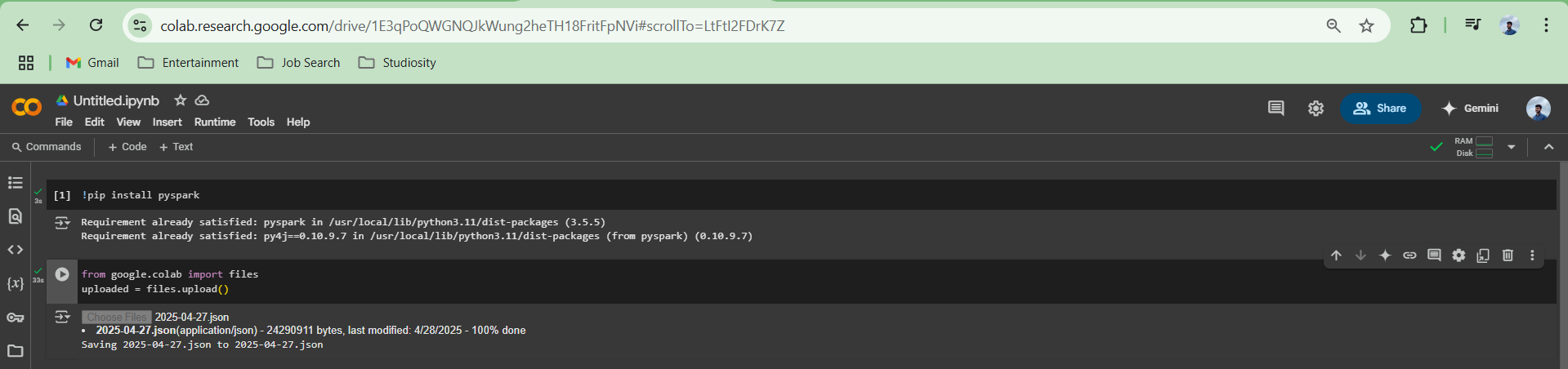
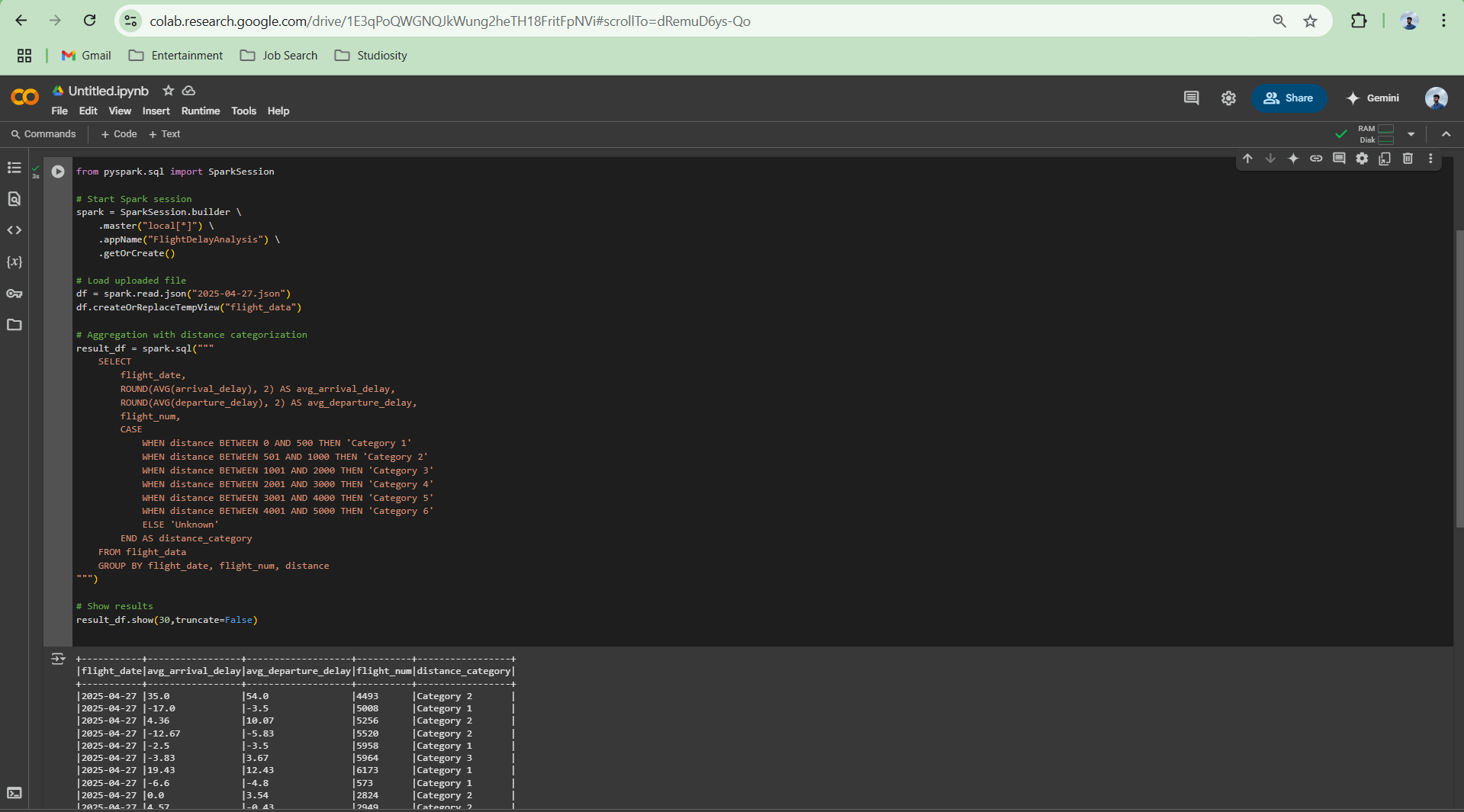
**Flight Delay Analysis**

