Adrien Protzel
3/5/2025

Group Homework: Spark Plane Distances Part 2

For this part of the assignment, I created a Dataproc cluster in the us-east1-b region, configuring both master and worker nodes with 2 CPUs each. I executed a job using a modified Python script, removing references to FSeen, and utilized the provided JAR file. The output had Lat and Long fields converted to Float, PosTime as Long, and ICAO as String. Despite encountering issues with the Dataprep recipe during the Transform step, I demonstrated my ability to edit the data by adding new PosTime and FSeen columns.

Additionally, I set up a compute engine for a cloud VM, created an instance in the west Oregon region, and uploaded the necessary files to a cloud bucket. I used Dataprep to import and prepare the data, although I faced challenges exporting it. I then created a new cluster with specific configurations for the manager and worker nodes, cleaned the data in BigQuery, and submitted the job using the main Python script and the required JAR file.

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. Despite this challenge, I have included a screenshot demonstrating my ability to edit the data to add the new PosTime and FSeen columns, showcasing my understanding of the data transformation process.

Part 2 Detailed Steps:

1. Set up compute engine - for cloud vm instead of local
    a. Create instance
    b. Region - west Oregon
    c. Set memory to 100GB
2. Connect to VM
    a. Click SSH
    b. Make directory: mkdir plane_data
    c. cd plane_data
    d. sudo apt-get install wget
    e. sudo apt-get install unzip
    f. wget https://web.engr.oregonstate.edu/~wolfordj/plane_data.zip
    g. unzip <tab>
3. Upload files to cloud bucket
    a. Cloud storage > create bucket
        i. cs512_aircraft
        ii. <Change nothing>
    b. <in SSH window>
    c. gcloud init
    d. Create new account: 2

        i. Copy link
        ii. Copy key code
        iii. Create project (or select project)
        iv. Move zip up one directory: mv plane_data.zip ../
        v. cd ..
        vi. gsutil -m cp -r plane_data/ gs://cs512-aircraft-protzela

4. Load data on dataprep
    a. Open dataprep
    b. Import data
        i. Google cloud
        ii. Select plane_data folder
        iii. Add description
            1. If import button does not show, click continue
            2. Remove structure of imported data folder
            3. Use in new flow
            4. Edit recipe to break on ‘}, ‘
            5. Add step to add suffix } to column 1
        iv. import
    c. Add recipe steps, ‘filter contains’ out data

<Note, the data from above steps is not used due to DataPrep not being able to export data>

5. Edited provided py file to remove reference to FSeen
    a. Uploaded data to flight data project bucket
6. Created new cluster
    a. Region us-east1 and zone us-east1-b
    b. Manager Node
        i. Series N2, machine type n2-standard-2 (2 vCPU, 1 core, 8 GB memory)
        ii. Primary disk size = 100GB
        iii. Primary disk type = Balanced Persistent Disk
        iv. Number of local SSDs = 0 x 375GB
        v. Local SSD Interface = SCSI
    c. Worker Node
        i. Series N2
        ii. Machine type = n2-standard-2 (2 vCPU, 1 core, 8 GB memory)
        iii. Number of worker nodes = 2
        iv. Primary disk size = 200GB
        v. Primary disk type = Balanced Persistent Disk
        vi. Number of local SSDs = 0x 375GB
        vii. Local SSD Interface = SCSI
7. <Data is already stored in BigQuery from previous assignment (No FSeen)>
8. Add Query to data

