Hive Case Study Assignment

[DS C29 - 2021]

Ecommerce Sales Data Analysis

By-

Shahad.Riyaz.Shaikh

And

Hanumant.Vaidya

Problem Statement

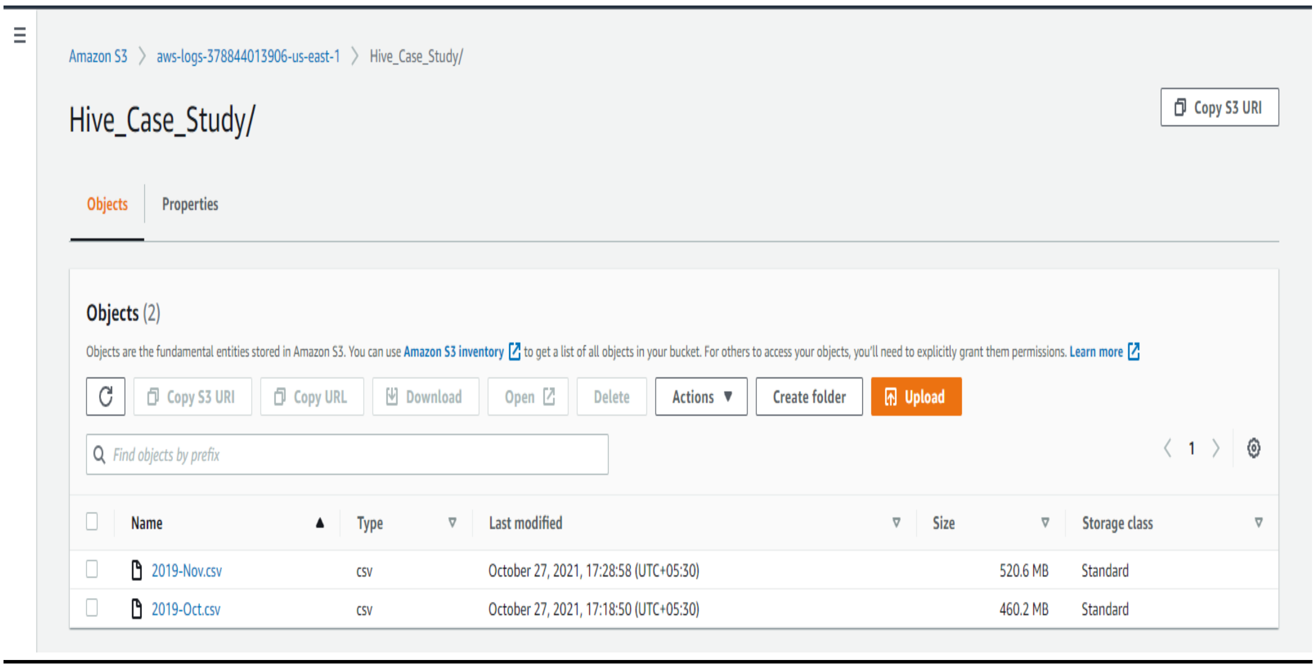
With online sales gaining popularity, tech companies are exploring ways to improve their sales by analyzing customer behavior and gaining insights about product trends. Furthermore, the websites make it easier for customers to find the products they require without much scavenging. Needless to say, the role of big data analysts is among the most sought-after job profiles of this decade. Therefore, as part of this assignment, we will be challenging you, as a big data analyst, to extract data and gather insights from a real-life data set of an e-commerce company.

The implementation phase can be divided into the following parts:

- Copying the data set into the HDFS:
- Launch an EMR cluster that utilizes the Hive services, and
- Move the data from the S3 bucket into the HDFS
- Creating the database and launching Hive queries on your EMR cluster:
- Create the structure of your database,
- Use optimized techniques to run your queries as efficiently as possible
- Show the improvement of the performance after using optimization on any single query.
- Run Hive queries to answer the questions given below.
- Cleaning up -:
- Drop your database, and
- Terminate your cluster

Data Collection and Processing

1. Uploading the data files 2019-Nov.csv & 2019-Oct.csv in AWS S3 platform.



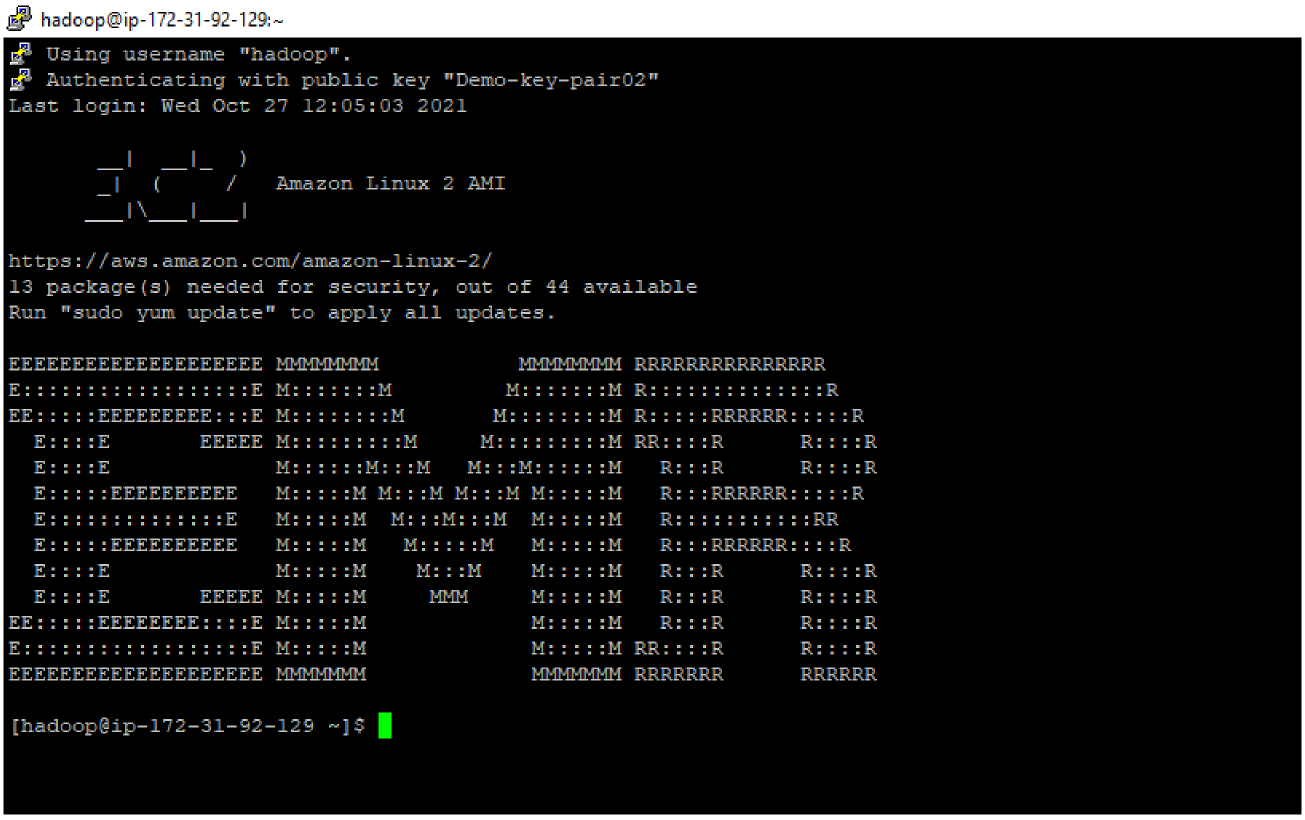
Hive_Case_Study/

Objects (2)

