

Spark Plane Distances Part 1 Report

Rahul Kumar Nalubandhu and Sandra Estrada

Rahul: During this assignment, I encountered several errors while working on the DataPrep stage. Initially, I used all the datasets and recipes from assignment 6 and made only two updates to the recipe by adding the FSeen and PosTime columns. However, when I ran the job, I received an error message stating that 'no files were found'. Despite attempting to resolve the issue with my partner, Sandra, we were unsuccessful in finding a solution.

After some debugging, I discovered that my storage settings in the DataPrep had been changed from the default setting to my current bucket. This change prevented the upload and running of jobs. Although I attempted to resolve the issue by referring to Google's documentation, none of the suggested solutions worked. Therefore, I decided to clear everything by deleting all the buckets, BigQuery, and instances to start from scratch. Despite trying various methods, I was unable to export data successfully after a job run. Upon reviewing the values, I found that they had been replaced, and there were numerous duplicate values. Additionally, the query value was different, making it impossible to continue. Therefore, I had no choice but to delete the current project and create a new one to start afresh.

After creating a new project, I followed all the steps from the previous assignment and exported the table to BigQuery. When I ran the query from the previous assignment, I obtained the expected value, which indicated that the data was being exported correctly. I then edited the recipe by adding the FSeen and PosTime columns and used delimiters to format the FSeen column to include only numerical values. After completing these steps, I ran the job again, and this time it was successful.

Rahul and Sandra: Once those issues were resolved, our experience with completing this assignment was consistent. We both created a cluster in DataProc as instructed in the Exploration: Hello Spark video 1. Next, we updated the plane data recipe to include PosTime and FSeen and reran the DataPrep job, which created a new table with the new run. In addition, we downloaded the python file provided, made changes to the code reflecting our project's specifications and uploaded the file to our cloud storage bucket.

Once we had completed this, we submitted a job in DataProc under the cluster we had created for PySpark with the python's file URI and the jar connector link provided. We both received a job failure error message and referenced ED Discussion for help in resolving this issue. Joshua Magana referenced to extract between delimiters, delete the column and rename the new one created. There were some concerns with whether that was to be completed in DataPrep or in the python file. After connecting with Rahul, he explained this was something that had to be completed in DataPrep during the recipe stage before running the job with the new columns. We made the necessary changes and reran the job in Dataproc. We got another error message and had to delete PySpark from the directory as instructed in the assignment. After deleting it from the directory, the job ran successfully.

We were both able to complete all the steps in the assignment. We completed the assignment at our own pace on Friday night and connected on Saturday afternoon via a Teams call to help each other overcome the issues we were running into.

Rahul's Screenshots:

The screenshot displays the Google Cloud Data Studio interface. On the left, a data table is shown with columns: Lat, Long, A_E, Icao, PosTime, and FSeen. The table contains 3,992 rows of data. On the right, a 'Recipe' panel is visible, showing a list of 11 steps for data transformation. The steps include deleting rows, replacing matches, changing column types, creating new columns, deleting columns, deleting rows with missing values, splitting FSeen on delimiters, deleting FSeen1, splitting FSeen2 on delimiters, deleting FSeen3, and renaming FSeen1 to 'FSeen'.

Lat	Long	A _E	Icao	PosTime	FSeen
-43.6 - 228.0	-129.2 - 174.8	2,477 Categories	152T - 152T	152T - 152T	
28.461868	-81.321804	A8754D	1515974431837	1515974487537	
-33.982818	151.177994	7C752D	1515974426333	1515974486333	
-33.948315	150.596537	7C671B	1515974430912	1515974399193	
33.953293	-117.66761	A3E6EA	1515974427998	1515974392349	
35.478249	140.393646	780A54	1515974431498	1515974391458	
41.328186	-106.958618	A88D47	1515974429837	1515974381899	
34.7953	135.3871	862226	1515974423693	1515974388536	
37.354523	-121.998434	A100E6	1515974430638	1515974379998	
37.039855	-122.889393	A8E35D	1515974431498	1515974379521	
51.689238	-114.622368	C888C0	1515974481787	1515974377852	
36.857437	148.871472	861886	1515974423365	1515974374849	
58.72197	25.617599	40881B	1515974427208	1515974374882	
49.836358	-1.715168	486441	1515974426583	1515974372146	
-33.875702	151.27272	C88804	1515974432177	1515974371788	
41.946888	-87.825581	A26793	1515974399521	1515974378693	
-43.447586	172.574456	C8173A	1515974438646	1515974369333	
33.434189	-111.917778	A7668A	1515974431474	1515974367536	
49.169769	-123.388856	C80108	1515974431412	1515974365568	
38.362628	-88.475391	A89E27	1515974398162	1515974365885	
33.398839	-116.765771	A408E2	1515974427599	1515974364288	
38.903549	-77.875361	A9CD47	1515974438865	1515974362388	
33.147889	-111.88433	A81B04	1515974429412	1515974361943	
35.952164	-115.134829	AC1EF0	1515974378396	1515974361388	
-36.088621	148.172491	7C1C5C	1515974482443	1515974361271	
28.675527	-83.236461	A8D249	1515974428865	1515974359661	
-31.86771	115.973888	7C42DC	1515974432177	1515974359365	
40.741516	-74.148442	A948FC	1515974438646	1515974358458	
14.337753	121.147788	758227	1515974429537	1515974355927	
38.688527	-77.823857	A77E8F	1515974381724	1515974349286	
32.823331	-98.143111	AA7E79	1515974412162	1515974346583	
39.314346	-119.493156	A97CDF	1515974431177	1515974344614	

