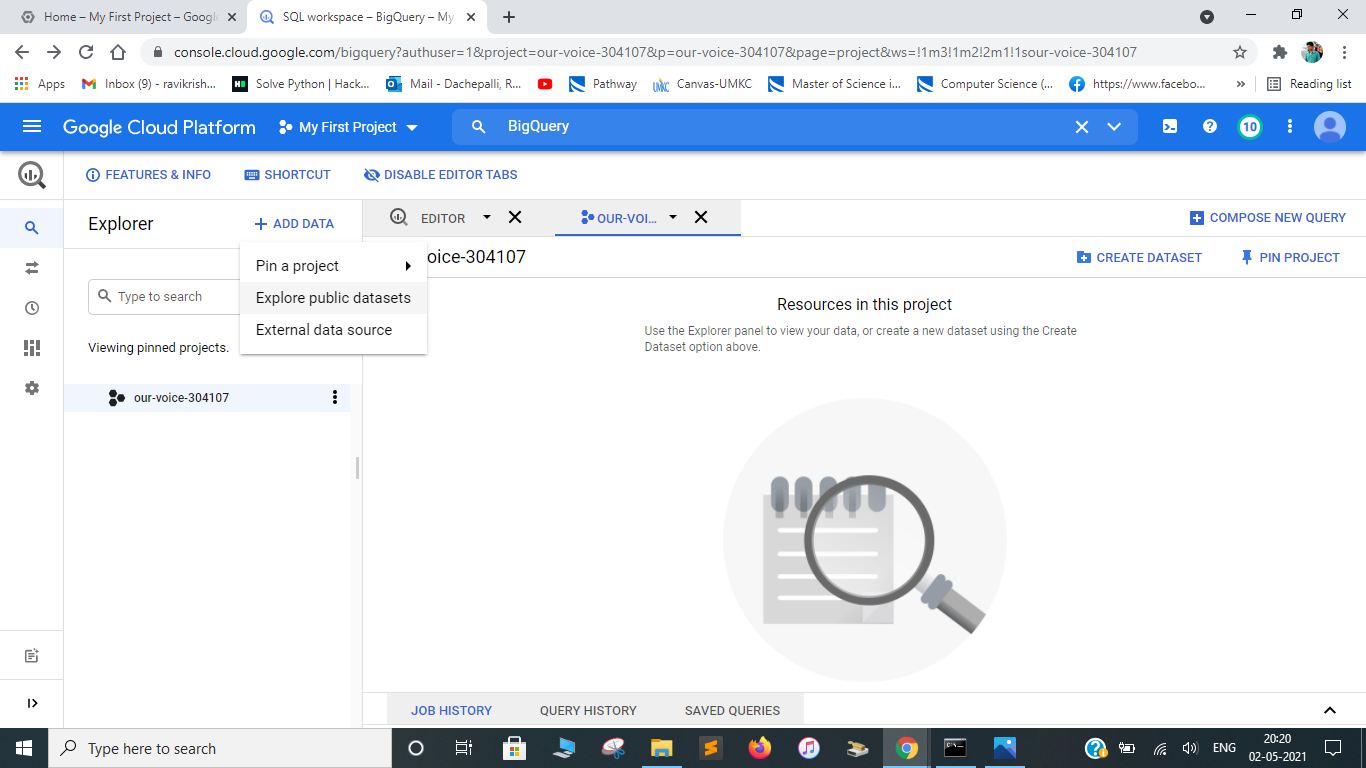
**Step-02:** Learning how to load data into BigQuery.