Objects are the fundamental entities stored in Amazon S3. You can use Amazon S3 inventory to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. Learn more

□	Name ▲	Type	Last modified	Size	Storage class
□	2019-Nov.csv	csv	October 27, 2021, 17:28:58 (UTC+05:30)	520.6 MB	Standard
□	2019-Oct.csv	csv	October 27, 2021, 17:18:50 (UTC+05:30)	460.2 MB	Standard

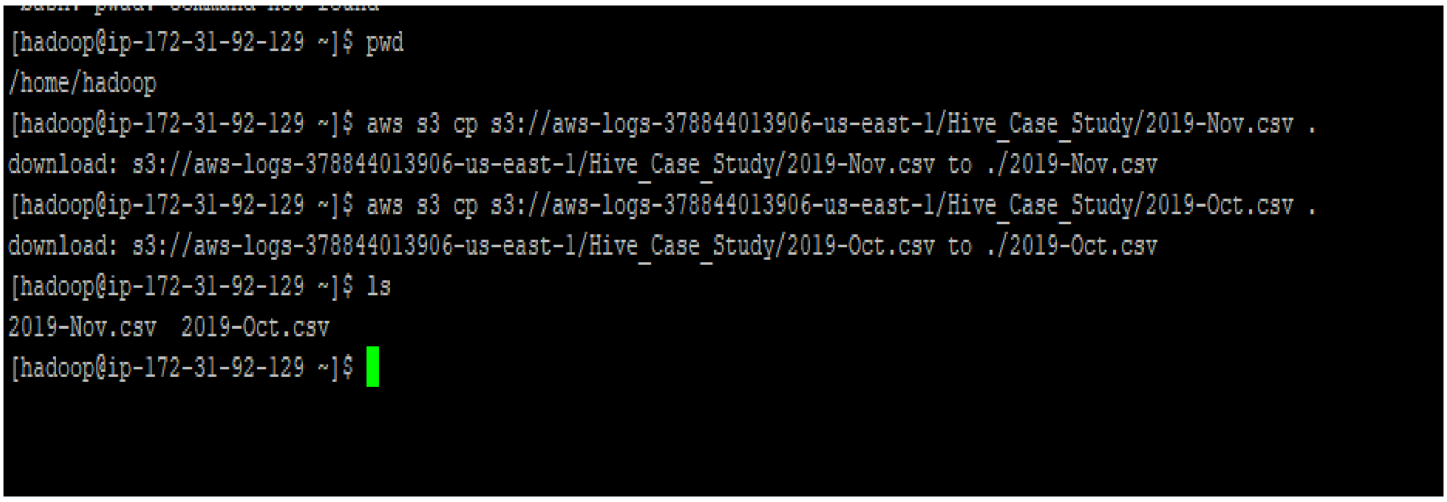
2. Launching the AWS EMR cluster via putty.exe.



hadoop@ip-172-31-92-129:~

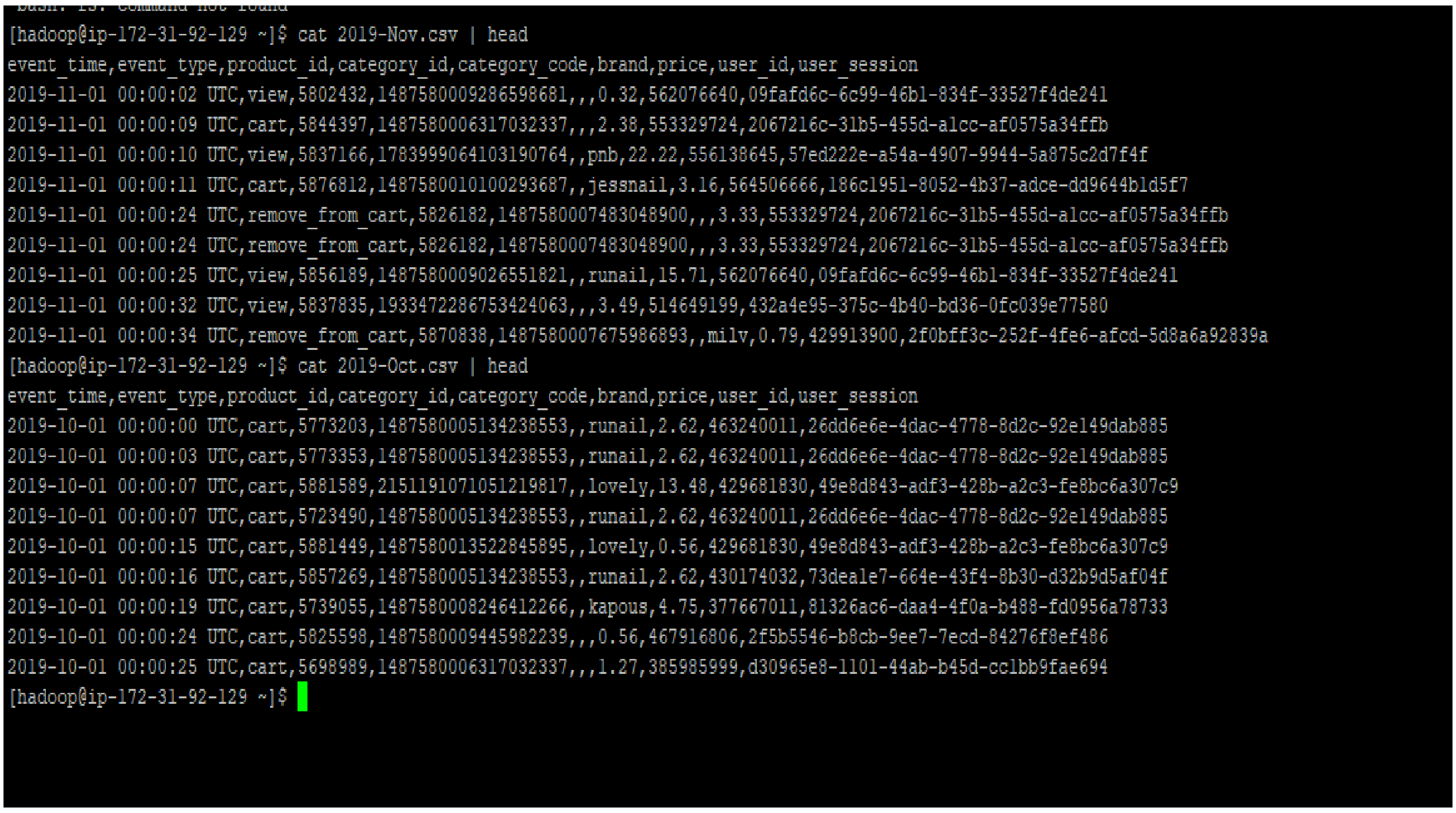
```
Using username "hadoop".
Authenticating with public key "Demo-key-pair02"
Last login: Wed Oct 27 12:05:03 2021

       __|  __|_  )
       _|  (     /   Amazon Linux 2 AMI
      ___|\___|___|

https://aws.amazon.com/amazon-linux-2/
13 package(s) needed for security, out of 44 available
Run "sudo yum update" to apply all updates.

EEEEEEEEEEEEEEEEEEEE MMMMMMMM           MMMMMMMM RRRRRRRRRRRRRRR
E::::::::::::::::::E M:::::::M         M:::::::M R::::::::::::::R
EE:::::EEEEEEEEE:::E M::::::::M       M::::::::M R:::::RRRRRR:::::R
  E::::E       EEEEE M:::::::::M     M:::::::::M RR::::R      R::::R
  E::::E             M::::::M:::M   M:::M::::::M   R:::R      R::::R
  E:::::EEEEEEEEEE   M:::::M M:::M M:::M M:::::M   R:::RRRRRR:::::R
  E::::::::::::::E   M:::::M  M:::M:::M  M:::::M   R:::::::::::RR
  E:::::EEEEEEEEEE   M:::::M   M:::::M   M:::::M   R:::RRRRRR::::R
  E::::E             M:::::M    M:::M    M:::::M   R:::R      R::::R
  E::::E       EEEEE M:::::M     MMM     M:::::M   R:::R      R::::R
EE:::::EEEEEEEEE::::E M:::::M             M:::::M   R:::R      R::::R
E::::::::::::::::::E M:::::M             M:::::M RR::::R      R::::R
EEEEEEEEEEEEEEEEEEEE MMMMMMM             MMMMMMM RRRRRRR      RRRRRR

[hadoop@ip-172-31-92-129 ~]$
```

