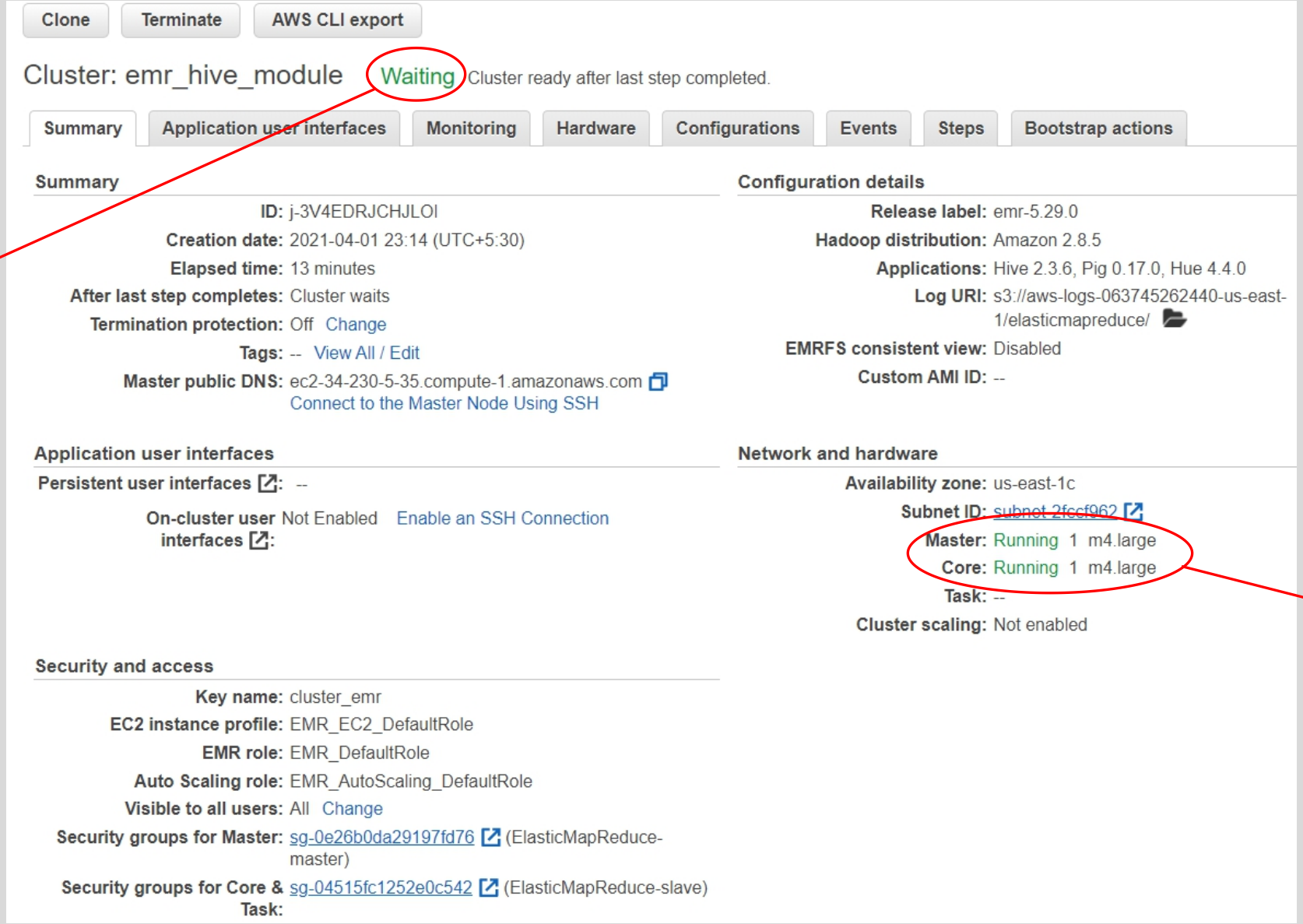
Hive Case Study

Working with a public clickstream dataset of a cosmetics store and extracting valuable insights

1) Launching EMR cluster -

The cluster is currently **active** and **waiting for steps to run**.



Both **“m4.large”** nodes **(master and core)** are now **running**.

