Large dataset: https://storage.googleapis.com/met-cs-777-data/taxi-data-sorted-large.csv.bz2

Top-10 Active Taxis

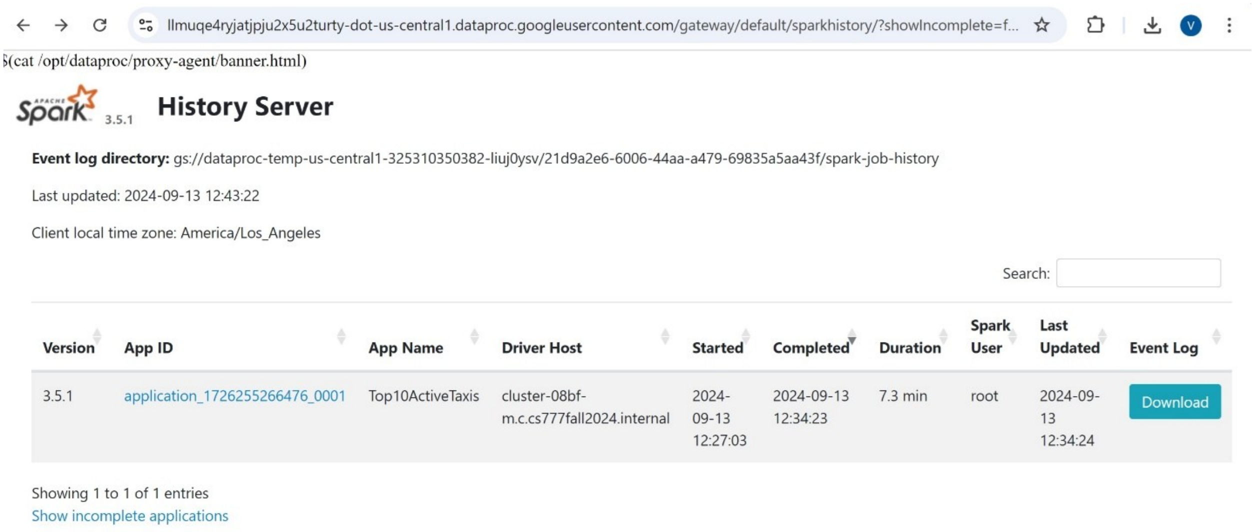
Many different taxis have had multiple drivers. Write and execute a Spark Python program that computes the top ten taxis that have had the largest number of drivers. Your output should be a set of (medallion, number of drivers) pairs.

Large dataset: https://storage.googleapis.com/met-cs-777-data/taxi-data-sorted-large.csv.bz2

***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)
('B07944BF31699A169091D2B16597A4A9', 333)
('738A62EEE9EC371689751A864C5EF811', 333)
('7D93E7FC4A7E4615A34B8286D92FF57F', 333)
```



llmuqe4ryjatjpju2x5u2turty-dot-us-central1.dataproc.googleusercontent.com/gateway/default/sparkhistory/?showIncomplete=f...
V

$(cat /opt/dataproc/proxy-agent/banner.html)

History Server

Event log directory: gs://dataproc-temp-us-central1-325310350382-liuj0ysv/21d9a2e6-6006-44aa-a479-69835a5aa43f/spark-job-history

Last updated: 2024-09-13 12:43:22

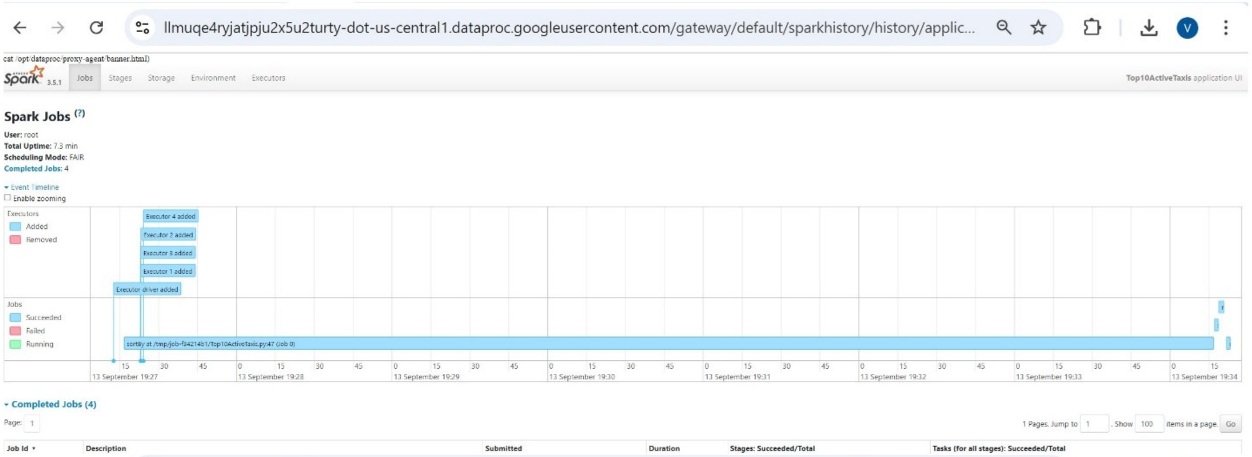
Client local time zone: America/Los_Angeles

Search:

Version	App ID	App Name	Driver Host	Started	Completed	Duration	Spark User	Last Updated	Event Log
3.5.1	application_1726255266476_0001	Top10ActiveTaxis	cluster-08bf-m.c.cs777fall2024.internal	2024-09-13 12:27:03	2024-09-13 12:34:23	7.3 min	root	2024-09-13 12:34:24	Download

Showing 1 to 1 of 1 entries

Show incomplete applications



llmuqe4ryjatjpju2x5u2turty-dot-us-central1.dataproc.googleusercontent.com/gateway/default/sparkhistory/history/applic...
V
3.5.1
Jobs
Stages
Storage
Environment
Executors
Top10ActiveTaxis application UI

cat /opt/dataproc/proxy-agent/banner.html)

Spark Jobs (?)

User: root
Total Uptime: 7.3 min
Scheduling Mode: FAIR
Completed Jobs: 4

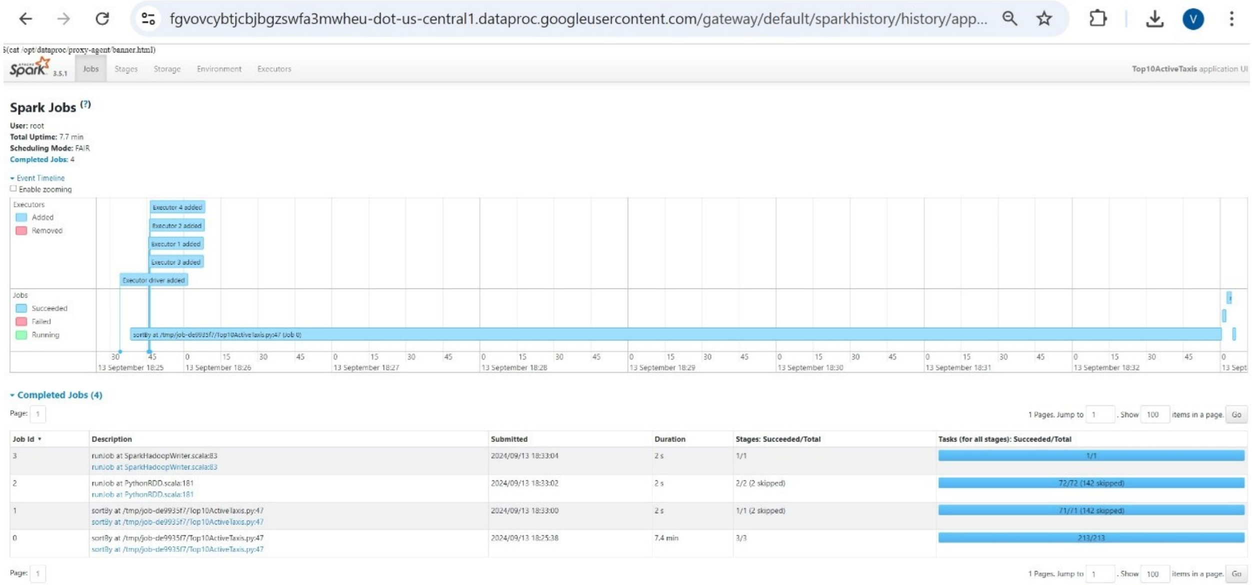
▾ Event Timeline
□ Enable zooming

▾ Completed Jobs (4)

Page: 1

1 Pages. Jump to 1 . Show 100 items in a page. Go

Job Id ▾	Description	Submitted	Duration	Stages: Succeeded/Total	Tasks (for all stages): Succeeded/Total



fgvovcybtjcbjbgzswfa3mwheu-dot-us-central1.dataproc.googleusercontent.com/gateway/default/sparkhistory/history/app...
V
3.5.1
Jobs
Stages
Storage
Environment
Executors
Top10ActiveTaxis application UI

$(cat /opt/dataproc/proxy-agent/banner.html)

Spark Jobs (?)

User: root
Total Uptime: 7.7 min
Scheduling Mode: FAIR
Completed Jobs: 4

▾ Event Timeline
□ Enable zooming

▾ Completed Jobs (4)

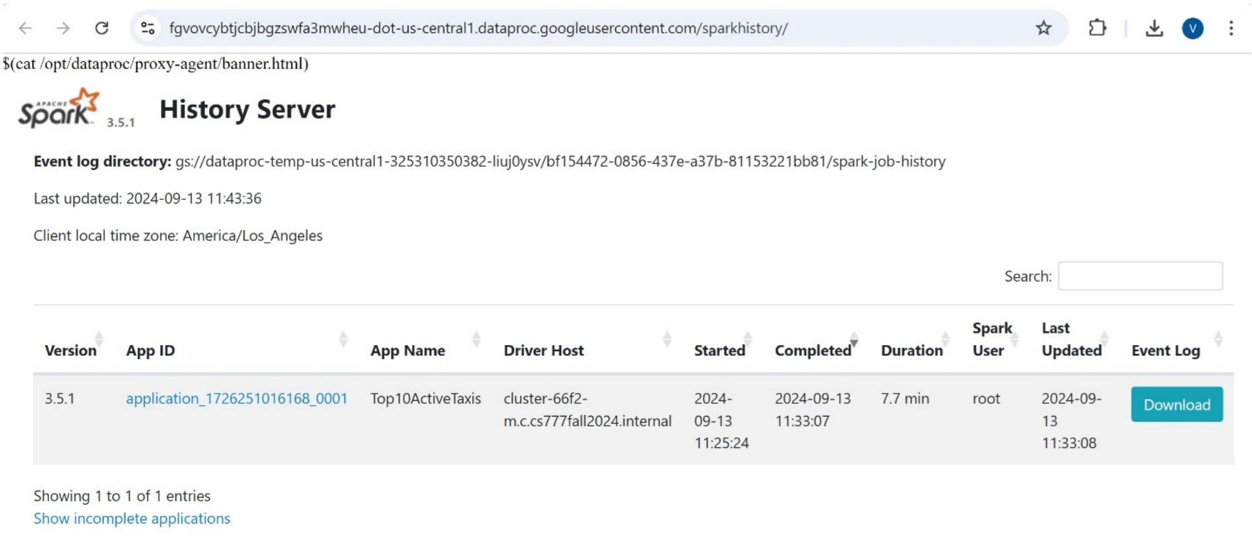
Page: 1

1 Pages. Jump to 1 . Show 100 items in a page. Go

Job Id ▾	Description	Submitted	Duration	Stages: Succeeded/Total	Tasks (for all stages): Succeeded/Total
3	runJob at SparkHadoopWriter.scala:83 runJob at SparkHadoopWriter.scala:83	2024/09/13 18:33:04	2 s	1/1	1/1
2	runJob at PythonRDD.scala:181 runJob at PythonRDD.scala:181	2024/09/13 18:33:02	2 s	2/2 (2 skipped)	72/72 (142 skipped)
1	sortBy at /tmp/job-de9935f7/Top10ActiveTaxis.py:47 sortBy at /tmp/job-de9935f7/Top10ActiveTaxis.py:47	2024/09/13 18:33:00	2 s	1/1 (2 skipped)	71/71 (142 skipped)
0	sortBy at /tmp/job-de9935f7/Top10ActiveTaxis.py:47 sortBy at /tmp/job-de9935f7/Top10ActiveTaxis.py:47	2024/09/13 18:25:38	7.4 min	3/3	213/213

Page: 1

1 Pages. Jump to 1 . Show 100 items in a page. Go


fgvovcybtjcbjbgzswfa3mwheu-dot-us-central1.dataproc.googleusercontent.com/sparkhistory/
V
3.5.1
Search:
Download

$(cat /opt/dataproc/proxy-agent/banner.html)

History Server

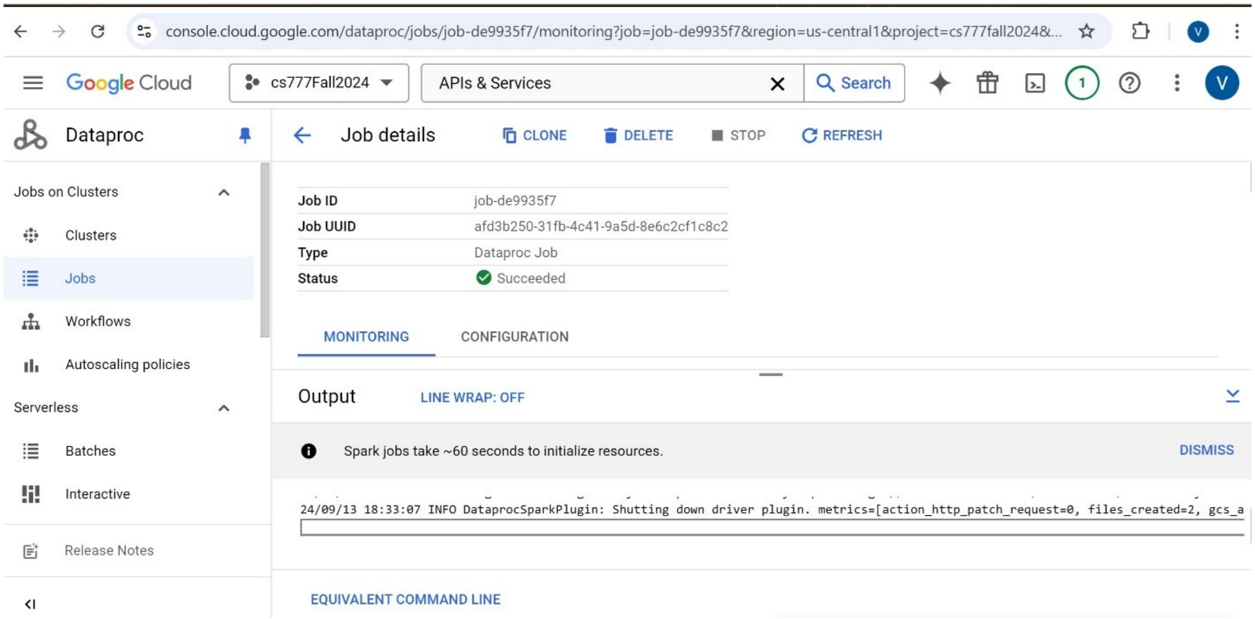
Event log directory: gs://dataproc-temp-us-central1-325310350382-liuj0ysv/bf154472-0856-437e-a37b-81153221bb81/spark-job-history

Last updated: 2024-09-13 11:43:36

Client local time zone: America/Los_Angeles

Version	App ID	App Name	Driver Host	Started	Completed	Duration	Spark User	Last Updated	Event Log
3.5.1	application_1726251016168_0001	Top10ActiveTaxis	cluster-66f2-m.c.cs777fall2024.internal	2024-09-13 11:25:24	2024-09-13 11:33:07	7.7 min	root	2024-09-13 11:33:08	Download

Showing 1 to 1 of 1 entries
Show incomplete applications



Job ID	job-de9935f7
Job UUID	afd3b250-31fb-4c41-9a5d-8e6c2cf1c8c2
Type	Dataproc Job
Status	✓ Succeeded

Spark jobs take ~60 seconds to initialize resources.