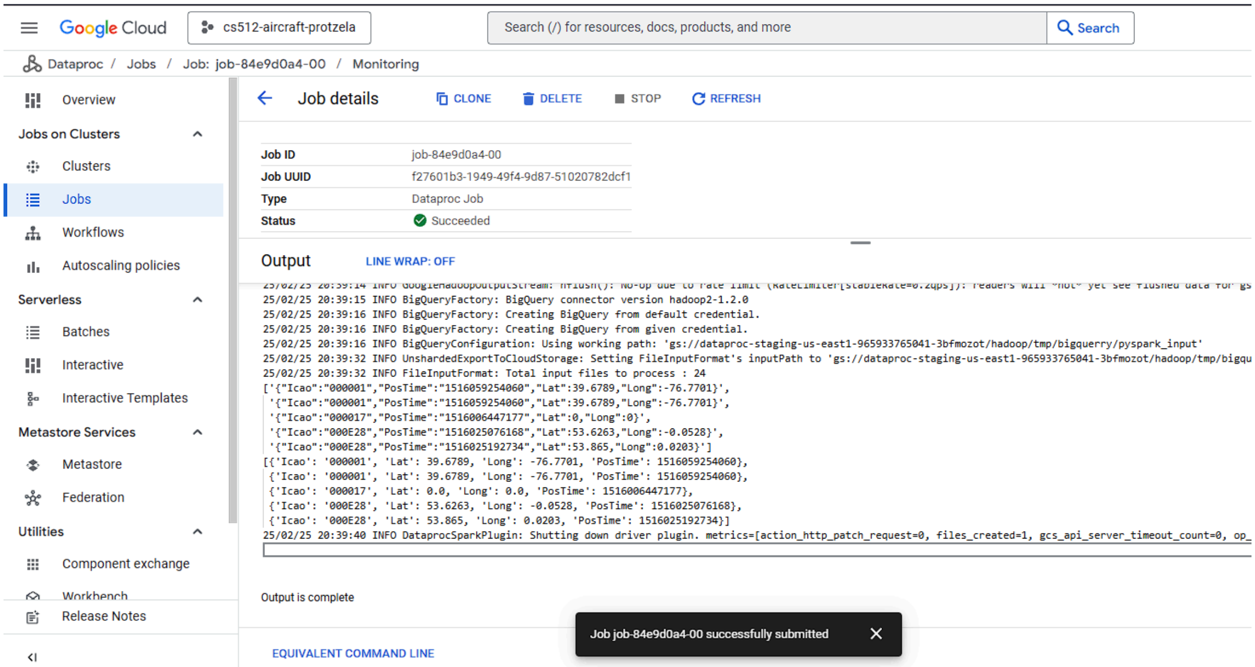
```
DELETE FROM aircraft_data.plane_loc
WHERE Long = 0
  OR Lat = 0
  OR Icao IS NULL
  OR Lat IS NULL
  OR Long IS NULL
```

OR PosTime IS NULL
OR Alt IS NULL
OR Lat NOT BETWEEN -90 AND 90
OR Long NOT BETWEEN -180 AND 180
OR Alt NOT BETWEEN 30000 AND 45000;

9. Submit job
    a. Main Python: gs://cs512-aircraft-protzela/Window_spark_planes_Solution-1.py
    b. Jar: gs://hadoop-lib/bigquery/bigquery-connector-hadoop2-latest.jar

Used files:
gs://cs512-aircraft-protzela/Window_spark_planes_Solution-1.py
gs://hadoop-lib/bigquery/bigquery-connector-hadoop2-latest.jar

Part 1: Screen snips of the DataProc output showing that you successfully ran the pyspark.



