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.

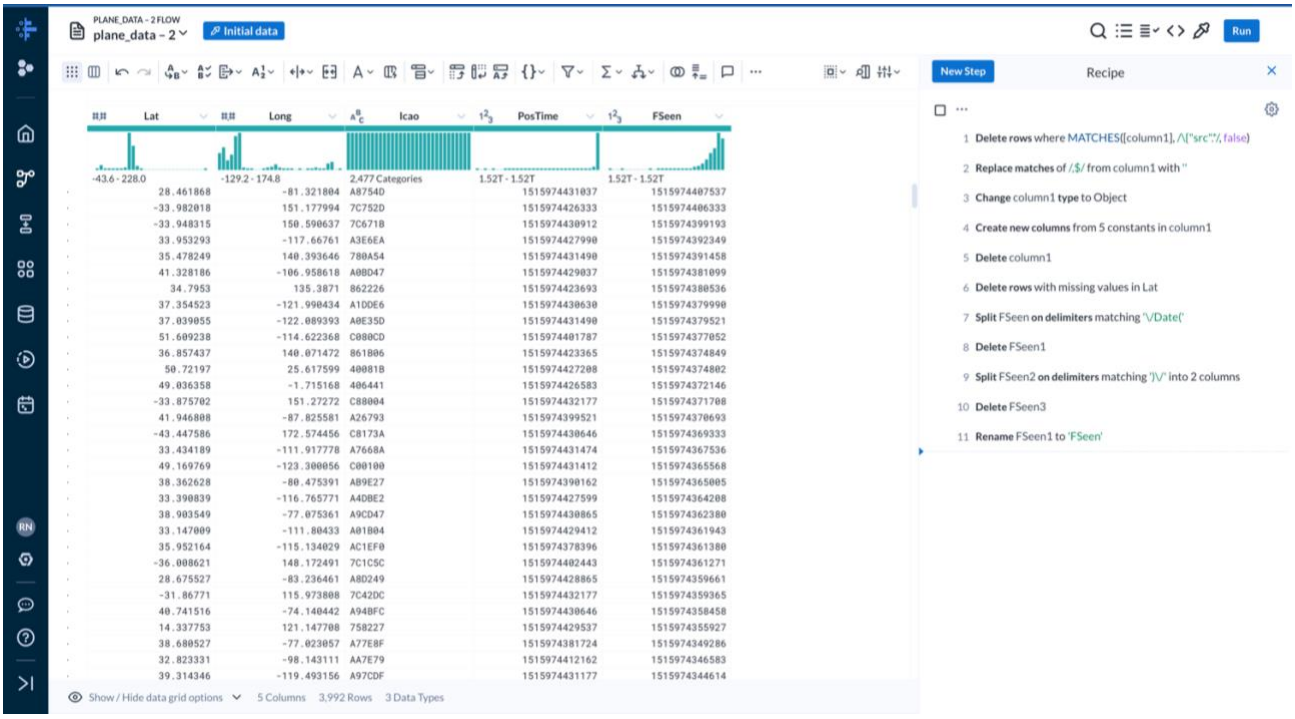
After creating a new project, I followed all the steps from the previous assignment and exported the table to BigQuery. When I ran the query from the previous assignment, I obtained the expected value, which indicated that the data was being exported correctly. I then edited the recipe by adding the 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.