```
24/09/13 18:33:07 INFO DataprocSparkPlugin: Shutting down driver plugin. metrics=[action_http_patch_request=0, files_created=2, gcs_a
```

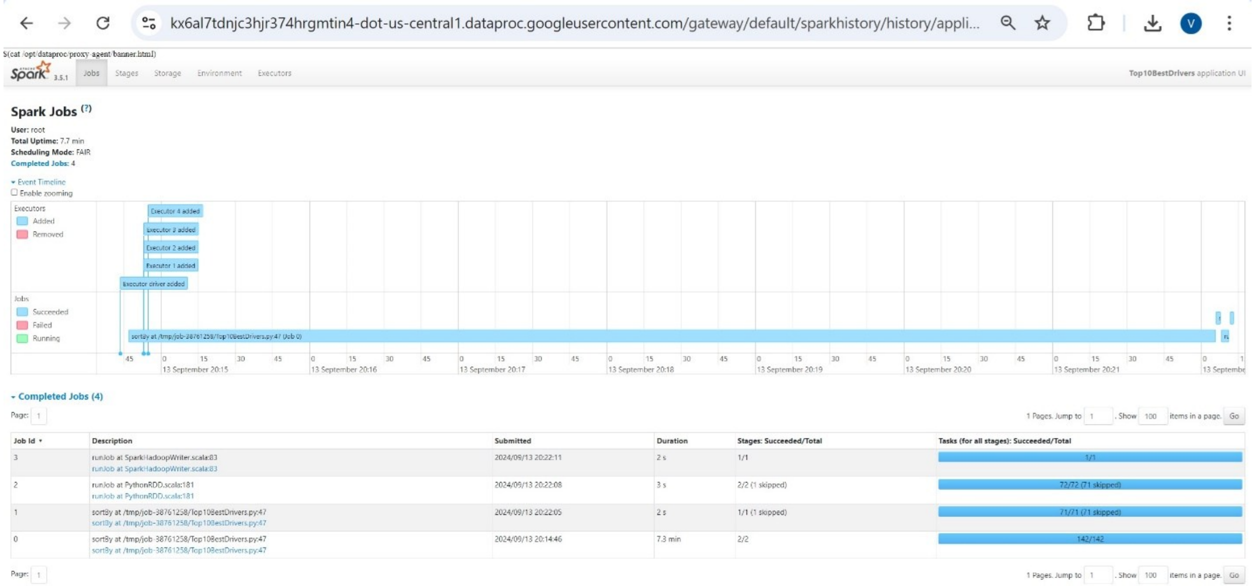
Top-10 Best Drivers

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)

```
('E4F99C9ABE9861F18BCD38BC63D007A9', 62103.18601982026)
('664927CDE376A32789BA48BF55DFB7E3', 32075.6481781048)
('51C1BE97280A80EBFA8DAD34E1956CF6', 22412.519037907998)
```

```
('D85749E8852FCC66A990E40605607B2F', 21838.61138628341)
('23DF80C977D15141F11DD713C523C311', 20791.02908617619)
('3D757E111C78F5CAC83D44A92885D490', 20555.115544088956)
('BE047851D97506885B99BDDFA7A13360', 19584.744671992667)
('3AAB94CA53FE93A64811F65690654649', 19575.05448827395)
('74CC809D28AE726DDB32249C044DA4F8', 19558.588651912727)
('3578782ABD6492CEB927B2A8EBCF1402', 19321.71774748813)
```



Spark Jobs (?)

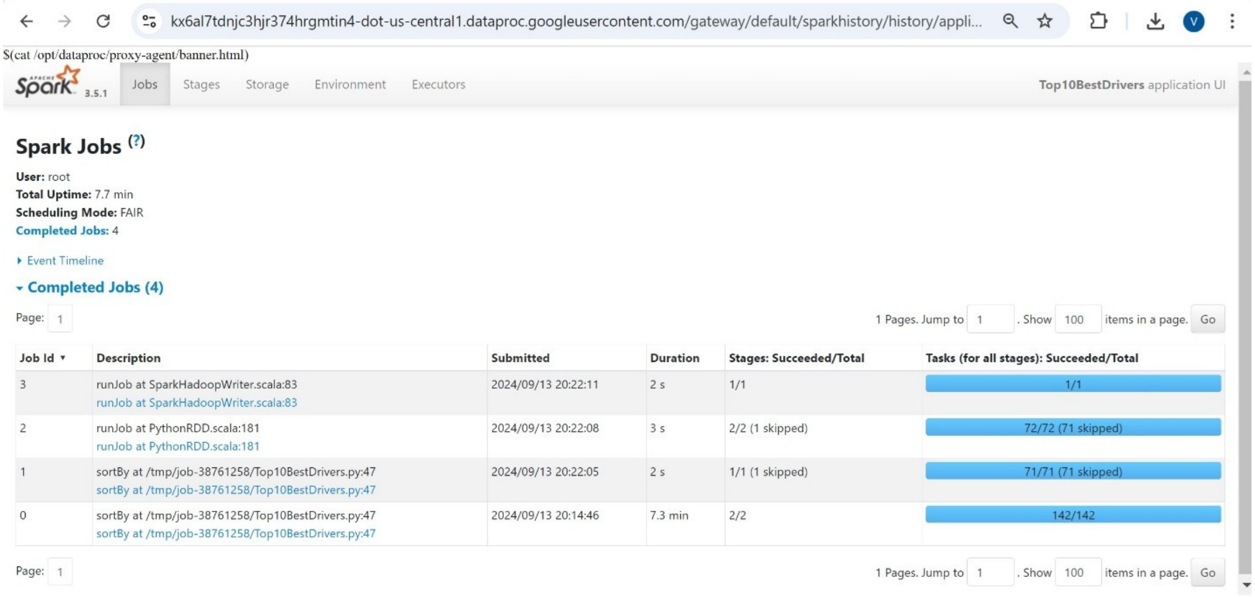
User: root
Total Uptime: 7.7 min
Scheduling Mode: FAIR
Completed Jobs: 4

Completed Jobs (4)

Page: 1 — 1 Pages. Jump to 1 . Show 100 items in a page. Go

Job Id ▾	Description	Submitted	Duration	Stages: Succeeded/Total	Tasks (for all stages): Succeeded/Total
3	runJob at SparkHadoopWriter.scala:83 runJob at SparkHadoopWriter.scala:83	2024/09/13 20:22:11	2 s	1/1	1/1
2	runJob at PythonRDD.scala:181 runJob at PythonRDD.scala:181	2024/09/13 20:22:08	3 s	2/2 (1 skipped)	72/72 (71 skipped)
1	sortBy at /tmp/job-38761258/Top10BestDrivers.py:47 sortBy at /tmp/job-38761258/Top10BestDrivers.py:47	2024/09/13 20:22:05	2 s	1/1 (1 skipped)	71/71 (71 skipped)
0	sortBy at /tmp/job-38761258/Top10BestDrivers.py:47 sortBy at /tmp/job-38761258/Top10BestDrivers.py:47	2024/09/13 20:14:46	7.3 min	2/2	142/142

Page: 1 — 1 Pages. Jump to 1 . Show 100 items in a page. Go


kx6al7tdnjc3hjr374hrgmtin4-dot-us-central1.dataproc.googleusercontent.com/gateway/default/sparkhistory/history/appli...
$(cat /opt/dataproc/proxy-agent/banner.html)
Spark 3.5.1
Jobs
Stages
Storage
Environment
Executors
Top10BestDrivers application UI
Event Timeline

Spark Jobs (?)

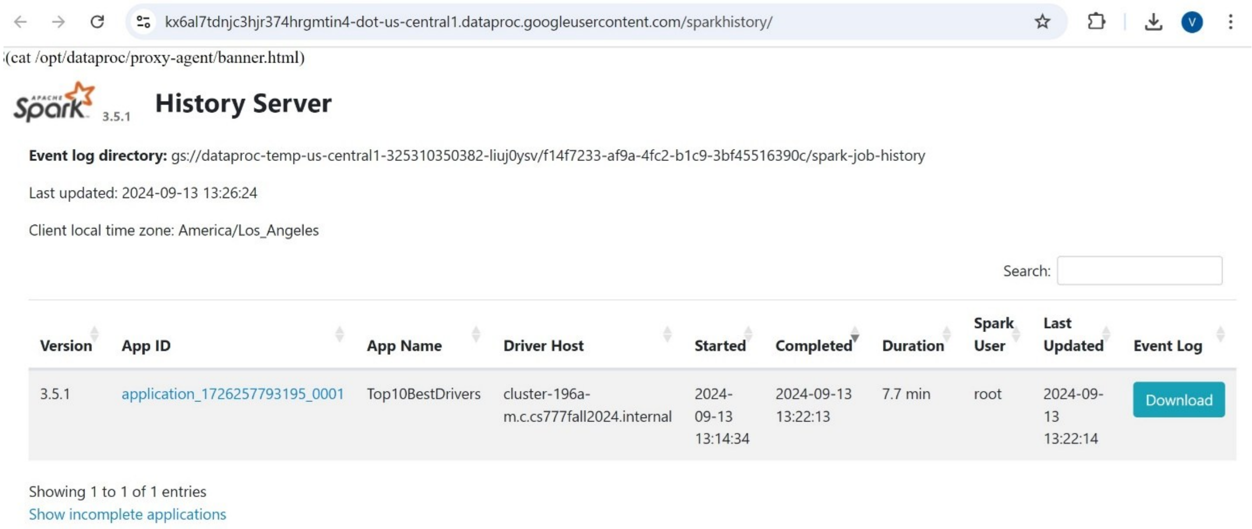
User: root
Total Uptime: 7.7 min
Scheduling Mode: FAIR
Completed Jobs: 4

Completed Jobs (4)

Page: 1 — 1 Pages. Jump to 1 . Show 100 items in a page. Go

Job Id ▾	Description	Submitted	Duration	Stages: Succeeded/Total	Tasks (for all stages): Succeeded/Total
3	runJob at SparkHadoopWriter.scala:83 runJob at SparkHadoopWriter.scala:83	2024/09/13 20:22:11	2 s	1/1	1/1
2	runJob at PythonRDD.scala:181 runJob at PythonRDD.scala:181	2024/09/13 20:22:08	3 s	2/2 (1 skipped)	72/72 (71 skipped)
1	sortBy at /tmp/job-38761258/Top10BestDrivers.py:47 sortBy at /tmp/job-38761258/Top10BestDrivers.py:47	2024/09/13 20:22:05	2 s	1/1 (1 skipped)	71/71 (71 skipped)
0	sortBy at /tmp/job-38761258/Top10BestDrivers.py:47 sortBy at /tmp/job-38761258/Top10BestDrivers.py:47	2024/09/13 20:14:46	7.3 min	2/2	142/142

Page: 1 — 1 Pages. Jump to 1 . Show 100 items in a page. Go



kx6al7tdnjc3hjr374hrgmtin4-dot-us-central1.dataproc.googleusercontent.com/sparkhistory/

:(cat /opt/dataproc/proxy-agent/banner.html)

3.5.1 History Server

Event log directory: gs://dataproc-temp-us-central1-325310350382-liuj0ysv/f14f7233-af9a-4fc2-b1c9-3bf45516390c/spark-job-history

Last updated: 2024-09-13 13:26:24

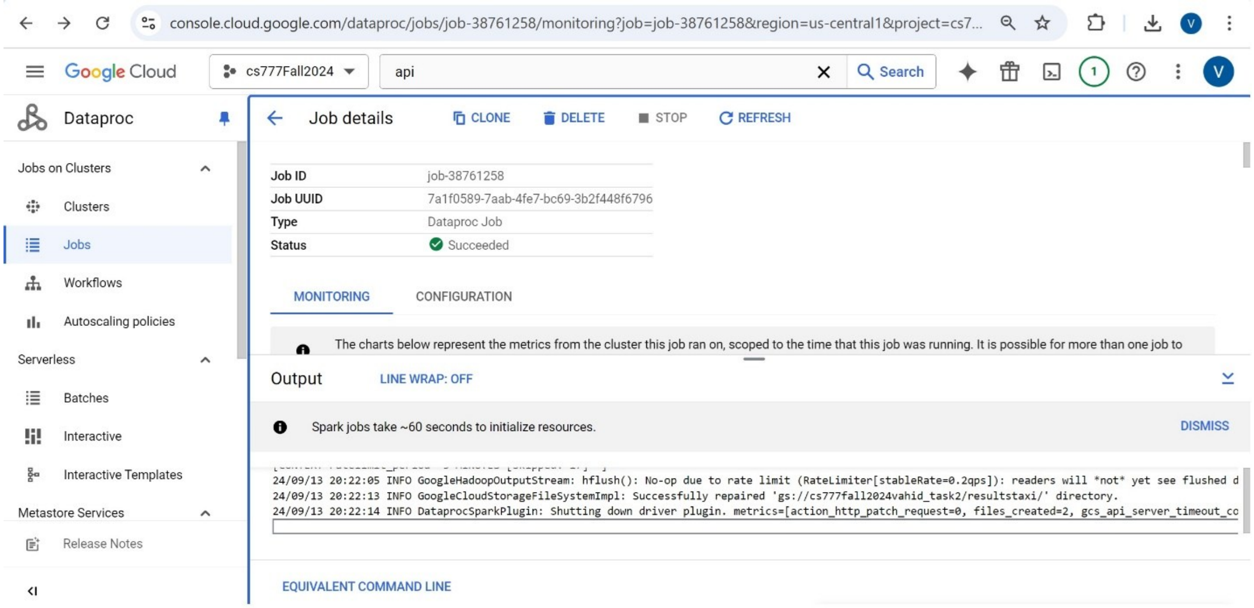
Client local time zone: America/Los_Angeles

Search:

Version	App ID	App Name	Driver Host	Started	Completed	Duration	Spark User	Last Updated	Event Log
3.5.1	application_1726257793195_0001	Top10BestDrivers	cluster-196a-m.c.cs777fall2024.internal	2024-09-13 13:14:34	2024-09-13 13:22:13	7.7 min	root	2024-09-13 13:22:14	Download

Showing 1 to 1 of 1 entries

Show incomplete applications



console.cloud.google.com/dataproc/jobs/job-38761258/monitoring?job=job-38761258&region=us-central1&project=cs7...

Job ID	job-38761258
Job UUID	7a1f0589-7aab-4fe7-bc69-3b2f448f6796
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

Spark jobs take ~60 seconds to initialize resources.

