Adrien Protzel
2/26/2025

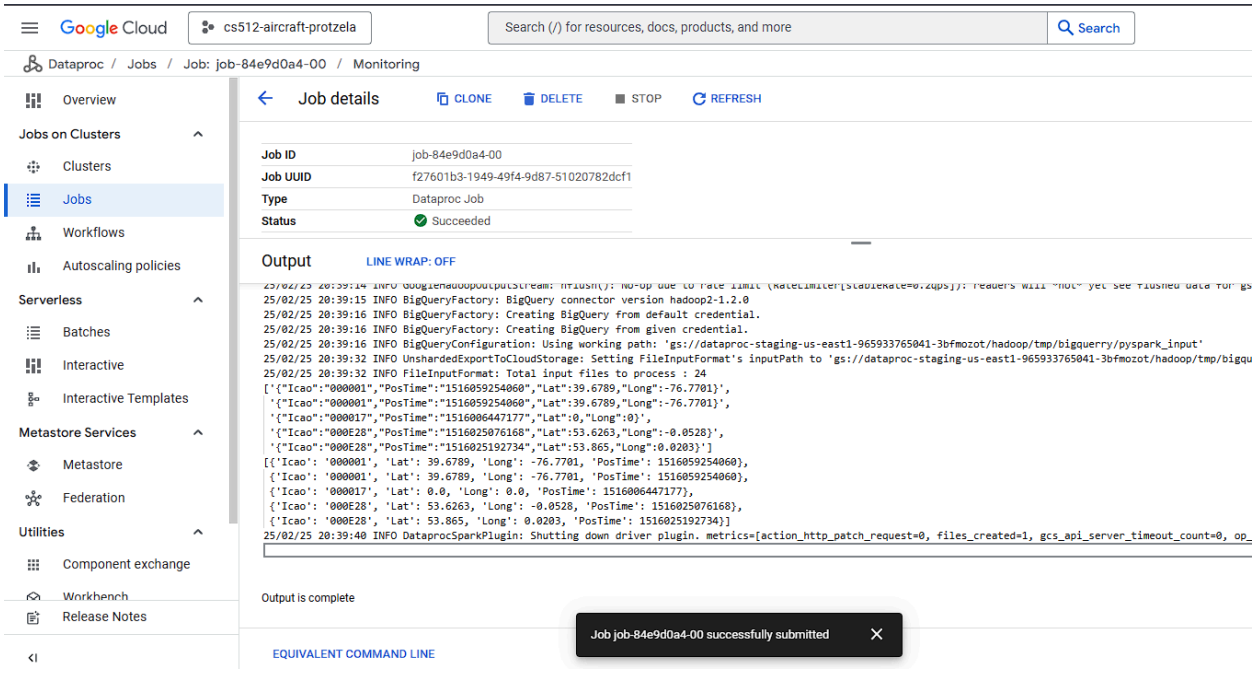
Group Homework: Spark Plane Distances Part 1

For this part of the assignment, I had to create a cluster on Dataproc. I set the region to us-east1-b and configured both the master and worker nodes to have 2 CPUs each. This setup resulted in the creation of multiple VM instances, which were utilized for running the jobs. I executed a job on the cluster using a modified Python script, where I removed all references to FSeen, and used the provided JAR file. Upon completion, the program generated the output, with the Lat and Long fields correctly converted to type Float, while PosTime was of type Long and ICAO was of type String.

In the screenshot below, you will notice that FSeen is not present, as I had removed it from the Python script. This modification was necessary because I encountered issues running the Dataprep recipe to generate a new CSV file or BigQuery dataset with the modified columns. Specifically, the process would unexpectedly error out during the Transform step. Despite this challenge, I have included a screenshot demonstrating my ability to edit the data to add the new PosTime and FSeen columns. This shows my understanding of the data transformation process, even though I faced technical difficulties with the Dataprep tool.

Used files:
gs://cs512-aircraft-protzela/Window_spark_planes_0.py
gs://cs512-aircraft-protzela/Window_spark_planes_1.py
gs://hadoop-lib/bigquery/bigquery-connector-hadoop2-latest.jar
wolford-cs512-aircraft-data/BQ_Table.csv

Screen snips of the DataProc output showing that you successfully ran the pyspark code.



Recipe to further edit data.

```
1 Delete rows where column1 contains 'src'
2 Create new columns from 5 constants in column1
3 Extract characters between 6 to 19 from FSeen
4 Delete column1, FSeen
5 Change FSeen1 type to Integer
6 Rename FSeen1 to 'Fseen'
7 Delete rows with missing values in Lat
```