CS777 – Week 1 Homework Submission Template

Poom Chantarapornrat

Task 1 – Top-10 Active Taxis

Many different taxis have had multiple drivers. Write and execute a Spark Python program that computes the top ten taxis that have had the largest number of drivers. Your output should be a set of (medallion, number of drivers) pairs.

***Note**: You should consider that this is a real-world data set that might include wrongly formatted data lines. You should clean up the data before the main processing, a line might not include all of the fields. If a data line is not correctly formatted, you should drop that line and do not consider it.*

- Print a list of top 10 taxis having the largest number of drivers, and the amount of drivers (taxi ID and count)

```
('11DC93DD66D8A9A3DD9223122CF99EFD', 352)

('EE06BD8A621CAC3B608ACFDF0585A76A', 348)

('6C1132EF70BC0A7DB02174592F9A64A1', 341)

('A10A65AFD9F401BF3BDB79C84D3549E7', 340)

('23DB792D3F7EBA03004E470B684F2738', 339)

('7DA8DF1E4414F81EBD3A0140073B2630', 337)

('0318F7BBB8FF48688698F04016E67F49', 335)

('738A62EEE9EC371689751A864C5EF811', 333)

('7D93E7FC4A7E4615A34B8286D92FF57F', 333)

('B07944BF31699A169091D2B16597A4A9', 333)
```

Task 2 – Top-10 Best Drivers

We would like to figure out who the top 10 best drivers are in terms of their average earned money per minute spent carrying a customer. The total amount field is the total money earned on a trip. In the end, we are interested in computing a set of (driver, money per minute) pairs.

- Print a list of top 10 best drivers based on earned money per minute carrying a customer (Driver ID and average earning)

```
('CC664699259C9867E735976611A82F64', 60159.17333333334)
```

('89EDAF45090C74611B52AFFC3E10A69D', 28579.504166666666)

('D4CA68ECC21536DE406F3D58C7813241', 12274.920895522388)

('155EBAC6C5A22D6CFD3518F6A0E9190C', 9094.441944709246)

('21166F1430E0F4A28F3AA39B53D4A935', 5681.818153846153)

('E2DF7E5E63A654B9A655FBDCC0B029BF', 2356.452307692308)

('E523F84E85708BCEB9FEB6F7825C0E08', 1440.104)

('E2DF7E5E63A654B9A655FBDCC0B029BF', 513.4639106145252)

('6166E80B7F3ABEE89D9A7CC1C248A8E0', 476.6666666666667)

('170C59E5D93B2A2FA739B30F5848CE02', 465.78688524590166)

Task 3 – The best time of the day to Work on Taxi (For Advanced Students - no points)

We would like to know which hour of the day is the best time for drivers that has the highest profit per mile. Consider the surcharge amount in dollars for each taxi ride (without tip amount) and the distance in miles, and sum up the rides for each hour of the day (24 hours) – consider the pickup time for your calculation. The profit ratio is the ration surcharge in dollars divided by the travel distance in miles for each specific time of the day.

Profit Ratio = (Surcharge Amount in US Dollar) / (Travel Distance in miles)

We are interested to know the time of the day that has the highest profit ratio.

- Print the profit ratio for the best hour of the day exhibiting the highest profit per mile

Your output should go here.

Task 4 - (For Advanced Students - no points)

Here are some further tasks for advanced groups.

a) How many percent of taxi customers pay with cash, and how many percent use electronic cards? Analyze these payment methods for different times of the day and provide a list of percent for each time of the day? As a result, provide two numbers for total percentages and a list like: hour of the day, percent paid card.

- For each hour of the day, provide percentages of customers paying with cash and cards

Your output should go here.

b) We would like to measure the efficiency of taxi drivers by finding out their average earned money per mile. (Consider the total amount, which includes tips, as their earned money) Implement a Spark job that can find out the top-10 efficient taxi drivers.

- Find the top 10 taxi drivers with highest average earned money per mile

Your output should go here.

c) What are the mean, median, first, and third quantiles of tip amount? How do you find the median?

- Print the mean, median, first and third quantiles of the tip amount

Your output should go here.

- Explain how do you find the median

Discuss your reasoning here.

d) Using the IQR outlier detection method, find out the top-10 outliers.