```
24/09/13 20:22:05 INFO GoogleHadoopOutputStream: hflush(): No-op due to rate limit (RateLimiter[stableRate=0.2qps]): readers will *not* yet see flushed d
24/09/13 20:22:13 INFO GoogleCloudStorageFileSystemImpl: Successfully repaired 'gs://cs777fall2024vahid_task2/resultstaxi/' directory.
24/09/13 20:22:14 INFO DataprocSparkPlugin: Shutting down driver plugin. metrics=[action_http_patch_request=0, files_created=2, gcs_api_server_timeout_co
```

The best time of the day to Work on Taxi

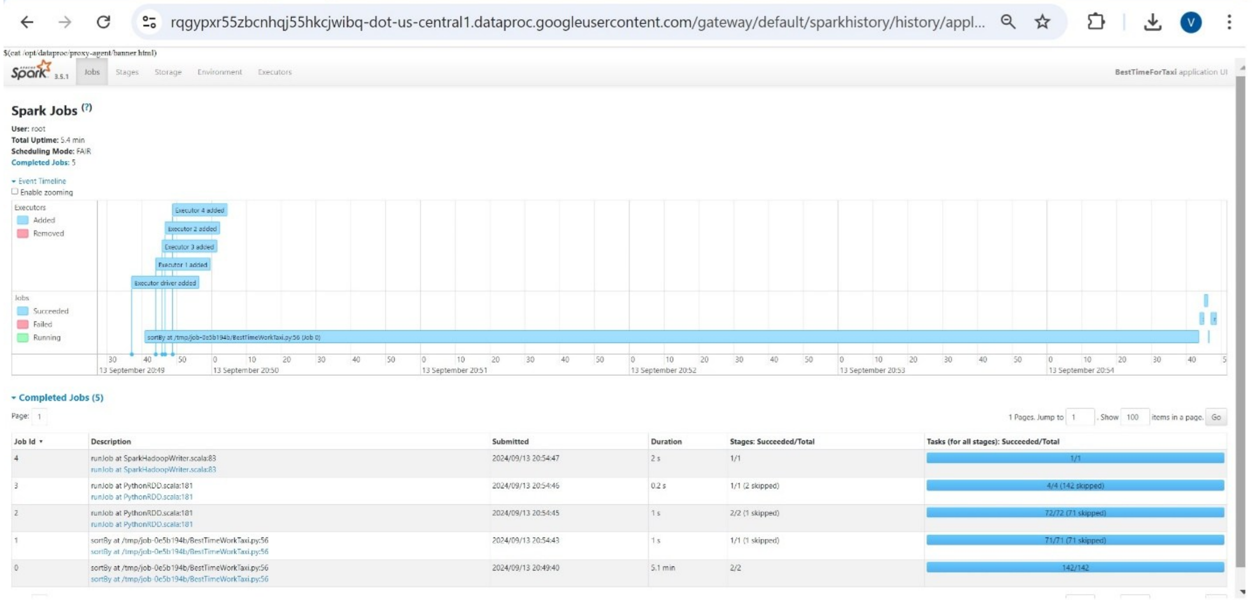
We would like to know which hour of the day is the best time for drivers that has the highest profit per mile. Consider the surcharge amount in dollars for each taxi ride (without tip amount) and the distance in miles, and sum up the rides for each hour of the day (24 hours) – consider the pickup time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

```
(18, 0.59983188861259467)
```



rqgypxr55zbcnhqj55hkcjwibq-dot-us-central1.dataproc.googleusercontent.com/gateway/default/sparkhistory/history/appl...

$(cat /opt/dataproc/proxy-agent/banner.html)

Spark Jobs (?)

User: root
Total Uptime: 5.4 min
Scheduling Mode: FAIR
Completed Jobs: 5

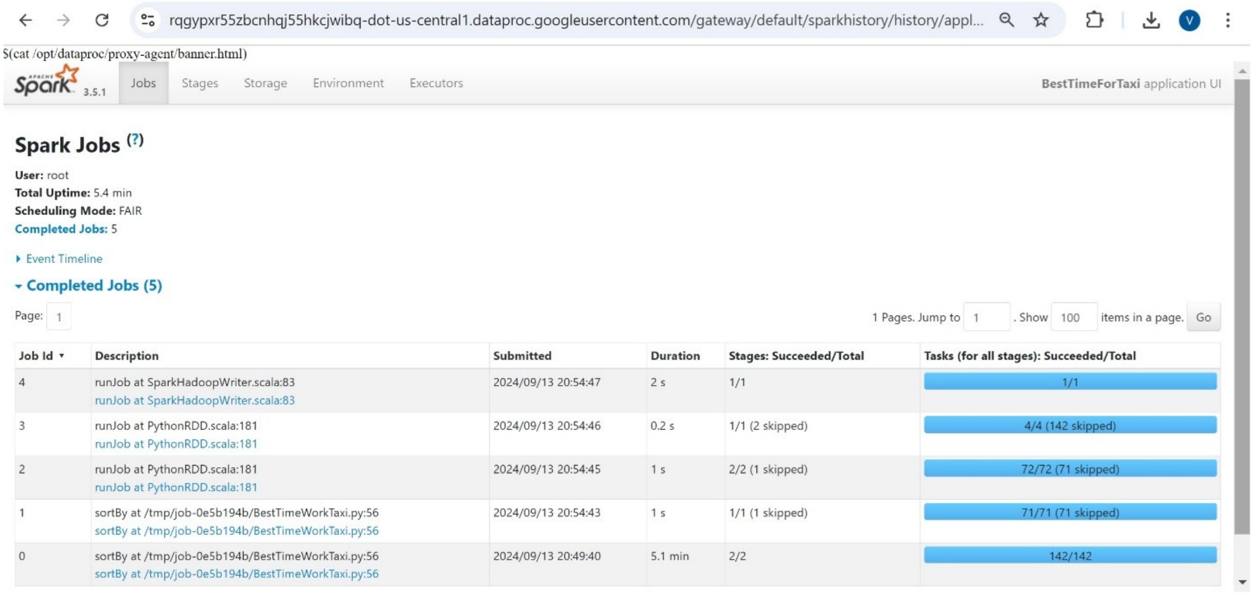
▾ Event Timeline
□ Enable zooming

▾ Completed Jobs (5)

Page: 1

1 Pages. Jump to 1 . Show 100 items in a page. Go

Job Id ▾	Description	Submitted	Duration	Stages: Succeeded/Total	Tasks (for all stages): Succeeded/Total
4	runJob at SparkHadoopWriter.scala:83 runJob at SparkHadoopWriter.scala:83	2024/09/13 20:54:47	2 s	1/1	1/1
3	runJob at PythonRDD.scala:181 runJob at PythonRDD.scala:181	2024/09/13 20:54:46	0.2 s	1/1 (2 skipped)	4/4 (142 skipped)
2	runJob at PythonRDD.scala:181 runJob at PythonRDD.scala:181	2024/09/13 20:54:45	1 s	2/2 (1 skipped)	72/72 (71 skipped)
1	sortBy at /tmp/job-0e5b194b/BestTimeWorkTaxi.py:56 sortBy at /tmp/job-0e5b194b/BestTimeWorkTaxi.py:56	2024/09/13 20:54:43	1 s	1/1 (1 skipped)	71/71 (71 skipped)
0	sortBy at /tmp/job-0e5b194b/BestTimeWorkTaxi.py:56 sortBy at /tmp/job-0e5b194b/BestTimeWorkTaxi.py:56	2024/09/13 20:49:40	5.1 min	2/2	142/142



rqgypxr55zbcnhqj55hkcjwibq-dot-us-central1.dataproc.googleusercontent.com/gateway/default/sparkhistory/history/appl...

$(cat /opt/dataproc/proxy-agent/banner.html)

Spark 3.5.1 Jobs Stages Storage Environment Executors BestTimeForTaxi application UI

Spark Jobs (?)

User: root
Total Uptime: 5.4 min
Scheduling Mode: FAIR
Completed Jobs: 5

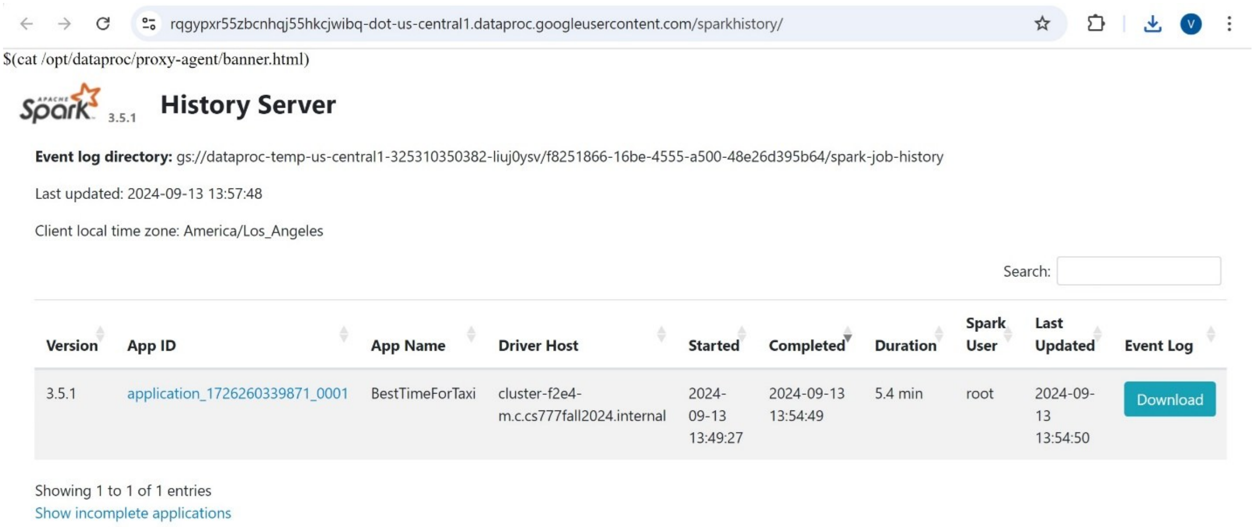
▸ Event Timeline

▾ Completed Jobs (5)

Page: 1

1 Pages. Jump to 1 . Show 100 items in a page. Go

Job Id ▾	Description	Submitted	Duration	Stages: Succeeded/Total	Tasks (for all stages): Succeeded/Total
4	runJob at SparkHadoopWriter.scala:83 runJob at SparkHadoopWriter.scala:83	2024/09/13 20:54:47	2 s	1/1	1/1
3	runJob at PythonRDD.scala:181 runJob at PythonRDD.scala:181	2024/09/13 20:54:46	0.2 s	1/1 (2 skipped)	4/4 (142 skipped)
2	runJob at PythonRDD.scala:181 runJob at PythonRDD.scala:181	2024/09/13 20:54:45	1 s	2/2 (1 skipped)	72/72 (71 skipped)
1	sortBy at /tmp/job-0e5b194b/BestTimeWorkTaxi.py:56 sortBy at /tmp/job-0e5b194b/BestTimeWorkTaxi.py:56	2024/09/13 20:54:43	1 s	1/1 (1 skipped)	71/71 (71 skipped)
0	sortBy at /tmp/job-0e5b194b/BestTimeWorkTaxi.py:56 sortBy at /tmp/job-0e5b194b/BestTimeWorkTaxi.py:56	2024/09/13 20:49:40	5.1 min	2/2	142/142



rqgypxr55zbcnhqj55hkcjwibq-dot-us-central1.dataproc.googleusercontent.com/sparkhistory/

$(cat /opt/dataproc/proxy-agent/banner.html)

Spark 3.5.1 **History Server**

Event log directory: gs://dataproc-temp-us-central1-325310350382-liuj0ysv/f8251866-16be-4555-a500-48e26d395b64/spark-job-history

Last updated: 2024-09-13 13:57:48

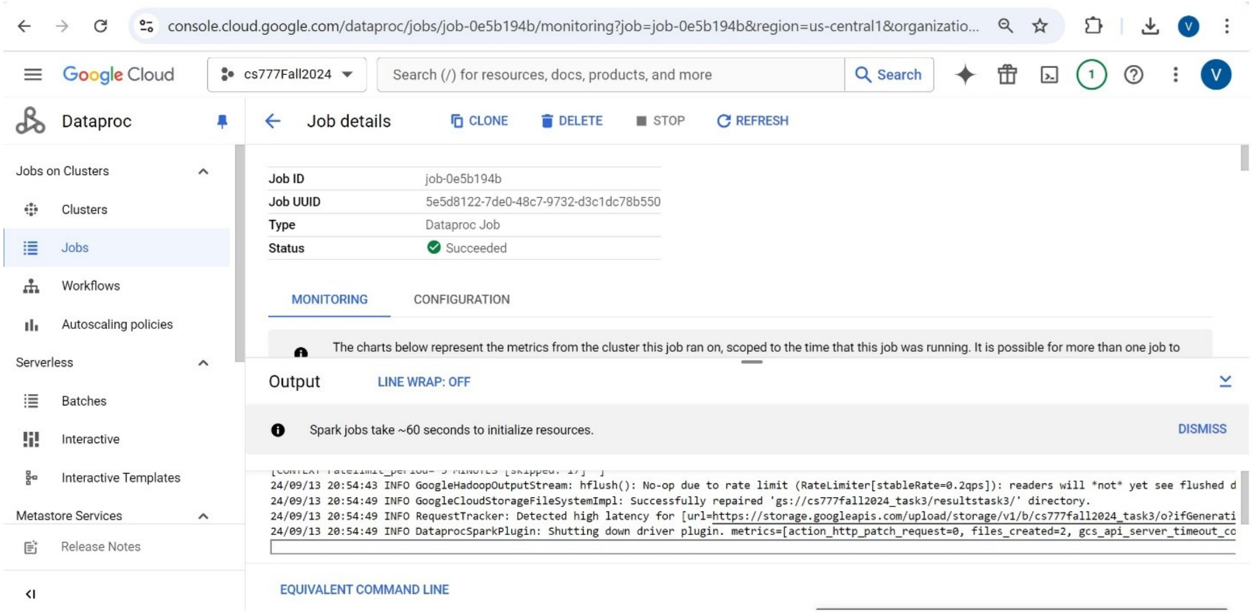
Client local time zone: America/Los_Angeles

Search:

Version	App ID	App Name	Driver Host	Started	Completed	Duration	Spark User	Last Updated	Event Log
3.5.1	application_1726260339871_0001	BestTimeForTaxi	cluster-f2e4-m.c.cs777fall2024.internal	2024-09-13 13:49:27	2024-09-13 13:54:49	5.4 min	root	2024-09-13 13:54:50	Download

Showing 1 to 1 of 1 entries

Show incomplete applications



console.cloud.google.com/dataproc/jobs/job-0e5b194b/monitoring?job=job-0e5b194b&region=us-central1&organizatio...

Job ID	job-0e5b194b
Job UUID	5e5d8122-7de0-48c7-9732-d3c1dc78b550
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

Spark jobs take ~60 seconds to initialize resources.