**Using local in Colab**

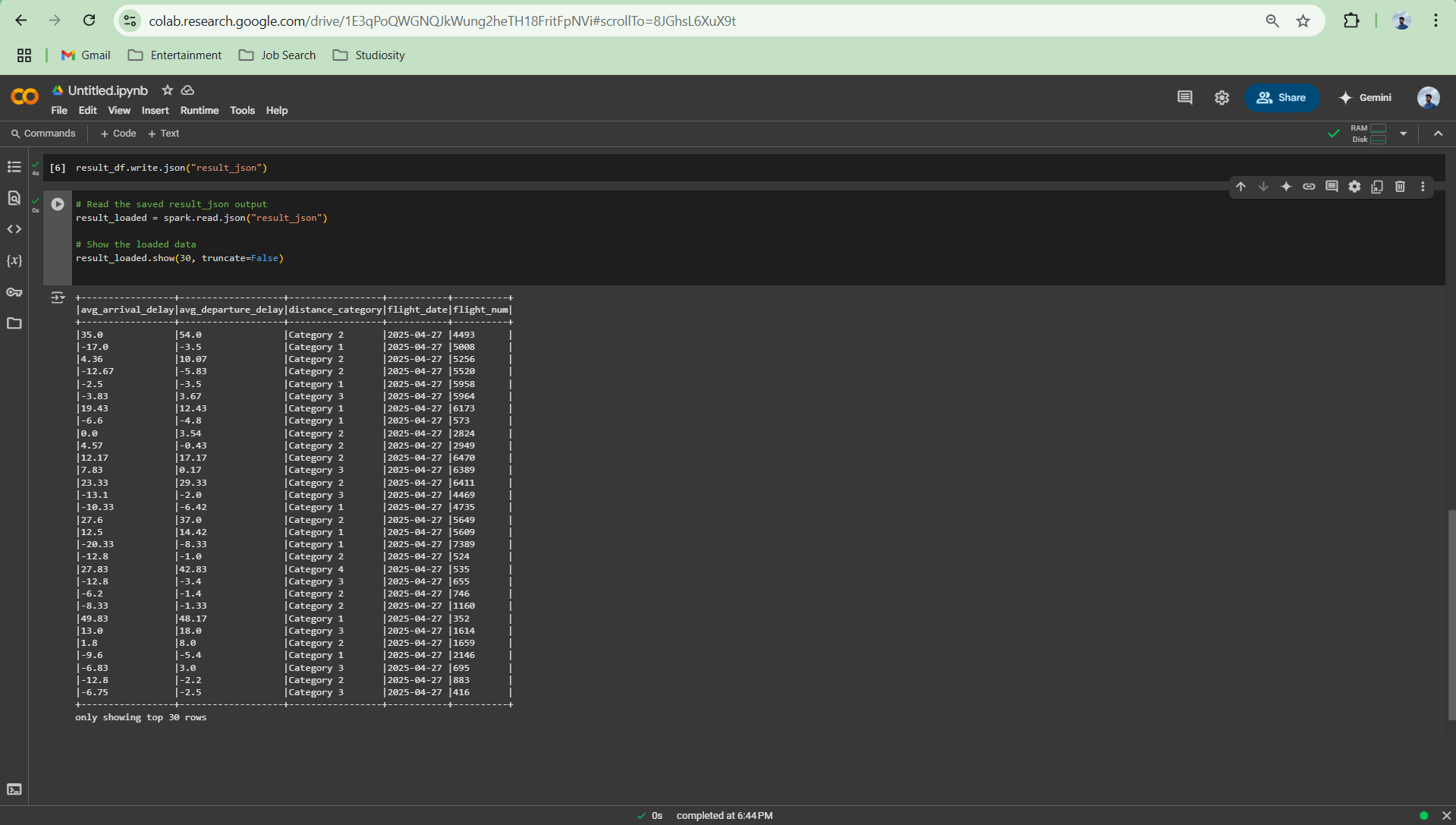
Install PySpark and upload json raw file