3. Loading both the given datasets in the HDFS.



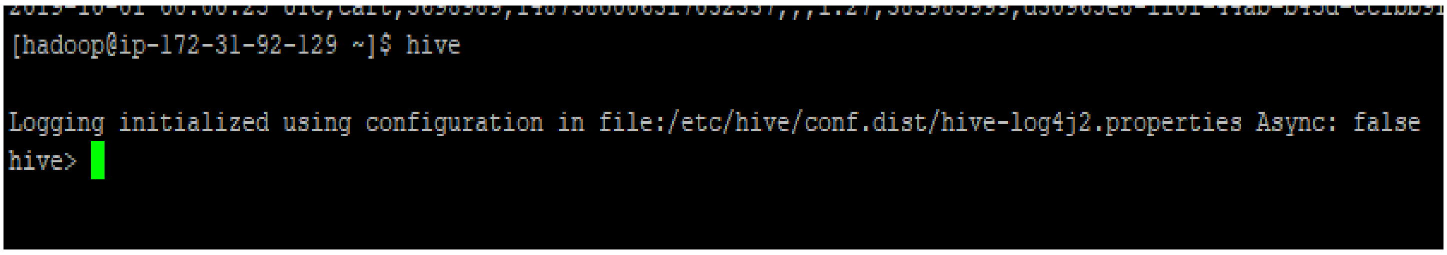
```
[hadoop@ip-172-31-92-129 ~]$ pwd
/home/hadoop
[hadoop@ip-172-31-92-129 ~]$ aws s3 cp s3://aws-logs-378844013906-us-east-1/Hive_Case_Study/2019-Nov.csv .
download: s3://aws-logs-378844013906-us-east-1/Hive_Case_Study/2019-Nov.csv to ./2019-Nov.csv
[hadoop@ip-172-31-92-129 ~]$ aws s3 cp s3://aws-logs-378844013906-us-east-1/Hive_Case_Study/2019-Oct.csv .
download: s3://aws-logs-378844013906-us-east-1/Hive_Case_Study/2019-Oct.csv to ./2019-Oct.csv
[hadoop@ip-172-31-92-129 ~]$ ls
2019-Nov.csv  2019-Oct.csv
[hadoop@ip-172-31-92-129 ~]$
```

4. Viewing both the datasets 2019-Nov.csv & 2019-Oct.csv in HDFS.



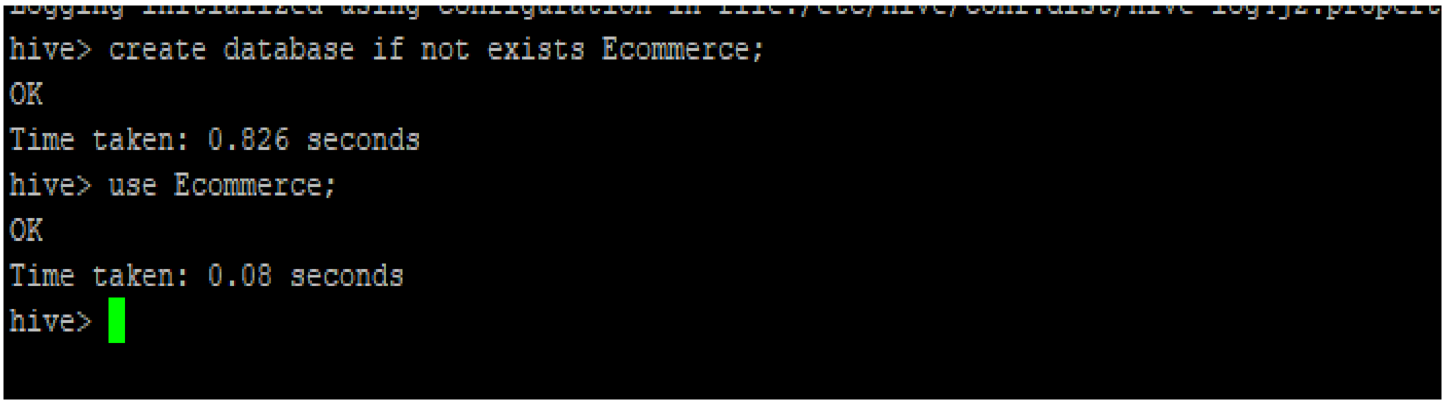
```
[hadoop@ip-172-31-92-129 ~]$ cat 2019-Nov.csv | head
event_time,event_type,product_id,category_id,category_code,brand,price,user_id,user_session
2019-11-01 00:00:02 UTC,view,5802432,1487580009286598681,,,0.32,562076640,09fafd6c-6c99-46b1-834f-33527f4de241
2019-11-01 00:00:09 UTC,cart,5844397,1487580006317032337,,,2.38,553329724,2067216c-31b5-455d-a1cc-af0575a34ffb
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
2019-11-01 00:00:25 UTC,view,5856189,1487580009026551821,,runail,15.71,562076640,09fafd6c-6c99-46b1-834f-33527f4de241
2019-11-01 00:00:32 UTC,view,5837835,1933472286753424063,,,3.49,514649199,432a4e95-375c-4b40-bd36-0fc039e77580
2019-11-01 00:00:34 UTC,remove_from_cart,5870838,1487580007675986893,,milv,0.79,429913900,2f0bff3c-252f-4fe6-afcd-5d8a6a92839a
[hadoop@ip-172-31-92-129 ~]$ cat 2019-Oct.csv | head
event_time,event_type,product_id,category_id,category_code,brand,price,user_id,user_session
2019-10-01 00:00:00 UTC,cart,5773203,1487580005134238553,,runail,2.62,463240011,26dd6e6e-4dac-4778-8d2c-92e149dab885
2019-10-01 00:00:03 UTC,cart,5773353,1487580005134238553,,runail,2.62,463240011,26dd6e6e-4dac-4778-8d2c-92e149dab885
2019-10-01 00:00:07 UTC,cart,5881589,2151191071051219817,,lovely,13.48,429681830,49e8d843-adf3-428b-a2c3-fe8bc6a307c9
2019-10-01 00:00:07 UTC,cart,5723490,1487580005134238553,,runail,2.62,463240011,26dd6e6e-4dac-4778-8d2c-92e149dab885
2019-10-01 00:00:15 UTC,cart,5881449,1487580013522845895,,lovely,0.56,429681830,49e8d843-adf3-428b-a2c3-fe8bc6a307c9
2019-10-01 00:00:16 UTC,cart,5857269,1487580005134238553,,runail,2.62,430174032,73deale7-664e-43f4-8b30-d32b9d5af04f
2019-10-01 00:00:19 UTC,cart,5739055,1487580008246412266,,kapous,4.75,377667011,81326ac6-daa4-4f0a-b488-fd0956a78733
2019-10-01 00:00:24 UTC,cart,5825598,1487580009445982239,,,0.56,467916806,2f5b5546-b8cb-9ee7-7ecd-84276f8ef486
2019-10-01 00:00:25 UTC,cart,5698989,1487580006317032337,,,1.27,385985999,d30965e8-1101-44ab-b45d-cc1bb9fae694
[hadoop@ip-172-31-92-129 ~]$
```