The screenshot displays the Google Cloud Data Studio interface, showing the details of a job named 'plane_data - 2 Flow'. The job is completed today at 7:10 PM. The 'Overview' tab is selected, showing a table of data with columns: Lat, Long, A_E, Icao, PosTime, and FSeen. The table contains 5 columns and 16,92M rows. The 'Execution stages' section shows the job's progress, including schema validation and transform with profile. The 'Job summary' section provides details about the job, including the job ID, status, flow, and output. The 'Execution summary' section shows the job type, user, start time, finish time, last update, duration, memory usage, and environment. The 'Optimization summary' section shows the optimization status, which is enabled.

Lat	Long	A _E	Icao	PosTime	FSeen
45.64698	11.7882	0888EE	1516047814538	1516046721288	
38.937688	-77.457886	0881F9	1516058364253	1516058172111	
35.74855	-5.288892	0288F2	1516048161786	151603921988	
51.999734	4.187047	02018D	1516036115786	151603813978	
39.436389	-8.513123	020124	1516017832251	1516016383687	
58.613689	3.964951	02012B	1516042297818	1516041864314	
41.87352	31.357852	06A816	1516035834921	1516032666106	
44.543468	26.845971	06A852	1516089233178	1516088428234	
42.764511	27.285956	06A89B	1516088152428	1516087885528	
48.954181	17.338138	06A88A	1516017581867	1516011377345	

Google Cloud Dataproc Job details

Output

Spark jobs take ~60 seconds to initialize resources.

```
23/02/26 03:24:22 INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat: Total input files to process : 14
[{"Lat": 45.64698, "Long": 11.7082, "Icao": "0000EE", "PosTime": "1516047814530", "FSeen": "1516046721280"},
{"Lat": 38.937688, "Long": -77.457886, "Icao": "0001F9", "PosTime": "1516058364253", "FSeen": "1516058172111"},
{"Lat": 35.74855, "Long": -5.208092, "Icao": "0200F2", "PosTime": "1516040161786", "FSeen": "1516038921988"},
{"Lat": 51.999734, "Long": 4.187047, "Icao": "020100", "PosTime": "1516036115706", "FSeen": "1516033013978"},
{"Lat": 39.436309, "Long": -0.513123, "Icao": "020124", "PosTime": "1516017032251", "FSeen": "1516016383607"}]
[{"FSeen": 1516046721280,
  "Icao": "0000EE",
  "Lat": 45.64698,
  "Long": 11.7082,
  "PosTime": 1516047814530},
{"FSeen": 1516058172111,
  "Icao": "0001F9",
  "Lat": 38.937688,
  "Long": -77.457886,
  "PosTime": 1516058364253},
{"FSeen": 1516038921988,
  "Icao": "0200F2",
  "Lat": 35.74855,
  "Long": -5.208092,
  "PosTime": 1516040161786},
{"FSeen": 1516033013978,
  "Icao": "020100",
  "Lat": 51.999734,
  "Long": 4.187047,
  "PosTime": 1516036115706},
{"FSeen": 1516016383607,
  "Icao": "020124",
  "Lat": 39.436309,
  "Long": -0.513123,
  "PosTime": 1516017032251}]
23/02/26 03:24:29 INFO com.google.cloud.hadoop.repackaged.gcs.com.google.cloud.hadoop.gcsio.GoogleCloudStorageFileSystem: Successfully repaired 'gs://cs512_aircraft_1/hadoop/tmp/big
23/02/26 03:24:29 INFO org.sparkproject.jetty.server.AbstractConnector: Stopped SparkK4cd49a38(HTTP/1.1, {http/1.1})(0.0.0.0:0)
```

Sandra's Screenshots:

PLANE_DATA - 4 FLOW

Initial data

Run

New Step Recipe

- Delete rows where MATCHES([column1], A["src":*, false])
- Replace matches of /\$/ from column1 with "
- Change column1 type to Object
- Create new columns from 5 constants in column1
- Delete column1
- Delete rows with missing values in Lat
- Split FSeen on delimiters matching "\VDate"
- Delete FSeen1
- Replace matches of "V" from FSeen2 with "
- Rename FSeen2 to 'FSeen'