Now run the aggregation code, and you can see the output as shown below.



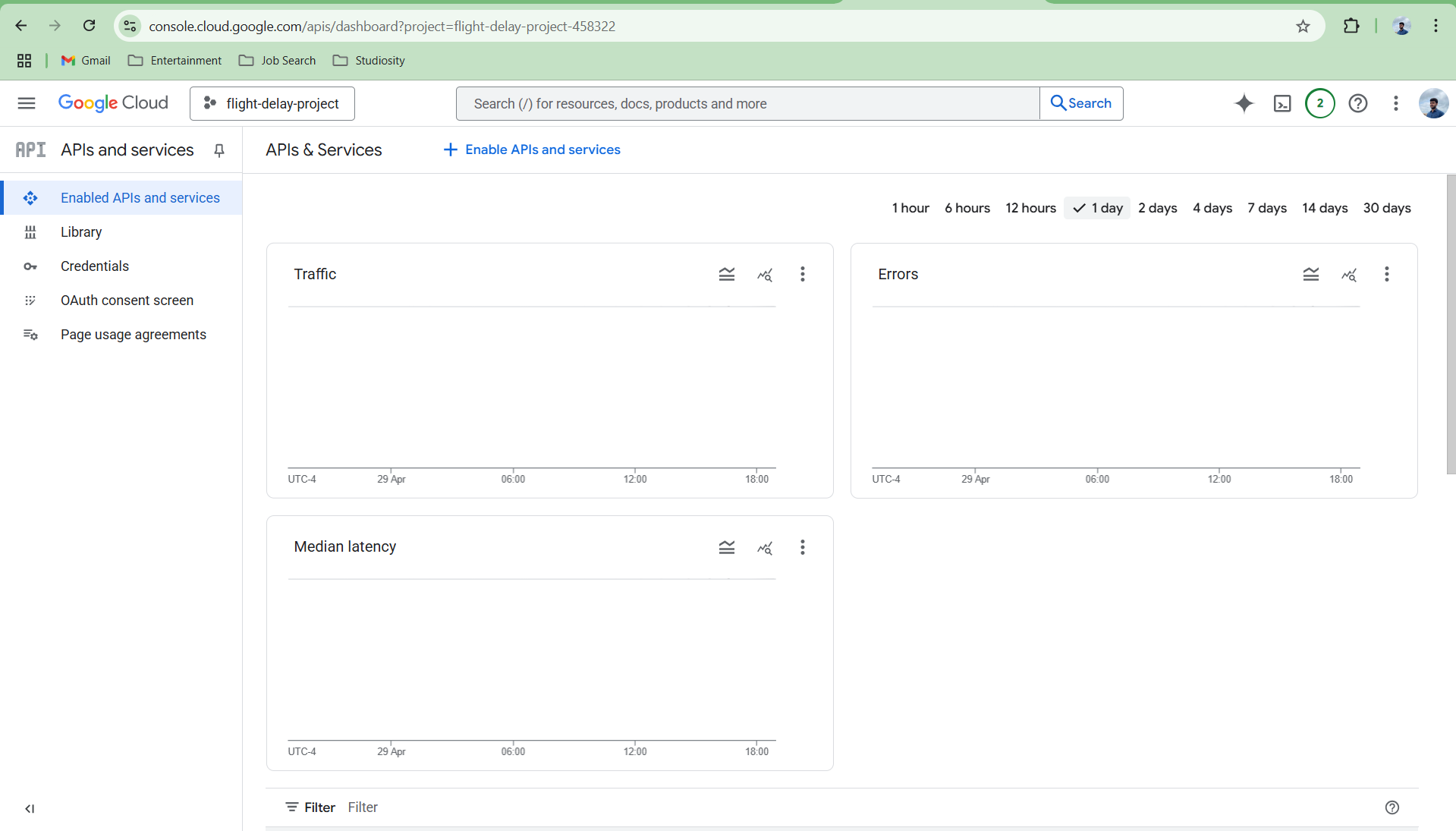
Saving the result and reading the result