5. Launching Hive



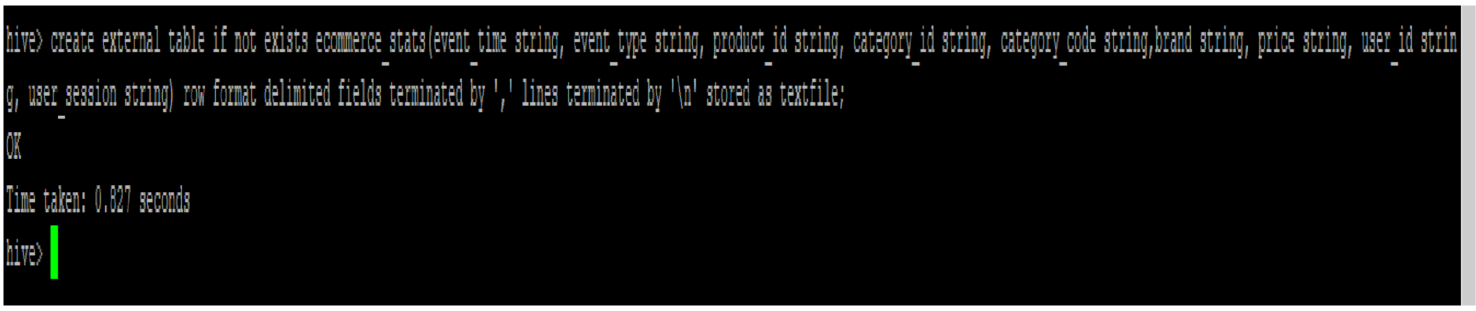
```
[hadoop@ip-172-31-92-129 ~]$ hive

Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j2.properties Async: false
hive>
```

6. Creating the database ‘Ecommerce’ and using it in Hive.



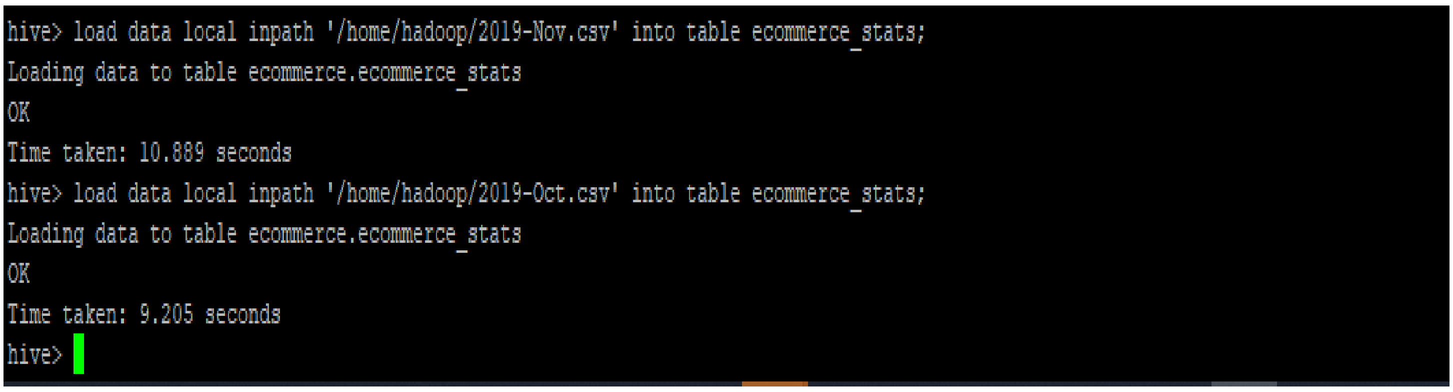
```
Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j2.propert
hive> create database if not exists Ecommerce;
OK
Time taken: 0.826 seconds
hive> use Ecommerce;
OK
Time taken: 0.08 seconds
hive>
```

7. Creating an External table ‘ecommerce_stats’.



```
hive> create external table if not exists ecommerce_stats(event_time string, event_type string, product_id string, category_id string, category_code string,brand string, price string, user_id strin
g, user_session string) row format delimited fields terminated by ',' lines terminated by '\n' stored as textfile;
OK
Time taken: 0.827 seconds
hive>
```