```
[CONTEXT ratelimit_period="5 MINUTES" [skipped: 17] ]
24/09/13 20:54:43 INFO GoogleHadoopOutputStream: hflush(): No-op due to rate limit (RateLimiter[stableRate=0.2qps]): readers will *not* yet see flushed d
24/09/13 20:54:49 INFO GoogleCloudStorageFileSystemImpl: Successfully repaired 'gs://cs777fall2024_task3/resultstask3/' directory.
24/09/13 20:54:49 INFO RequestTracker: Detected high latency for [url=https://storage.googleapis.com/upload/storage/v1/b/cs777fall2024_task3/o?ifGenerati
24/09/13 20:54:49 INFO DataprocSparkPlugin: Shutting down driver plugin. metrics=[action_http_patch_request=0, files_created=2, gcs_api_server_timeout_co
```

Advanced Part

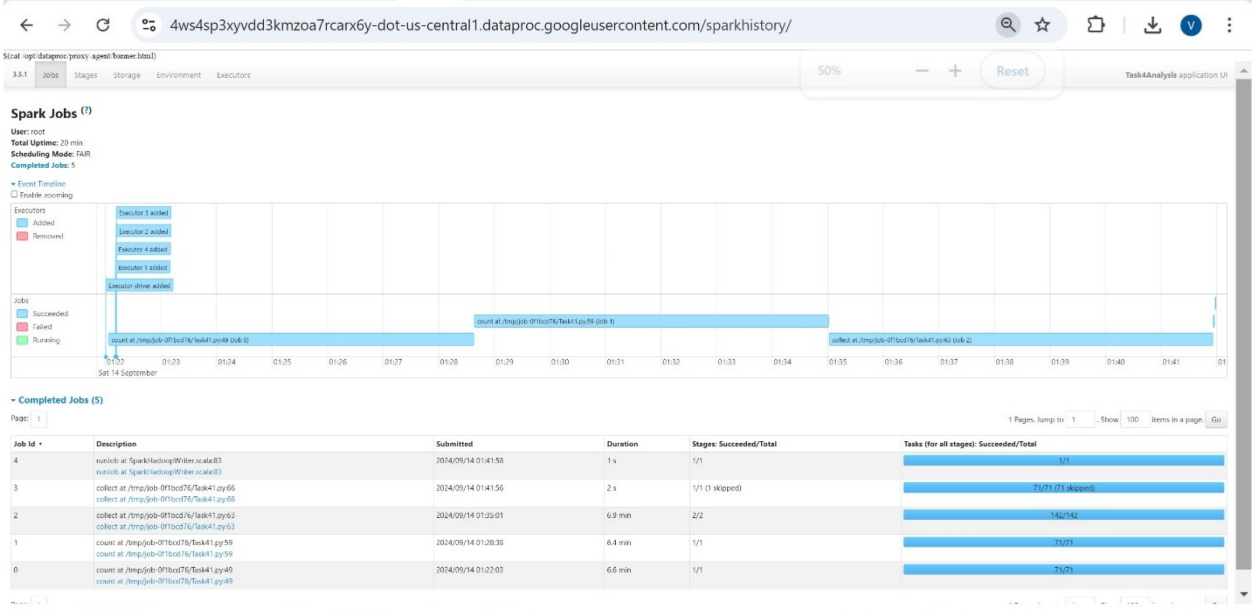
Here are some further tasks for advanced groups.

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

(0, 56.56)
(1, 56.16)

```
(2, 55.82)
(3, 54.43)
(4, 48.83)
(5, 51.68)
(6, 54.85)
(7, 57.59)
(8, 59.93)
(9, 57.88)
(10, 52.75)
(11, 50.95)
(12, 50.63)
(13, 50.07)
(14, 49.75)
(15, 49.19)
(16, 49.46)
(17, 51.5)
(18, 54.12)
(19, 55.76)
(20, 57.8)
(21, 57.52)
(22, 57.34)
(23, 57.18)
```



$(cat /opt/dataproc/proxy-agent/banner.html)

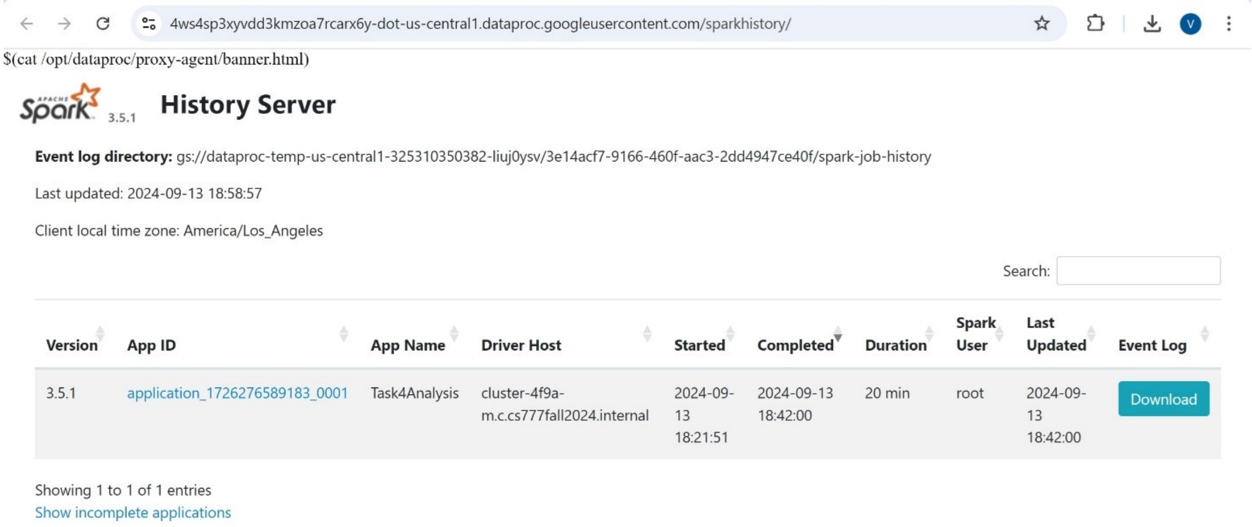
Spark Jobs (?)

User: root
Total Uptime: 20 min
Scheduling Mode: FAIR
Completed Jobs: 5

Completed Jobs (5)

Page: 1 1 Pages. Jump to 1 . Show 100 items in a page. Go

Job Id ▾	Description	Submitted	Duration	Stages: Succeeded/Total	Tasks (for all stages): Succeeded/Total
4	runJob at SparkHadoopWriter.scala:83 runJob at SparkHadoopWriter.scala:83	2024/09/14 01:41:58	1 s	1/1	1/1
3	collect at /tmp/job-0f1bcd76/Task41.py:66 collect at /tmp/job-0f1bcd76/Task41.py:66	2024/09/14 01:41:56	2 s	1/1 (1 skipped)	71/71 (71 skipped)
2	collect at /tmp/job-0f1bcd76/Task41.py:63 collect at /tmp/job-0f1bcd76/Task41.py:63	2024/09/14 01:35:01	6.9 min	2/2	142/142
1	count at /tmp/job-0f1bcd76/Task41.py:59 count at /tmp/job-0f1bcd76/Task41.py:59	2024/09/14 01:28:38	6.4 min	1/1	71/71
0	count at /tmp/job-0f1bcd76/Task41.py:49 count at /tmp/job-0f1bcd76/Task41.py:49	2024/09/14 01:22:03	6.6 min	1/1	71/71


4ws4sp3xyvdd3kmzoa7rcarx6y-dot-us-central1.dataproc.googleusercontent.com/sparkhistory/
Apache Spark 3.5.1

$(cat /opt/dataproc/proxy-agent/banner.html)

History Server

Event log directory: gs://dataproc-temp-us-central1-325310350382-liuj0ysv/3e14acf7-9166-460f-aac3-2dd4947ce40f/spark-job-history

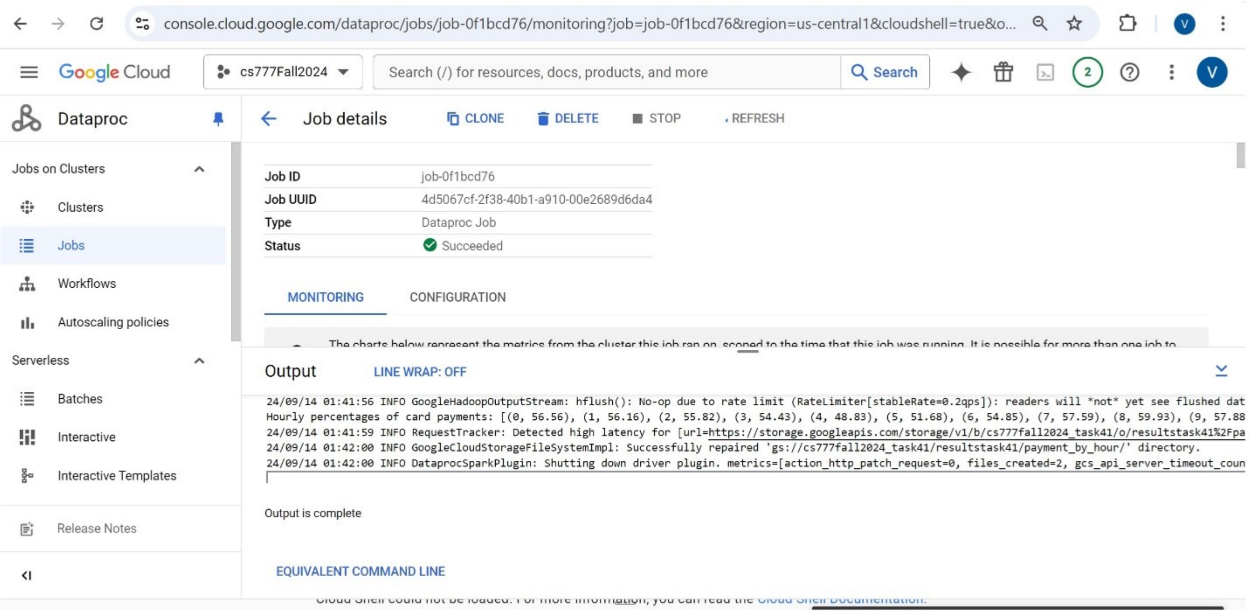
Last updated: 2024-09-13 18:58:57

Client local time zone: America/Los_Angeles

Search:

Version	App ID	App Name	Driver Host	Started	Completed	Duration	Spark User	Last Updated	Event Log
3.5.1	application_1726276589183_0001	Task4Analysis	cluster-4f9a-m.c.cs777fall2024.internal	2024-09-13 18:21:51	2024-09-13 18:42:00	20 min	root	2024-09-13 18:42:00	Download

Showing 1 to 1 of 1 entries
Show incomplete applications



Job ID job-0f1bcd76
Job UUID 4d5067cf-2f38-40b1-a910-00e2689d6da4
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

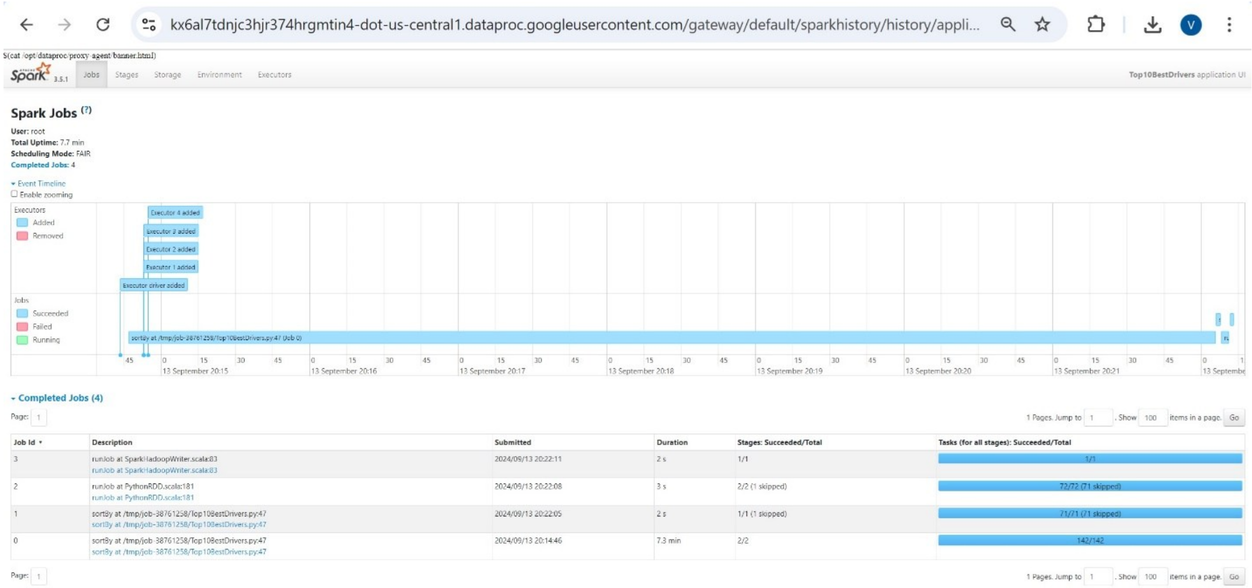
```
24/09/14 01:41:56 INFO GoogleHadoopOutputStream: hflush(): No-op due to rate limit (RateLimiter[stableRate=0.2qps]): readers will *not* yet see flushed dat
Hourly percentages of card payments: [(0, 56.56), (1, 56.16), (2, 55.82), (3, 54.43), (4, 48.83), (5, 51.68), (6, 54.85), (7, 57.59), (8, 59.93), (9, 57.88
24/09/14 01:41:59 INFO RequestTracker: Detected high latency for [url=https://storage.googleapis.com/storage/v1/b/cs777fall2024_task41/o/resultstask41%2Fpa
24/09/14 01:42:00 INFO GoogleCloudStorageFileSystemImpl: Successfully repaired 'gs://cs777fall2024_task41/resultstask41/payment_by_hour/' directory.
24/09/14 01:42:00 INFO DataprocSparkPlugin: Shutting down driver plugin. metrics=[action_http_patch_request=0, files_created=2, gcs_api_server_timeout_coun
```

Output is complete

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

