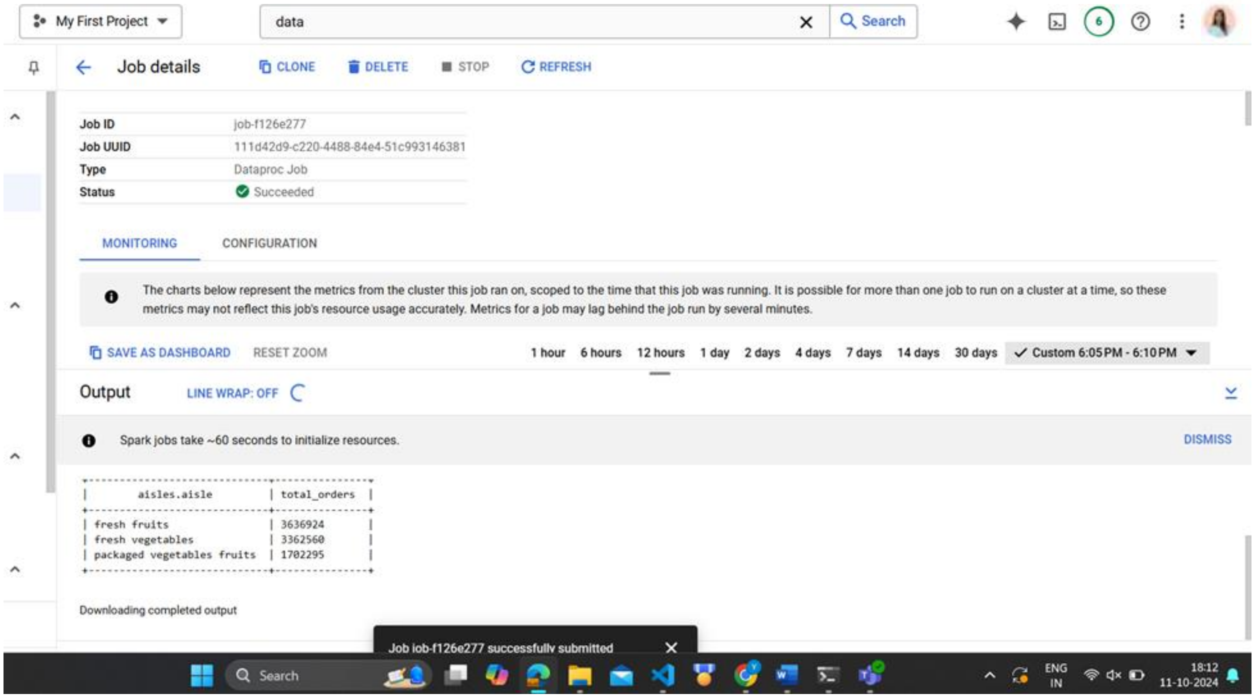
Solutions

1. Name the top 3 aisles with the most products ordered

Screenshot - Hive:



Job ID job-f126e277

Job UUID 111d42d9-c220-4488-84e4-51c993146381

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 run on a cluster at a time, so these metrics may not reflect this job's resource usage accurately. Metrics for a job may lag behind the job run by several minutes.

✓ Custom 6:05 PM - 6:10 PM

Output

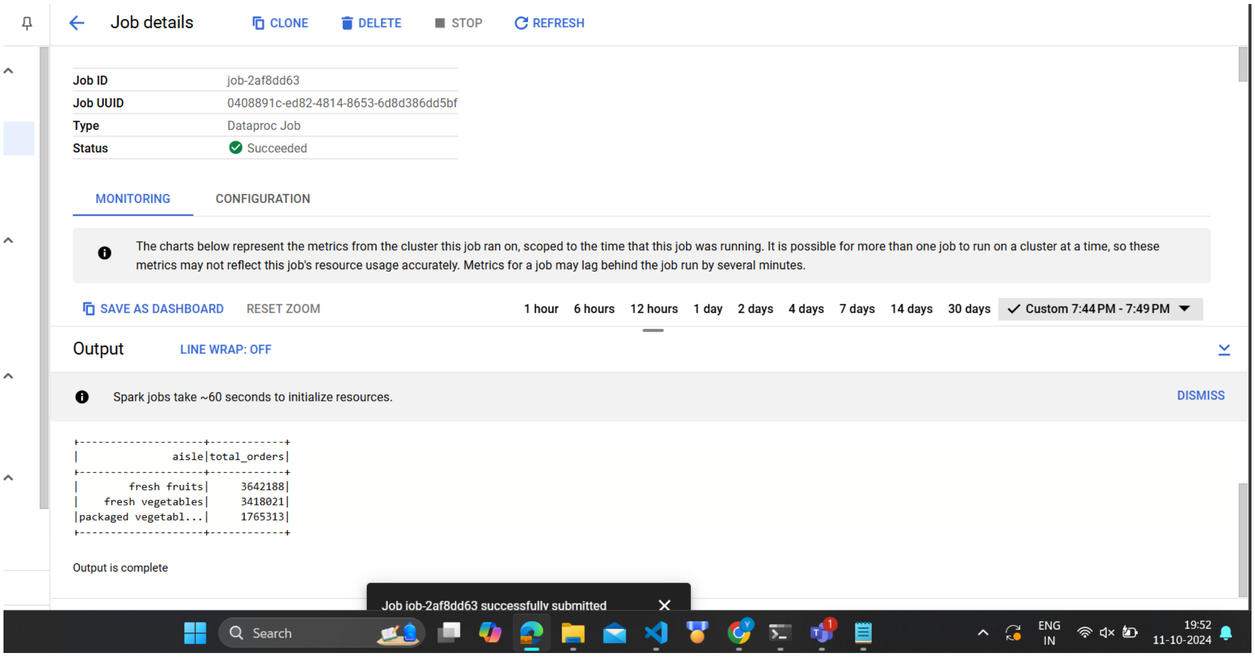
Spark jobs take ~60 seconds to initialize resources.

```
+-----------------------------+---------------+
|         aisles.aisle        | total_orders  |
+-----------------------------+---------------+
| fresh fruits                | 3636924       |
| fresh vegetables            | 3362560       |
| packaged vegetables fruits  | 1702295       |
+-----------------------------+---------------+
```

Downloading completed output

Job iob-f126e277 successfully submitted

Screenshot - PySpark:



Job ID job-2af8dd63

Job UUID 0408891c-ed82-4814-8653-6d8d386dd5bf

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 run on a cluster at a time, so these metrics may not reflect this job's resource usage accurately. Metrics for a job may lag behind the job run by several minutes.

✓ Custom 7:44 PM - 7:49 PM

Output

Spark jobs take ~60 seconds to initialize resources.

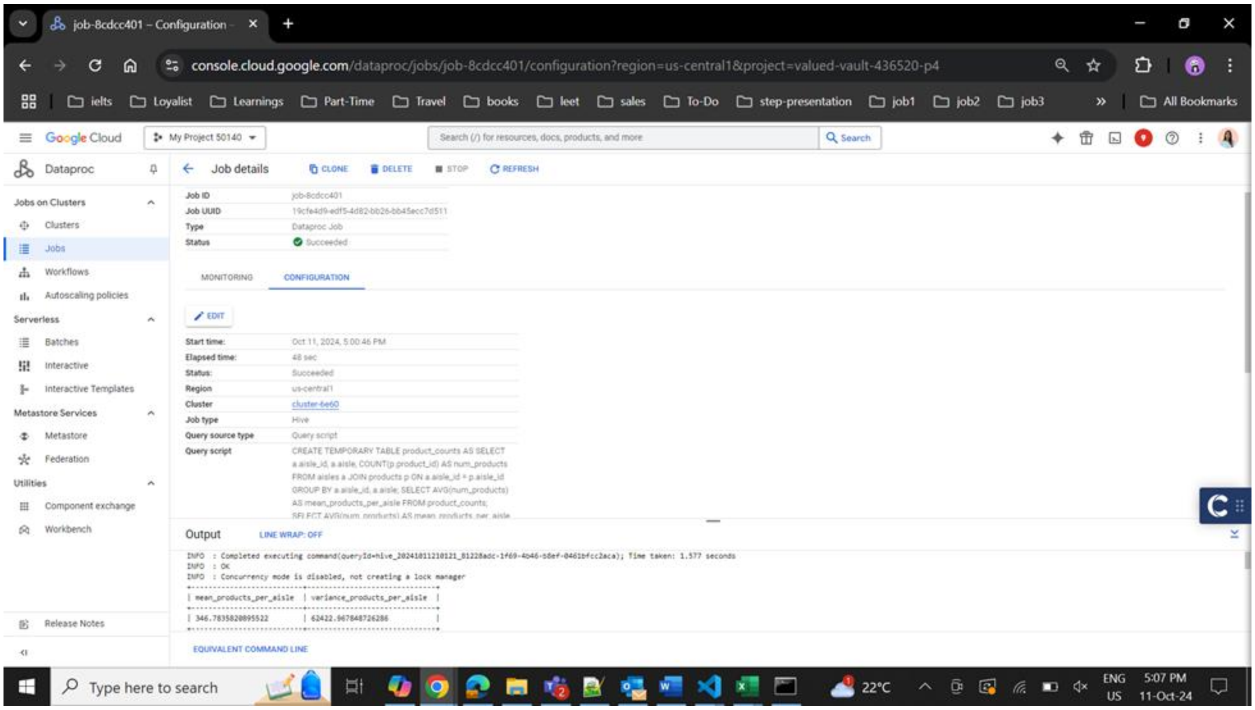
```
+--------------------+------------+
|               aisle|total_orders|
+--------------------+------------+
|        fresh fruits|     3642188|
|    fresh vegetables|     3418021|
|packaged vegetabl...|     1765313|
+--------------------+------------+
```

Output is complete

Job iob-2af8dd63 successfully submitted

2. What is the mean and the variance of number of products in an aisle?

Screenshot - Hive:



console.cloud.google.com/dataproc/jobs/job-8cdcc401/configuration?region=us-central1&project=valued-vault-436520-p4

Job details

Job ID: job-8cdcc401
Job UUID: 19cfe4d9-edf5-4d82-bb26-bb45ecc7d511
Type: Dataproc Job
Status: Succeeded