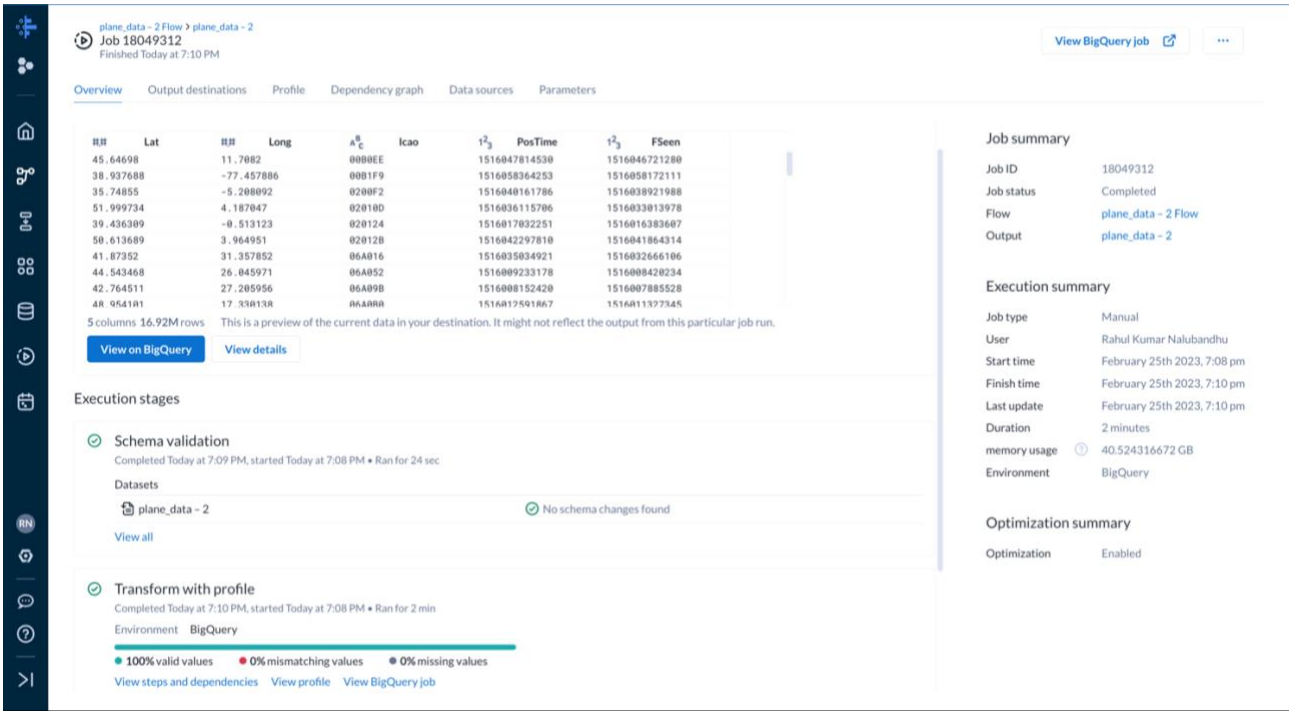
Rahul and Sandra: Once those issues were resolved, our experience with completing this assignment was consistent. We both created a cluster in DataProc as instructed in the Exploration: Hello Spark video 1. Next, we updated the plane data recipe to include PosTime and FSeen and reran the DataPrep job, 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.

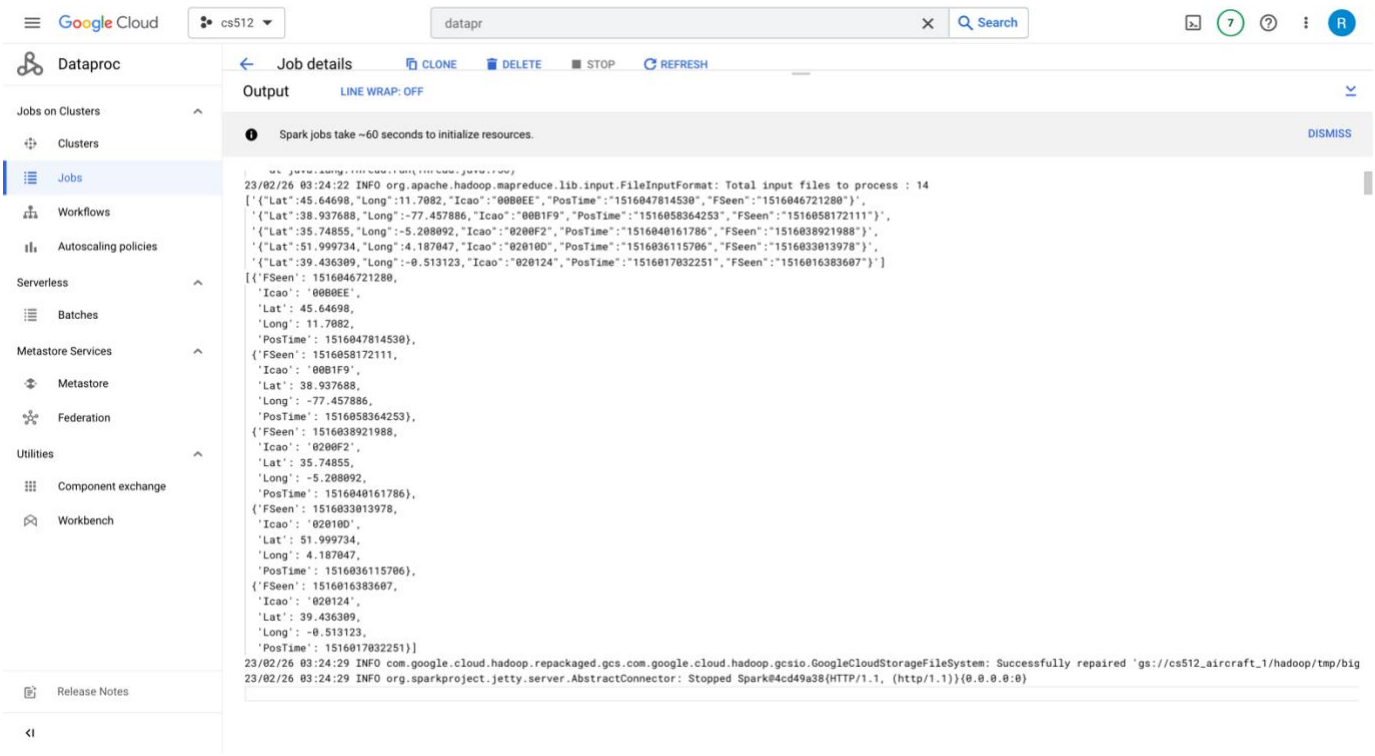
Once we had completed this, we submitted a job in DataProc under the cluster we had created for PySpark with the python’s file URI and the jar connector link provided. We both received a job failure error message and referenced ED Discussion for help in resolving this issue. Joshua Magana referenced 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.