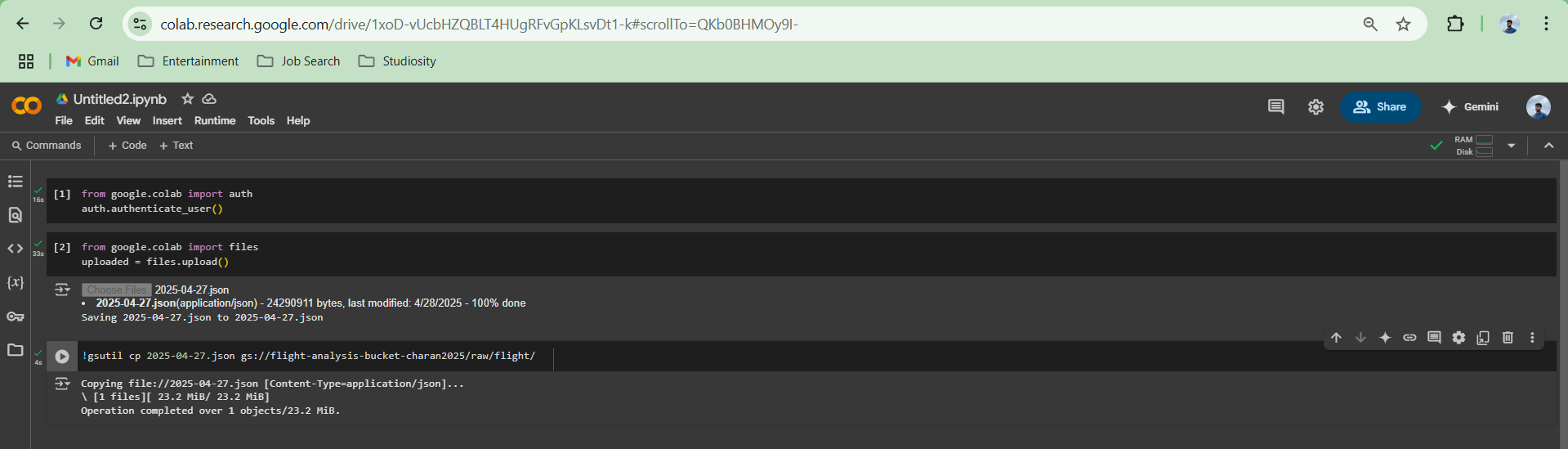
**Now let’s make the same using GCP Dataproc Adaption**