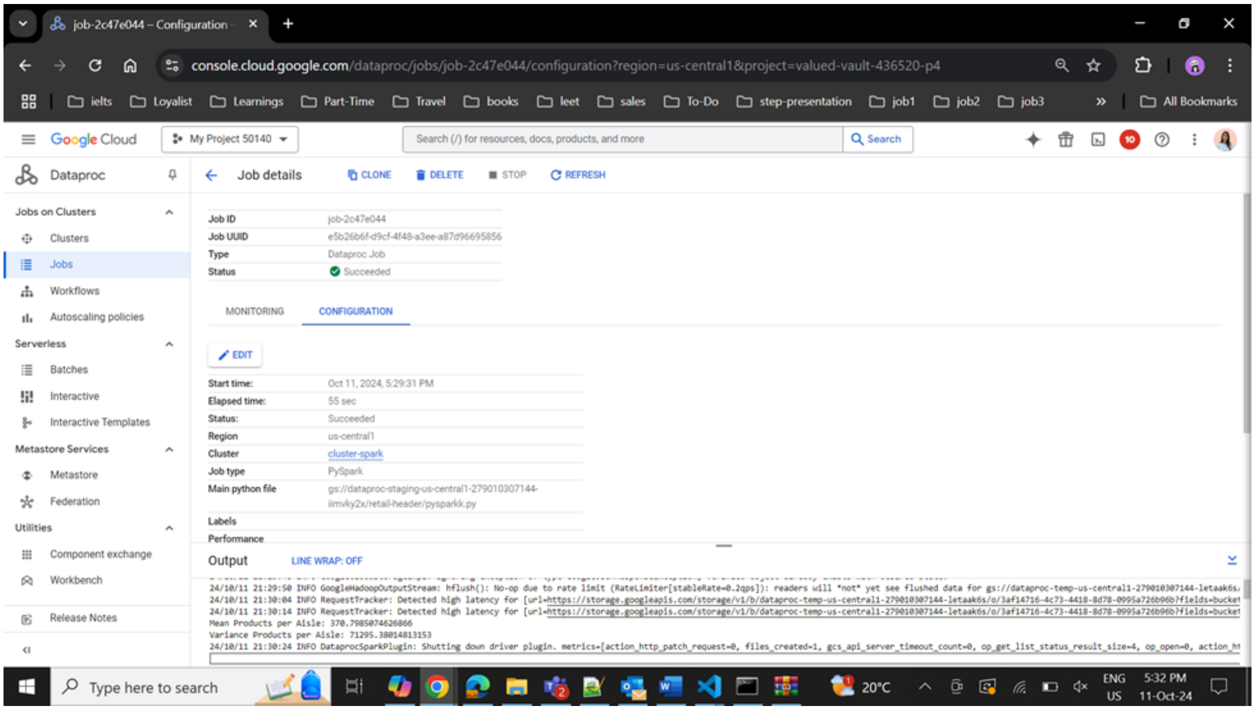
Start time: Oct 11, 2024, 5:00:46 PM
Elapsed time: 48 sec
Status: Succeeded
Region: us-central1
Cluster: cluster-6e60
Job type: Hive
Query source type: Query script
Query script:

```
CREATE TEMPORARY TABLE product_counts AS SELECT
a.aisle_id, a.aisle, COUNT(p.product_id) AS num_products
FROM aisles a JOIN products p ON a.aisle_id = p.aisle_id
GROUP BY a.aisle_id, a.aisle; SELECT AVG(num_products)
AS mean_products_per_aisle FROM product_counts;
SELECT AVG(num_products) AS mean_products_per_aisle
```

Output

```
INFO  : Completed executing command(queryId=hive_20241011210121_81228adc-1f69-4b46-b8ef-0461bfcc2aca); Time taken: 1.577 seconds
INFO  : OK
INFO  : Concurrency mode is disabled, not creating a lock manager
+--------------------------+------------------------------+
| mean_products_per_aisle  | variance_products_per_aisle  |
+--------------------------+------------------------------+
| 346.7835820895522        | 62422.967848726286           |
+--------------------------+------------------------------+
```

Screenshot - PySpark:



console.cloud.google.com/dataproc/jobs/job-2c47e044/configuration?region=us-central1&project=valued-vault-436520-p4

Job details

Job ID: job-2c47e044
Job UUID: e5b26b6f-d9cf-4f48-a3ee-a87d96695856
Type: Dataproc Job
Status: Succeeded

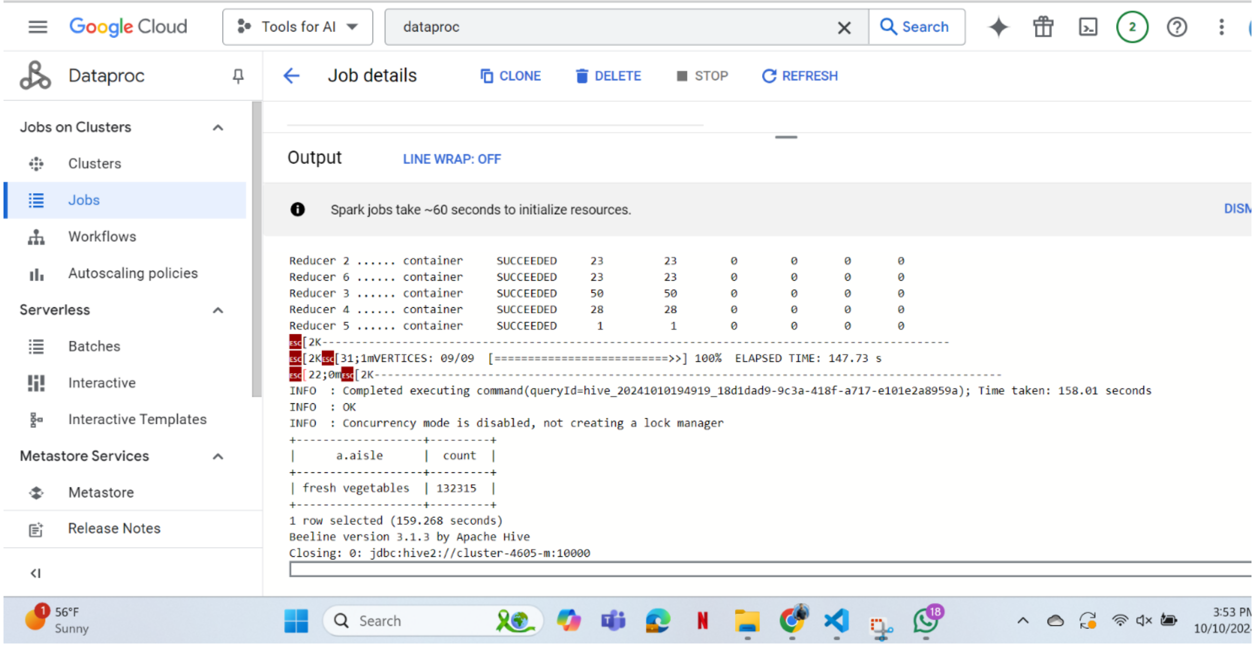
Start time: Oct 11, 2024, 5:29:31 PM
Elapsed time: 55 sec
Status: Succeeded
Region: us-central1
Cluster: cluster-spark
Job type: PySpark
Main python file: gs://dataproc-staging-us-central1-279010307144-iimvky2x/retail-header/pysparkk.py

Output

```
24/10/11 21:29:50 INFO GoogleHadoopOutputStream: hflush(): No-op due to rate limit (RateLimiter[stableRate=0.2qps]): readers will *not* yet see flushed data for gs://dataproc-temp-us-central1-279010307144-letaak6s/
24/10/11 21:30:04 INFO RequestTracker: Detected high latency for [url=https://storage.googleapis.com/storage/v1/b/dataproc-temp-us-central1-279010307144-letaak6s/o/3af14716-4c73-4418-8d78-0995a726b96b?fields=bucket
24/10/11 21:30:14 INFO RequestTracker: Detected high latency for [url=https://storage.googleapis.com/storage/v1/b/dataproc-temp-us-central1-279010307144-letaak6s/o/3af14716-4c73-4418-8d78-0995a726b96b?fields=bucket
Mean Products per Aisle: 370.7985074626866
Variance Products per Aisle: 71295.38014813153
24/10/11 21:30:24 INFO DataprocSparkPlugin: Shutting down driver plugin. metrics=[action_http_patch_request=0, files_created=1, gcs_api_server_timeout_count=0, op_get_list_status_result_size=4, op_open=0, action_ht
```

3. Which aisle products are most bought with products from the "speciality cheeses" aisle?

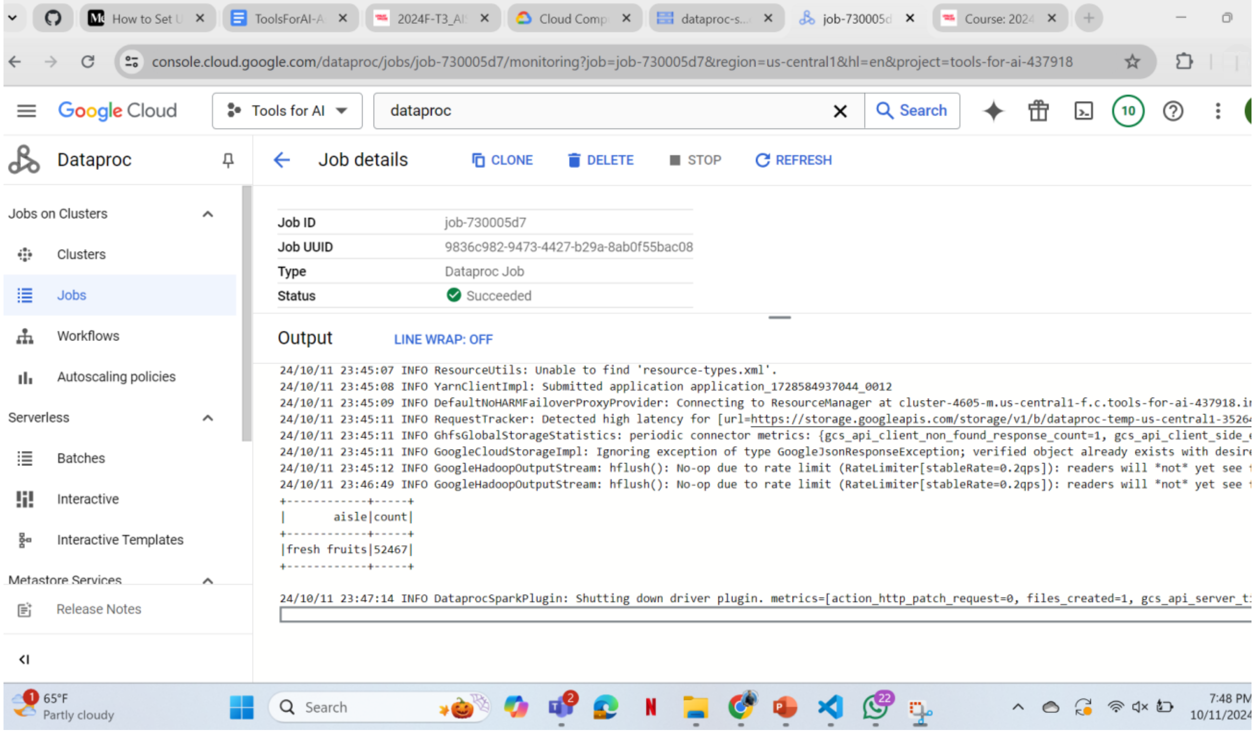
Screenshot - Hive:



Spark jobs take ~60 seconds to initialize resources.

```
Reducer 2 ...... container     SUCCEEDED     23         23        0        0       0       0
Reducer 6 ...... container     SUCCEEDED     23         23        0        0       0       0
Reducer 3 ...... container     SUCCEEDED     50         50        0        0       0       0
Reducer 4 ...... container     SUCCEEDED     28         28        0        0       0       0
Reducer 5 ...... container     SUCCEEDED      1          1        0        0       0       0
ESC[2K----------------------------------------------------------------------------------------------------
ESC[2KESC[31;1mVERTICES: 09/09  [==========================>>] 100%  ELAPSED TIME: 147.73 s
ESC[22;0mESC[2K----------------------------------------------------------------------------------------------------
INFO  : Completed executing command(queryId=hive_20241010194919_18d1dad9-9c3a-418f-a717-e101e2a8959a); Time taken: 158.01 seconds
INFO  : OK
INFO  : Concurrency mode is disabled, not creating a lock manager
+-------------------+---------+
|      a.aisle      |  count  |
+-------------------+---------+
| fresh vegetables  | 132315  |
+-------------------+---------+
1 row selected (159.268 seconds)
Beeline version 3.1.3 by Apache Hive
Closing: 0: jdbc:hive2://cluster-4605-m:10000
```

Screenshot - PySpark:

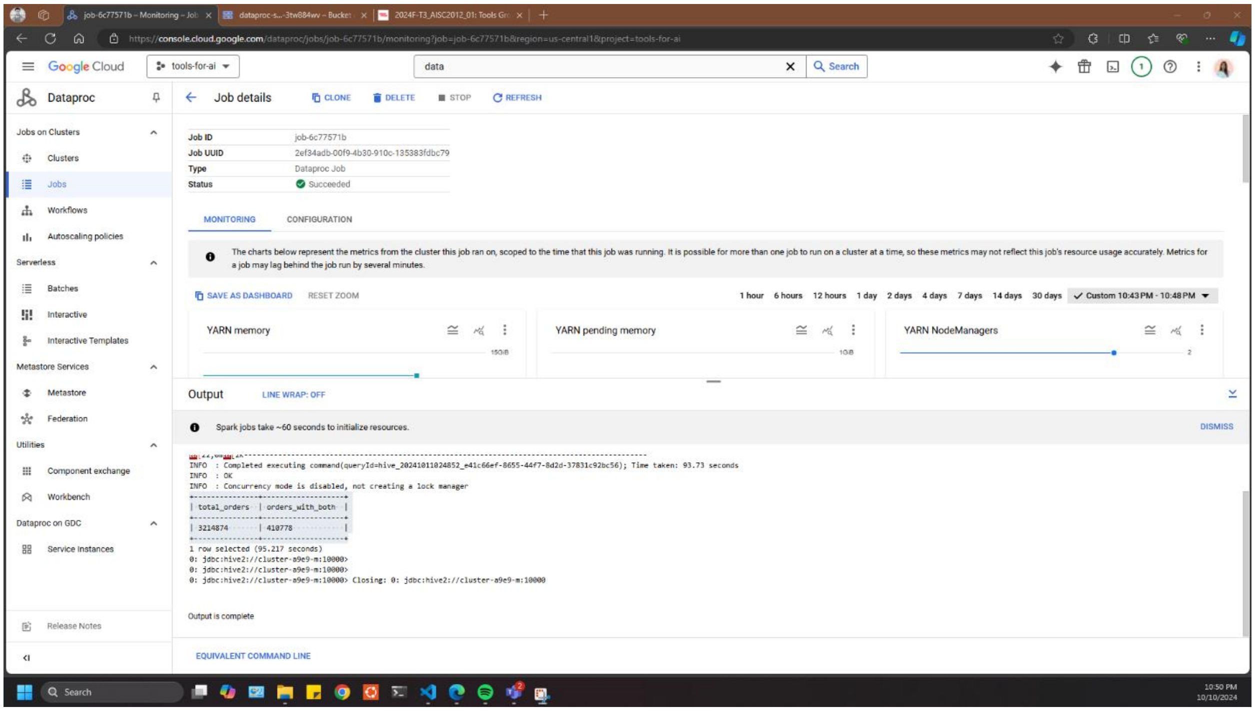


Job ID	job-730005d7
Job UUID	9836c982-9473-4427-b29a-8ab0f55bac08
Type	Dataproc Job
Status	✓ Succeeded

```
24/10/11 23:45:07 INFO ResourceUtils: Unable to find 'resource-types.xml'.
24/10/11 23:45:08 INFO YarnClientImpl: Submitted application application_1728584937044_0012
24/10/11 23:45:09 INFO DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at cluster-4605-m.us-central1-f.c.tools-for-ai-437918.in
24/10/11 23:45:11 INFO RequestTracker: Detected high latency for [url=https://storage.googleapis.com/storage/v1/b/dataproc-temp-us-central1-3526
24/10/11 23:45:11 INFO GhfsGlobalStorageStatistics: periodic connector metrics: {gcs_api_client_non_found_response_count=1, gcs_api_client_side_
24/10/11 23:45:11 INFO GoogleCloudStorageImpl: Ignoring exception of type GoogleJsonResponseException; verified object already exists with desir
24/10/11 23:45:12 INFO GoogleHadoopOutputStream: hflush(): No-op due to rate limit (RateLimiter[stableRate=0.2qps]): readers will *not* yet see
24/10/11 23:46:49 INFO GoogleHadoopOutputStream: hflush(): No-op due to rate limit (RateLimiter[stableRate=0.2qps]): readers will *not* yet see
+------------+-----+
|       aisle|count|
+------------+-----+
|fresh fruits|52467|
+------------+-----+

24/10/11 23:47:14 INFO DataprocSparkPlugin: Shutting down driver plugin. metrics=[action_http_patch_request=0, files_created=1, gcs_api_server_t
```

4. Sales department is making a recommendation that frozen products should be placed next to bakery products. Write 1-2 SQL queries to get the data that would support/oppose this recommendation.

Screenshot - Hive:



Job ID	job-6c77571b
Job UUID	2ef34adb-00f9-4b30-910c-135383fdbc79
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 run on a cluster at a time, so these metrics may not reflect this job's resource usage accurately. Metrics for a job may lag behind the job run by several minutes.

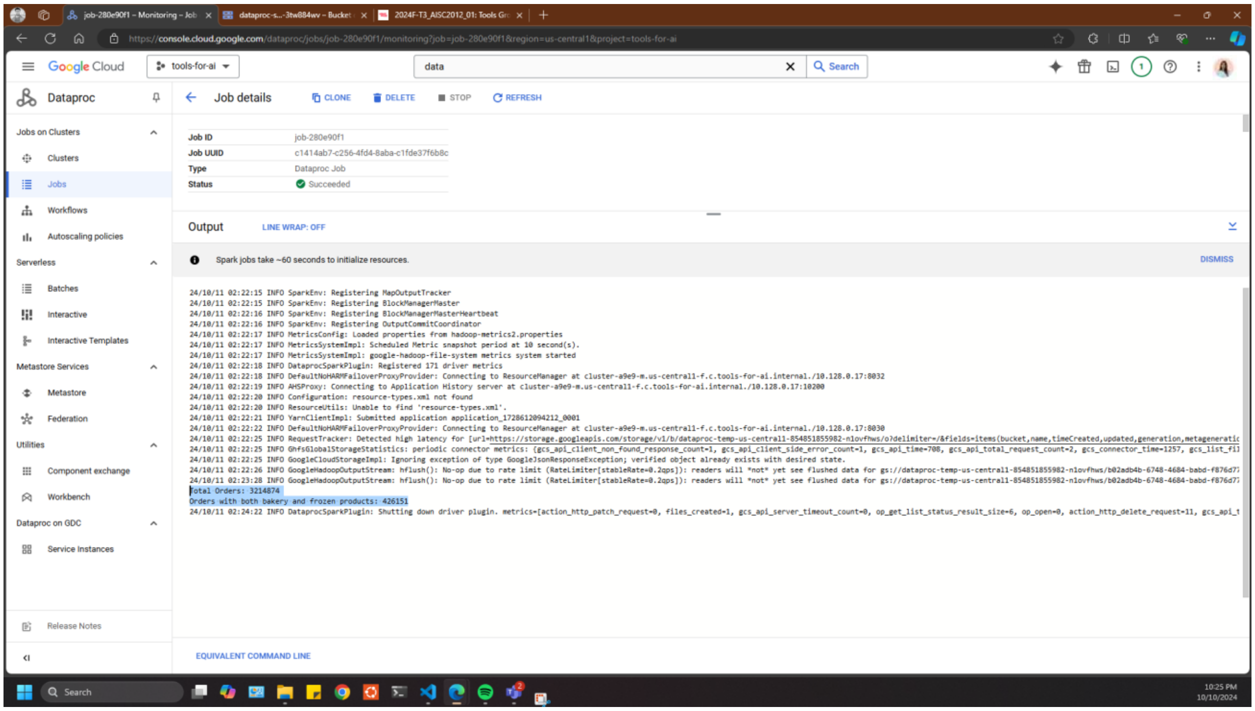
Output

Spark jobs take ~60 seconds to initialize resources.

```
INFO  : Completed executing command(queryId=hive_20241011024852_e41c66ef-8655-44f7-8d2d-37831c92bc56); Time taken: 93.73 seconds
INFO  : OK
INFO  : Concurrency mode is disabled, not creating a lock manager
+---------------+-------------------+
| total_orders  | orders_with_both  |
+---------------+-------------------+
| 3214874       | 410778            |
+---------------+-------------------+
1 row selected (95.217 seconds)
0: jdbc:hive2://cluster-a9e9-m:10000>
0: jdbc:hive2://cluster-a9e9-m:10000>
0: jdbc:hive2://cluster-a9e9-m:10000> Closing: 0: jdbc:hive2://cluster-a9e9-m:10000
```