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

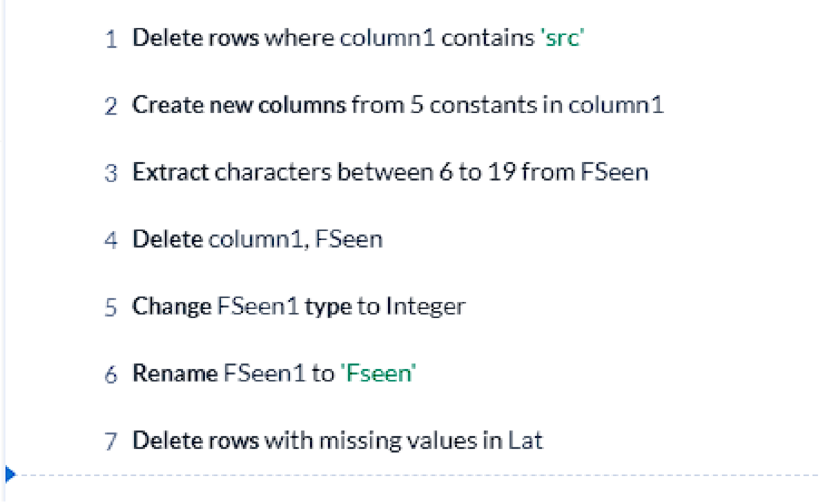
Output

```
25/02/25 20:39:14 INFO GoogleHadoopOutputStream: hflush(): No-op due to rate limit (RateLimiter[stableRate=0.2qps]): readers will *not* yet see flushed data for gs
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: Creating BigQuery from given credential.
25/02/25 20:39:16 INFO BigQueryConfiguration: Using working path: 'gs://dataproc-staging-us-east1-965933765041-3bfmozot/hadoop/tmp/bigquerry/pyspark_input'
25/02/25 20:39:32 INFO UnshardedExportToCloudStorage: Setting FileInputFormat's inputPath to 'gs://dataproc-staging-us-east1-965933765041-3bfmozot/hadoop/tmp/bigqu
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,"Long":0}',
 '{"Icao":"000E28","PosTime":"1516025076168","Lat":53.6263,"Long":-0.0528}',
 '{"Icao":"000E28","PosTime":"1516025192734","Lat":53.865,"Long":0.0203}']
[{'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': '000E28', 'Lat': 53.6263, 'Long': -0.0528, 'PosTime': 1516025076168},
 {'Icao': '000E28', 'Lat': 53.865, 'Long': 0.0203, 'PosTime': 1516025192734}]
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

Job job-84e9d0a4-00 successfully submitted

Part 1: Recipe to further edit data, include Long and Lat, FSeen.



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

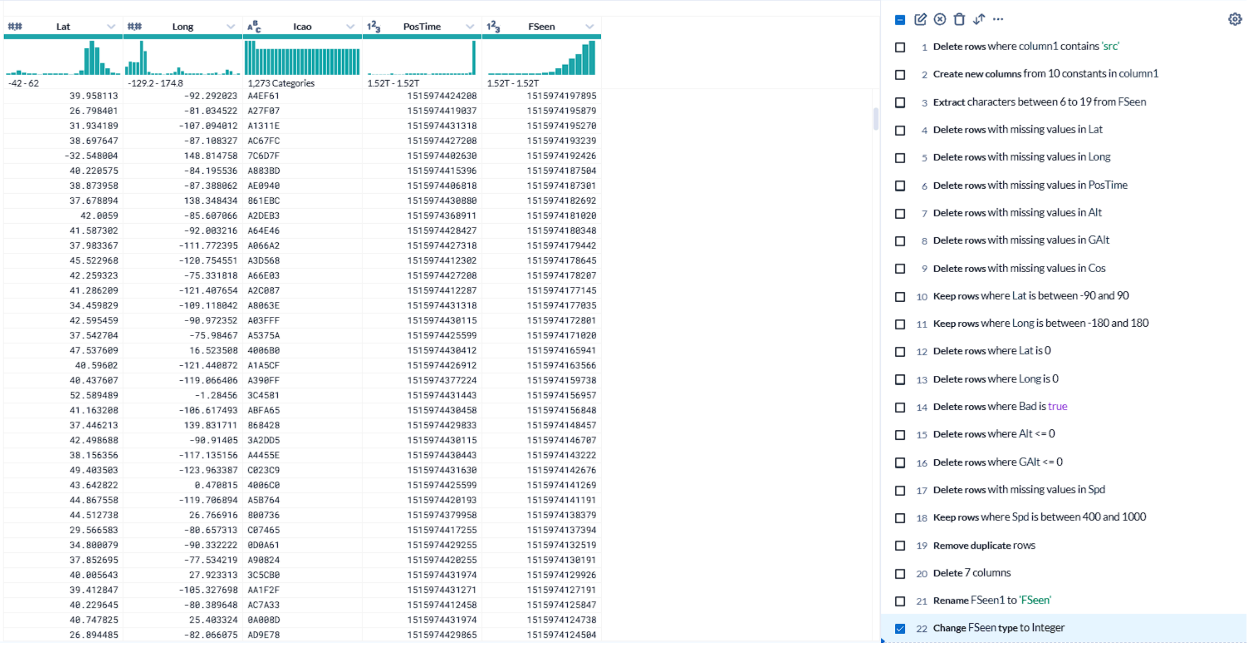
4 Delete column1, FSeen

5 Change FSeen1 type to Integer

6 Rename FSeen1 to 'Fseen'

7 Delete rows with missing values in Lat

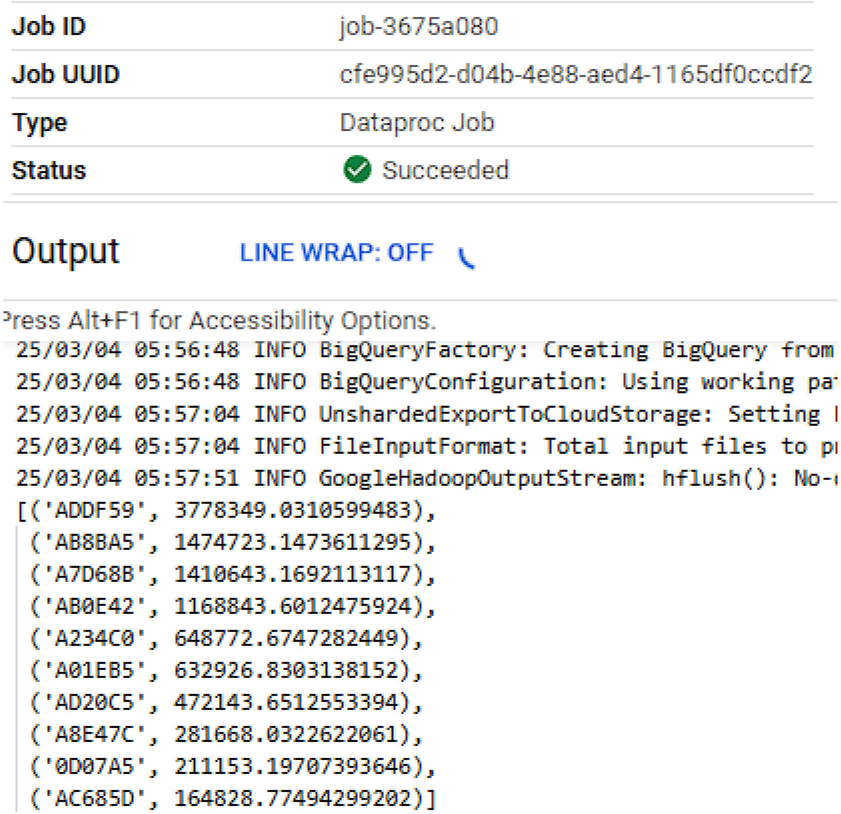
Part 2: DataPrep Recipe Filtering



Lat	Long	Icao	PosTime	FSeen
-42 - 62	-129.2 - 174.8	1,273 Categories	1.52T - 1.52T	1.52T - 1.52T
39.958113	-92.292023	A4EF61	1515974424208	1515974197895
26.798401	-81.034522	A27F07	1515974419037	1515974195879
31.934189	-107.094012	A1311E	1515974431318	1515974195270