#	Lat	Long	A ⁰ _C	Icao	PosTime	FSeen
-	28.461868	-81.321804	A8754D		1515974431037	1515974407537
-	-33.982018	151.177994	7C752D		1515974426333	1515974406333
-	-33.948315	150.590637	7C671B		1515974430912	1515974399193
-	33.953293	-117.66761	A3E6EA		1515974427990	1515974392349
-	35.478249	140.393646	780A54		1515974431490	1515974391458
-	41.328186	-106.958618	A08D47		1515974429037	1515974381099
-	34.7953	135.3871	862226		1515974423693	1515974380536
-	37.354523	-121.990434	A1DDE6		1515974430630	1515974379990
-	37.039055	-122.089393	A0E35D		1515974431490	1515974379521
-	51.609238	-114.622368	C080CD		1515974401787	1515974377052
-	36.857437	140.071472	861B06		1515974423365	1515974374849
-	50.72197	25.617599	A0081B		1515974427208	1515974374802
-	49.036358	-1.715168	A06441		1515974426583	1515974372146
-	-33.875702	151.27272	C80004		1515974432177	1515974371708
-	41.946808	-87.825581	A26793		1515974399521	1515974370693
-	-43.447586	172.574456	C0173A		1515974430646	1515974369333
-	33.434189	-111.917778	A7668A		1515974431474	1515974367536
-	49.169769	-123.300056	C00100		1515974431412	1515974365568
-	38.362628	-80.475391	A09E27		1515974390162	1515974365005
-	33.390839	-116.765771	A4DBE2		1515974427599	1515974364208
-	38.903549	-77.075361	A9CD47		1515974430865	1515974362380
-	33.147009	-111.80433	A01B04		1515974429412	1515974361943
-	35.952164	-115.134029	AC1EF0		1515974378396	1515974361380

Show / Hide data grid options 5 Columns 3,992 Rows 3 Data Types

plane_data - 4 Flow > plane_data - 4

Job 18046831

Finished Yesterday at 5:09 PM

View BigQuery job

...

Overview

Output destinations

Profile

Dependency graph

Data sources

Parameters

Execution stages

Schema validation

Completed Yesterday at 5:08 PM, started Yesterday at 5:07 PM • Ran for 23 sec

Datasets

plane_data - 4

No schema changes found

View all

Transform with profile

Completed Yesterday at 5:09 PM, started Yesterday at 5:08 PM • Ran for 2 min

Environment BigQuery

100% valid values

0% mismatching values

0% missing values

View steps and dependencies

View profile

View BigQuery job

Publish

Completed Yesterday at 5:09 PM, started Yesterday at 5:09 PM • Ran for <1 sec

Activity

plane_data__4_20230225_230759

Completed

View all

Job summary

Job ID

18046831

Job status

Completed

Flow

plane_data - 4 Flow

Output

plane_data - 4

Execution summary

Job type

Manual

User

Sandra Estrada

Start time

February 25th 2023, 5:08 pm

Finish time

February 25th 2023, 5:09 pm

Last update

February 25th 2023, 5:09 pm

Duration

2 minutes

memory usage

40.524316672 GB

Environment

BigQuery

Optimization summary

Optimization

Enabled

Google Cloud

CS512 Project

Search (/) for resources, docs, products, and more

Search

Dataproc

Jobs on Clusters

Clusters

Jobs

Workflows

Autoscaling policies

Serverless

Batches

Metastore Services

Metastore

Federation

Utilities

Component exchange

Release Notes

Job details

CLONE

DELETE

STOP

REFRESH

Job ID

job-2cc56280

Job UUID

46dd1227-d7a4-4de1-86fe-5f45052f8539

Type

Dataproc Job

Status

Succeeded

Output

LINE WRAP: OFF

Spark jobs take ~60 seconds to initialize resources.

{'PosTime': 1516008025419},

{'FSeen': 1516024612368,

'Icao': '01010C',

'Lat': 47.526204,

'Long': 12.782381,

'PosTime': 1516025012338},

{'FSeen': 1516018713572,

'Icao': '01013D',

'Lat': 45.616196,

'Long': -77.65197,

'PosTime': 1516019131871},

{'FSeen': 1515969500010,

'Icao': '01018C',

'Lat': 46.47757,

'Long': 13.456041,

'PosTime': 1515974491802},

{'FSeen': 1516024065027,

'Icao': '020026',

'Lat': 50.760609,

'Long': 3.14827,

'PosTime': 1516026156581}]}

Dataproc

Jobs on Clusters

Clusters

Jobs

Workflows

Autoscaling policies

Serverless

Batches

Metastore Services

Metastore

Cluster details

SUBMIT JOB

REFRESH

START

STOP

DELETE

VIEW LOGS

Bucket name 'cs512_aircraftdata' contains underscore, which may cause job failures.

MORE

Name

cs512-hello-spark

Cluster UUID

eca3b637-0cc6-42b6-b348-326ae4de2492

Type

Dataproc Cluster

Status

Stopped

MONITORING

JOBS

VM INSTANCES

CONFIGURATION

WEB INTERFACES

Filter

Filter jobs

Job ID	Status	Region	Type	Start time	Elapsed time	Labels
job-2cc56280	Succeeded	us-east1	PySpark	Feb 25, 2023, 5:23:21 PM	36 sec	None