Output is complete

Screenshot - PySpark:



Job ID	job-280e90f1
Job UUID	c1414ab7-c256-4fd4-8aba-c1fde37f6b8c
Type	Dataproc Job
Status	Succeeded

Output

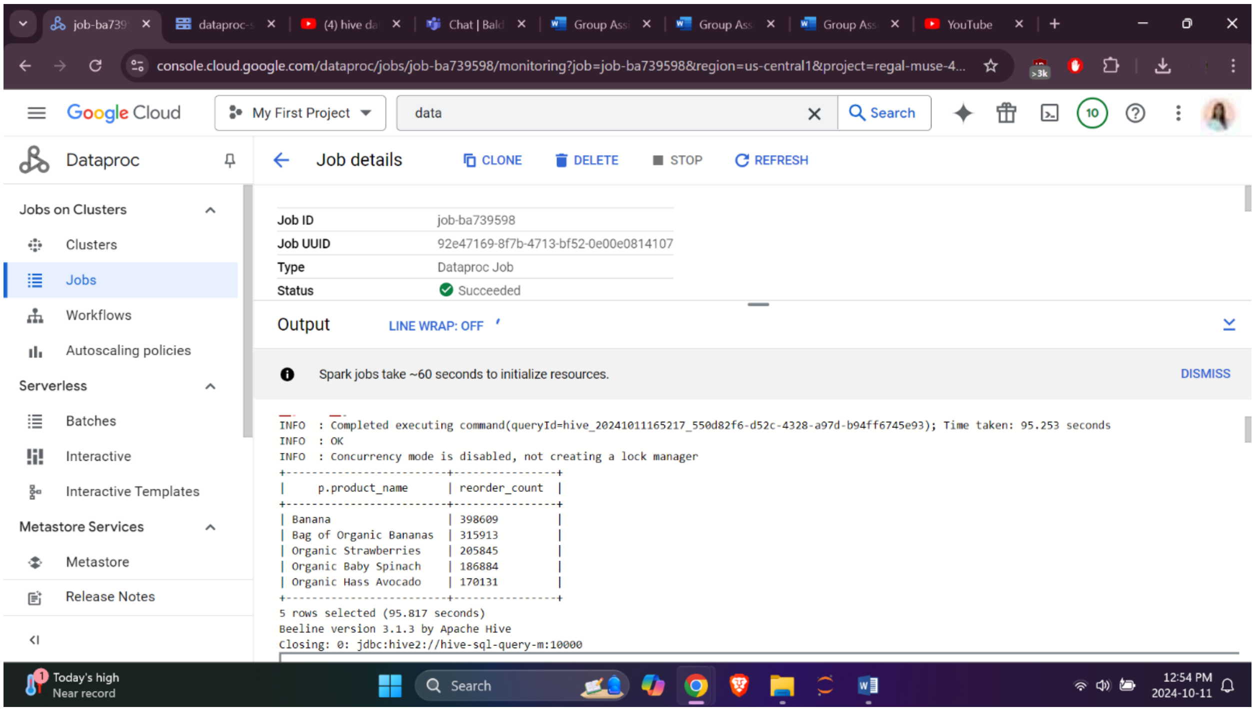
Spark jobs take ~60 seconds to initialize resources.

```
24/10/11 02:22:15 INFO SparkEnv: Registering MapOutputTracker
24/10/11 02:22:15 INFO SparkEnv: Registering BlockManagerMaster
24/10/11 02:22:16 INFO SparkEnv: Registering BlockManagerMasterHeartbeat
24/10/11 02:22:16 INFO SparkEnv: Registering OutputCommitCoordinator
24/10/11 02:22:17 INFO MetricsConfig: Loaded properties from hadoop-metrics2.properties
24/10/11 02:22:17 INFO MetricsSystemImpl: Scheduled Metric snapshot period at 10 second(s).
24/10/11 02:22:17 INFO MetricsSystemImpl: google-hadoop-file-system metrics system started
24/10/11 02:22:18 INFO DataprocSparkPlugin: Registered 171 driver metrics
24/10/11 02:22:18 INFO DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at cluster-a9e9-m.us-central1-f.c.tools-for-ai.internal./10.128.0.17:8032
24/10/11 02:22:19 INFO AHSProxy: Connecting to Application History server at cluster-a9e9-m.us-central1-f.c.tools-for-ai.internal./10.128.0.17:10200
24/10/11 02:22:20 INFO Configuration: resource-types.xml not found
24/10/11 02:22:20 INFO ResourceUtils: Unable to find 'resource-types.xml'.
24/10/11 02:22:21 INFO YarnClientImpl: Submitted application application_1728612094212_0001
24/10/11 02:22:22 INFO DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at cluster-a9e9-m.us-central1-f.c.tools-for-ai.internal./10.128.0.17:8030
24/10/11 02:22:25 INFO RequestTracker: Detected high latency for [url=https://storage.googleapis.com/storage/v1/b/dataproc-temp-us-central1-854851855982-n1ovfhws/o?delimiter=/&fields=items(bucket,name,timeCreated,updated,generation,metageneratic
24/10/11 02:22:25 INFO GhfsGlobalStorageStatistics: periodic connector metrics: {gcs_api_client_non_found_response_count=1, gcs_api_client_side_error_count=1, gcs_api_time=708, gcs_api_total_request_count=2, gcs_connector_time=1257, gcs_list_fil
24/10/11 02:22:25 INFO GoogleCloudStorageImpl: Ignoring exception of type GoogleJsonResponseException; verified object already exists with desired state.
24/10/11 02:22:26 INFO GoogleHadoopOutputStream: hflush(): No-op due to rate limit (RateLimiter[stableRate=0.2qps]): readers will *not* yet see flushed data for gs://dataproc-temp-us-central1-854851855982-n1ovfhws/b02adb4b-6748-4684-babd-f876d7
24/10/11 02:23:28 INFO GoogleHadoopOutputStream: hflush(): No-op due to rate limit (RateLimiter[stableRate=0.2qps]): readers will *not* yet see flushed data for gs://dataproc-temp-us-central1-854851855982-n1ovfhws/b02adb4b-6748-4684-babd-f876d7
Total Orders: 3214874
Orders with both bakery and frozen products: 426151
24/10/11 02:24:22 INFO DataprocSparkPlugin: Shutting down driver plugin. metrics=[action_http_patch_request=0, files_created=1, gcs_api_server_timeout_count=0, op_get_list_status_result_size=6, op_open=0, action_http_delete_request=11, gcs_api_t
```

5. Name the top 5 products which are the most reordered

Ravi

Screenshot - Hive:



console.cloud.google.com/dataproc/jobs/job-ba739598/monitoring?job=job-ba739598&region=us-central1&project=regal-muse-4...

data

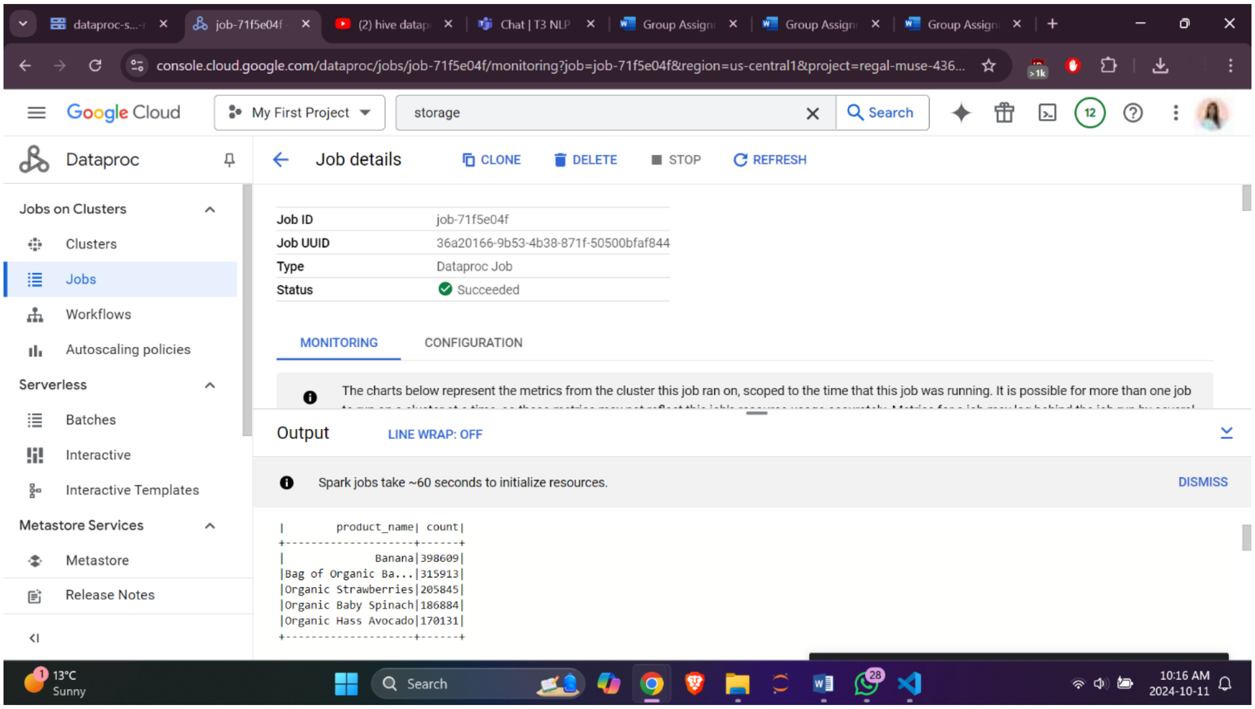
Job ID	job-ba739598
Job UUID	92e47169-8f7b-4713-bf52-0e00e0814107
Type	Dataproc Job
Status	Succeeded