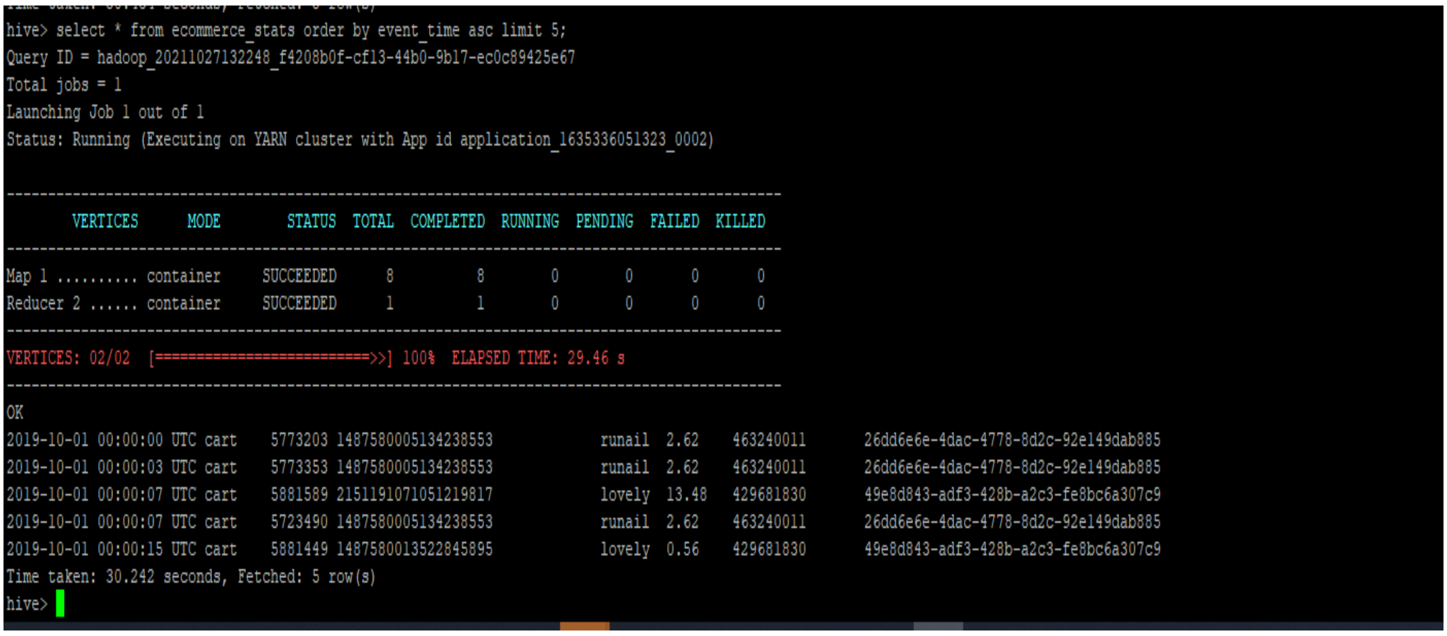
8. Loading and inserting the data 2019-Nov.csv & 2019-Oct.csv in the ‘ecommerce_stats’ table.



```
hive> load data local inpath '/home/hadoop/2019-Nov.csv' into table ecommerce_stats;
Loading data to table ecommerce.ecommerce_stats
OK
Time taken: 10.889 seconds
hive> load data local inpath '/home/hadoop/2019-Oct.csv' into table ecommerce_stats;
Loading data to table ecommerce.ecommerce_stats
OK
Time taken: 9.205 seconds
hive>
```

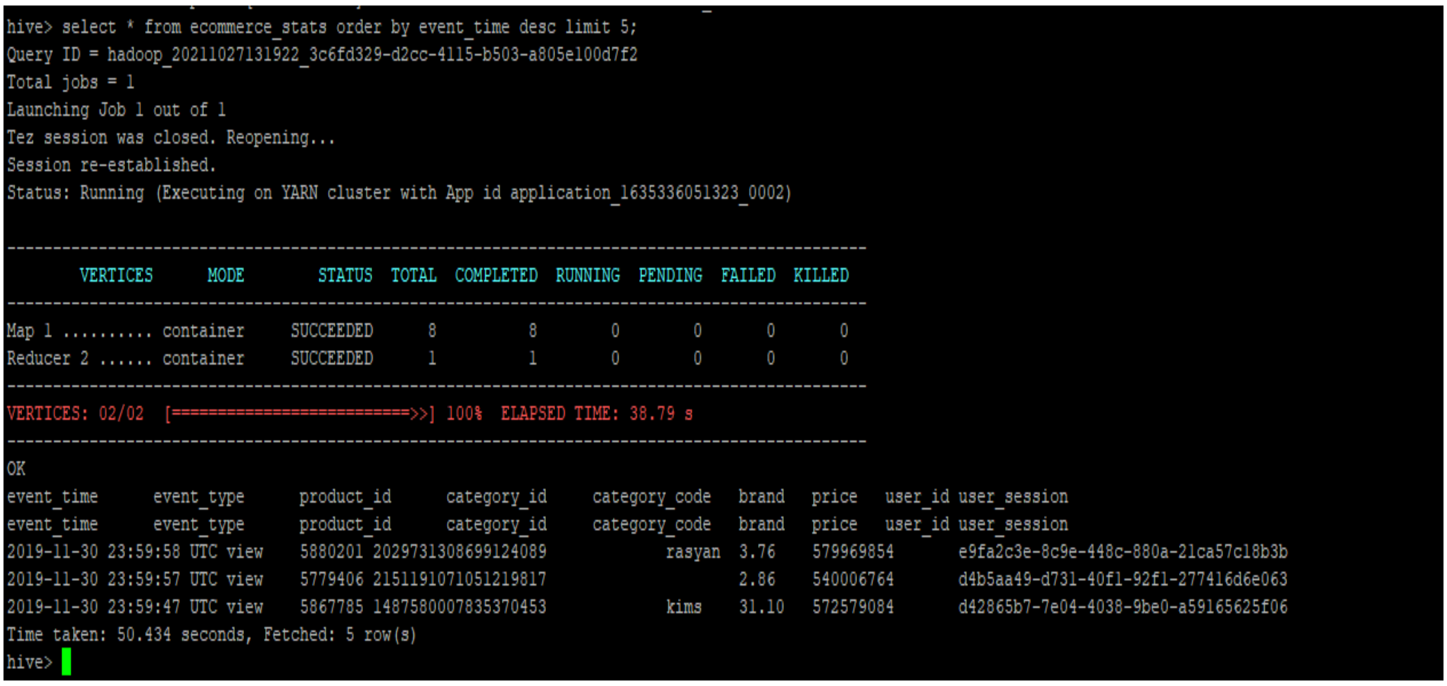
9. Viewing the table records in month – wise manner.

[Oct-2019]



```
hive> select * from ecommerce_stats order by event_time asc limit 5;
Query ID = hadoop_20211027132248_f4208b0f-cf13-44b0-9b17-ec0c89425e67
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1635336051323_0002)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      8          8        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 29.46 s
----------------------------------------------------------------------------------------------
OK
2019-10-01 00:00:00 UTC cart    5773203 1487580005134238553             runail  2.62    463240011       26dd6e6e-4dac-4778-8d2c-92e149dab885
2019-10-01 00:00:03 UTC cart    5773353 1487580005134238553             runail  2.62    463240011       26dd6e6e-4dac-4778-8d2c-92e149dab885
2019-10-01 00:00:07 UTC cart    5881589 2151191071051219817             lovely  13.48   429681830       49e8d843-adf3-428b-a2c3-fe8bc6a307c9
2019-10-01 00:00:07 UTC cart    5723490 1487580005134238553             runail  2.62    463240011       26dd6e6e-4dac-4778-8d2c-92e149dab885
2019-10-01 00:00:15 UTC cart    5881449 1487580013522845895             lovely  0.56    429681830       49e8d843-adf3-428b-a2c3-fe8bc6a307c9
Time taken: 30.242 seconds, Fetched: 5 row(s)
hive>
```

[Nov-2019]

