

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.

The screenshot shows the Google Cloud Dataproc console interface. The left sidebar contains navigation links for Overview, Jobs on Clusters, Clusters, Jobs, Workflows, Autoscaling policies, Serverless, Batches, Interactive, Interactive Templates, Metastore Services, Metastore, Federation, Utilities, Component exchange, Workbench, and Release Notes. The main panel displays the 'Job details' for job-84e9d0a4-00, which is a Dataproc Job that has succeeded. The output section shows the execution of a Python script using BigQuery connector and Hadoop2, resulting in a successful completion. A toast notification at the bottom right confirms 'Job job-84e9d0a4-00 successfully submitted'.

Job ID	job-84e9d0a4-00
Job UUID	f27601b3-1949-49f4-9d87-51020782dcf1
Type	Dataproc Job
Status	Succeeded

```
25/02/25 20:39:14 INFO org.apache.hadoop.mapreduce.lib.input: Total input files to process : 24
25/02/25 20:39:15 INFO BigQueryFactory: BigQuery connector version hadoop2-1.2.0
25/02/25 20:39:16 INFO BigQueryFactory: Creating BigQuery from default credential.
25/02/25 20:39:16 INFO BigQueryFactory: Using working path: 'gs://dataproc-staging-us-east1-965933765041-3bfmozot/hadoop/tmp/bigquery/pyspark_input'
25/02/25 20:39:32 INFO UnshardedExportToCloudStorage: Setting FileInputFormat's inputPath to 'gs://dataproc-staging-us-east1-965933765041-3bfmozot/hadoop/tmp/bigquery/pyspark_input'
25/02/25 20:39:32 INFO FileInputFormat: Total input files to process : 24
[{"Icao": "000001", "PosTime": "1516059254060", "Lat": 39.6789, "Long": -76.7701}, {"Icao": "000001", "PosTime": "1516059254060", "Lat": 39.6789, "Long": -76.7701}, {"Icao": "000017", "PosTime": "1516006447177", "Lat": 0.0, "Long": 0.0}, {"Icao": "000017", "PosTime": "1516006447177", "Lat": 0.0, "Long": 0.0}, {"Icao": "000028", "PosTime": "1516025076168", "Lat": 53.6263, "Long": -0.0528}, {"Icao": "000028", "PosTime": "1516025076168", "Lat": 53.6263, "Long": -0.0528}, {"Icao": "000001", "Lat": 39.6789, "Long": -76.7701, "PosTime": 1516059254060}, {"Icao": "000001", "Lat": 39.6789, "Long": -76.7701, "PosTime": 1516059254060}, {"Icao": "000017", "Lat": 0.0, "Long": 0.0, "PosTime": 1516006447177}, {"Icao": "000017", "Lat": 0.0, "Long": 0.0, "PosTime": 1516006447177}, {"Icao": "000028", "Lat": 53.6263, "Long": -0.0528, "PosTime": 1516025076168}, {"Icao": "000028", "Lat": 53.6263, "Long": -0.0528, "PosTime": 1516025076168}
25/02/25 20:39:40 INFO DataprocSparkPlugin: Shutting down driver plugin. metrics={action_http_patch_request=0, files_created=1, gcs_api_server_timeout_count=0, op...
```

Output is complete

EQUIVALENT COMMAND LINE

Job job-84e9d0a4-00 successfully submitted

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