Output

Spark jobs take ~60 seconds to initialize resources.

```
INFO  : Completed executing command(queryId=hive_20241011165217_550d82f6-d52c-4328-a97d-b94ff6745e93); Time taken: 95.253 seconds
INFO  : OK
INFO  : Concurrency mode is disabled, not creating a lock manager
+-------------------------+-----------------+
|     p.product_name      | reorder_count   |
+-------------------------+-----------------+
| Banana                  | 398609          |
| Bag of Organic Bananas  | 315913          |
| Organic Strawberries    | 205845          |
| Organic Baby Spinach    | 186884          |
| Organic Hass Avocado    | 170131          |
+-------------------------+-----------------+
5 rows selected (95.817 seconds)
Beeline version 3.1.3 by Apache Hive
Closing: 0: jdbc:hive2://hive-sql-query-m:10000
```

Screenshot - PySpark:



console.cloud.google.com/dataproc/jobs/job-71f5e04f/monitoring?job=job-71f5e04f&region=us-central1&project=regal-muse-436...

storage

Job ID	job-71f5e04f
Job UUID	36a20166-9b53-4b38-871f-50500bfaf844
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

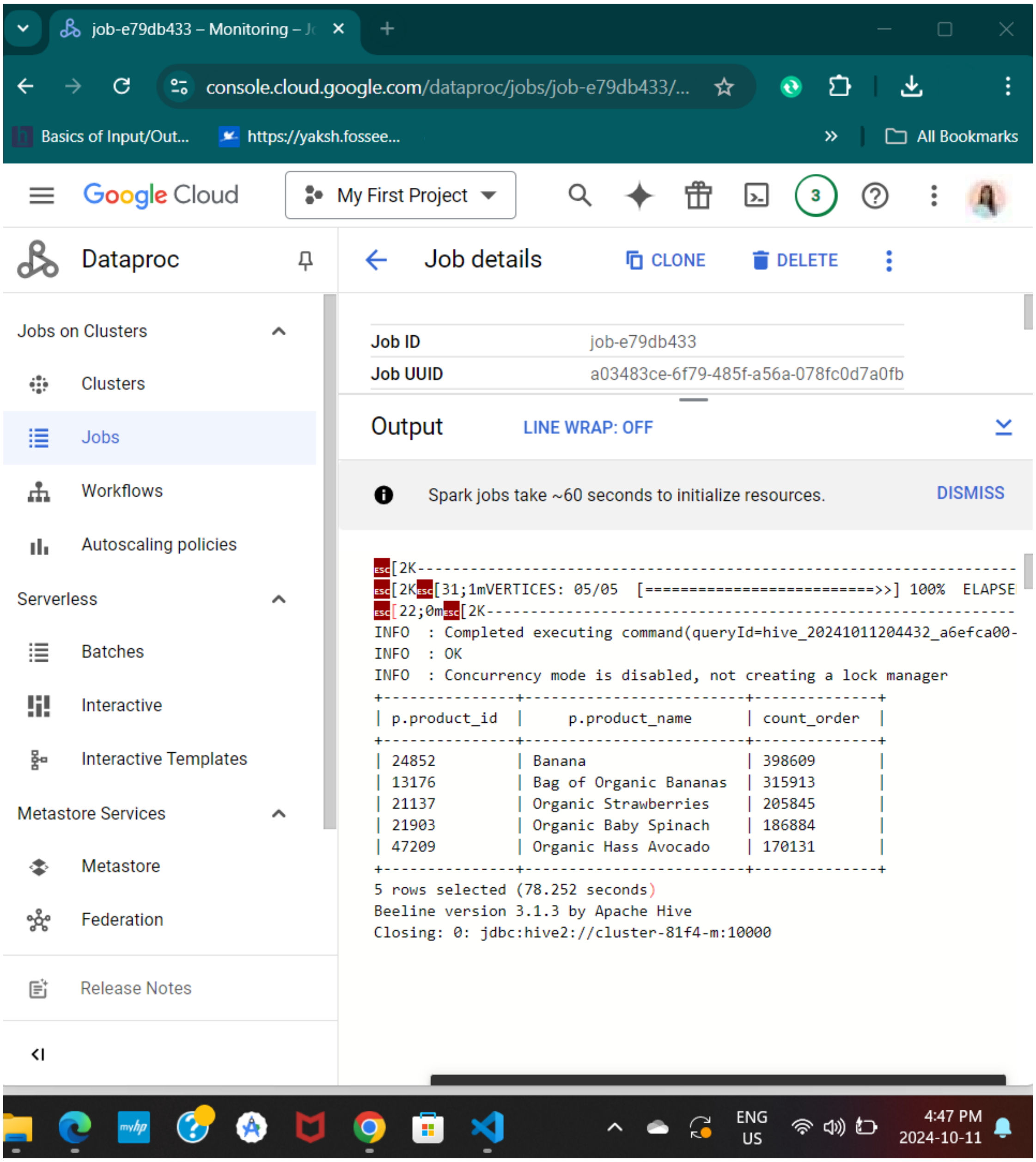
Output

Spark jobs take ~60 seconds to initialize resources.

```
|        product_name| count|
+--------------------+------+
|              Banana|398609|
|Bag of Organic Ba...|315913|
|Organic Strawberries|205845|
|Organic Baby Spinach|186884|
|Organic Hass Avocado|170131|
+--------------------+------+
```

Aman Belwal

Screenshot – Hive:



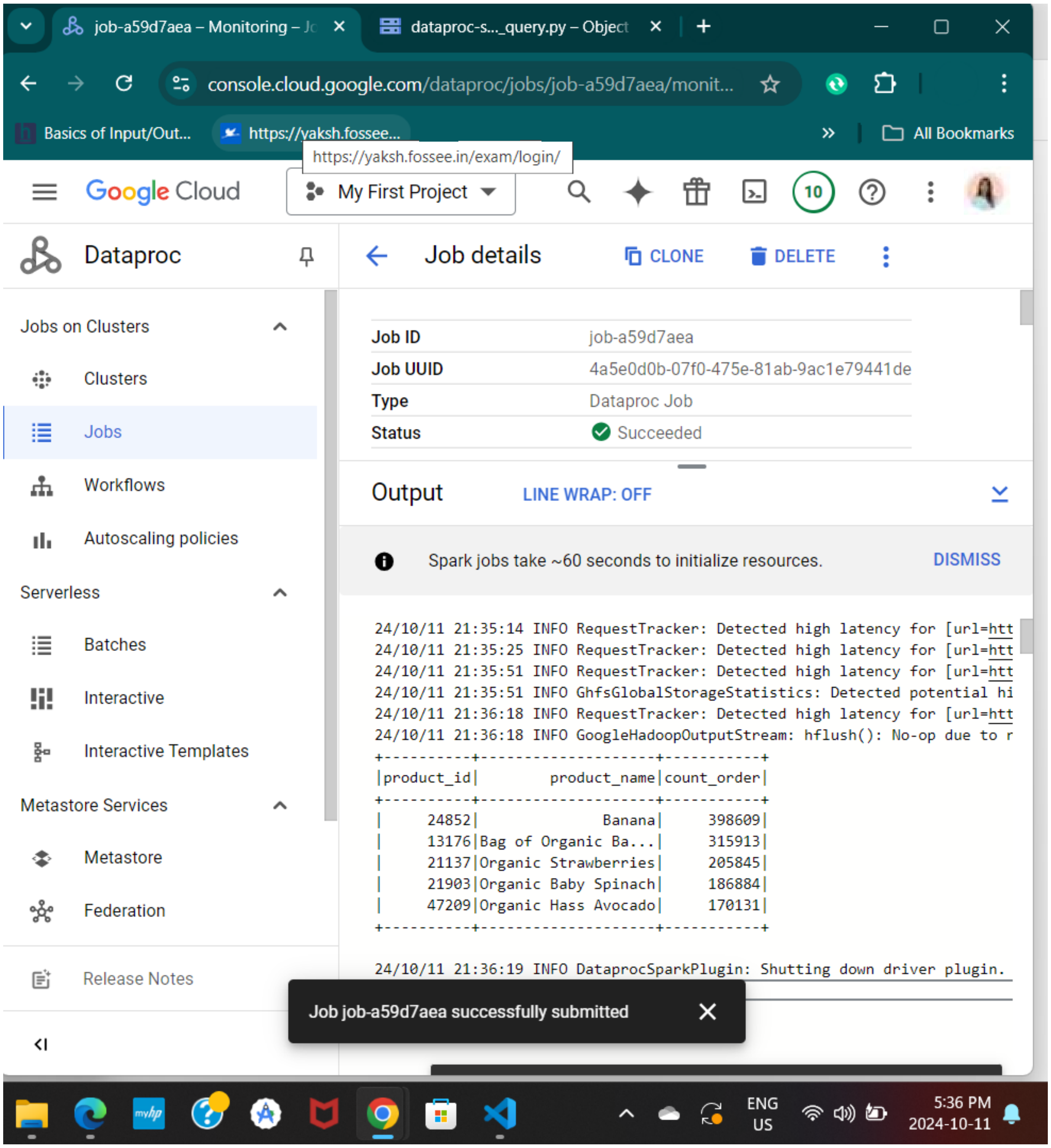
Job ID job-e79db433

Job UUID a03483ce-6f79-485f-a56a-078fc0d7a0fb

Spark jobs take ~60 seconds to initialize resources.

```
ESC[2K---------------------------------------------------------------------
ESC[2KESC[31;1mVERTICES: 05/05  [==========================>>] 100%  ELAPSE
ESC[22;0mESC[2K-------------------------------------------------------------------
INFO  : Completed executing command(queryId=hive_20241011204432_a6efca00-
INFO  : OK
INFO  : Concurrency mode is disabled, not creating a lock manager
+---------------+--------------------------+--------------+
| p.product_id  |      p.product_name      | count_order  |
+---------------+--------------------------+--------------+
| 24852         | Banana                   | 398609       |
| 13176         | Bag of Organic Bananas   | 315913       |
| 21137         | Organic Strawberries     | 205845       |
| 21903         | Organic Baby Spinach     | 186884       |
| 47209         | Organic Hass Avocado     | 170131       |
+---------------+--------------------------+--------------+
5 rows selected (78.252 seconds)
Beeline version 3.1.3 by Apache Hive
Closing: 0: jdbc:hive2://cluster-81f4-m:10000
```