38.697647	-87.108327	AC67FC	1515974427208	1515974193239
-32.548004	148.814758	7C6D7F	1515974402630	1515974192426
40.220575	-84.195536	A883BD	1515974415396	1515974187504
38.873958	-87.388062	AE0940	1515974406818	1515974187301
37.678894	138.348434	861EBC	1515974430880	1515974182692
42.0059	-85.607066	A2DEB3	1515974368911	1515974181020
41.587302	-92.003216	A64E46	1515974428427	1515974180348
37.983367	-111.772395	A066A2	1515974427318	1515974179442
45.522968	-120.754551	A3D568	1515974412302	1515974178645
42.259323	-75.331818	A66E03	1515974427208	1515974178207
41.286209	-121.407654	A2C087	1515974412287	1515974177145
34.459829	-109.118042	A8063E	1515974431318	1515974177035
42.595459	-90.972352	A03FFF	1515974430115	1515974172801
37.542704	-75.98467	A5375A	1515974425599	1515974171020
47.537609	16.523508	4006B0	1515974430412	1515974165941
40.59602	-121.440872	A1A5CF	1515974426912	1515974163566
40.437607	-119.066406	A390FF	1515974377224	1515974159738
52.589489	-1.28456	3C4581	1515974431443	1515974156957
41.163208	-106.617493	ABFA65	1515974430458	1515974156848
37.446213	139.831711	868428	1515974429833	1515974148457
42.498688	-90.91405	3A2DD5	1515974430115	1515974146707
38.156356	-117.135156	A4455E	1515974430443	1515974143222
49.403503	-123.963387	C023C9	1515974431630	1515974142676
43.642822	0.470815	4006C0	1515974425599	1515974141269
44.867558	-119.706894	A5B764	1515974420193	1515974141191
44.512738	26.766916	800736	1515974379958	1515974138379
29.566583	-80.657313	C07465	1515974417255	1515974137394
34.800079	-90.332222	0D0A61	1515974429255	1515974132519
37.852695	-77.534219	A90824	1515974420255	1515974130191
40.005643	27.923313	3C5CB0	1515974431974	1515974129926
39.412847	-105.327698	AA1F2F	1515974431271	1515974127191
40.229645	-80.389648	AC7A33	1515974412458	1515974125847
40.747825	25.403324	0A008D	1515974431974	1515974124738