2) Moving data from S3 to HDFS

Query to move data from **S3 to HDFS**.

```
[hadoop@ip-172-31-27-255 ~]$ hadoop distcp 's3://upgradprak123/oct_nov_data/*' '/user/hive/datafiles/'
21/03/28 13:32:35 INFO tools.DistCp: Input Options: DistCpOptions{atomicCommit=false, syncFolder=false, deleteMissing=false, ignoreFailures=false, overwrite=false, skipCRC=false, blocking=true, numListstatusThreads=0, maxMaps=20, mapBandwidth=100, sslConfigurationFile='null', copyStrategy='uniformsize', preserveStatus=[], preserveRawXattrs=false, atomicWorkPath=null, logPath=null, sourceFileListing=null, sourcePaths=[s3://upgradprak123/oct_nov_data/*], targetPath=/user/hive/datafiles, targetPathExists=false, filtersFile='null'}
21/03/28 13:32:36 INFO client.RMProxy: Connecting to ResourceManager at ip-172-31-27-255.ec2.internal/172.31.27.255:8032
21/03/28 13:32:40 INFO tools.SimpleCopyListing: Paths (files+dirs) cnt = 2; dirCnt = 0
21/03/28 13:32:40 INFO tools.SimpleCopyListing: Build file listing completed.
21/03/28 13:32:40 INFO Configuration.deprecation: io.sort.mb is deprecated. Instead, use mapreduce.task.io.sort.mb
21/03/28 13:32:40 INFO Configuration.deprecation: io.sort.factor is deprecated. Instead, use mapreduce.task.io.sort.factor
21/03/28 13:32:40 INFO tools.DistCp: Number of paths in the copy list: 2
21/03/28 13:32:40 INFO tools.DistCp: Number of paths in the copy list: 2
21/03/28 13:32:40 INFO client.RMProxy: Connecting to ResourceManager at ip-172-31-27-255.ec2.internal/172.31.27.255:8032
21/03/28 13:32:40 INFO mapreduce.JobSubmitter: number of splits:2
21/03/28 13:32:41 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1616937844129_0001
21/03/28 13:32:41 INFO impl.YarnClientImpl: Submitted application application_1616937844129_0001
21/03/28 13:32:42 INFO mapreduce.Job: The url to track the job: http://ip-172-31-27-255.ec2.internal:20888/proxy/application_1616937844129_0001/
21/03/28 13:32:42 INFO tools.DistCp: DistCp job-id: job_1616937844129_0001
21/03/28 13:32:42 INFO mapreduce.Job: Running job: job_1616937844129_0001
21/03/28 13:32:51 INFO mapreduce.Job: Job job_1616937844129_0001 running in uber mode : false
21/03/28 13:32:51 INFO mapreduce.Job:  map 0% reduce 0%
21/03/28 13:33:12 INFO mapreduce.Job:  map 100% reduce 0%
21/03/28 13:33:26 INFO mapreduce.Job: Job job_1616937844129_0001 completed successfully
21/03/28 13:33:26 INFO mapreduce.Job: Counters: 38
        File System Counters
                FILE: Number of bytes read=0
                FILE: Number of bytes written=344948
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=916
                HDFS: Number of bytes written=1028381690
                HDFS: Number of read operations=24
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=8
                S3: Number of bytes read=1028381690
                S3: Number of bytes written=0
                S3: Number of read operations=0
                S3: Number of large read operations=0
                S3: Number of write operations=0
        Job Counters
                Launched map tasks=2
                Other local map tasks=2
                Total time spent by all maps in occupied slots (ms)=2055040
```