Screenshot – PySpark:



Job ID	job-a59d7aea
Job UUID	4a5e0d0b-07f0-475e-81ab-9ac1e79441de
Type	Dataproc Job
Status	✓ Succeeded

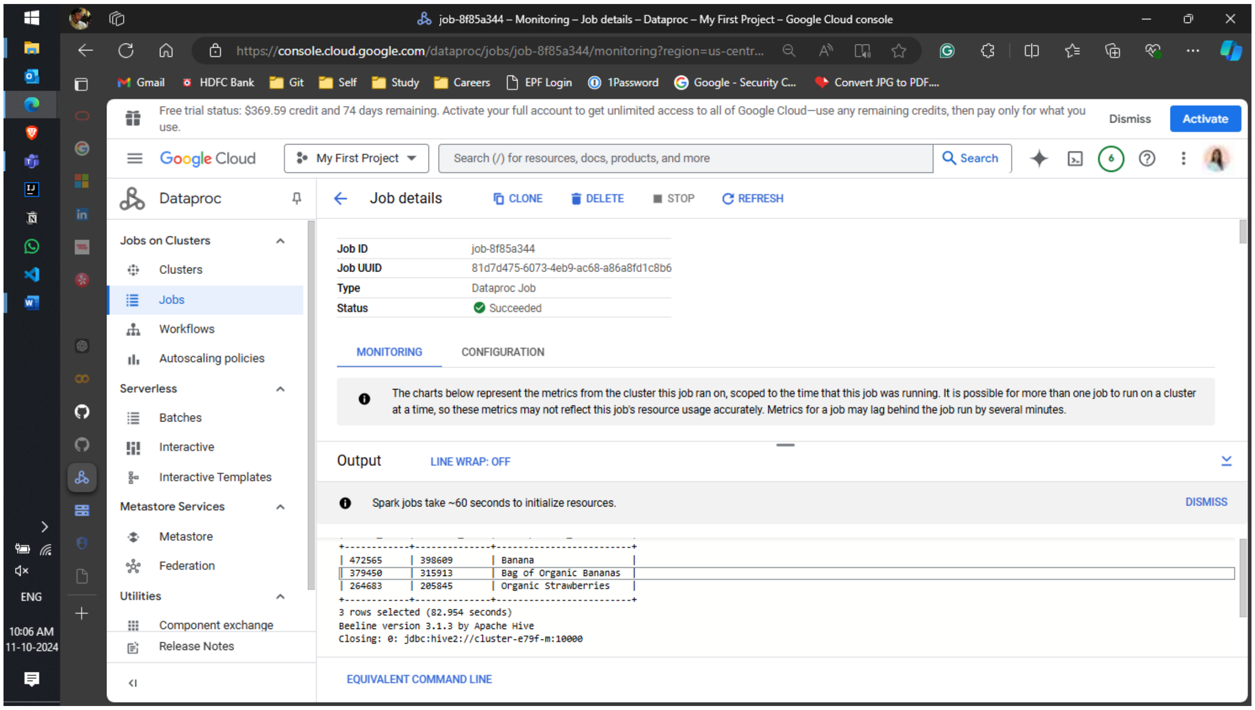
Spark jobs take ~60 seconds to initialize resources.

```
24/10/11 21:35:14 INFO RequestTracker: Detected high latency for [url=htt
24/10/11 21:35:25 INFO RequestTracker: Detected high latency for [url=htt
24/10/11 21:35:51 INFO RequestTracker: Detected high latency for [url=htt
24/10/11 21:35:51 INFO GhfsGlobalStorageStatistics: Detected potential hi
24/10/11 21:36:18 INFO RequestTracker: Detected high latency for [url=htt
24/10/11 21:36:18 INFO GoogleHadoopOutputStream: hflush(): No-op due to r
+----------+--------------------+-----------+
|product_id|        product_name|count_order|
+----------+--------------------+-----------+
|     24852|              Banana|     398609|
|     13176|Bag of Organic Ba...|     315913|
|     21137|Organic Strawberries|     205845|
|     21903|Organic Baby Spinach|     186884|
|     47209|Organic Hass Avocado|     170131|
+----------+--------------------+-----------+

24/10/11 21:36:19 INFO DataprocSparkPlugin: Shutting down driver plugin.
```

Job job-a59d7aea successfully submitted

6. Which 3 products should a customer retargeting (to bring customers back) campaign offer discounts on?

Screenshot - Hive:



Job details

Job ID	job-8f85a344
Job UUID	81d7d475-6073-4eb9-ac68-a86a8fd1c8b6
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 run on a cluster at a time, so these metrics may not reflect this job's resource usage accurately. Metrics for a job may lag behind the job run by several minutes.

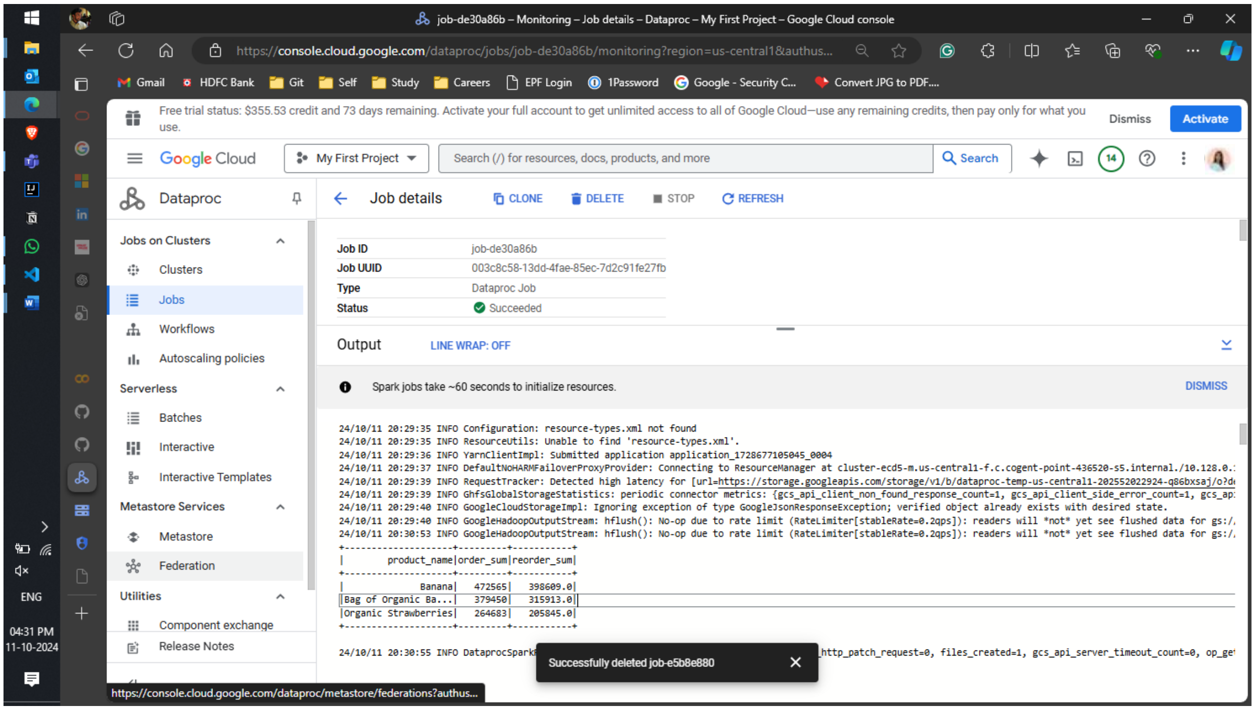
Output

Spark jobs take ~60 seconds to initialize resources.

```
+------------+-------------+------------------------+
| 472565     | 398609      | Banana                 |
| 379450     | 315913      | Bag of Organic Bananas |
| 264683     | 205845      | Organic Strawberries   |
+------------+-------------+------------------------+
3 rows selected (82.954 seconds)
Beeline version 3.1.3 by Apache Hive
Closing: 0: jdbc:hive2://cluster-e79f-m:10000
```

job-8f85a344 – Monitoring – Job details – Dataproc – My First Project – Google Cloud console
https://console.cloud.google.com/dataproc/jobs/job-8f85a344/monitoring?region=us-centr...
Free trial status: $369.59 credit and 74 days remaining. Activate your full account to get unlimited access to all of Google Cloud–use any remaining credits, then pay only for what you use.
Dismiss
Activate
Google Cloud
My First Project
Search (/) for resources, docs, products, and more
Search
Dataproc
CLONE
DELETE
STOP
REFRESH
Jobs on Clusters
Clusters
Jobs
Workflows
Autoscaling policies
Serverless
Batches
Interactive
Interactive Templates
Metastore Services
Metastore
Federation
Utilities
Component exchange
Release Notes
MONITORING
CONFIGURATION
LINE WRAP: OFF
DISMISS
EQUIVALENT COMMAND LINE
ENG
10:06 AM
11-10-2024

Screenshot - PySpark:



Job details

Job ID	job-de30a86b
Job UUID	003c8c58-13dd-4fae-85ec-7d2c91fe27fb
Type	Dataproc Job
Status	Succeeded

Output

Spark jobs take ~60 seconds to initialize resources.