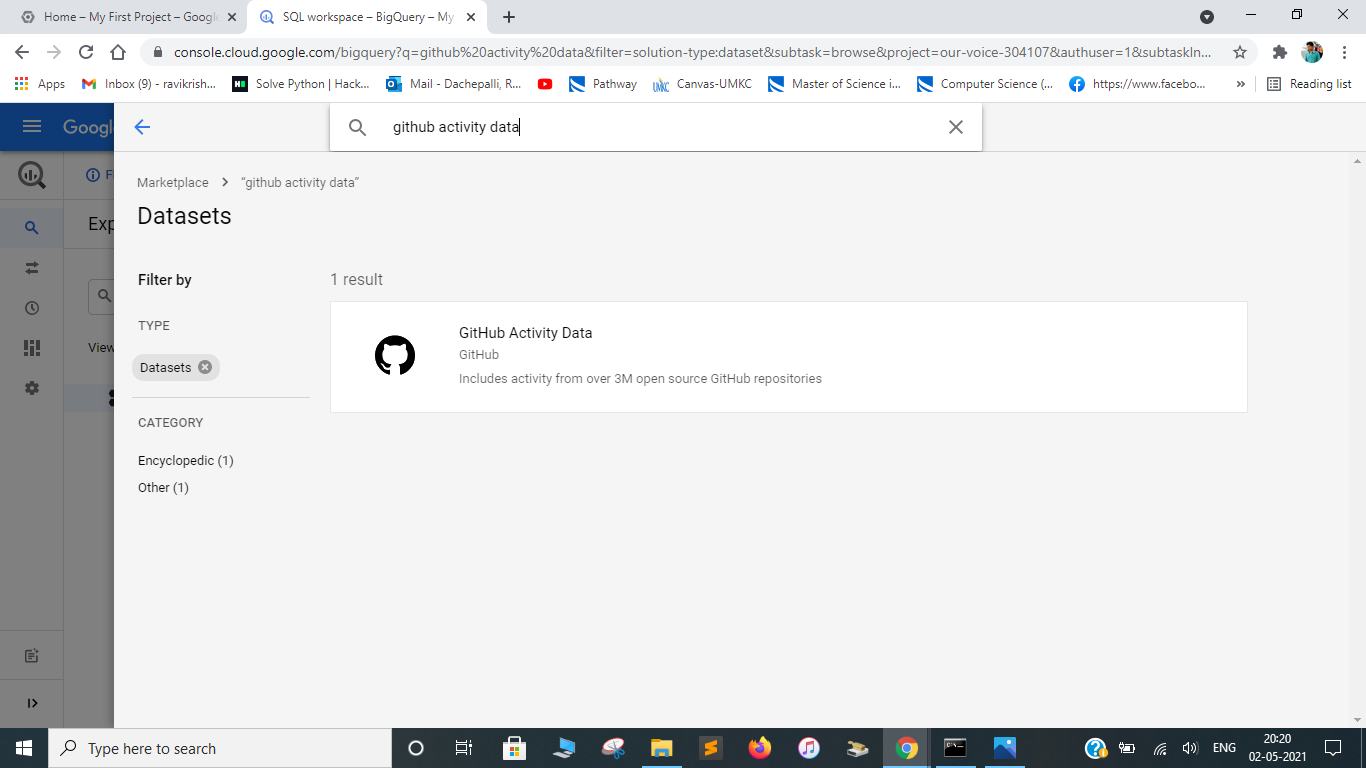


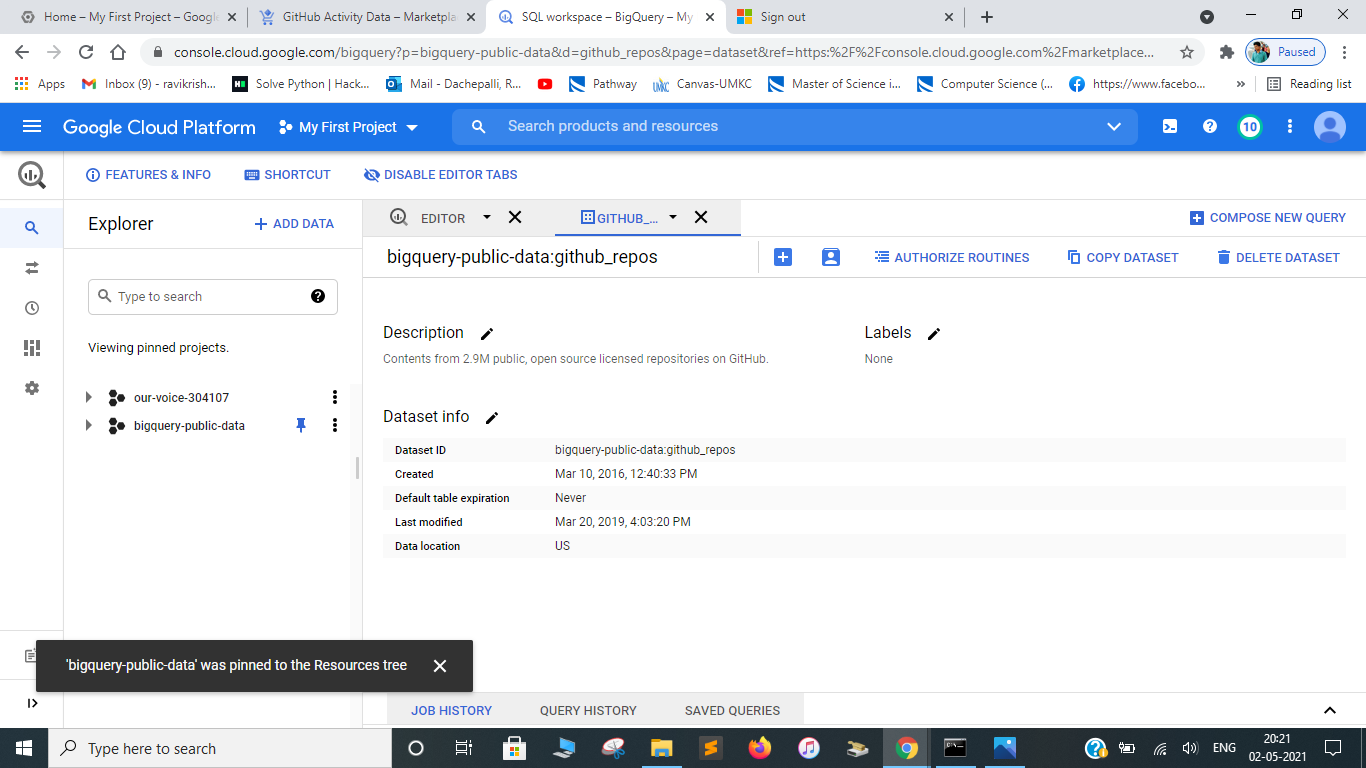
**Step-03:** Checking out the BigQuery SQL documentation as a reference for query syntax.