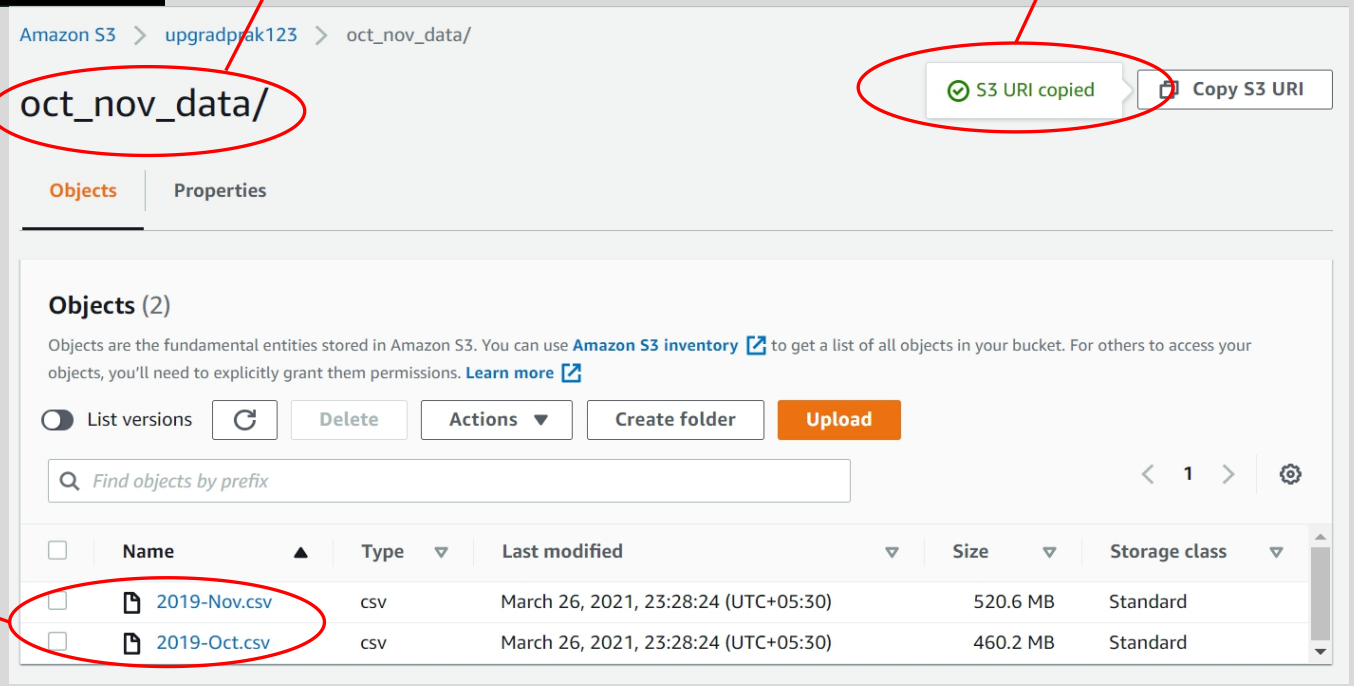
```
                Total time spent by all reduces in occupied slots (ms)=0
                Total time spent by all map tasks (ms)=64220
                Total vcore-milliseconds taken by all map tasks=64220
                Total megabyte-milliseconds taken by all map tasks=65761280
        Map-Reduce Framework
                Map input records=2
                Map output records=0
                Input split bytes=272
                Spilled Records=0
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=1078
                CPU time spent (ms)=41760
                Physical memory (bytes) snapshot=1164541952
                Virtual memory (bytes) snapshot=6581919744
                Total committed heap usage (bytes)=981467136
        File Input Format Counters
                Bytes Read=644
        File Output Format Counters
                Bytes Written=0
        DistCp Counters
                Bytes Copied=1028381690
                Bytes Expected=1028381690
                Files Copied=2
[hadoop@ip-172-31-27-255 ~]$
```

2

An S3 bucket named **“oct_nov_data”** has been made here, consisting of both the required files to be moved to HDFS.

URL for this particular S3 bucket is used in the command to move **files from S3 to HDFS**.

Both files **“2019-Nov.csv”** and **“2019-Oct.csv”** copied



3) Creating Database

```
[hadoop@ip-172-31-25-232 ~]$ hive

Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j2.properties Async: false
hive> create database hive;
OK
Time taken: 0.793 seconds
hive> show databases;
OK
default
hive
Time taken: 0.198 seconds, Fetched: 2 row(s)
hive> use  hive;
OK
Time taken: 0.047 seconds
hive>
```

- Enter Hive CLI by giving **“Hive”** command.
- Create database **“hive”**.
- We can see the database has been created using **“show databases”** command.
- Set the database to **use**.

4) Creating Tables

- We will be making 2 tables for our data -
  - One normal table with no optimization - **“dump_data”**
  - Optimized table for faster and efficient query processing - **“optimize_data”**

4.1) “dump_data” table -

```
hive> CREATE EXTERNAL TABLE IF NOT EXISTS dump_data
    > (event_time timestamp, event_type string, product_id string, category_id string, category_code string,brand string,
    >  price float, user_id int, user_session string)
    > ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.OpenCSVSerde'
    > STORED AS TEXTFILE
    > LOCATION '/user/hive/datafiles'
    > tblproperties ("skip.header.line.count"="1");
OK
Time taken: 0.067 seconds
```

To avoid the header line from being included into the dataset.

Pointer provided towards the **“datafiles”** folder made in **HDFS** containing the data.

CSVSerde with the default properties value for loading the dataset.

Describing “dump_data” table -

```
hive> desc dump_data;
OK
col_name        data_type       comment
event_time              string                  from deserializer
event_type              string                  from deserializer
product_id              string                  from deserializer
category_id             string                  from deserializer
category_code           string                  from deserializer
brand                   string                  from deserializer
price                   string                  from deserializer
user_id                 string                  from deserializer
user_session            string                  from deserializer
Time taken: 0.11 seconds, Fetched: 9 row(s)
```