```
('E4F99C9ABE9861F18BCD38BC63D007A9', 62103.18601982026)
('664927CDE376A32789BA48BF55DFB7E3', 32075.6481781048)
('51C1BE97280A80EBFA8DAD34E1956CF6', 22412.519037907998)
('D85749E8852FCC66A990E40605607B2F', 21838.61138628341)
('23DF80C977D15141F11DD713C523C311', 20791.02908617619)
('3D757E111C78F5CAC83D44A92885D490', 20555.115544088956)
('BE047851D97506885B99BDDFA7A13360', 19584.744671992667)
('3AAB94CA53FE93A64811F65690654649', 19575.05448827395)
('74CC809D28AE726DDB32249C044DA4F8', 19558.588651912727)
('3578782ABD6492CEB927B2A8EBCF1402', 19321.71774748813)
```



kx6al7tdnjc3hjr374hrgmtin4-dot-us-central1.dataproc.googleusercontent.com/gateway/default/sparkhistory/history/appli...
$(cat /opt/dataproc/proxy-agent/banner.html)
Spark 3.5.1
Jobs
Stages
Storage
Environment
Executors
Top10BestDrivers application UI

Spark Jobs (?)

User: root
Total Uptime: 7.7 min
Scheduling Mode: FAIR
Completed Jobs: 4

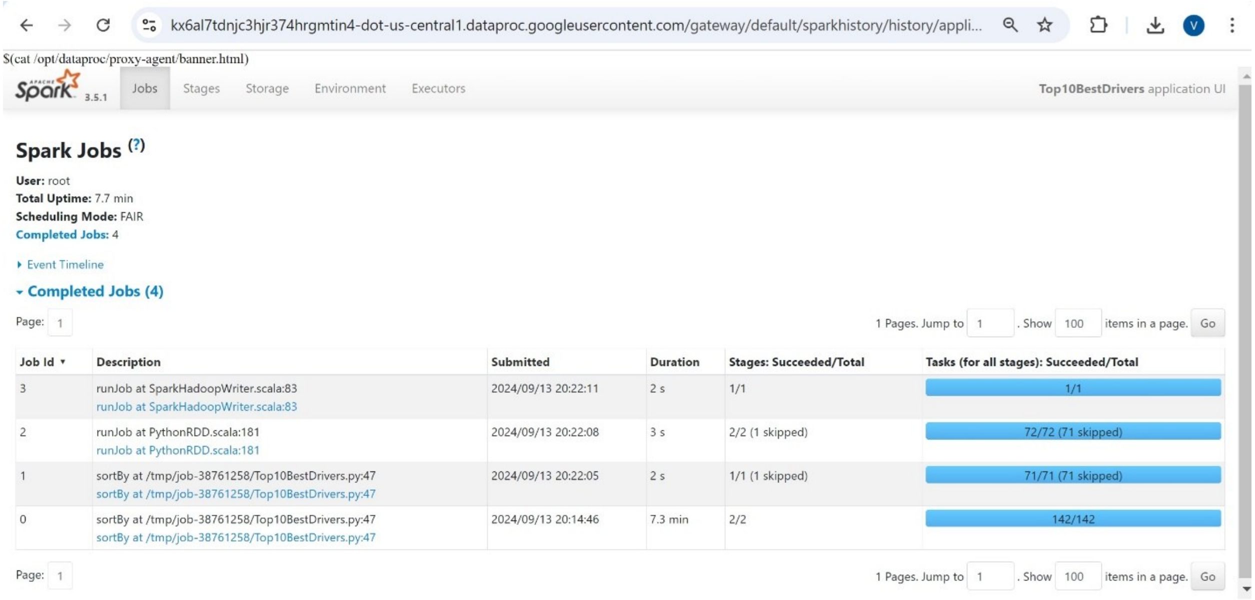
▾ Event Timeline
□ Enable zooming

▾ Completed Jobs (4)

Page: 1 — 1 Pages. Jump to 1 . Show 100 items in a page. Go

Job Id ▾	Description	Submitted	Duration	Stages: Succeeded/Total	Tasks (for all stages): Succeeded/Total
3	runJob at SparkHadoopWriter.scala:83 runJob at SparkHadoopWriter.scala:83	2024/09/13 20:22:11	2 s	1/1	1/1
2	runJob at PythonRDD.scala:181 runJob at PythonRDD.scala:181	2024/09/13 20:22:08	3 s	2/2 (1 skipped)	72/72 (71 skipped)
1	sortBy at /tmp/job-38761258/Top10BestDrivers.py:47 sortBy at /tmp/job-38761258/Top10BestDrivers.py:47	2024/09/13 20:22:05	2 s	1/1 (1 skipped)	71/71 (71 skipped)
0	sortBy at /tmp/job-38761258/Top10BestDrivers.py:47 sortBy at /tmp/job-38761258/Top10BestDrivers.py:47	2024/09/13 20:14:46	7.3 min	2/2	142/142

Page: 1 — 1 Pages. Jump to 1 . Show 100 items in a page. Go



kx6al7tdnjc3hjr374hrgmtin4-dot-us-central1.dataproc.googleusercontent.com/gateway/default/sparkhistory/history/appli...

$(cat /opt/dataproc/proxy-agent/banner.html)

Spark 3.5.1 Jobs Stages Storage Environment Executors — Top10BestDrivers application UI

Spark Jobs (?)

User: root
Total Uptime: 7.7 min
Scheduling Mode: FAIR
Completed Jobs: 4

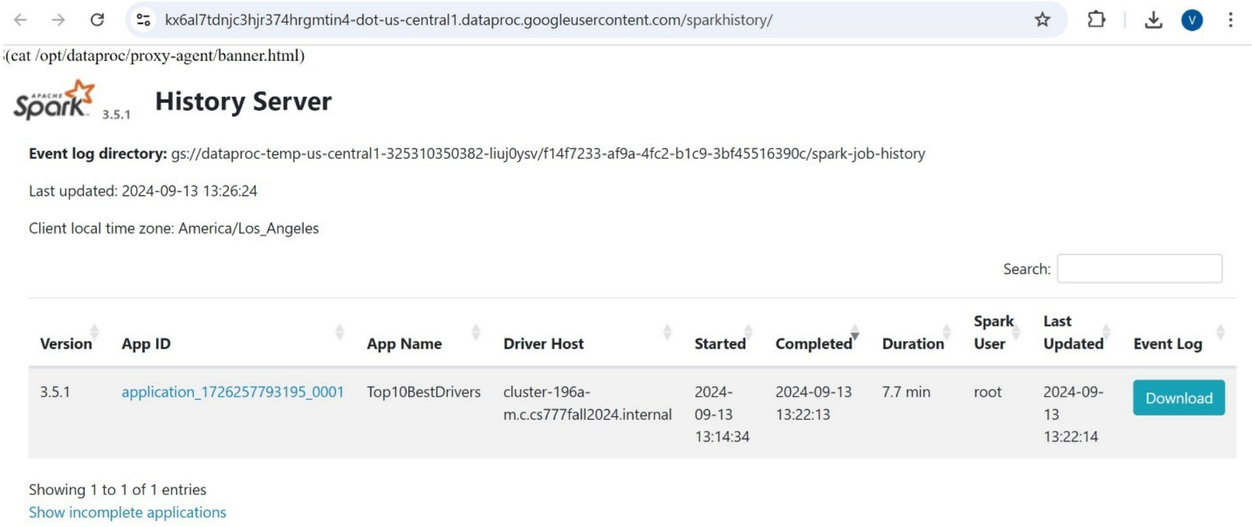
▸ Event Timeline

▾ Completed Jobs (4)

Page: 1 — 1 Pages. Jump to 1 . Show 100 items in a page. Go

Job Id ▾	Description	Submitted	Duration	Stages: Succeeded/Total	Tasks (for all stages): Succeeded/Total
3	runJob at SparkHadoopWriter.scala:83 runJob at SparkHadoopWriter.scala:83	2024/09/13 20:22:11	2 s	1/1	1/1
2	runJob at PythonRDD.scala:181 runJob at PythonRDD.scala:181	2024/09/13 20:22:08	3 s	2/2 (1 skipped)	72/72 (71 skipped)
1	sortBy at /tmp/job-38761258/Top10BestDrivers.py:47 sortBy at /tmp/job-38761258/Top10BestDrivers.py:47	2024/09/13 20:22:05	2 s	1/1 (1 skipped)	71/71 (71 skipped)
0	sortBy at /tmp/job-38761258/Top10BestDrivers.py:47 sortBy at /tmp/job-38761258/Top10BestDrivers.py:47	2024/09/13 20:14:46	7.3 min	2/2	142/142

Page: 1 — 1 Pages. Jump to 1 . Show 100 items in a page. Go



kx6al7tdnjc3hjr374hrgmtin4-dot-us-central1.dataproc.googleusercontent.com/sparkhistory/

:(cat /opt/dataproc/proxy-agent/banner.html)

3.5.1 History Server

Event log directory: gs://dataproc-temp-us-central1-325310350382-liuj0ysv/f14f7233-af9a-4fc2-b1c9-3bf45516390c/spark-job-history

Last updated: 2024-09-13 13:26:24

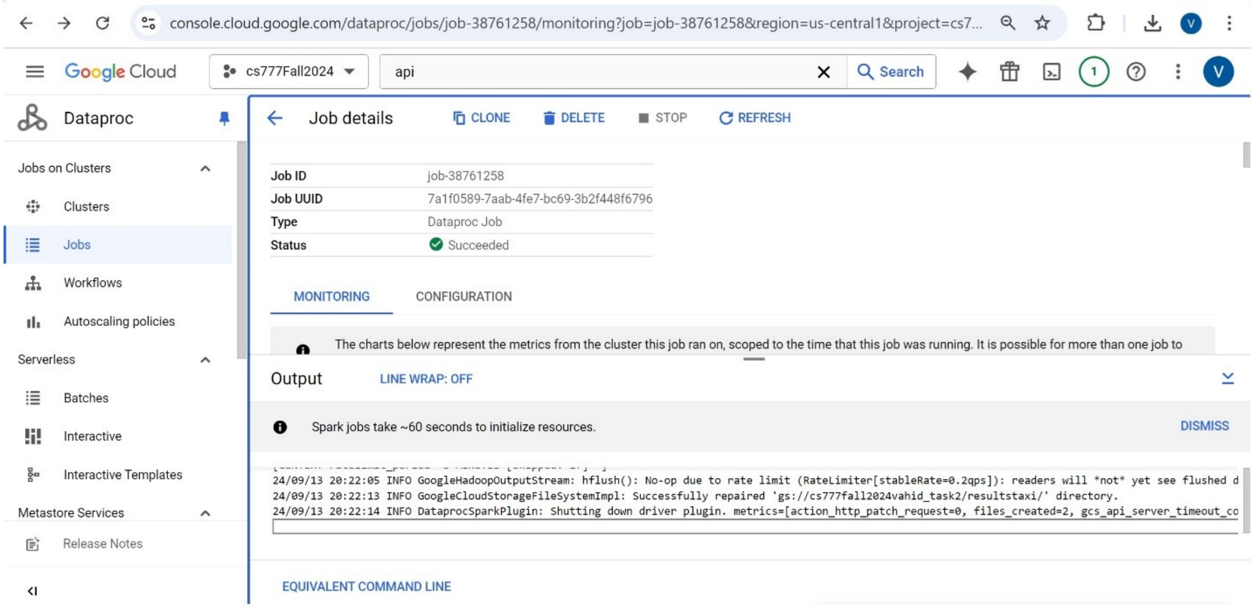
Client local time zone: America/Los_Angeles

Search:

Version	App ID	App Name	Driver Host	Started	Completed	Duration	Spark User	Last Updated	Event Log
3.5.1	application_1726257793195_0001	Top10BestDrivers	cluster-196a-m.c.cs777fall2024.internal	2024-09-13 13:14:34	2024-09-13 13:22:13	7.7 min	root	2024-09-13 13:22:14	Download

Showing 1 to 1 of 1 entries

Show incomplete applications



console.cloud.google.com/dataproc/jobs/job-38761258/monitoring?job=job-38761258&region=us-central1&project=cs7...

Job ID	job-38761258
Job UUID	7a1f0589-7aab-4fe7-bc69-3b2f448f6796
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

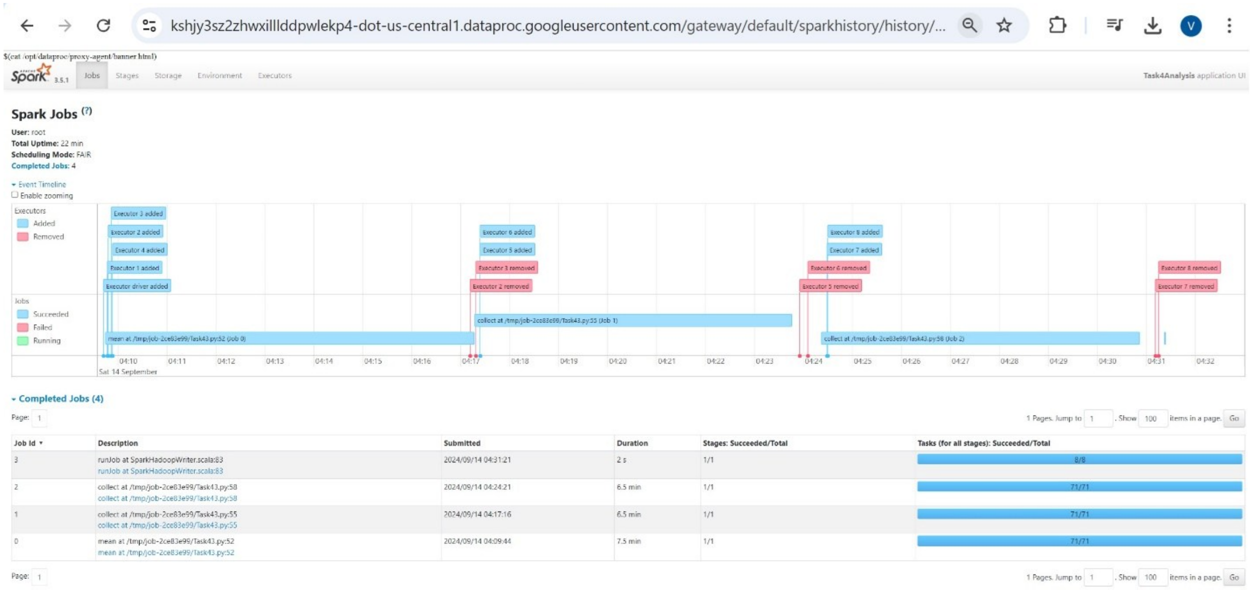
Spark jobs take ~60 seconds to initialize resources.