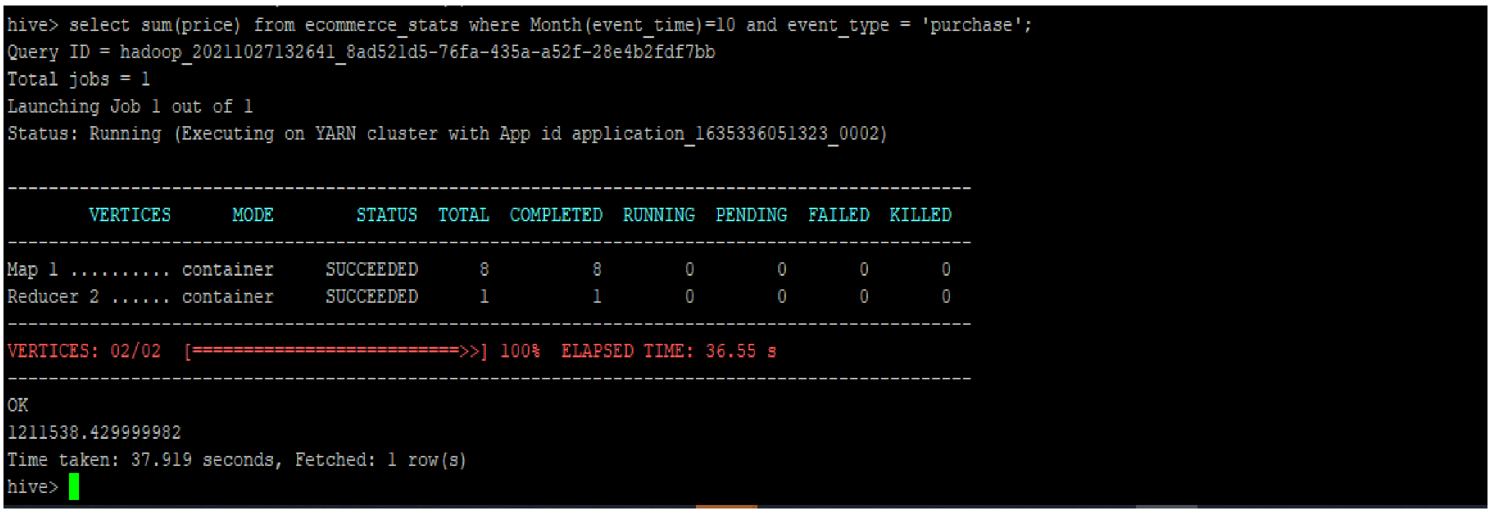


```
hive> select * from ecommerce_stats order by event_time desc limit 5;
Query ID = hadoop_20211027131922_3c6fd329-d2cc-4115-b503-a805e100d7f2
Total jobs = 1
Launching Job 1 out of 1
Tez session was closed. Reopening...
Session re-established.
Status: Running (Executing on YARN cluster with App id application_1635336051323_0002)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      8          8        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 38.79 s
----------------------------------------------------------------------------------------------
OK
event_time      event_type      product_id      category_id     category_code   brand   price   user_id user_session
event_time      event_type      product_id      category_id     category_code   brand   price   user_id user_session
2019-11-30 23:59:58 UTC view    5880201 2029731308699124089             rasyan  3.76    579969854       e9fa2c3e-8c9e-448c-880a-21ca57c18b3b
2019-11-30 23:59:57 UTC view    5779406 2151191071051219817                     2.86    540006764       d4b5aa49-d731-40f1-92f1-277416d6e063
2019-11-30 23:59:47 UTC view    5867785 1487580007835370453             kims    31.10   572579084       d42865b7-7e04-4038-9be0-a59165625f06
Time taken: 50.434 seconds, Fetched: 5 row(s)
hive>
```

Querying and Data Analysis

Q.1> Find the total revenue generated due to purchases made in October.

Ans> SELECT SUM(price) FROM ecommerce_stats WHERE Month(event_time) = 10 AND event_type = 'purchase';



```
hive> select sum(price) from ecommerce_stats where Month(event_time)=10 and event_type = 'purchase';
Query ID = hadoop_20211027132641_8ad521d5-76fa-435a-a52f-28e4b2fdf7bb
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1635336051323_0002)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      8          8        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 36.55 s
----------------------------------------------------------------------------------------------
OK
1211538.429999982
Time taken: 37.919 seconds, Fetched: 1 row(s)
hive>
```

Q.2> Write a query to yield the total sum of purchases per month in a single output.

Ans> SELECT Month(event_time) AS pur_month,

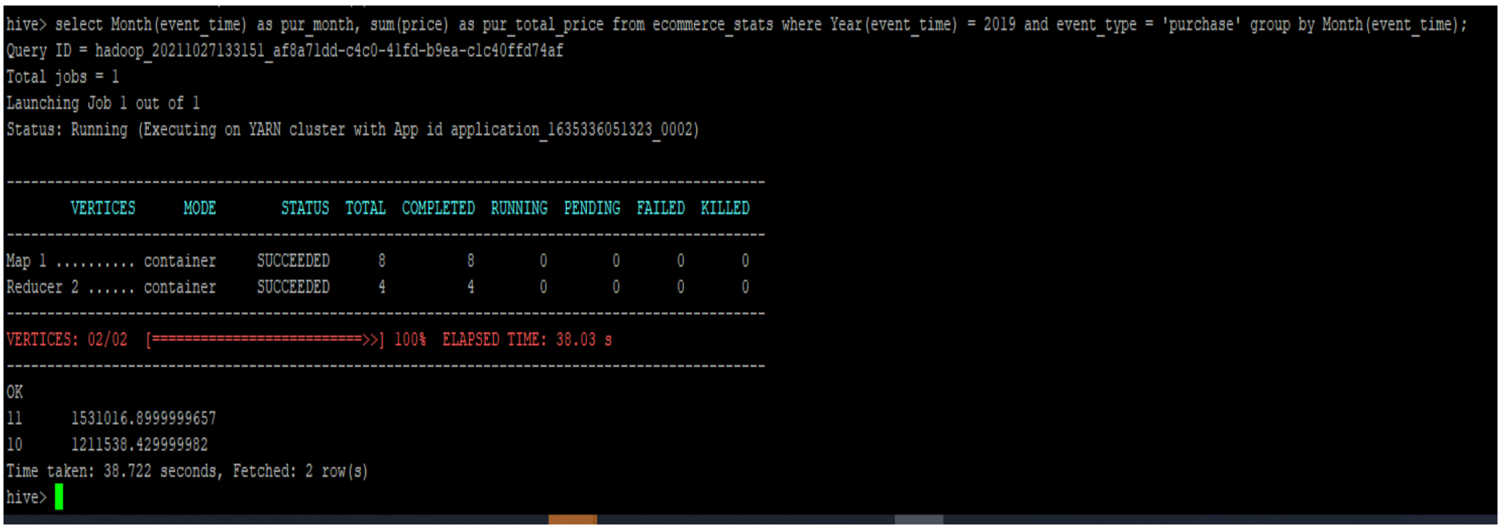
SUM(price) AS pur_total_price

FROM ecommerce_stats

WHERE Year(event_time) = 2019

AND event_type = 'purchase'

GROUP BY Month(event_time);



```
hive> select Month(event_time) as pur_month, sum(price) as pur_total_price from ecommerce_stats where Year(event_time) = 2019 and event_type = 'purchase' group by Month(event_time);
Query ID = hadoop_20211027133151_af8a71dd-c4c0-41fd-b9ea-c1c40ffd74af
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1635336051323_0002)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      8          8        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      4          4        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 38.03 s
----------------------------------------------------------------------------------------------
OK
11      1531016.8999999657
10      1211538.429999982
Time taken: 38.722 seconds, Fetched: 2 row(s)
hive>
```

Q.3> Write a query to find the change in revenue generated due to purchases from October to November.