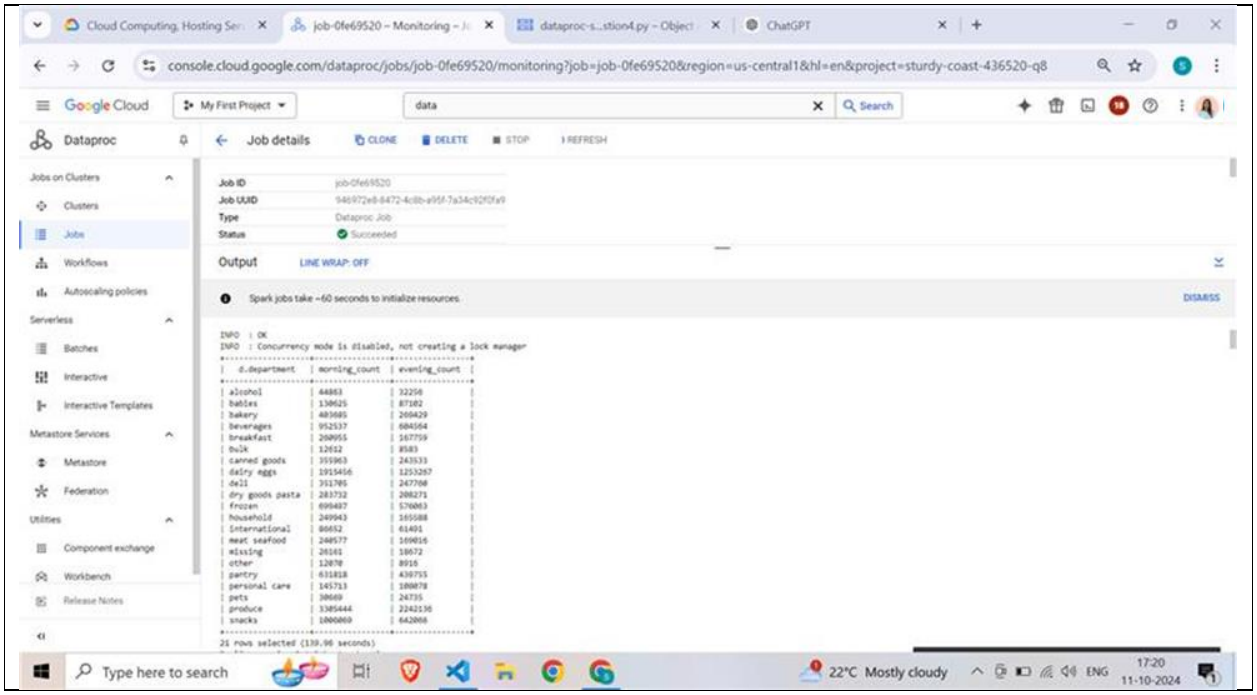
```
24/10/11 20:29:35 INFO Configuration: resource-types.xml not found
24/10/11 20:29:35 INFO ResourceUtils: Unable to find 'resource-types.xml'.
24/10/11 20:29:36 INFO YarnClientImpl: Submitted application application_1728677105045_0004
24/10/11 20:29:37 INFO DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at cluster-ecd5-m.us-central1-f.c.cogent-point-436520-s5.internal./10.128.0.
24/10/11 20:29:39 INFO RequestTracker: Detected high latency for [url=https://storage.googleapis.com/storage/v1/b/dataproc-temp-us-central1-202552022924-q86bxsaj/o?d
24/10/11 20:29:39 INFO GhfsGlobalStorageStatistics: periodic connector metrics: {gcs_api_client_non_found_response_count=1, gcs_api_client_side_error_count=1, gcs_ap
24/10/11 20:29:40 INFO GoogleCloudStorageImpl: Ignoring exception of type GoogleJsonResponseException; verified object already exists with desired state.
24/10/11 20:29:40 INFO GoogleHadoopOutputStream: hflush(): No-op due to rate limit (RateLimiter[stableRate=0.2qps]): readers will *not* yet see flushed data for gs:/
24/10/11 20:30:53 INFO GoogleHadoopOutputStream: hflush(): No-op due to rate limit (RateLimiter[stableRate=0.2qps]): readers will *not* yet see flushed data for gs:/
+--------------------+---------+-----------+
|        product_name|order_sum|reorder_sum|
+--------------------+---------+-----------+
|              Banana|   472565|   398609.0|
|Bag of Organic Ba...|   379450|   315913.0|
|Organic Strawberries|   264683|   205845.0|
+--------------------+---------+-----------+

24/10/11 20:30:55 INFO DataprocSpark                    http_patch_request=0, files_created=1, gcs_api_server_timeout_count=0, op_ge
```

Successfully deleted job-e5b8e880

job-de30a86b – Monitoring – Job details – Dataproc – My First Project – Google Cloud console
https://console.cloud.google.com/dataproc/jobs/job-de30a86b/monitoring?region=us-central1&authus...
Free trial status: $355.53 credit and 73 days remaining. Activate your full account to get unlimited access to all of Google Cloud–use any remaining credits, then pay only for what you use.
Dismiss
Activate
Google Cloud
My First Project
Search (/) for resources, docs, products, and more
Search
Dataproc
CLONE
DELETE
STOP
REFRESH
Jobs on Clusters
Clusters
Jobs
Workflows
Autoscaling policies
Serverless
Batches
Interactive
Interactive Templates
Metastore Services
Metastore
Federation
Utilities
Component exchange
Release Notes
LINE WRAP: OFF
DISMISS
https://console.cloud.google.com/dataproc/metastore/federations?authus...
ENG
04:31 PM
11-10-2024

7. Do orders from different departments vary over time of the day? Are there morning and evening departments (popular in the morning and popular in the evening)?

Screenshot - Hive:



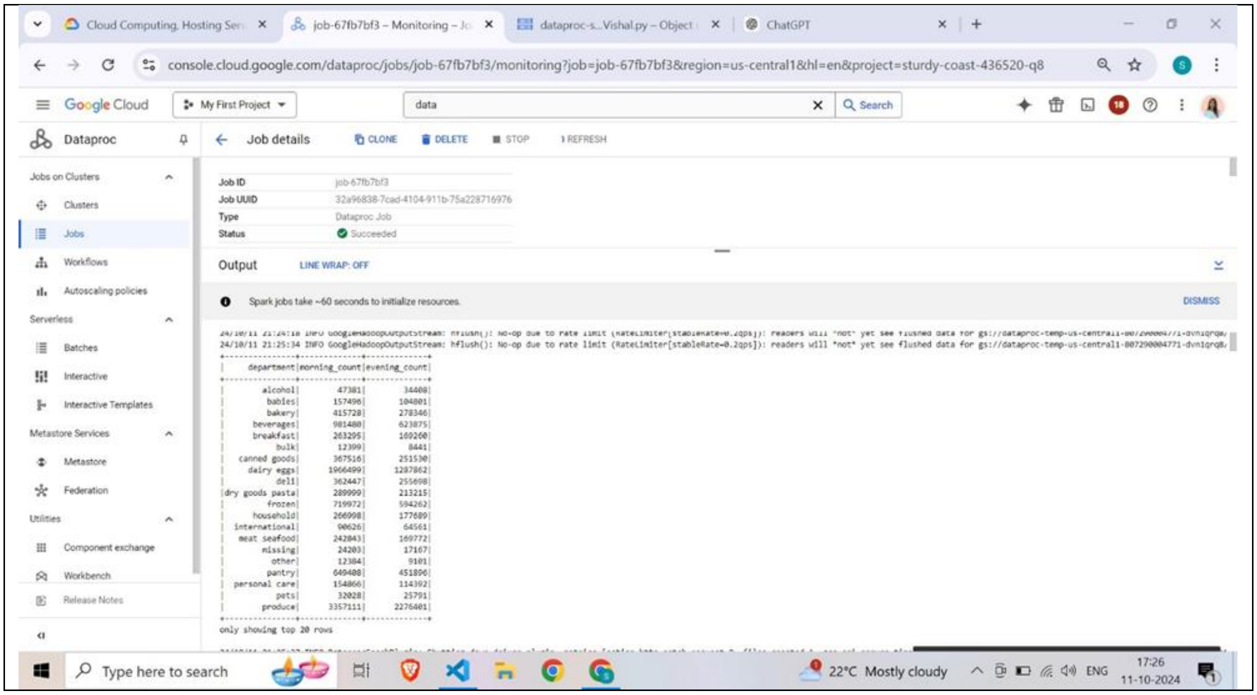
Job ID job-0fe69520
Job UUID 946972e8-8472-4c6b-a95f-7a34c92f0fa9
Type Dataproc Job
Status Succeeded

Output LINE WRAP: OFF

Spark jobs take ~60 seconds to initialize resources.

```
INFO  : OK
INFO  : Concurrency mode is disabled, not creating a lock manager
+-------------------+----------------+----------------+
|   d.department    | morning_count  | evening_count  |
+-------------------+----------------+----------------+
| alcohol           | 44863          | 32256          |
| babies            | 130625         | 87102          |
| bakery            | 403605         | 260429         |
| beverages         | 952537         | 604564         |
| breakfast         | 260955         | 167759         |
| bulk              | 12612          | 8583           |
| canned goods      | 355963         | 243533         |
| dairy eggs        | 1915456        | 1253267        |
| deli              | 351705         | 247700         |
| dry goods pasta   | 283732         | 208271         |
| frozen            | 699487         | 576063         |
| household         | 249943         | 165588         |
| international     | 86652          | 61401          |
| meat seafood      | 240577         | 169016         |
| missing           | 26161          | 18672          |
| other             | 12070          | 8916           |
| pantry            | 631818         | 430755         |
| personal care     | 145713         | 100078         |
| pets              | 30669          | 24735          |
| produce           | 3305444        | 2242136        |
| snacks            | 1000069        | 642008         |
+-------------------+----------------+----------------+
21 rows selected (139.96 seconds)
```

Screenshot - PySpark:



Job ID job-67fb7bf3
Job UUID 32a96838-7cad-4104-911b-75a228716976
Type Dataproc Job
Status Succeeded