```
24/09/13 20:22:05 INFO GoogleHadoopOutputStream: hflush(): No-op due to rate limit (RateLimiter[stableRate=0.2qps]): readers will *not* yet see flushed d
24/09/13 20:22:13 INFO GoogleCloudStorageFileSystemImpl: Successfully repaired 'gs://cs777fall2024vahid_task2/resultstaxi/' directory.
24/09/13 20:22:14 INFO DataprocSparkPlugin: Shutting down driver plugin. metrics=[action_http_patch_request=0, files_created=2, gcs_api_server_timeout_co
```

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

Mean Tip: 1.365121487449525
Median Tip: 1.0
Q1: 0.0
Q3: 2.0



Spark Jobs (?)

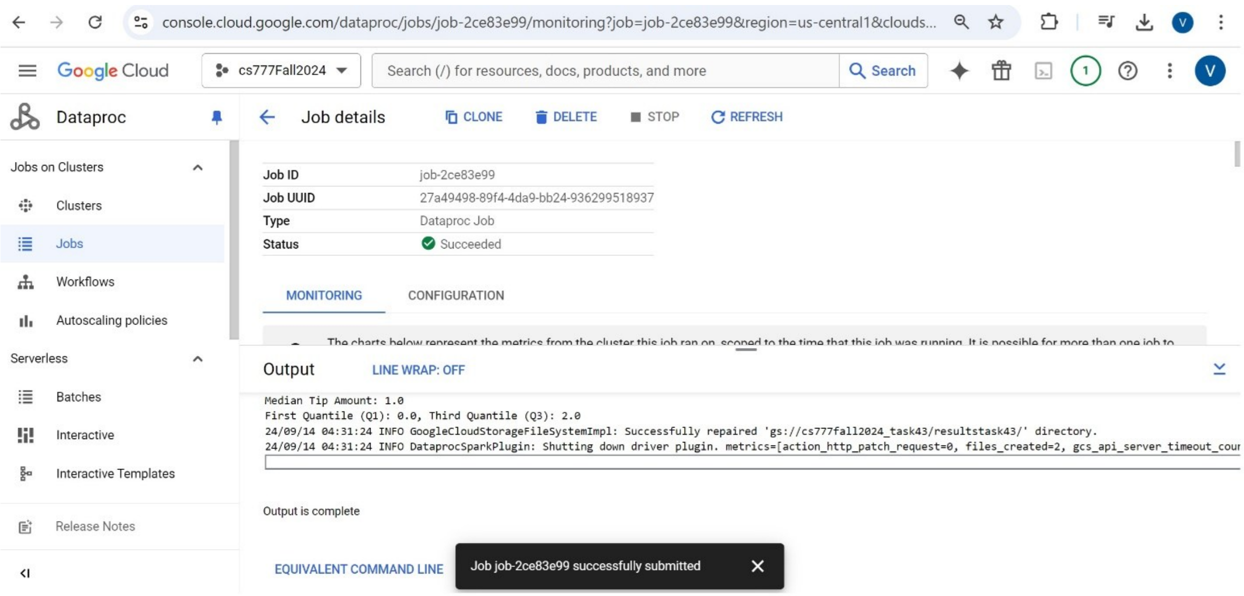
User: root
Total Uptime: 22 min
Scheduling Mode: FAIR
Completed Jobs: 4

Completed Jobs (4)

Page: 1 — 1 Pages. Jump to 1 . Show 100 items in a page. Go

Job Id ▾	Description	Submitted	Duration	Stages: Succeeded/Total	Tasks (for all stages): Succeeded/Total
3	runJob at SparkHadoopWriter.scala:83 runJob at SparkHadoopWriter.scala:83	2024/09/14 04:31:21	2 s	1/1	8/8
2	collect at /tmp/job-2ce83e99/Task43.py:58 collect at /tmp/job-2ce83e99/Task43.py:58	2024/09/14 04:24:21	6.5 min	1/1	71/71
1	collect at /tmp/job-2ce83e99/Task43.py:55 collect at /tmp/job-2ce83e99/Task43.py:55	2024/09/14 04:17:16	6.5 min	1/1	71/71
0	mean at /tmp/job-2ce83e99/Task43.py:52 mean at /tmp/job-2ce83e99/Task43.py:52	2024/09/14 04:09:44	7.5 min	1/1	71/71

Page: 1 — 1 Pages. Jump to 1 . Show 100 items in a page. Go



Job ID	job-2ce83e99
Job UUID	27a49498-89f4-4da9-bb24-936299518937
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

```
Median Tip Amount: 1.0
First Quantile (Q1): 0.0, Third Quantile (Q3): 2.0
24/09/14 04:31:24 INFO GoogleCloudStorageFileSystemImpl: Successfully repaired 'gs://cs777fall2024_task43/resultstask43/' directory.
24/09/14 04:31:24 INFO DataprocSparkPlugin: Shutting down driver plugin. metrics=[action_http_patch_request=0, files_created=2, gcs_api_server_timeout_cour
```

Output is complete

Job job-2ce83e99 successfully submitted ✕

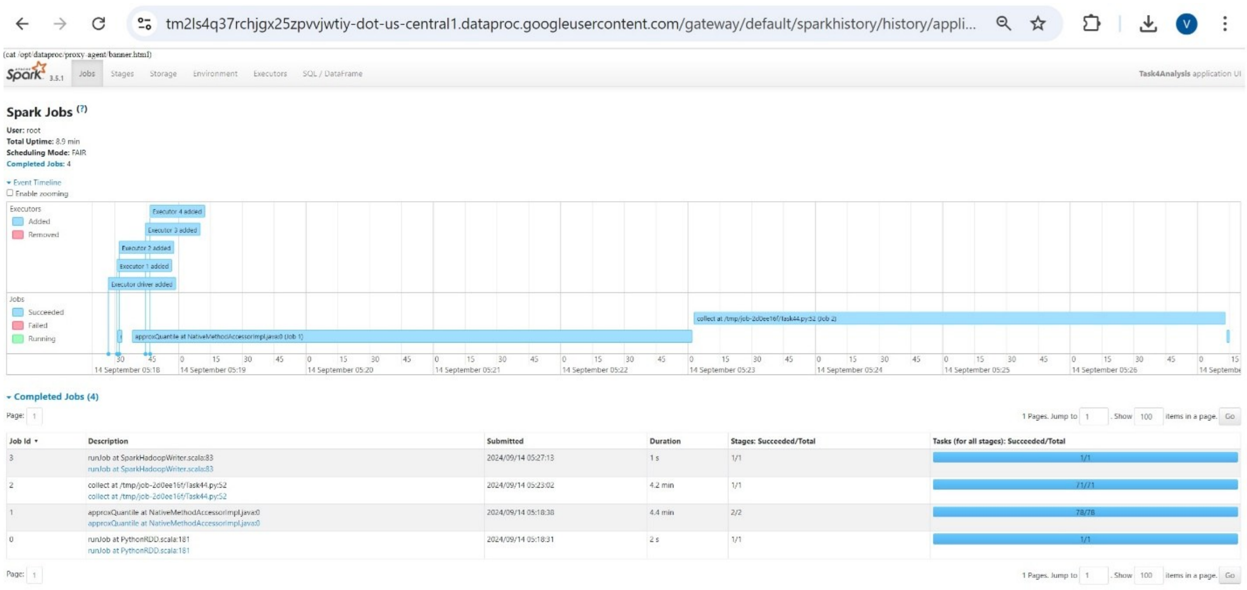
- Explain how do you find the median

the median is calculated using PySpark's approxQuantile function.
The median is calculated as the 50th percentile (or second quartile, Q2) of the "tip_amount" column using the approxQuantile function. This method is efficient for large datasets and provides an approximate result with a specified error tolerance (0.01 in this case, meaning a 1% tolerance).

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

- Find the top 10 outliers

Top 10 Outliers: [685908.1, 541432.56, 159001.16, 82241.97, 61553.03, 9000.65, 7683.47, 7659.17, 7658.47, 5661.47]



tm2ls4q37rchjgx25zpvvjwtiy-dot-us-central1.dataproc.googleusercontent.com/gateway/default/sparkhistory/history/appli...

(cat /opt/dataproc/proxy-agent/banner.html)

Spark Jobs (?)

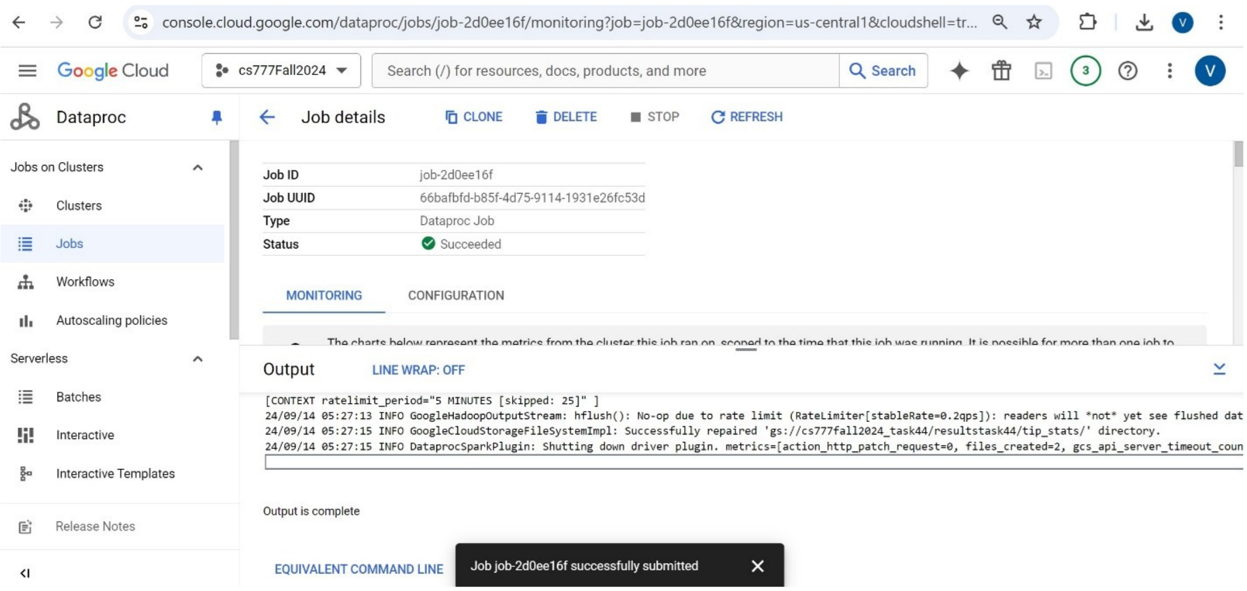
User: root
Total Uptime: 8.9 min
Scheduling Mode: FAIR
Completed Jobs: 4

▾ Completed Jobs (4)

Page: 1 — 1 Pages. Jump to 1 . Show 100 items in a page. Go

Job Id ▾	Description	Submitted	Duration	Stages: Succeeded/Total	Tasks (for all stages): Succeeded/Total
3	runJob at SparkHadoopWriter.scala:83 runJob at SparkHadoopWriter.scala:83	2024/09/14 05:27:13	1 s	1/1	1/1
2	collect at /tmp/job-2d0ee16f/Task44.py:52 collect at /tmp/job-2d0ee16f/Task44.py:52	2024/09/14 05:23:02	4.2 min	1/1	71/71
1	approxQuantile at NativeMethodAccessorImpl.java:0 approxQuantile at NativeMethodAccessorImpl.java:0	2024/09/14 05:18:38	4.4 min	2/2	78/78
0	runJob at PythonRDD.scala:181 runJob at PythonRDD.scala:181	2024/09/14 05:18:31	2 s	1/1	1/1

Page: 1 — 1 Pages. Jump to 1 . Show 100 items in a page. Go



console.cloud.google.com/dataproc/jobs/job-2d0ee16f/monitoring?job=job-2d0ee16f&region=us-central1&cloudshell=tr...

Job ID	job-2d0ee16f
Job UUID	66bafbfd-b85f-4d75-9114-1931e26fc53d
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

Output