Checking “dump_data” table -

```
hive> select * from dump_data limit 6;
OK
dump_data.event_time    dump_data.event_type    dump_data.product_id    dump_data.category_id   dump_data.category_code dump_data.brand dump_data.price dump_data.user_id       dump_data.use
r_session
2019-11-01 00:00:02 UTC view    5802432 1487580009286598681                     0.32    562076640       09fafd6c-6c99-46b1-834f-33527f4de241
2019-11-01 00:00:09 UTC cart    5844397 1487580006317032337                     2.38    553329724       2067216c-31b5-455d-a1cc-af0575a34ffb
2019-11-01 00:00:10 UTC view    5837166 1783999064103190764             pnb     22.22   556138645       57ed222e-a54a-4907-9944-5a875c2d7f4f
2019-11-01 00:00:11 UTC cart    5876812 1487580010100293687             jessnail        3.16    564506666       186c1951-8052-4b37-adce-dd9644b1d5f7
2019-11-01 00:00:24 UTC remove_from_cart        5826182 1487580007483048900                     3.33    553329724       2067216c-31b5-455d-a1cc-af0575a34ffb
2019-11-01 00:00:24 UTC remove_from_cart        5826182 1487580007483048900                     3.33    553329724       2067216c-31b5-455d-a1cc-af0575a34ffb
Time taken: 2.178 seconds, Fetched: 6 row(s)
```

4.2) “optimize_data” table -

Partition by event_type string, event_month int
Clustered by category_code string, brand string

```
hive> create table if not exists optimize_data (event_time timestamp, product_id string, category_id string,category_code string,brand string, price float, user_id bigint, user_session stri
ng)
    > partitioned by (event_type string,event_month int) clustered by (category_code,brand) into 10 buckets
    > ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.OpenCSVSerde'
    > stored as textfile;
OK
```

5) Queries

Question 1: **Find the total revenue generated due to purchases made in October**

Note: *For question 1, we will compare the query processing time in case of "dump_data" table and that of "optimize_data" table to see the performance improvement.*

Query 1 -

```
SELECT ROUND(SUM(price),2) AS Total_Revenue
FROM dump_data
WHERE event_type='purchase' AND month(event_time)=10;
```

Query 2 -

```
SELECT ROUND(SUM(price),2) AS Total_Revenue
FROM optimize_data
WHERE event_type='purchase' AND month(event_time)=10;
```

a) Using “dump_data” table -

Query time: 53.475 seconds

```
hive> SELECT ROUND(SUM(price),2)AS Total_Revenue
    > FROM dump_data
    > WHERE event_type='purchase' AND month(event_time)=10;
Query ID = hadoop_20210331121111_9a92ffb2-cb97-4152-88ce-a3d15848daab
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1617191194187_0003)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      2          2        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 52.65 s
--------------------------------------------------------------------------------
OK
total_revenue
1211538.43
Time taken: 53.475 seconds, Fetched: 1 row(s)
hive>
```

b) Using “optimize_data” table -

Query time: 24.635 seconds

As we can clearly see, when we use the optimized table, the query time is reduced to half from around 53 seconds to around 25 seconds. Hence, for further queries, we will continue with optimized table “optimize_data” for better results.

```
hive> SELECT ROUND(SUM(price),2)AS Total_Revenue
    > FROM optimize_data
    > WHERE event_type='purchase' AND month(event_time)=10;
Query ID = hadoop_20210331120906_a3ab0be7-c84d-4975-95aa-61342939f4c3
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1617191194187_0003)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      3          3        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 23.89 s
--------------------------------------------------------------------------------
OK
total_revenue
1211538.43
Time taken: 24.635 seconds, Fetched: 1 row(s)
hive>
```

Question 2: **Write a query to yield the total sum of purchases per month in a single output**

Query -

```
SELECT event_month AS Purchase_Month, ROUND(SUM(price),2) AS Sum_Of_Purchases
FROM optimize_data
WHERE event_type = ‘purchase’
GROUP BY event_month;
```

Screenshot for Question 2 -

```
hive> SELECT event_month AS Purchase_Month, ROUND(SUM(price),2) AS Sum_Of_Purchases
    > FROM optimize_data
    > WHERE event_type = 'purchase'
    > GROUP BY event_month;
Query ID = hadoop_20210331121527_c6987330-47ac-4808-a764-82ced9d4092b
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1617191194187_0003)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      3          3        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 20.23 s
--------------------------------------------------------------------------------
OK
purchase_month  sum_of_purchases
10      1211538.43
11      1531016.9
Time taken: 20.901 seconds, Fetched: 2 row(s)
hive>
```

As we can see here, total purchases for **October** are **1211538.43** while for **November** are **1531016.9**

Question 3: **Write a query to find the change in revenue generated due to purchases from October to November**

Query -

```
SELECT Oct_Rev, Nov_Rev, Nov_Rev - Oct_Rev as Rev_Change
FROM
(
SELECT round(sum(case when month(event_time)=10 then price else 0 end), 2) AS Oct_Rev,
        round(sum(case when month(event_time)=11 then price else 0 end), 2) AS Nov_Rev
FROM optimize_data
WHERE event_type=’purchase’
) a;
```