- Find the top 10 outliers

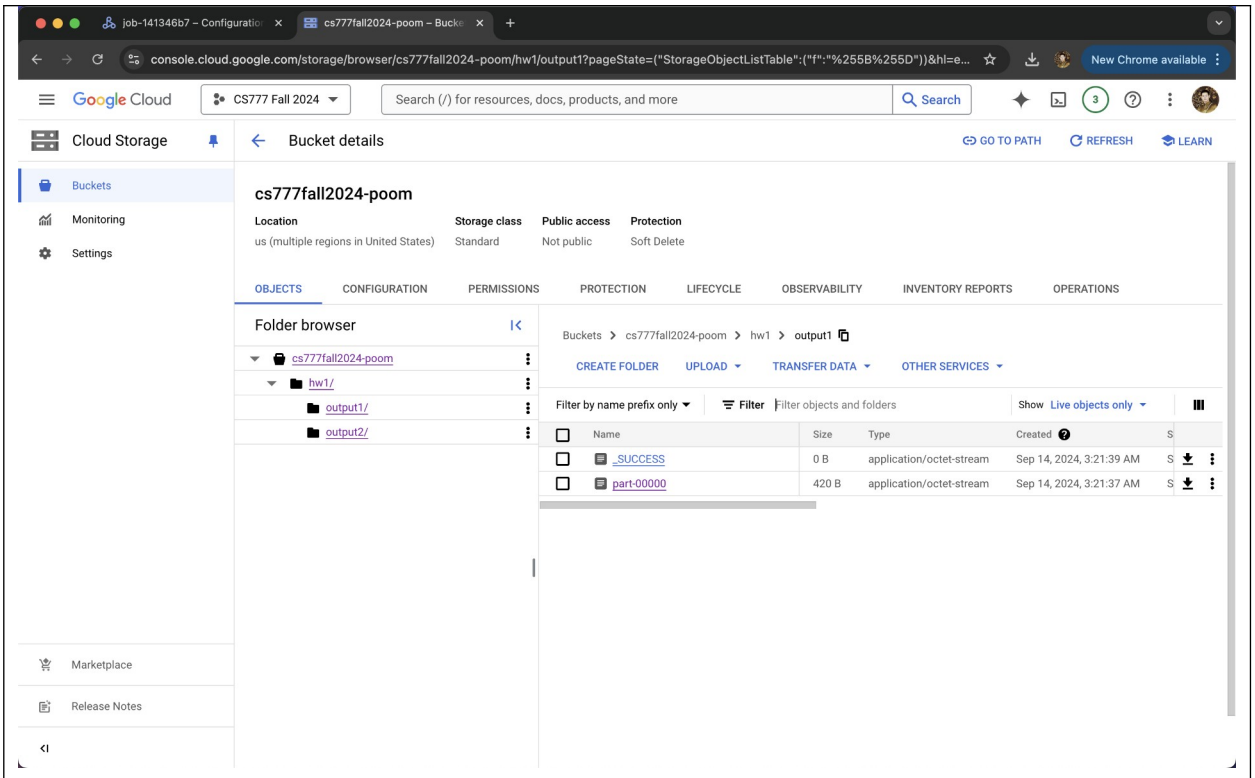
Your output should go here.