Output LINE WRAP: OFF

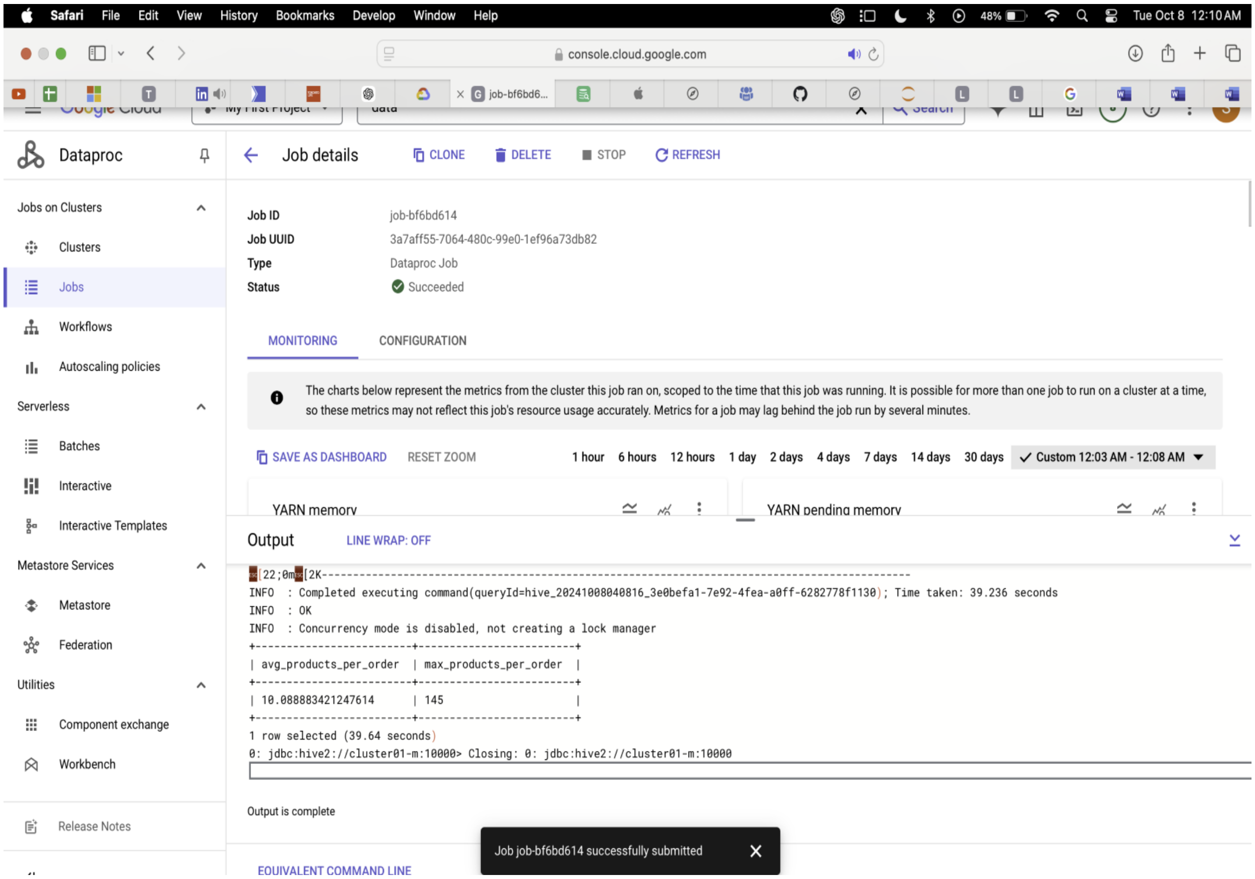
Spark jobs take ~60 seconds to initialize resources.

```
24/10/11 21:24:18 INFO GoogleHadoopOutputStream: hflush(): No-op due to rate limit (RateLimiter[stableRate=0.2qps]): readers will *not* yet see flushed data for gs://dataproc-temp-us-central1-807290004771-dvn1qrq8/
24/10/11 21:25:34 INFO GoogleHadoopOutputStream: hflush(): No-op due to rate limit (RateLimiter[stableRate=0.2qps]): readers will *not* yet see flushed data for gs://dataproc-temp-us-central1-807290004771-dvn1qrq8/
+---------------+-------------+-------------+
|     department|morning_count|evening_count|
+---------------+-------------+-------------+
|        alcohol|        47381|        34408|
|         babies|       157496|       104801|
|         bakery|       415728|       278346|
|      beverages|       981480|       623875|
|      breakfast|       263295|       169260|
|           bulk|        12399|         8441|
|   canned goods|       367516|       251530|
|     dairy eggs|      1966499|      1287862|
|           deli|       362447|       255698|
|dry goods pasta|       289999|       213215|
|         frozen|       719972|       594262|
|      household|       266998|       177689|
|  international|        90626|        64561|
|   meat seafood|       242843|       169772|
|        missing|        24203|        17167|
|          other|        12384|         9101|
|         pantry|       649408|       451896|
|  personal care|       154866|       114392|
|           pets|        32028|        25791|
|        produce|      3357111|      2276401|
+---------------+-------------+-------------+
only showing top 20 rows
```

—------- Only needed for 8 ppl groups

8. What is the average number of products in an order? What is the max?

Screenshot - Hive:



Job ID job-bf6bd614

Job UUID 3a7aff55-7064-480c-99e0-1ef96a73db82

Type Dataproc Job

Status ✓ Succeeded

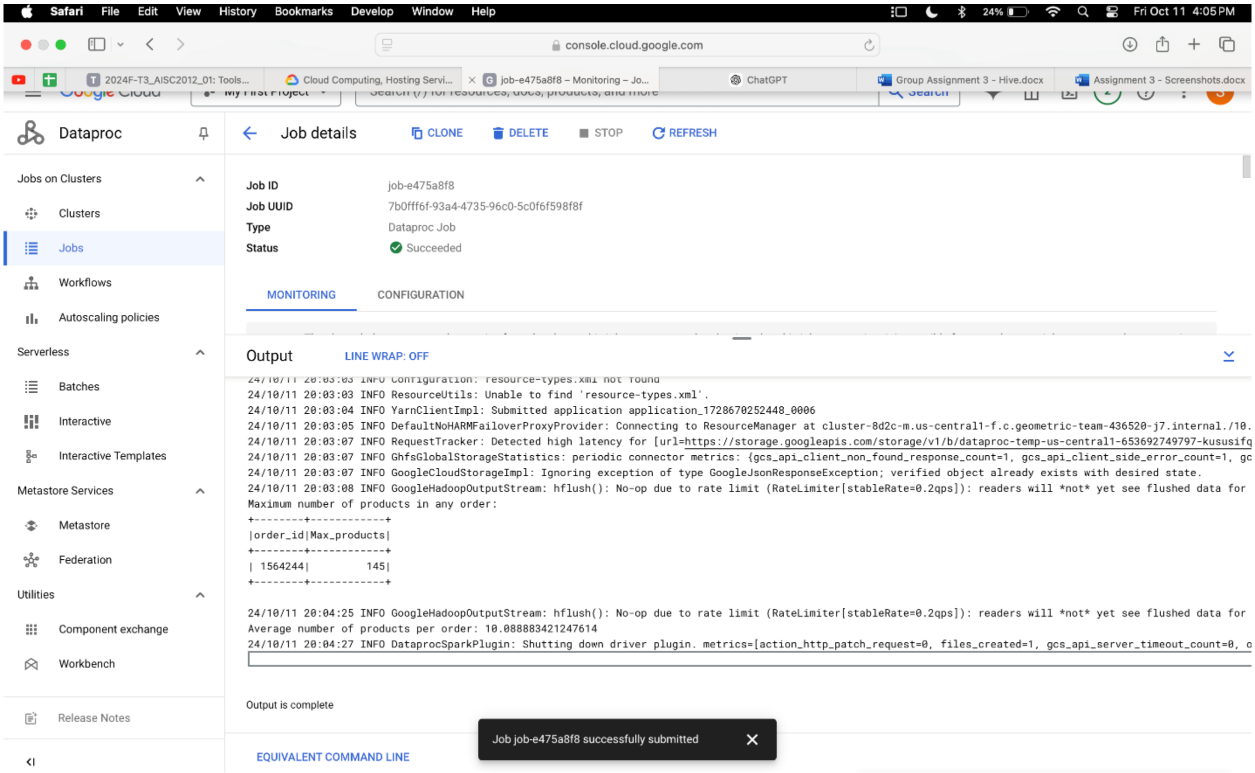
The charts below represent the metrics from the cluster this job ran on, scoped to the time that this job was running. It is possible for more than one job to run on a cluster at a time, so these metrics may not reflect this job's resource usage accurately. Metrics for a job may lag behind the job run by several minutes.

```
[22;0m[2K-----------------------------------------------------------------------------------------------
INFO  : Completed executing command(queryId=hive_20241008040816_3e0befa1-7e92-4fea-a0ff-6282778f1130); Time taken: 39.236 seconds
INFO  : OK
INFO  : Concurrency mode is disabled, not creating a lock manager
+-------------------------+-------------------------+
| avg_products_per_order  | max_products_per_order  |
+-------------------------+-------------------------+
| 10.088883421247614      | 145                     |
+-------------------------+-------------------------+
1 row selected (39.64 seconds)
0: jdbc:hive2://cluster01-m:10000> Closing: 0: jdbc:hive2://cluster01-m:10000
```

Output is complete

Job job-bf6bd614 successfully submitted

Screenshot - PySpark:



Job ID job-e475a8f8
Job UUID 7b0fff6f-93a4-4735-96c0-5c0f6f598f8f
Type Dataproc Job
Status ✓ Succeeded

```
24/10/11 20:03:03 INFO Configuration: resource-types.xml not found
24/10/11 20:03:03 INFO ResourceUtils: Unable to find 'resource-types.xml'.
24/10/11 20:03:04 INFO YarnClientImpl: Submitted application application_1728670252448_0006
24/10/11 20:03:05 INFO DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at cluster-8d2c-m.us-central1-f.c.geometric-team-436520-j7.internal./10.
24/10/11 20:03:07 INFO RequestTracker: Detected high latency for [url=https://storage.googleapis.com/storage/v1/b/dataproc-temp-us-central1-653692749797-kususifq
24/10/11 20:03:07 INFO GhfsGlobalStorageStatistics: periodic connector metrics: {gcs_api_client_non_found_response_count=1, gcs_api_client_side_error_count=1, gc
24/10/11 20:03:07 INFO GoogleCloudStorageImpl: Ignoring exception of type GoogleJsonResponseException; verified object already exists with desired state.
24/10/11 20:03:08 INFO GoogleHadoopOutputStream: hflush(): No-op due to rate limit (RateLimiter[stableRate=0.2qps]): readers will *not* yet see flushed data for
Maximum number of products in any order:
+--------+------------+
|order_id|Max_products|
+--------+------------+
| 1564244|         145|
+--------+------------+

24/10/11 20:04:25 INFO GoogleHadoopOutputStream: hflush(): No-op due to rate limit (RateLimiter[stableRate=0.2qps]): readers will *not* yet see flushed data for
Average number of products per order: 10.088883421247614
24/10/11 20:04:27 INFO DataprocSparkPlugin: Shutting down driver plugin. metrics=[action_http_patch_request=0, files_created=1, gcs_api_server_timeout_count=0, o
```

Output is complete

Job job-e475a8f8 successfully submitted