Create a project named flight-delay-project and enable

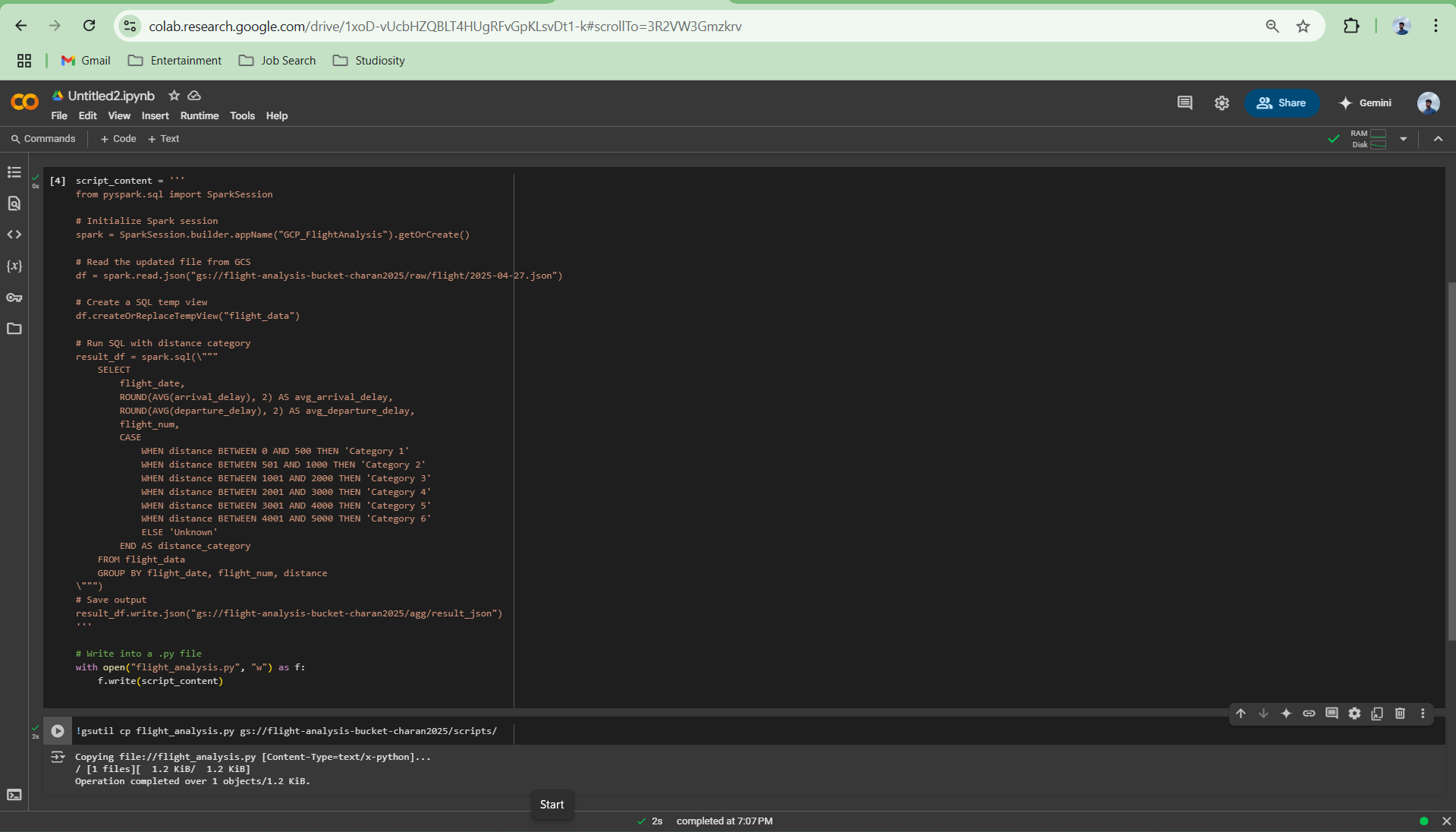
1. **Compute Engine API**
2. **Cloud Storage API**
3. **Dataproc API**

****

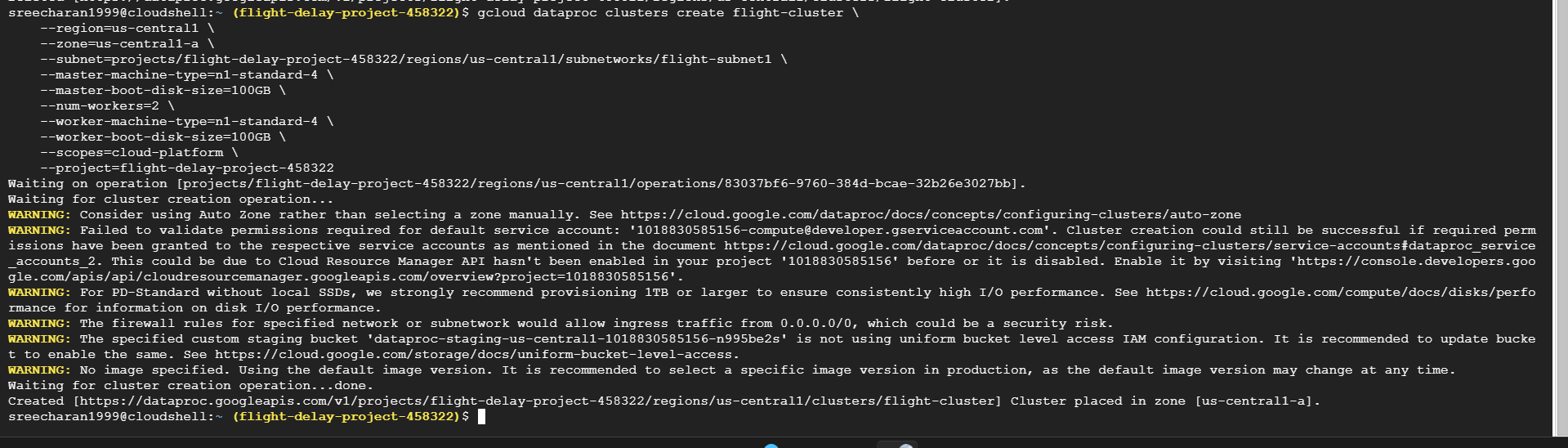
Create a bucket, now authenticate user, upload raw file and copy the file to bucket using colab as shown below