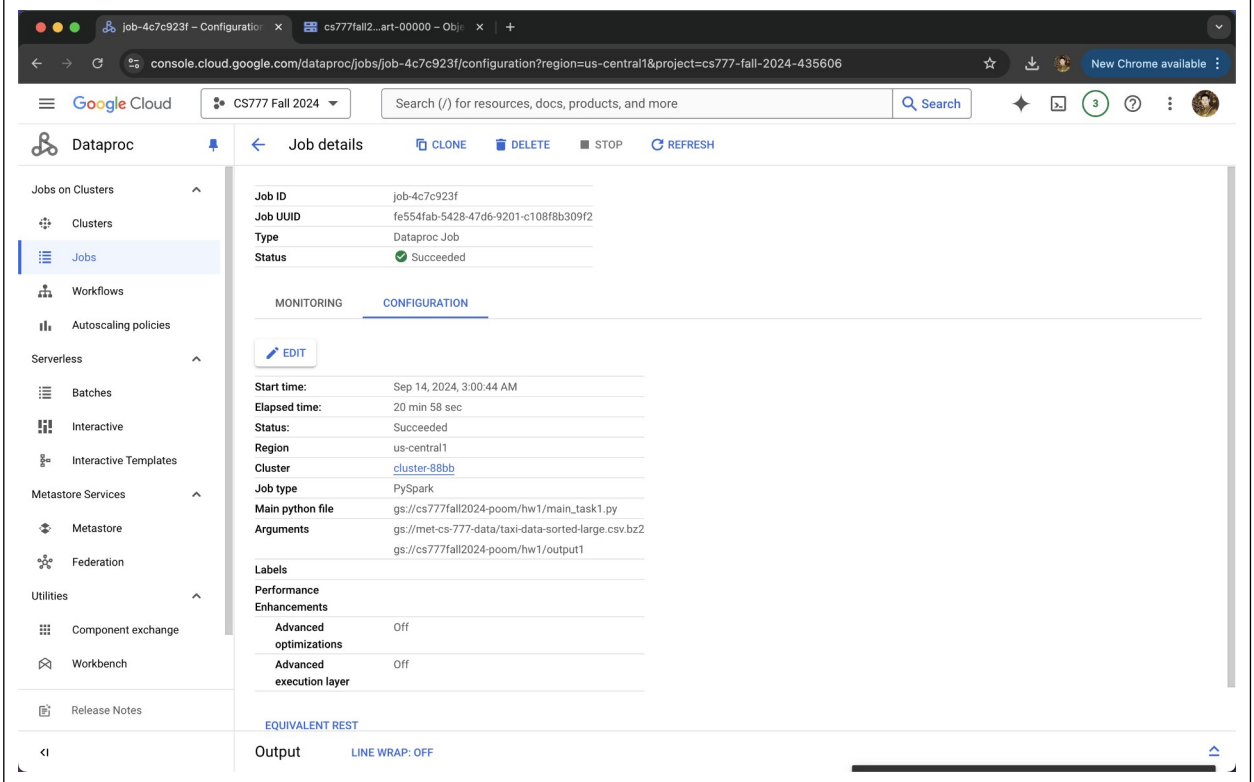
Spark History Output:

To demonstrate that you did execute your code on the cloud it is important to include URLs in the screenshots. Otherwise, there is no way for us to verify if the code was executed in your cloud account.

Task 1:

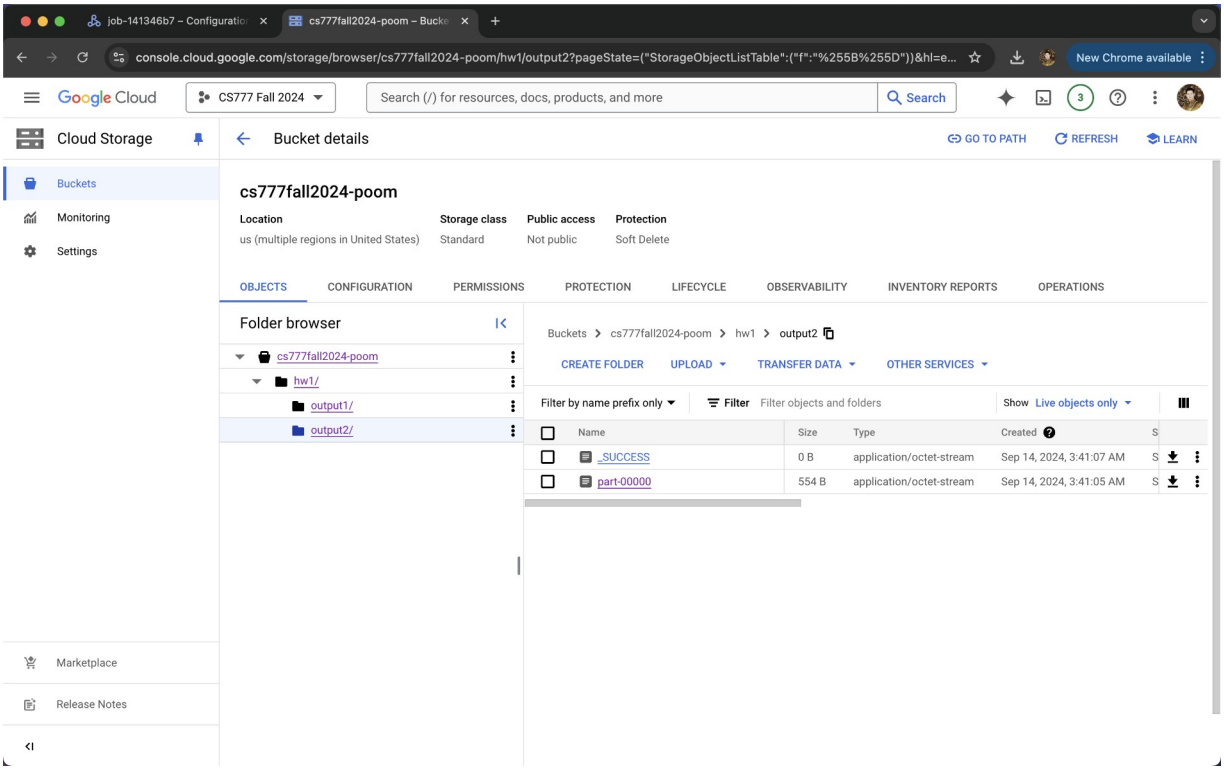
https://console.cloud.google.com/dataproc/jobs/job-4c7c923f/monitoring?region=us-central1&project=cs777-fall-2024-435606