```
[CONTEXT ratelimit_period="5 MINUTES [skipped: 25]" ]
24/09/14 05:27:13 INFO GoogleHadoopOutputStream: hflush(): No-op due to rate limit (RateLimiter[stableRate=0.2qps]): readers will *not* yet see flushed dat
24/09/14 05:27:15 INFO GoogleCloudStorageFileSystemImpl: Successfully repaired 'gs://cs777fall2024_task44/resultstask44/tip_stats/' directory.
24/09/14 05:27:15 INFO DataprocSparkPlugin: Shutting down driver plugin. metrics=[action_http_patch_request=0, files_created=2, gcs_api_server_timeout_coun
```

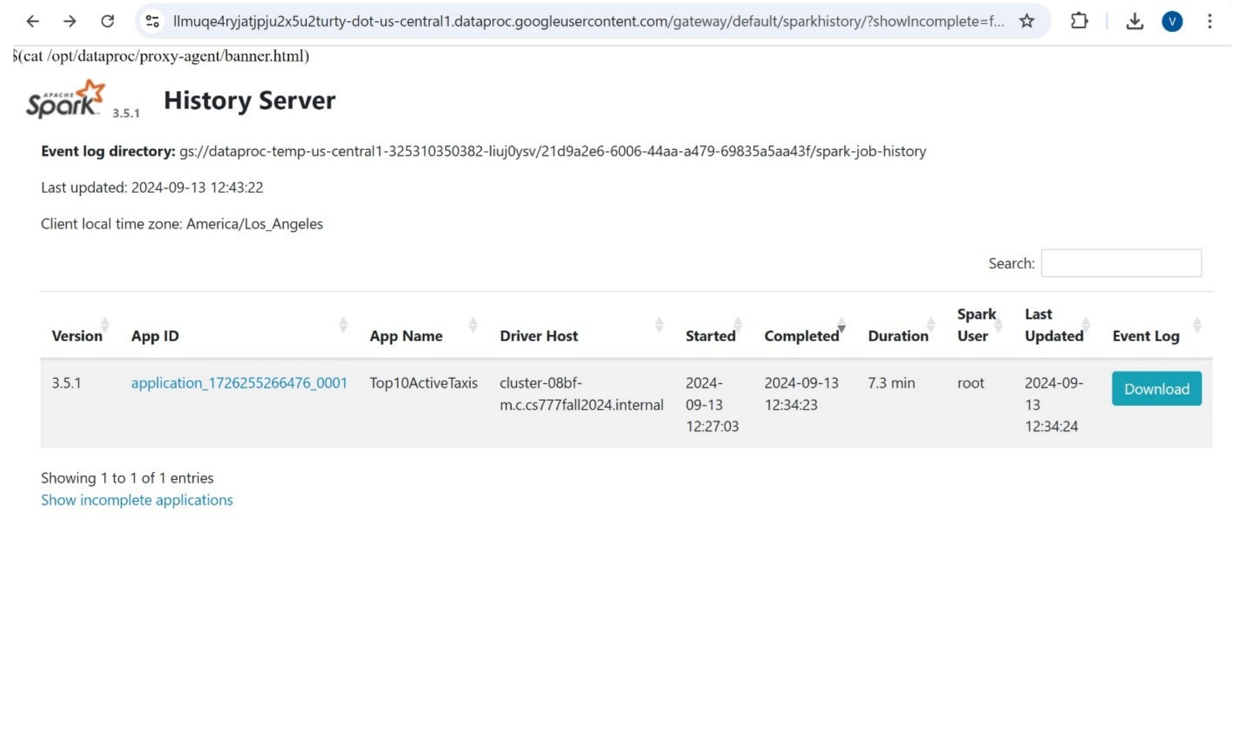
Output is complete

Job job-2d0ee16f successfully submitted ✕

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.

Top-10 Active Taxis


llmuqe4ryjatjpju2x5u2turty-dot-us-central1.dataproc.googleusercontent.com/gateway/default/sparkhistory/?showIncomplete=f...

$(cat /opt/dataproc/proxy-agent/banner.html)

3.5.1 **History Server**

Event log directory: gs://dataproc-temp-us-central1-325310350382-liuj0ysv/21d9a2e6-6006-44aa-a479-69835a5aa43f/spark-job-history

Last updated: 2024-09-13 12:43:22

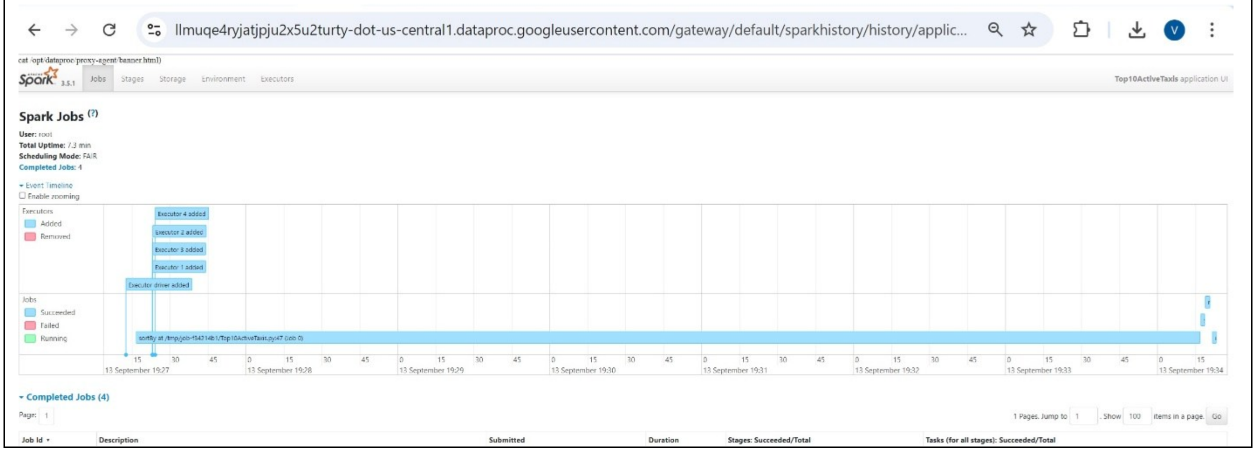
Client local time zone: America/Los_Angeles

Search:

Version	App ID	App Name	Driver Host	Started	Completed	Duration	Spark User	Last Updated	Event Log
3.5.1	application_1726255266476_0001	Top10ActiveTaxis	cluster-08bf-m.c.cs777fall2024.internal	2024-09-13 12:27:03	2024-09-13 12:34:23	7.3 min	root	2024-09-13 12:34:24	Download

Showing 1 to 1 of 1 entries

Show incomplete applications


llmuqe4ryjatjpju2x5u2turty-dot-us-central1.dataproc.googleusercontent.com/gateway/default/sparkhistory/history/applic...
cat /opt/dataproc/proxy-agent/banner.html)
3.5.1
Jobs
Stages
Storage
Environment
Executors
Top10ActiveTaxis application UI

Spark Jobs (?)

User: root
Total Uptime: 7.3 min
Scheduling Mode: FAIR
Completed Jobs: 4

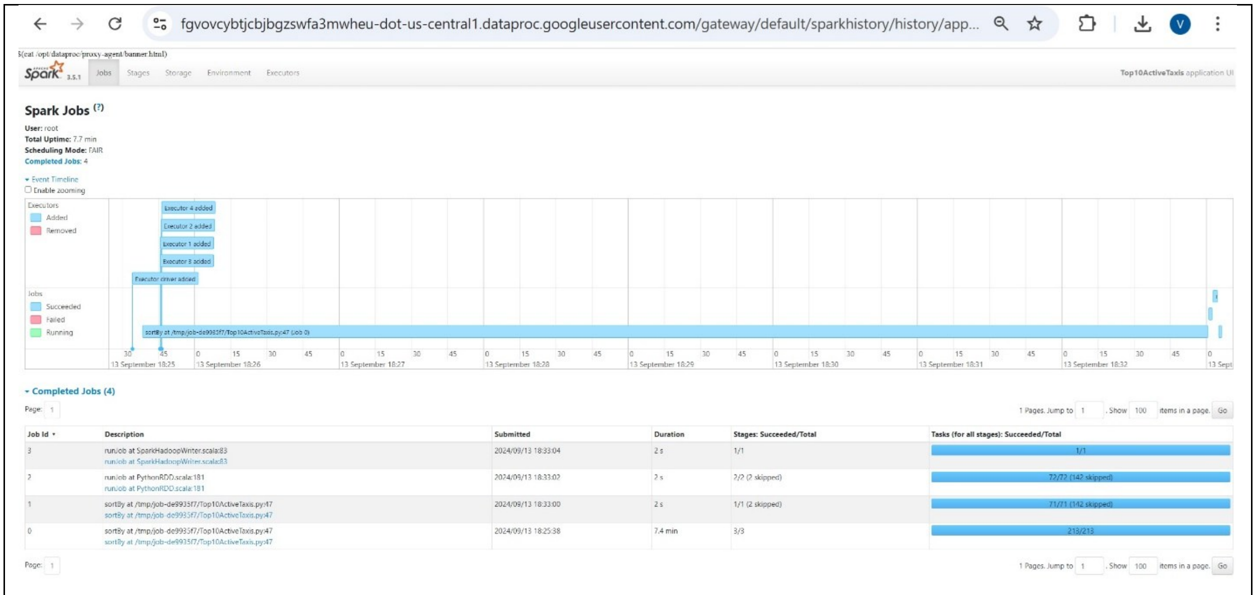
▾ Event Timeline
□ Enable zooming

▾ **Completed Jobs (4)**

Page: 1

1 Pages. Jump to 1 . Show 100 items in a page. Go

Job Id ▾	Description	Submitted	Duration	Stages: Succeeded/Total	Tasks (for all stages): Succeeded/Total



$(cat /opt/dataproc/proxy-agent/banner.html)

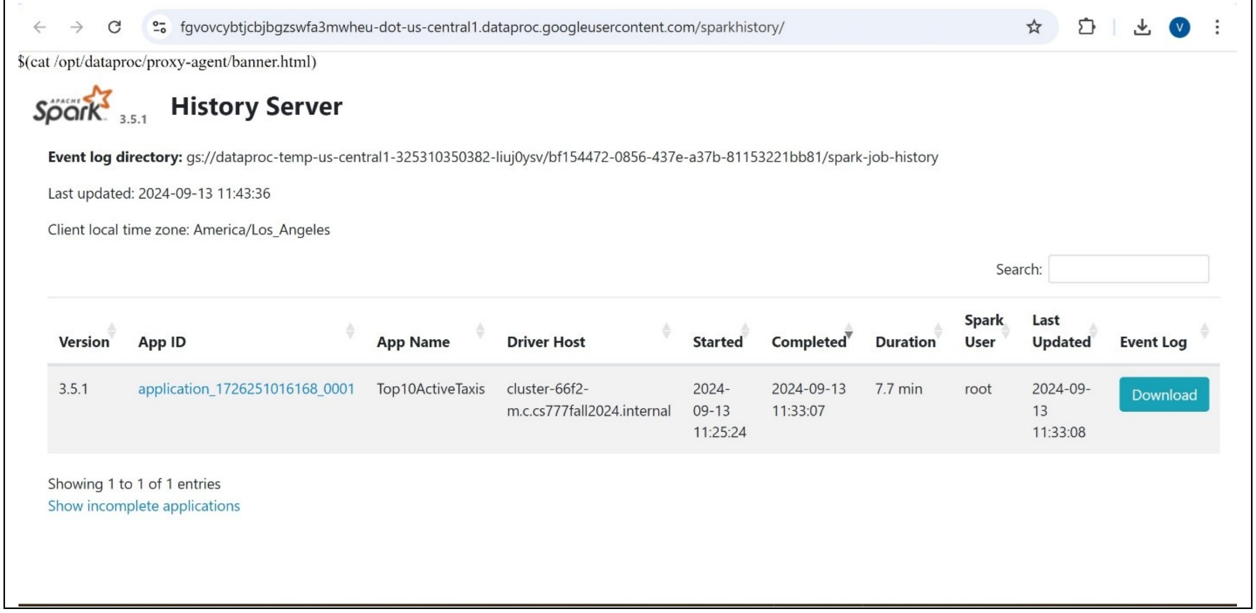
Spark Jobs (?)

User: root
Total Uptime: 7.7 min
Scheduling Mode: FAIR
Completed Jobs: 4

▾ Event Timeline
□ Enable zooming

▾ **Completed Jobs (4)**

Job Id ▾	Description	Submitted	Duration	Stages: Succeeded/Total	Tasks (for all stages): Succeeded/Total
3	runJob at SparkHadoopWriter.scala:83 runJob at SparkHadoopWriter.scala:83	2024/09/13 18:33:04	2 s	1/1	1/1
2	runJob at PythonRDD.scala:181 runJob at PythonRDD.scala:181	2024/09/13 18:33:02	2 s	2/2 (2 skipped)	72/72 (142 skipped)
1	sortBy at /tmp/job-de9935f7/Top10ActiveTaxis.py:47 sortBy at /tmp/job-de9935f7/Top10ActiveTaxis.py:47	2024/09/13 18:33:00	2 s	1/1 (2 skipped)	71/71 (142 skipped)
0	sortBy at /tmp/job-de9935f7/Top10ActiveTaxis.py:47 sortBy at /tmp/job-de9935f7/Top10ActiveTaxis.py:47	2024/09/13 18:25:38	7.4 min	3/3	213/213


fgvovcybtjcbjbgzswfa3mwheu-dot-us-central1.dataproc.googleusercontent.com/sparkhistory/
V
Spark 3.5.1
Search:
Download

$(cat /opt/dataproc/proxy-agent/banner.html)

History Server

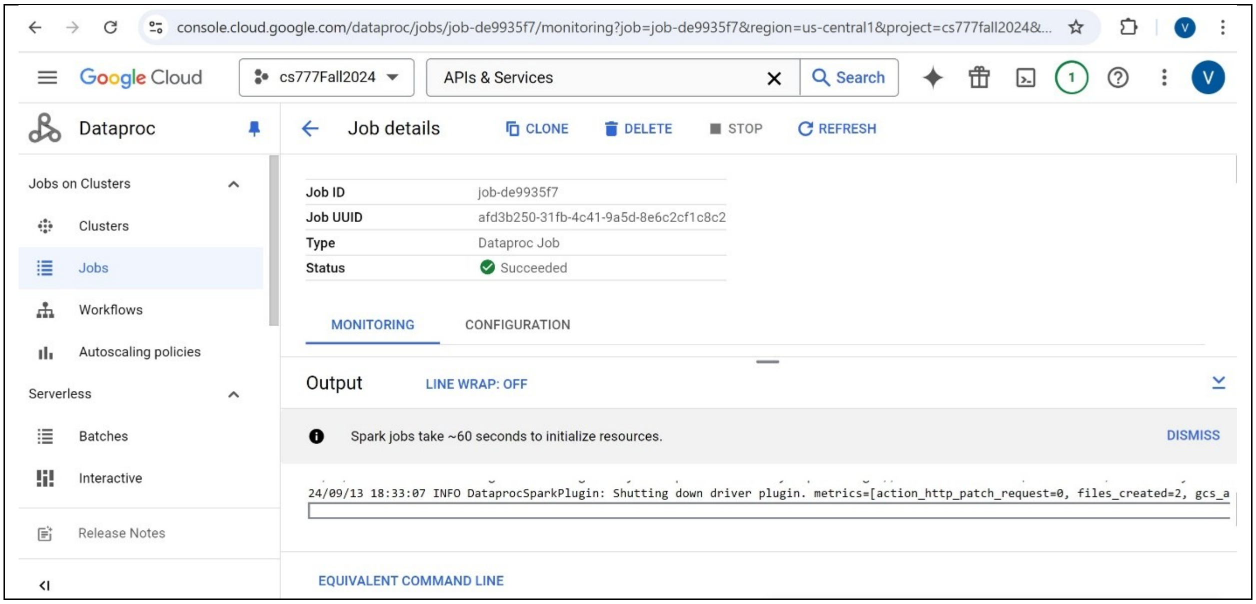
Event log directory: gs://dataproc-temp-us-central1-325310350382-liuj0ysv/bf154472-0856-437e-a37b-81153221bb81/spark-job-history

Last updated: 2024-09-13 11:43:36

Client local time zone: America/Los_Angeles

Version	App ID	App Name	Driver Host	Started	Completed	Duration	Spark User	Last Updated	Event Log
3.5.1	application_1726251016168_0001	Top10ActiveTaxis	cluster-66f2-m.c.cs777fall2024.internal	2024-09-13 11:25:24	2024-09-13 11:33:07	7.7 min	root	2024-09-13 11:33:08	Download

Showing 1 to 1 of 1 entries
Show incomplete applications

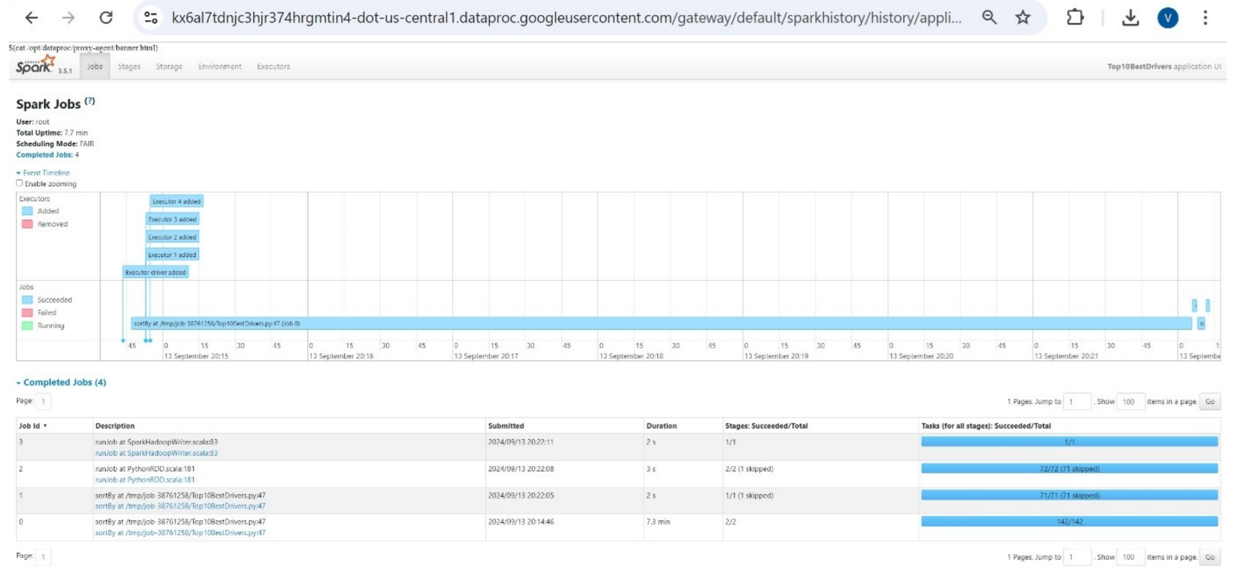


Job ID	job-de9935f7
Job UUID	afd3b250-31fb-4c41-9a5d-8e6c2cf1c8c2
Type	Dataproc Job
Status	✓ Succeeded

Spark jobs take ~60 seconds to initialize resources.

