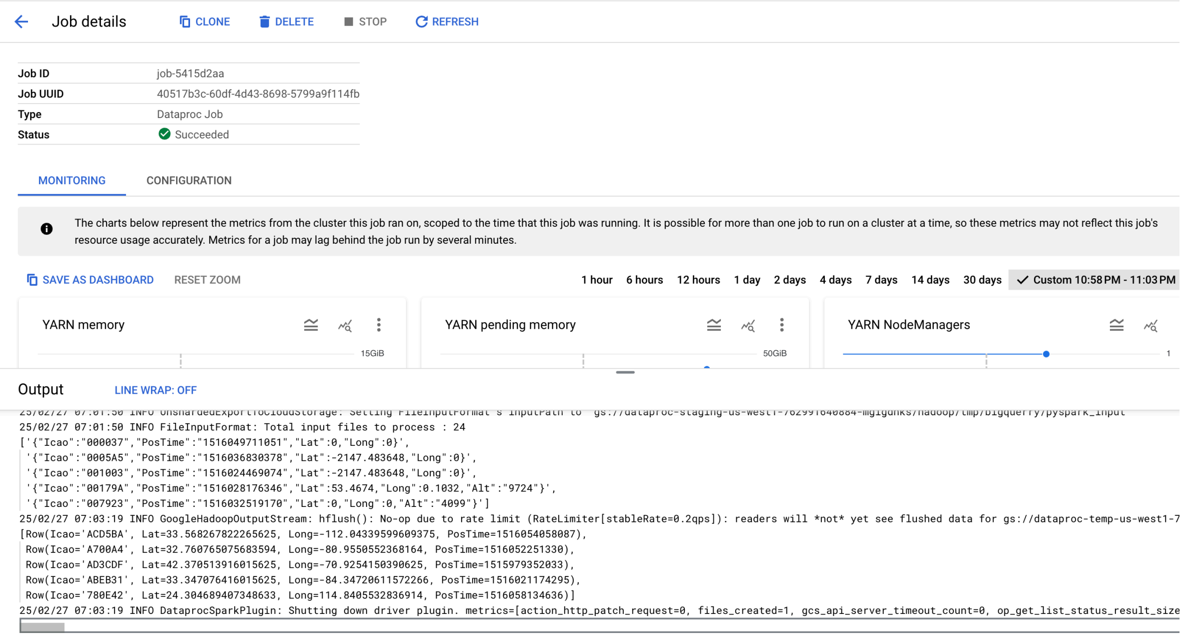
Spark Plane Distances Part 1

Brian Cervantes Alvarez
February 26, 2025

Working on this assignment was both challenging and satisfying. First, I set up a DataProc cluster on Google Cloud and configured my PySpark code so it could read data from BigQuery. I ran into a big problem called a "Conflict occurred creating export directory," which happened because an old folder in Google Cloud Storage was still there. This made Spark fail when trying to export data from BigQuery. After searching online and testing different fixes, I solved the issue by deleting the old folder before I tried to export again. That way, the folder was empty and ready to be used. Also, I started out trying to use the **FSeen** column in the plane data, but realized it wasn't needed for my main task. By removing it, my code ran smoothly without KeyError messages.

Once everything was set up, the DataProc job started up its worker nodes, accessed my BigQuery data, and showed me the first few records so I knew it worked. If you look at the screenshot, you'll see proof of my PySpark script running successfully on DataProc—no more errors this time. Along the way, I got valuable practice using Google Cloud Platform services—DataProc, GCS, and BigQuery—and I learned how important it is to carefully manage export folders so things don't crash or get stuck.



Job ID	job-5415d2aa
Job UUID	40517b3c-60df-4d43-8698-5799a9f114fb
Type	Dataproc Job
Status	✓ Succeeded

The charts below represent the metrics from the cluster this job ran on, scoped to the time that this job was running. It is possible for more than one job to run on a cluster at a time, so these metrics may not reflect this job's resource usage accurately. Metrics for a job may lag behind the job run by several minutes.

```
25/02/27 07:01:50 INFO UnshardedExportToCloudStorage: Setting FileInputFormat's inputPath to 'gs://dataproc-staging-us-west1-762991640884-mg1gdnks/hadoop/tmp/bigquerry/pyspark_input
25/02/27 07:01:50 INFO FileInputFormat: Total input files to process : 24
['{"Icao":"000037","PosTime":"1516049711051","Lat":0,"Long":0}',
 '{"Icao":"0005A5","PosTime":"1516036830378","Lat":-2147.483648,"Long":0}',
 '{"Icao":"001003","PosTime":"1516024469074","Lat":-2147.483648,"Long":0}',
 '{"Icao":"00179A","PosTime":"1516028176346","Lat":53.4674,"Long":0.1032,"Alt":"9724"}',
 '{"Icao":"007923","PosTime":"1516032519170","Lat":0,"Long":0,"Alt":"4099"}']
25/02/27 07:03:19 INFO GoogleHadoopOutputStream: hflush(): No-op due to rate limit (RateLimiter[stableRate=0.2qps]): readers will *not* yet see flushed data for gs://dataproc-temp-us-west1-7
[Row(Icao='ACD5BA', Lat=33.568267822265625, Long=-112.04339599609375, PosTime=1516054058087),
 Row(Icao='A700A4', Lat=32.760765075683594, Long=-80.9550552368164, PosTime=1516052251330),
 Row(Icao='AD3CDF', Lat=42.370513916015625, Long=-70.9254150390625, PosTime=1515979352033),
 Row(Icao='ABEB31', Lat=33.347076416015625, Long=-84.34720611572266, PosTime=1516021174295),
 Row(Icao='780E42', Lat=24.304689407348633, Long=114.8405532836914, PosTime=1516058134636)]
25/02/27 07:03:19 INFO DataprocSparkPlugin: Shutting down driver plugin. metrics=[action_http_patch_request=0, files_created=1, gcs_api_server_timeout_count=0, op_get_list_status_result_size
```