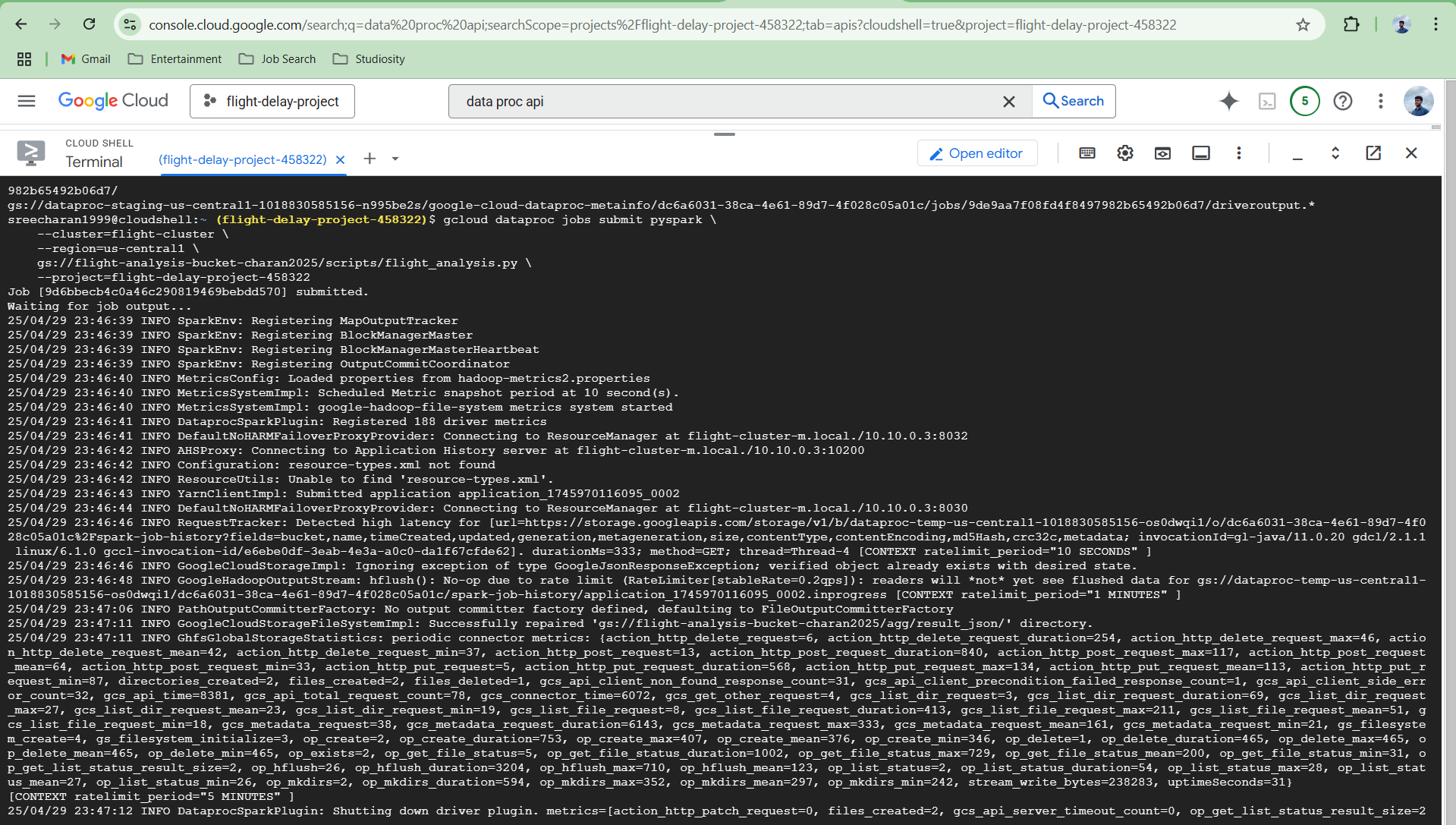
Create a PySpark Script and store in bucket



Create cluster



Submit the PySpark Job to Dataproc



You can see the result in bucket