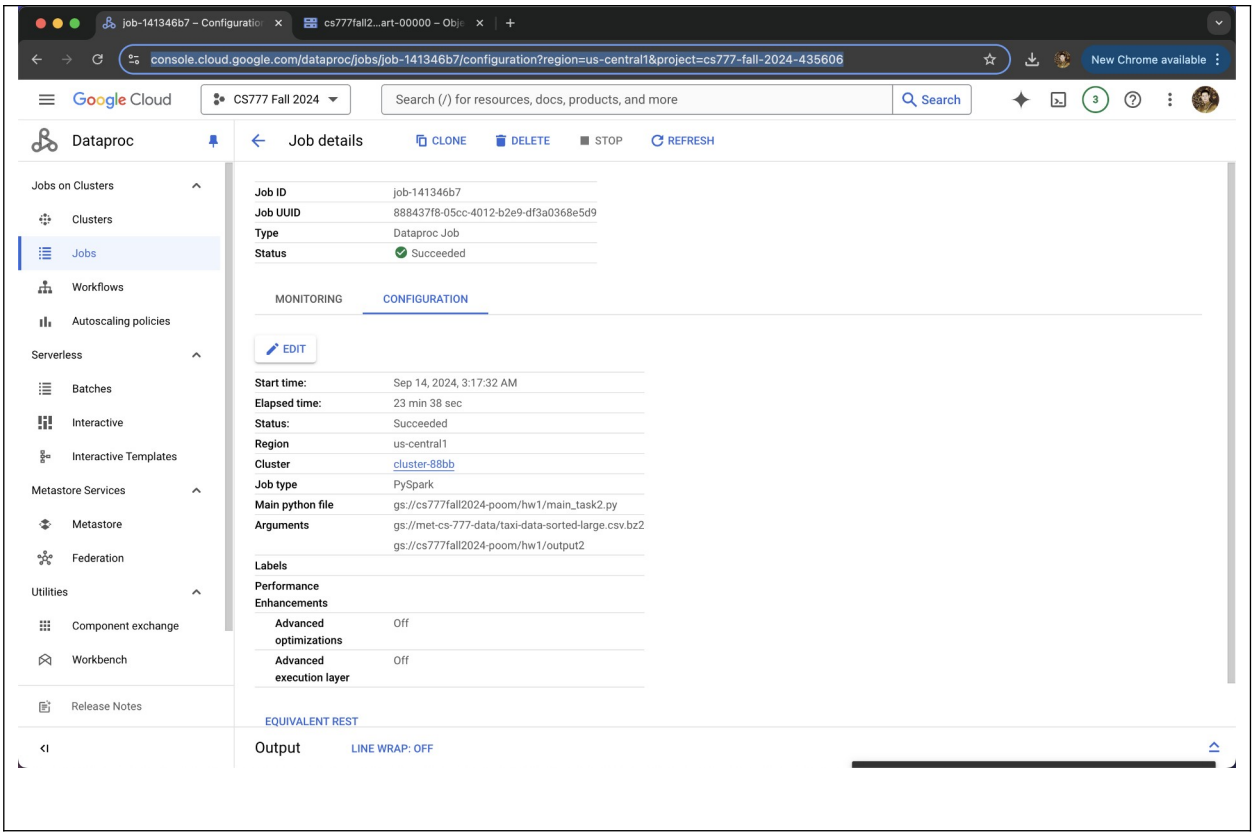

job-141346b7 – Configuration
cs777fall2024-poom – Bucket
console.cloud.google.com/storage/browser/cs777fall2024-poom/hw1/output1?pageState=("StorageObjectListTable":("f":"%255B%255D"))&hl=e...
New Chrome available
Google Cloud
CS777 Fall 2024
Search (/) for resources, docs, products, and more
Search
Cloud Storage
Bucket details
GO TO PATH
REFRESH
LEARN
Buckets
Monitoring
Settings
cs777fall2024-poom
Location
us (multiple regions in United States)
Storage class
Standard
Public access
Not public
Protection
Soft Delete
OBJECTS
CONFIGURATION
PERMISSIONS
PROTECTION
LIFECYCLE
OBSERVABILITY
INVENTORY REPORTS
OPERATIONS
Folder browser
cs777fall2024-poom
hw1/
output1/
output2/
Buckets > cs777fall2024-poom > hw1 > output1
CREATE FOLDER
UPLOAD
TRANSFER DATA
OTHER SERVICES
Filter by name prefix only
Filter
Filter objects and folders
Show
Live objects only
Name
Size
Type
Created
_SUCCESS
0 B
application/octet-stream
Sep 14, 2024, 3:21:39 AM
part-00000
420 B
application/octet-stream
Sep 14, 2024, 3:21:37 AM
Marketplace
Release Notes


job-4c7c923f – Configuration
cs777fall2...art-00000 – Obje
console.cloud.google.com/dataproc/jobs/job-4c7c923f/configuration?region=us-central1&project=cs777-fall-2024-435606
New Chrome available
Google Cloud
CS777 Fall 2024
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
Interactive
Interactive Templates
Metastore Services
Metastore
Federation
Utilities
Component exchange
Workbench
Release Notes
Job ID
job-4c7c923f
Job UUID
fe554fab-5428-47d6-9201-c108f8b309f2
Type
Dataproc Job