Ans> SELECT SUM (CASE

WHEN Month(event_time) = 10 THEN price

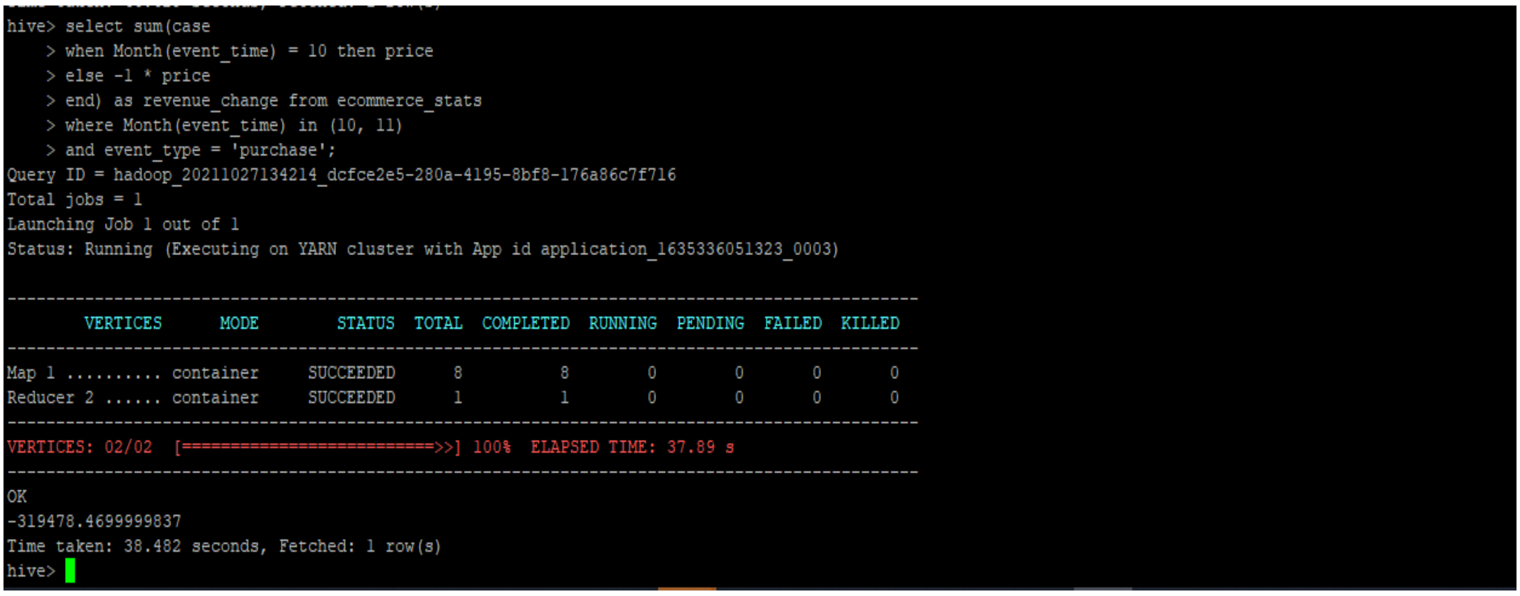
ELSE -1 * price

END) AS revenue_change

FROM ecommerce_stats

WHERE Month(event_time) IN (10, 11)

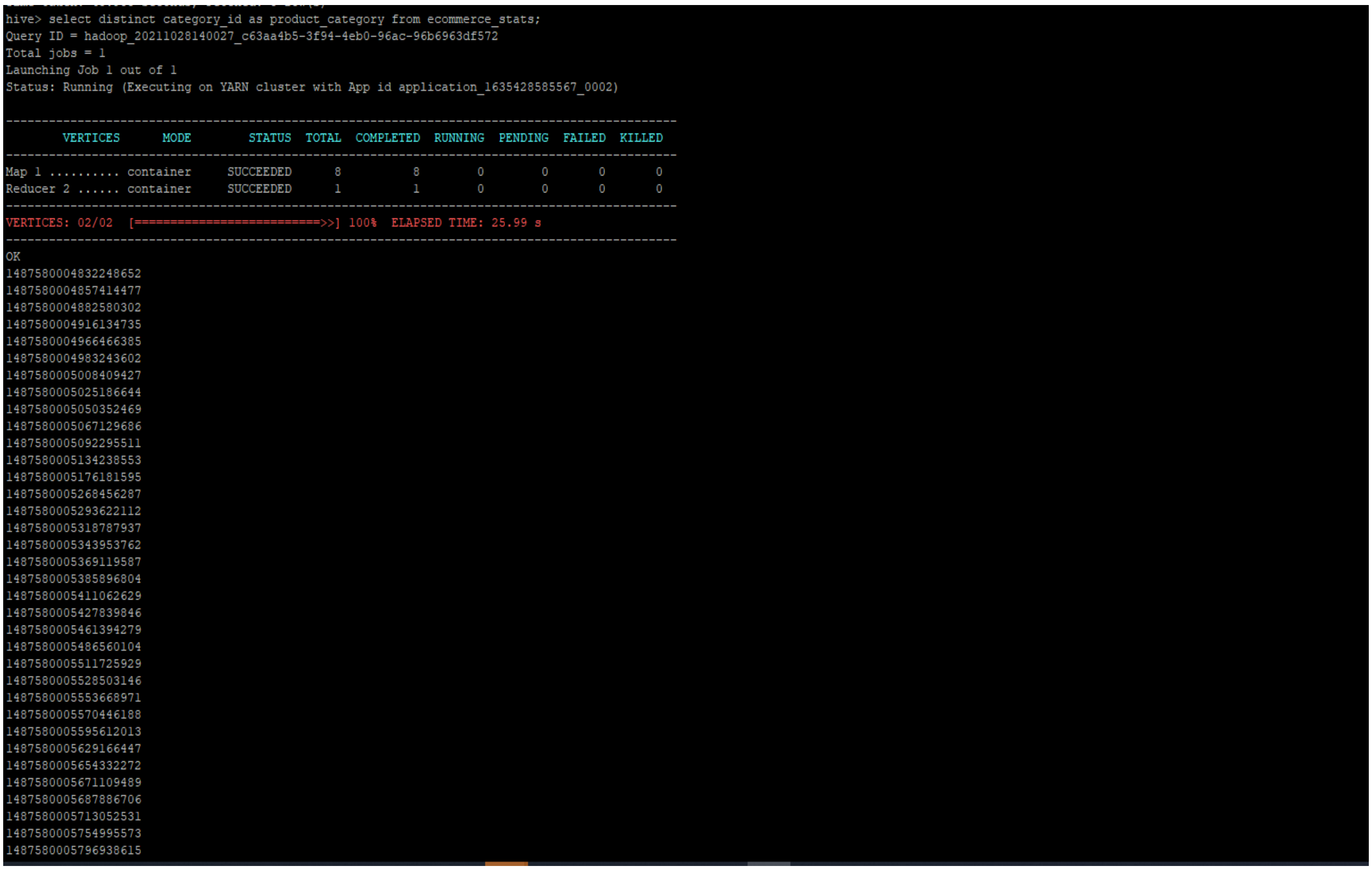
AND event_type = 'purchase';