Screenshot for Question 3 -

```
hive> SELECT Oct_Rev, Nov_Rev, Nov_Rev - Oct_Rev as Rev_Change
    > FROM
    > (
    > SELECT round(sum(case when month(event_time)=10 then price else 0 end),2) AS Oct_Rev,
    >        round(sum(case when month(event_time)=11 then price else 0 end),2) AS Nov_Rev
    > FROM optimize_data
    > WHERE event_type='purchase'
    > ) a;
Query ID = hadoop_20210401182015_5c3ea4c4-d48a-4250-8676-182c5dbc52b2
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1617299489868_0002)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      3          3        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 25.09 s
----------------------------------------------------------------------------------------------
OK
oct_rev nov_rev rev_change
1211538.43      1531016.9       319478.47
Time taken: 25.682 seconds, Fetched: 1 row(s)
hive>
```

As we can see here, the revenue in **November** has **increased by 319478.47**

Question 4: **Find distinct categories of products. Categories with null category code can be ignored**

Query -

```
SELECT distinct(category_code) AS category_name
FROM optimize_data
WHERE category_code!=’’;
```

Screenshot for Question 4 -

List of **distinct product categories**

```
hive> SELECT distinct(category_code) AS category_name
    > FROM optimize_data
    > WHERE category_code!='';
Query ID = hadoop_20210331122258_a52e3669-80f6-4854-a7fb-8b76ed13a08c
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1617191194187_0004)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      6          6        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      5          5        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 63.12 s
----------------------------------------------------------------------------------------------
OK
category_name
accessories.cosmetic_bag
stationery.cartrige
accessories.bag
appliances.environment.vacuum
furniture.living_room.chair
sport.diving
appliances.personal.hair_cutter
appliances.environment.air_conditioner
apparel.glove
furniture.bathroom.bath
furniture.living_room.cabinet
Time taken: 63.792 seconds, Fetched: 11 row(s)
hive>
```

Question 5: **Find the total number of products available under each category**

Query -

```
SELECT category_code, count(product_id) AS Total_Products
FROM optimize_data
WHERE category_code!=’’
GROUP BY category_code
ORDER BY Total_Products DESC;
```

Screenshot for Question 5 -

```
hive> SELECT category_code, count(product_id) AS Total_Products
    > FROM optimize_data
    > WHERE category_code!=''
    > GROUP BY category_code
    > ORDER BY Total_Products DESC;
Query ID = hadoop_20210401201351_d681fb46-fce4-469c-b2ec-9bec2189061c
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1617306740879_0002)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      6          6        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      5          5        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 63.43 s
--------------------------------------------------------------------------------
OK
category_code   total_products
appliances.environment.vacuum   59761
stationery.cartrige     26722
apparel.glove   18232
furniture.living_room.cabinet   13439
accessories.bag 11681
furniture.bathroom.bath 9857
appliances.personal.hair_cutter 1643
accessories.cosmetic_bag        1248
appliances.environment.air_conditioner  332
furniture.living_room.chair     308
sport.diving    2
Time taken: 64.187 seconds, Fetched: 11 row(s)
hive>
```

List of **product categories** and the **number of products** available in each.

Question 6: **Which brand had the maximum sales in October and November combined?**

Query -

```
SELECT brand AS Brand_With_Max_Sales, ROUND(SUM(price),2) AS Total_Sales
FROM optimize_data
WHERE event_type=’purchase’ and brand!=’’
GROUP BY brand
ORDER BY Total_Sales DESC
LIMIT 1;
```

Screenshot for Question 6 -

```
hive> SELECT brand AS Brand_With_Max_Sales, ROUND(SUM(price),2) AS Total_Sales
    > FROM optimize_data
    > WHERE event_type='purchase' and brand!=''
    > GROUP BY brand
    > ORDER BY Total_Sales DESC
    > LIMIT 1;
Query ID = hadoop_20210331130736_c9b3a1ee-62f7-47f7-8a1a-0e69f74be3bc
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1617191194187_0006)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      3          3        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 21.77 s
--------------------------------------------------------------------------------
OK
brand_with_max_sales    total_sales
runail  148297.94
Time taken: 22.36 seconds, Fetched: 1 row(s)
hive>
```

The brand to have **maximum sales** in October and November combined is **“Runail”**.