26.894485	-82.066075	AD9E78	1515974429865	1515974124504

☐ 1 **Delete rows** where column1 contains 'src'
☐ 2 **Create new columns** from 10 constants in column1
☐ 3 **Extract** characters between 6 to 19 from FSeen
☐ 4 **Delete rows** with missing values in Lat
☐ 5 **Delete rows** with missing values in Long
☐ 6 **Delete rows** with missing values in PosTime
☐ 7 **Delete rows** with missing values in Alt
☐ 8 **Delete rows** with missing values in GAlt
☐ 9 **Delete rows** with missing values in Cos
☐ 10 **Keep rows** where Lat is between -90 and 90
☐ 11 **Keep rows** where Long is between -180 and 180
☐ 12 **Delete rows** where Lat is 0
☐ 13 **Delete rows** where Long is 0
☐ 14 **Delete rows** where Bad is true
☐ 15 **Delete rows** where Alt <= 0
☐ 16 **Delete rows** where GAlt <= 0
☐ 17 **Delete rows** with missing values in Spd
☐ 18 **Keep rows** where Spd is between 400 and 1000
☐ 19 **Remove duplicate** rows
☐ 20 **Delete** 7 columns
☐ 21 **Rename** FSeen1 to 'FSeen'
☑ 22 **Change** FSeen **type** to Integer

Part 2: Dataproc output



Job ID	job-3675a080
Job UUID	cfe995d2-d04b-4e88-aed4-1165df0ccdf2
Type	Dataproc Job
Status	✓ Succeeded

Output

```
ress Alt+F1 for Accessibility Options.
25/03/04 05:56:48 INFO BigQueryFactory: Creating BigQuery from
25/03/04 05:56:48 INFO BigQueryConfiguration: Using working pa
25/03/04 05:57:04 INFO UnshardedExportToCloudStorage: Setting 
25/03/04 05:57:04 INFO FileInputFormat: Total input files to p
25/03/04 05:57:51 INFO GoogleHadoopOutputStream: hflush(): No-
[('ADDF59', 3778349.0310599483),
 ('AB8BA5', 1474723.1473611295),
 ('A7D68B', 1410643.1692113117),
 ('AB0E42', 1168843.6012475924),
 ('A234C0', 648772.6747282449),
 ('A01EB5', 632926.8303138152),
 ('AD20C5', 472143.6512553394),
 ('A8E47C', 281668.0322622061),
 ('0D07A5', 211153.19707393646),
 ('AC685D', 164828.77494299202)]
```