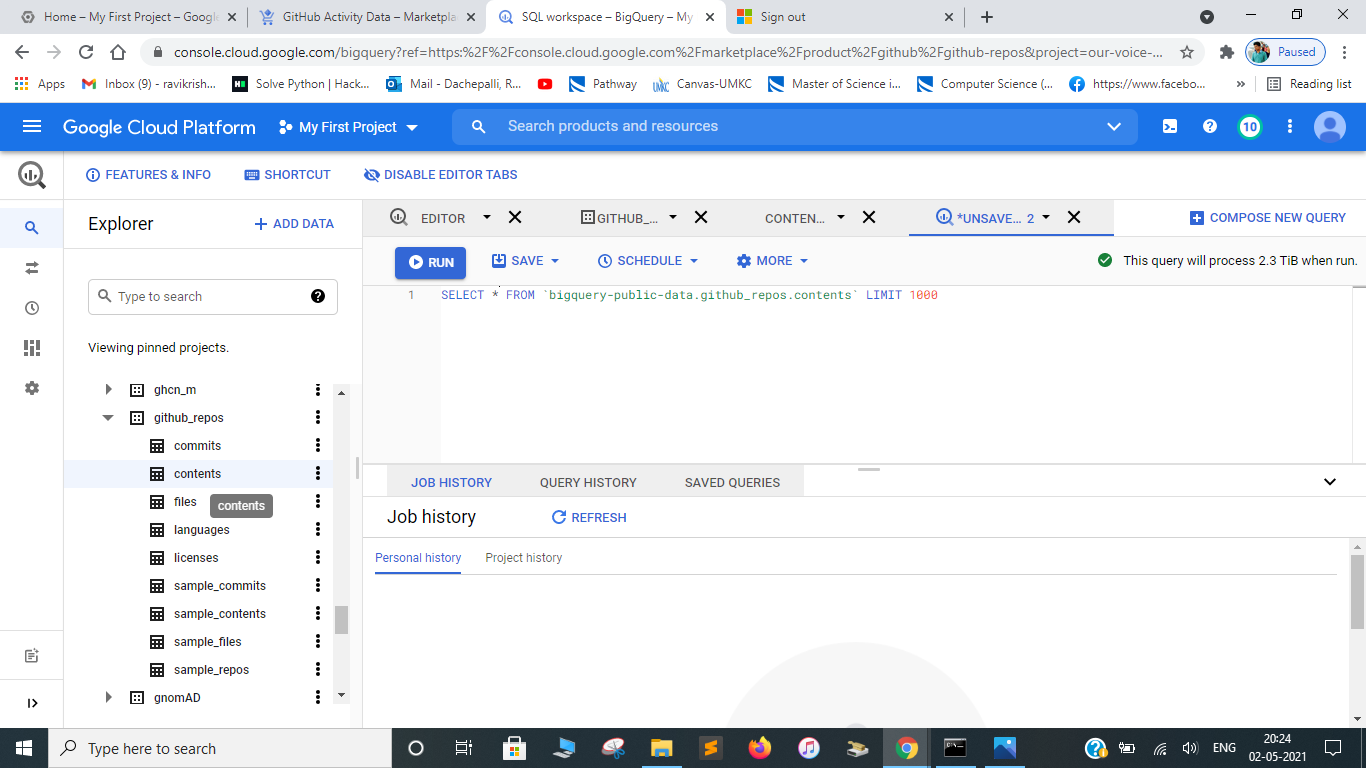


**Step-04:** Enabling BigQuery and loading github public data.



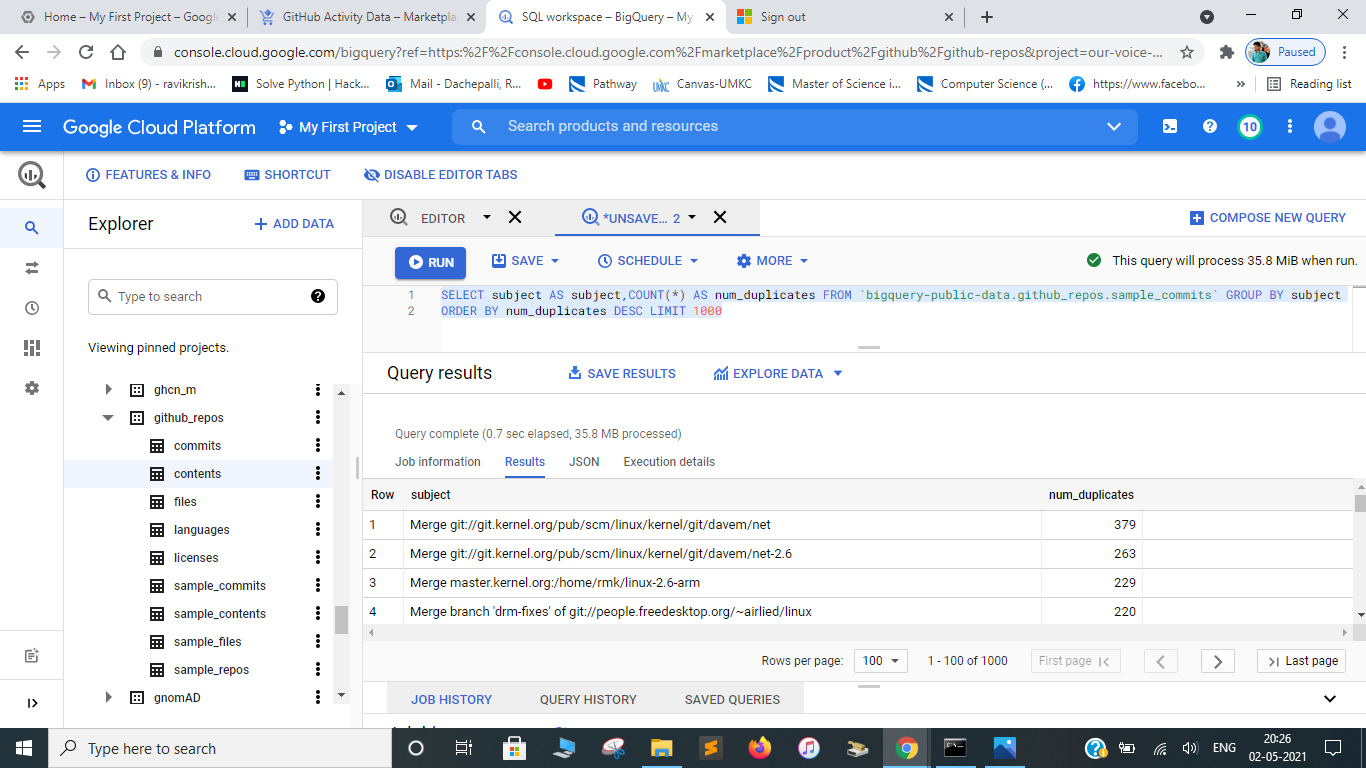






**Step-05:** Query github data through the following sql query and display the output.

**Query:** SELECT subject AS subject, COUNT(\*) AS num\_duplicates FROM `bigquery-public-data.github\_repos.sample\_commits` GROUP BY subject ORDER BY num\_duplicates DESC LIMIT 1000



**Step-06:** Delete the clusters create in dataproc to avoid charges and for cloud hygiene.

**Challenges:**

1. Got few errors while running the commands in SDK and rectified them after going through the gcp documentation.
2. A little bit confusion when loading the github public data in BigQuery.

**Fun Part:**

1. Going through documentation and experiencing the big data tools on cloud platform.
2. Gone through the optional resources provided and they are quite interesting where I have spent almost 3 hours reading them.