Question 7: **Which brands increased their sales from October to November**

Query -

```
SELECT a.brand as brands_name
FROM
(SELECT brand,
        ROUND(SUM(case when event_month=10 then price else 0 end), 2) AS Oct_Sale,
        ROUND(SUM(case when event_month=11 then price else 0 end), 2) AS Nov_Sale
FROM optimize_data
WHERE event_type='purchase' and brand!=''
GROUP BY brand
)a
WHERE Oct_Sale-Nov_Sale>0;
```

Screenshot for Question 7 -

```
hive> SELECT a.brand as brands_name
    > FROM
    > (SELECT brand,
    >         ROUND(SUM(case when event_month=10 then price else 0 end),2) AS Oct_Sale,
    >         ROUND(SUM(case when event_month=11 then price else 0 end),2) AS Nov_Sale
    > FROM optimize_data
    > WHERE event_type='purchase' and brand!=''
    > GROUP BY brand
    > )a
    > WHERE Oct_Sale-Nov_Sale>0;
Query ID = hadoop_20210331125146_7923141e-b9a1-4cfa-b358-79b66e58e3d8
Total jobs = 1
Launching Job 1 out of 1
Tez session was closed. Reopening...
Session re-established.
Status: Running (Executing on YARN cluster with App id application_1617191194187_0006)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 ..........   container     SUCCEEDED      3          3        0        0       0       0
Reducer 2 ......   container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 23.35 s
--------------------------------------------------------------------------------
OK
brands_name
almea
andrea
ardell
bergamo
bespecial
biofollica
bosnic
cnd
consly
coocla
coxir
cruset
dermal
dorena
dr.gloderm
emil
enas
enigma
eunyul
fancy
farmstay
frozen
```

1

```
i-laq
inoface
keune
koreatida
labay
lakme
lamixx
lebelage
litaline
lsanic
lunaris
max
meisterwerk
mielle
naturmed
nitrimax
parachute
petitfee
philips
pnb
pole
riche
rocknailstar
sawa
siberina
skipofit
sun
sunuv
tannymaxx
tazol
thuya
tosowoong
vosev
weaver
ypsed
zinger
Time taken: 31.125 seconds, Fetched: 58 row(s)
hive>
```

2

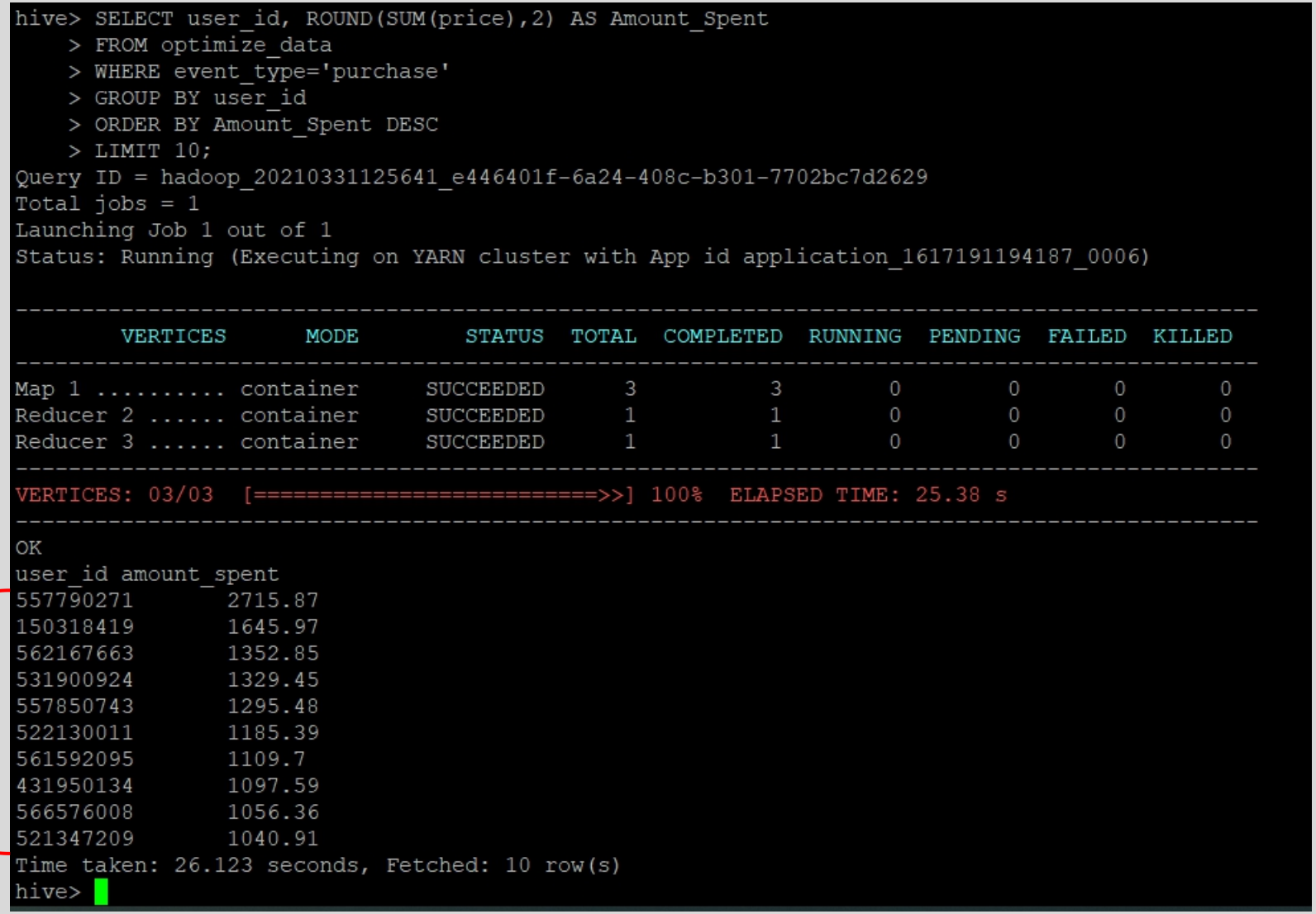
Question 8: **Your company wants to reward the top 10 users of its website with a Golden Customer plan. Write a query to generate a list of top 10 users who spend the most.**