```
24/09/13 18:33:07 INFO DataprocSparkPlugin: Shutting down driver plugin. metrics=[action_http_patch_request=0, files_created=2, gcs_a
```

Top-10 Best Drivers


kx6al7tdnjc3hjr374hrgmtin4-dot-us-central1.dataproc.googleusercontent.com/gateway/default/sparkhistory/history/appli...
$(cat /opt/dataproc/proxy-agent/banner.html)
Spark 3.5.1
Jobs
Stages
Storage
Environment
Executors
Top10BestDrivers application UI

Spark Jobs (?)

User: root
Total Uptime: 7.7 min
Scheduling Mode: FAIR
Completed Jobs: 4

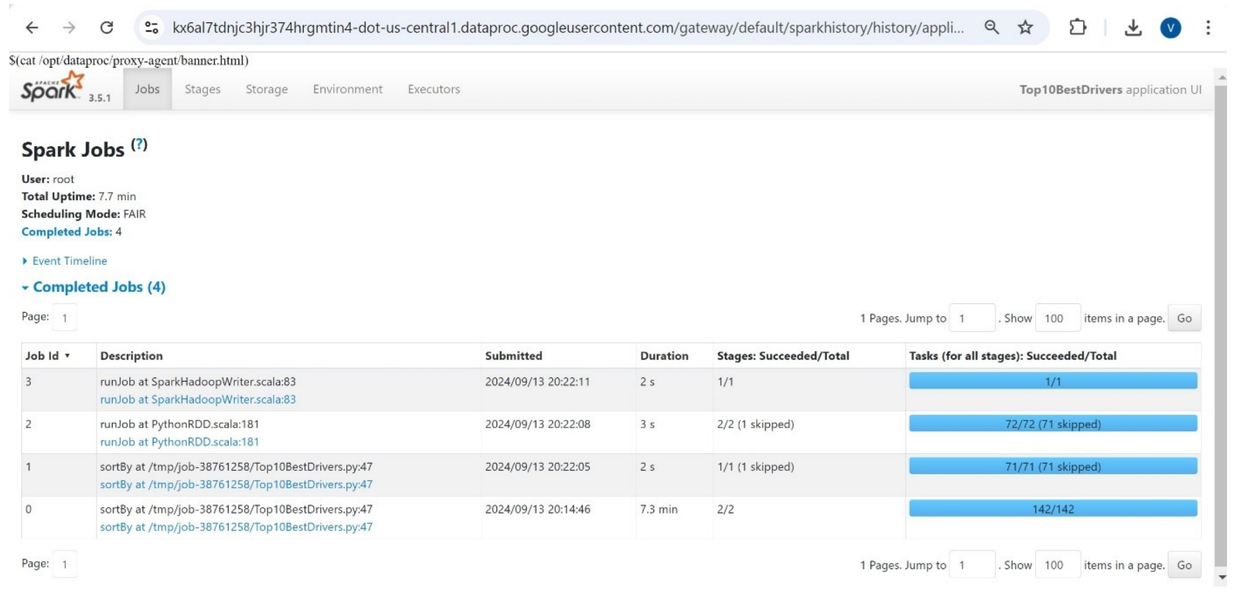
▾ Event Timeline
□ Enable zooming

▾ Completed Jobs (4)

Page: 1 | 1 Pages. Jump to 1 . Show 100 items in a page. Go

Job Id ▾	Description	Submitted	Duration	Stages: Succeeded/Total	Tasks (for all stages): Succeeded/Total
3	runJob at SparkHadoopWriter.scala:83 runJob at SparkHadoopWriter.scala:83	2024/09/13 20:22:11	2 s	1/1	1/1
2	runJob at PythonRDD.scala:181 runJob at PythonRDD.scala:181	2024/09/13 20:22:08	3 s	2/2 (1 skipped)	72/72 (71 skipped)
1	sortBy at /tmp/job-38761258/Top10BestDrivers.py:47 sortBy at /tmp/job-38761258/Top10BestDrivers.py:47	2024/09/13 20:22:05	2 s	1/1 (1 skipped)	71/71 (71 skipped)
0	sortBy at /tmp/job-38761258/Top10BestDrivers.py:47 sortBy at /tmp/job-38761258/Top10BestDrivers.py:47	2024/09/13 20:14:46	7.3 min	2/2	142/142

Page: 1 | 1 Pages. Jump to 1 . Show 100 items in a page. Go


kx6al7tdnjc3hjr374hrgmtin4-dot-us-central1.dataproc.googleusercontent.com/gateway/default/sparkhistory/history/appli...
$(cat /opt/dataproc/proxy-agent/banner.html)
Spark 3.5.1
Jobs
Stages
Storage
Environment
Executors
Top10BestDrivers application UI

Spark Jobs (?)

User: root
Total Uptime: 7.7 min
Scheduling Mode: FAIR
Completed Jobs: 4

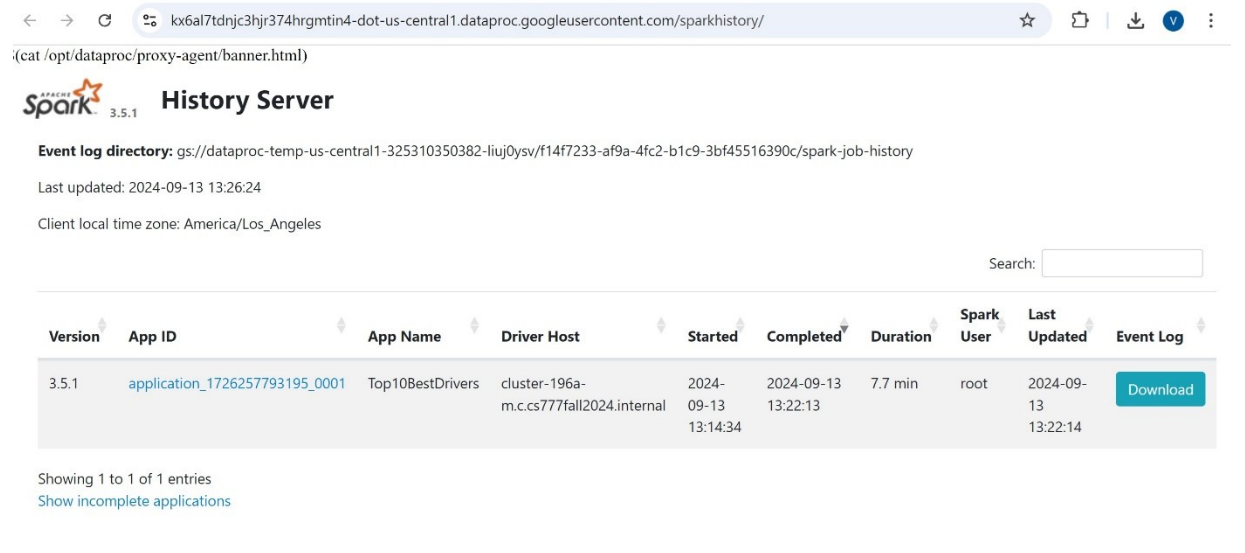
▸ Event Timeline

▾ Completed Jobs (4)

Page: 1 | 1 Pages. Jump to 1 . Show 100 items in a page. Go

Job Id ▾	Description	Submitted	Duration	Stages: Succeeded/Total	Tasks (for all stages): Succeeded/Total
3	runJob at SparkHadoopWriter.scala:83 runJob at SparkHadoopWriter.scala:83	2024/09/13 20:22:11	2 s	1/1	1/1
2	runJob at PythonRDD.scala:181 runJob at PythonRDD.scala:181	2024/09/13 20:22:08	3 s	2/2 (1 skipped)	72/72 (71 skipped)
1	sortBy at /tmp/job-38761258/Top10BestDrivers.py:47 sortBy at /tmp/job-38761258/Top10BestDrivers.py:47	2024/09/13 20:22:05	2 s	1/1 (1 skipped)	71/71 (71 skipped)
0	sortBy at /tmp/job-38761258/Top10BestDrivers.py:47 sortBy at /tmp/job-38761258/Top10BestDrivers.py:47	2024/09/13 20:14:46	7.3 min	2/2	142/142

Page: 1 | 1 Pages. Jump to 1 . Show 100 items in a page. Go



kx6al7tdnjc3hjr374hrgmtin4-dot-us-central1.dataproc.googleusercontent.com/sparkhistory/

:(cat /opt/dataproc/proxy-agent/banner.html)

History Server

Event log directory: gs://dataproc-temp-us-central1-325310350382-liuj0ysv/f14f7233-af9a-4fc2-b1c9-3bf45516390c/spark-job-history

Last updated: 2024-09-13 13:26:24

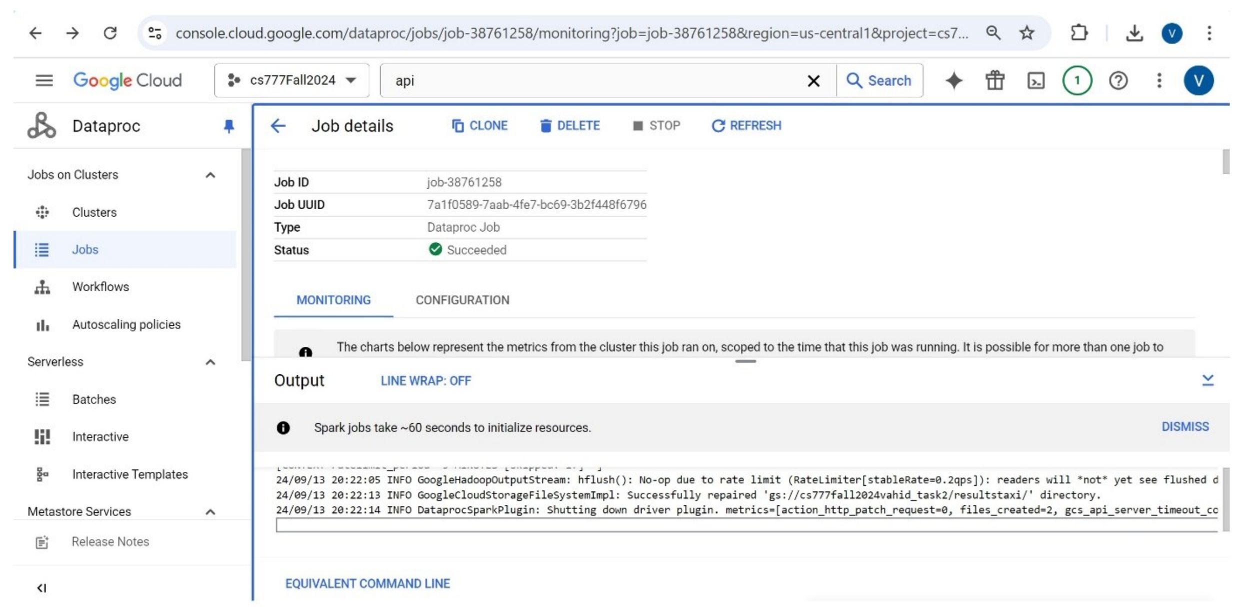
Client local time zone: America/Los_Angeles

Search:

Version	App ID	App Name	Driver Host	Started	Completed	Duration	Spark User	Last Updated	Event Log
3.5.1	application_1726257793195_0001	Top10BestDrivers	cluster-196a-m.c.cs777fall2024.internal	2024-09-13 13:14:34	2024-09-13 13:22:13	7.7 min	root	2024-09-13 13:22:14	Download

Showing 1 to 1 of 1 entries

Show incomplete applications



console.cloud.google.com/dataproc/jobs/job-38761258/monitoring?job=job-38761258&region=us-central1&project=cs7...

Job ID	job-38761258
Job UUID	7a1f0589-7aab-4fe7-bc69-3b2f448f6796
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

Spark jobs take ~60 seconds to initialize resources.

```
24/09/13 20:22:05 INFO GoogleHadoopOutputStream: hflush(): No-op due to rate limit (RateLimiter[stableRate=0.2qps]): readers will *not* yet see flushed d
24/09/13 20:22:13 INFO GoogleCloudStorageFileSystemImpl: Successfully repaired 'gs://cs777fall2024vahid_task2/resultstaxi/' directory.
24/09/13 20:22:14 INFO DataprocSparkPlugin: Shutting down driver plugin. metrics=[action_http_patch_request=0, files_created=2, gcs_api_server_timeout_cc
```