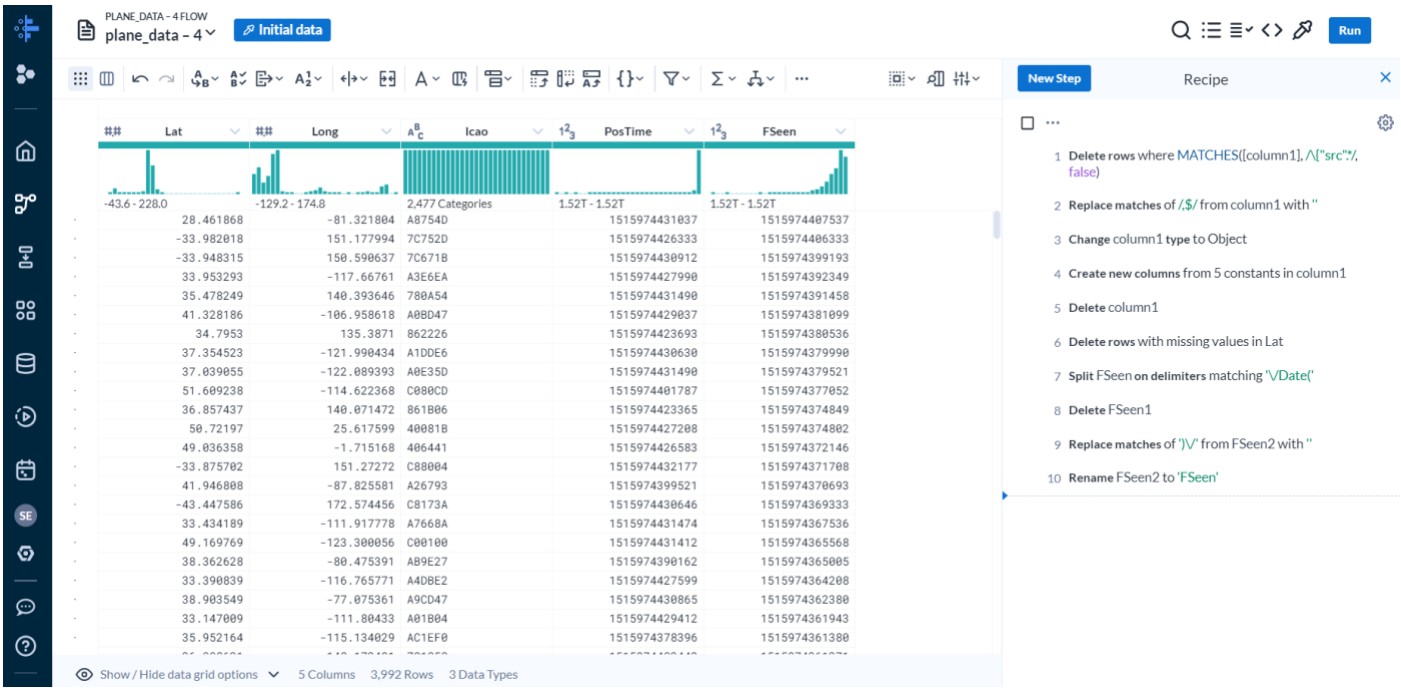
Rahul's Screenshots:

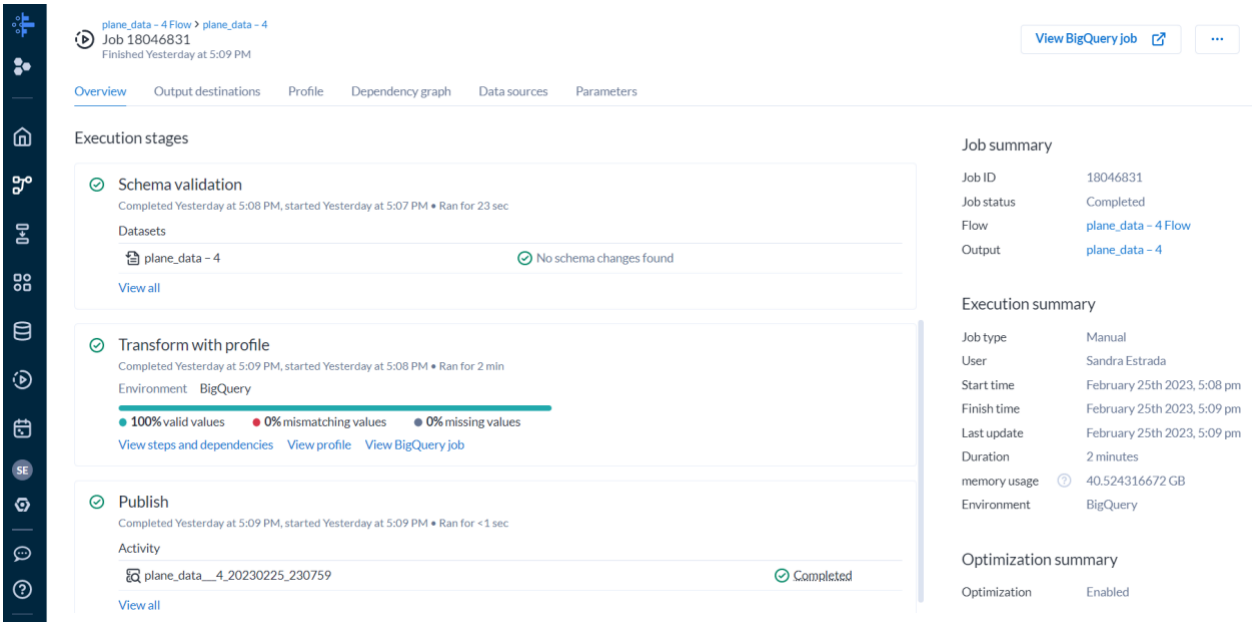


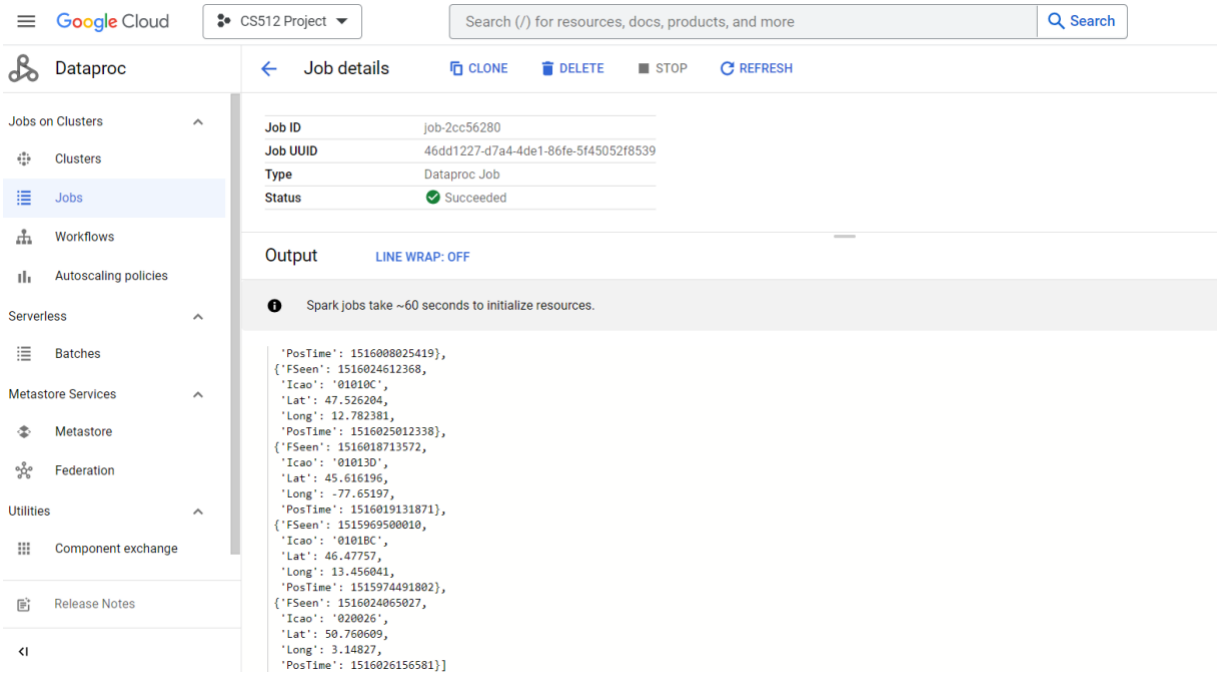


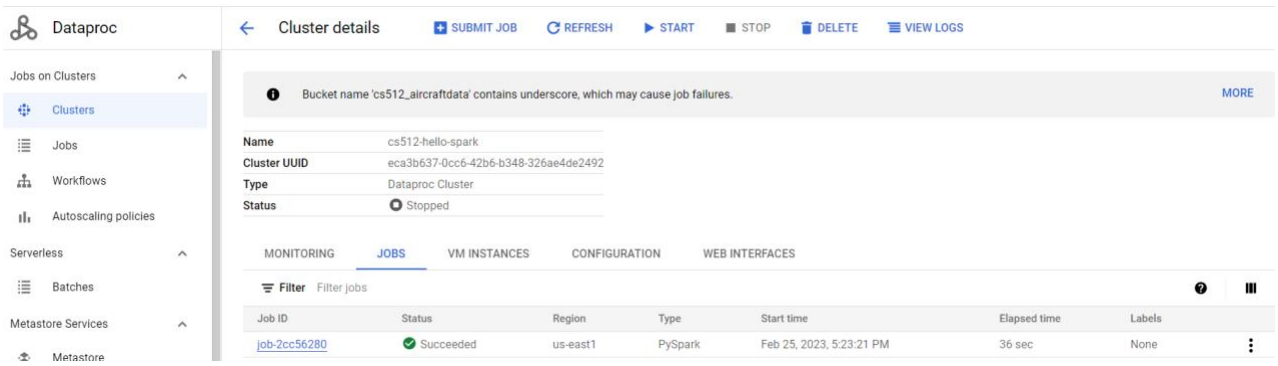


Sandra's Screenshots:




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
Job details
CLONE
DELETE
STOP
REFRESH
Jobs on Clusters
Clusters
Jobs
Workflows
Autoscaling policies
Serverless
Batches
Metastore Services
Metastore
Federation
Utilities
Component exchange
Release Notes
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': '0101BC',
'Lat': 46.47757,
'Long': 13.456041,
'PosTime': 1515974491802},
{'FSeen': 1516024065027,
'Icao': '020026',
'Lat': 50.760609,
'Long': 3.14827,
'PosTime': 1516026156581}]


Dataproc
Cluster details
SUBMIT JOB
REFRESH
START
STOP
DELETE
VIEW LOGS
Jobs on Clusters
Clusters
Jobs
Workflows
Autoscaling policies
Serverless
Batches
Metastore Services
Metastore
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