Query -

```
SELECT user_id, ROUND(SUM(price), 2) AS Amount_Spent
FROM optimize_data
WHERE event_type='purchase'
GROUP BY user_id
ORDER BY Amount_Spent DESC
LIMIT 10;
```

Screenshot for Question 8 -

List of User ID's of **top 10 users** of the website eligible for the **Golden Customer Plan**.

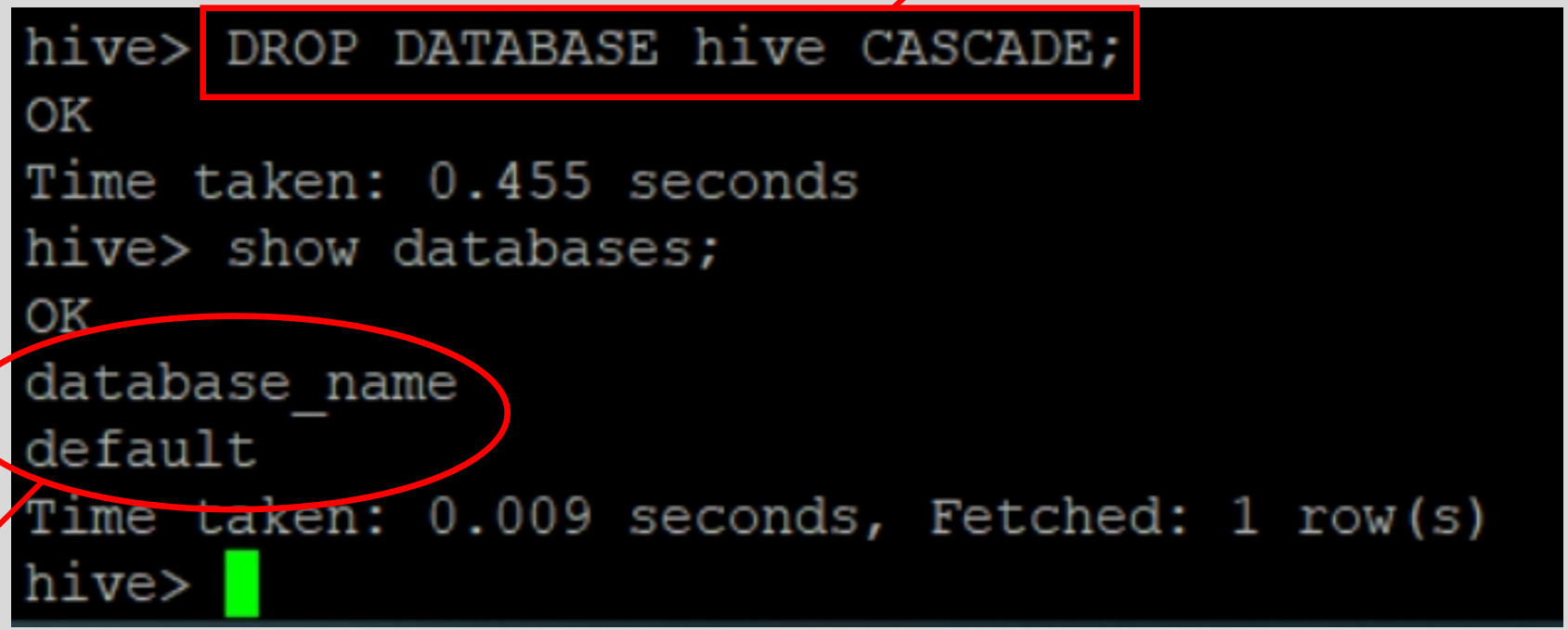


```
hive> SELECT user_id, ROUND(SUM(price),2) AS Amount_Spent
    > FROM optimize_data
    > WHERE event_type='purchase'
    > GROUP BY user_id
    > ORDER BY Amount_Spent DESC
    > LIMIT 10;
Query ID = hadoop_20210331125641_e446401f-6a24-408c-b301-7702bc7d2629
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1617191194187_0006)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      3          3        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 25.38 s
--------------------------------------------------------------------------------
OK
user_id amount_spent
557790271       2715.87
150318419       1645.97
562167663       1352.85
531900924       1329.45
557850743       1295.48
522130011       1185.39
561592095       1109.7
431950134       1097.59
566576008       1056.36
521347209       1040.91
Time taken: 26.123 seconds, Fetched: 10 row(s)
hive>
```