Status
Succeeded
MONITORING
CONFIGURATION
EDIT
Start time:
Sep 14, 2024, 3:00:44 AM
Elapsed time:
20 min 58 sec
Status:
Succeeded
Region
us-central1
Cluster
cluster-88bb
Job type
PySpark
Main python file
gs://cs777fall2024-poom/hw1/main_task1.py
Arguments
gs://met-cs-777-data/taxi-data-sorted-large.csv.bz2
gs://cs777fall2024-poom/hw1/output1
Labels
Performance Enhancements
Advanced optimizations
Off
Advanced execution layer
Off
EQUIVALENT REST
Output
LINE WRAP: OFF

Task 2:

https://console.cloud.google.com/dataproc/jobs/job-141346b7/configuration?region=us-central1&project=cs777-fall-2024-435606




job-141346b7 – Configuration
cs777fall2...art-00000 – Obje
console.cloud.google.com/dataproc/jobs/job-141346b7/configuration?region=us-central1&project=cs777-fall-2024-435606
New Chrome available
Google Cloud
CS777 Fall 2024
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
Interactive
Interactive Templates
Metastore Services
Metastore
Federation
Utilities
Component exchange
Workbench
Release Notes
Job ID job-141346b7
Job UUID 888437f8-05cc-4012-b2e9-df3a0368e5d9
Type Dataproc Job
Status Succeeded
MONITORING
CONFIGURATION
EDIT
Start time: Sep 14, 2024, 3:17:32 AM
Elapsed time: 23 min 38 sec
Status: Succeeded
Region us-central1
Cluster cluster-88bb
Job type PySpark
Main python file gs://cs777fall2024-poom/hw1/main_task2.py
Arguments gs://met-cs-777-data/taxi-data-sorted-large.csv.bz2
gs://cs777fall2024-poom/hw1/output2
Labels
Performance Enhancements
Advanced optimizations Off
Advanced execution layer Off
EQUIVALENT REST
Output
LINE WRAP: OFF