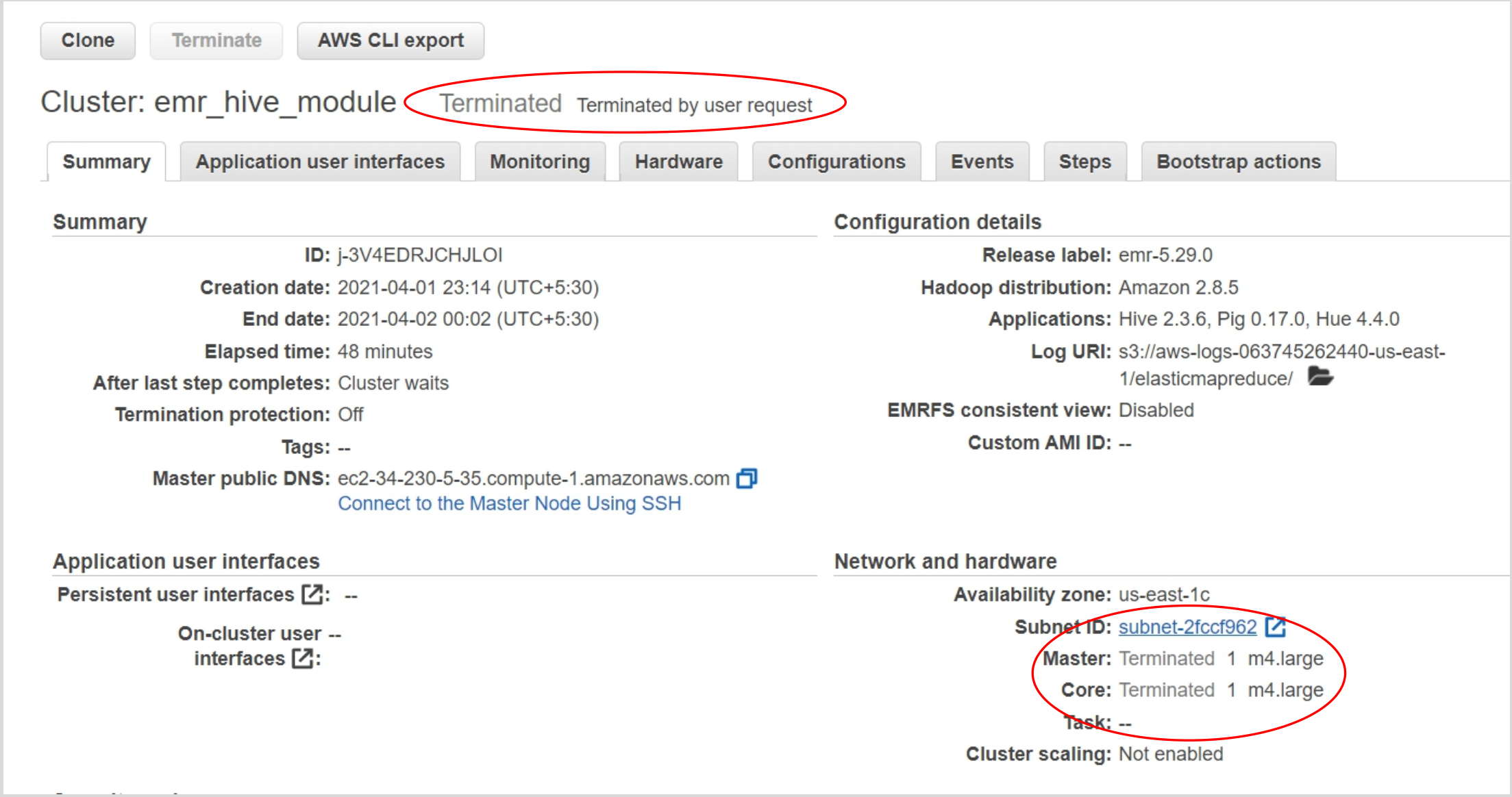
6) Dropping Database

Here, since we have our database with tables in it, we can use the **“CASCADE”** keyword, which **drops the databases loaded with tables**.



As we can see here, our database **“hive” is now gone**.

7) Terminating EMR cluster



THANK YOU!

Submitted By -
Prakhar Shrivastava (DS C22)
Milan Mithal (DS C22)