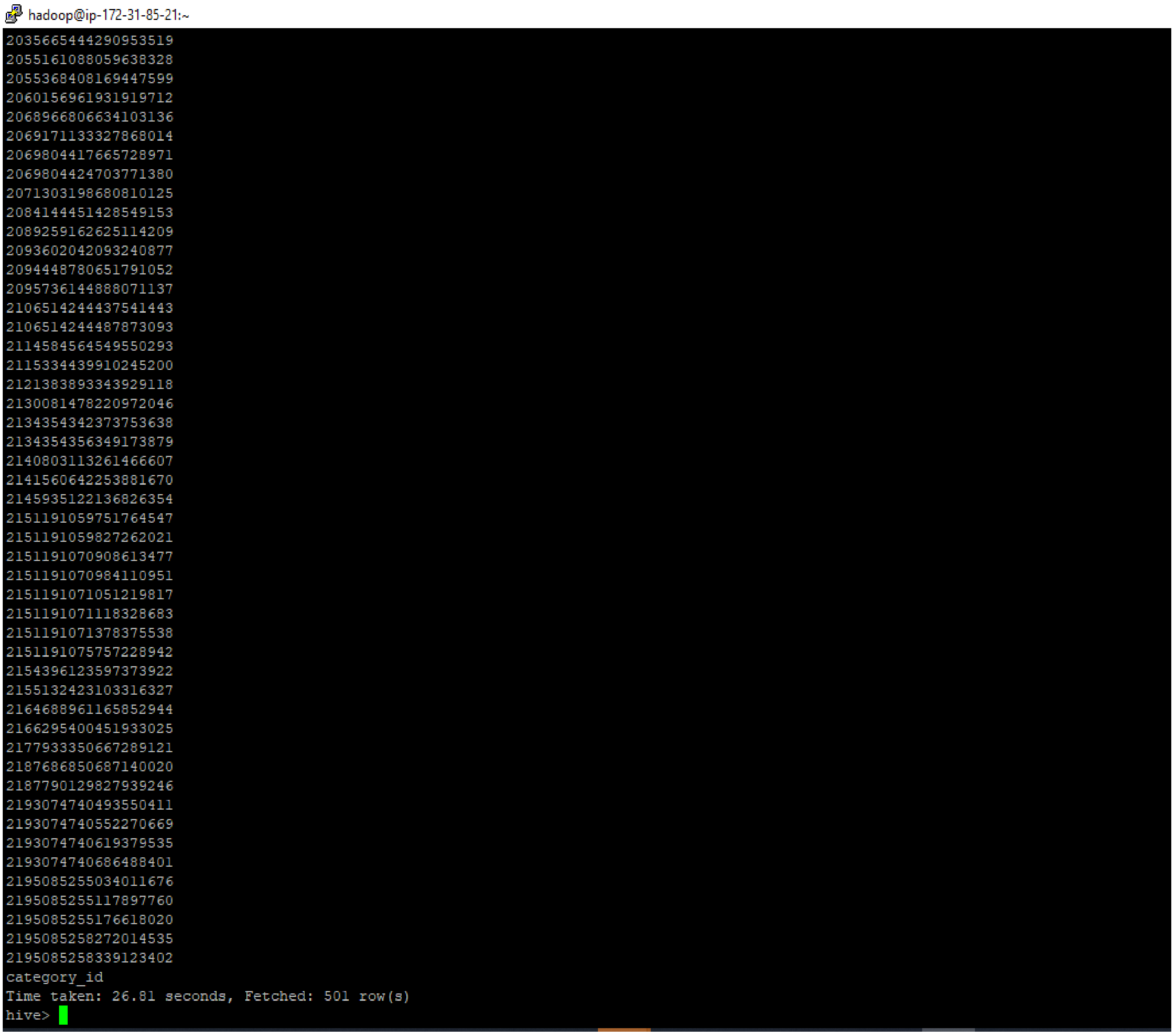
```
hive> select sum(case
    > when Month(event_time) = 10 then price
    > else -1 * price
    > end) as revenue_change from ecommerce_stats
    > where Month(event_time) in (10, 11)
    > and event_type = 'purchase';
Query ID = hadoop_20211027134214_dcfce2e5-280a-4195-8bf8-176a86c7f716
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1635336051323_0003)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      8          8        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 37.89 s
----------------------------------------------------------------------------------------------
OK
-319478.4699999837
Time taken: 38.482 seconds, Fetched: 1 row(s)
hive>
```

Q.4> Find distinct categories of products. Categories with null category code can be ignored.

Ans> SELECT DISTINCT category_id AS product_category FROM ecommerce_stats;



```
hive> select distinct category_id as product_category from ecommerce_stats;
Query ID = hadoop_20211028140027_c63aa4b5-3f94-4eb0-96ac-96b6963df572
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1635428585567_0002)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      8          8        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 25.99 s
----------------------------------------------------------------------------------------------
OK
1487580004832248652
1487580004857414477
1487580004882580302
1487580004916134735
1487580004966466385
1487580004983243602
1487580005008409427
1487580005025186644
1487580005050352469
1487580005067129686
1487580005092295511
1487580005134238553
1487580005176181595
1487580005268456287
1487580005293622112
1487580005318787937
1487580005343953762
1487580005369119587
1487580005385896804
1487580005411062629
1487580005427839846
1487580005461394279
1487580005486560104
1487580005511725929
1487580005528503146
1487580005553668971
1487580005570446188
1487580005595612013
1487580005629166447
1487580005654332272
1487580005671109489
1487580005687886706
1487580005713052531
1487580005754995573
1487580005796938615
```



hadoop@ip-172-31-85-21:~

```
2035665444290953519
2055161088059638328
2055368408169447599
2060156961931919712
2068966806634103136
2069171133327868014
2069804417665728971
2069804424703771380
2071303198680810125
2084144451428549153
2089259162625114209
2093602042093240877
2094448780651791052
2095736144888071137
2106514244437541443
2106514244487873093
2114584564549550293
2115334439910245200
2121383893343929118
2130081478220972046
2134354342373753638
2134354356349173879
2140803113261466607
2141560642253881670
2145935122136826354
2151191059751764547
2151191059827262021
2151191070908613477
2151191070984110951
2151191071051219817
2151191071118328683
2151191071378375538
2151191075757228942
2154396123597373922
2155132423103316327
2164688961165852944
2166295400451933025
2177933350667289121
2187686850687140020
2187790129827939246
2193074740493550411
2193074740552270669
2193074740619379535
2193074740686488401
2195085255034011676
2195085255117897760
2195085255176618020
2195085258272014535
2195085258339123402
category_id
Time taken: 26.81 seconds, Fetched: 501 row(s)
hive>
```

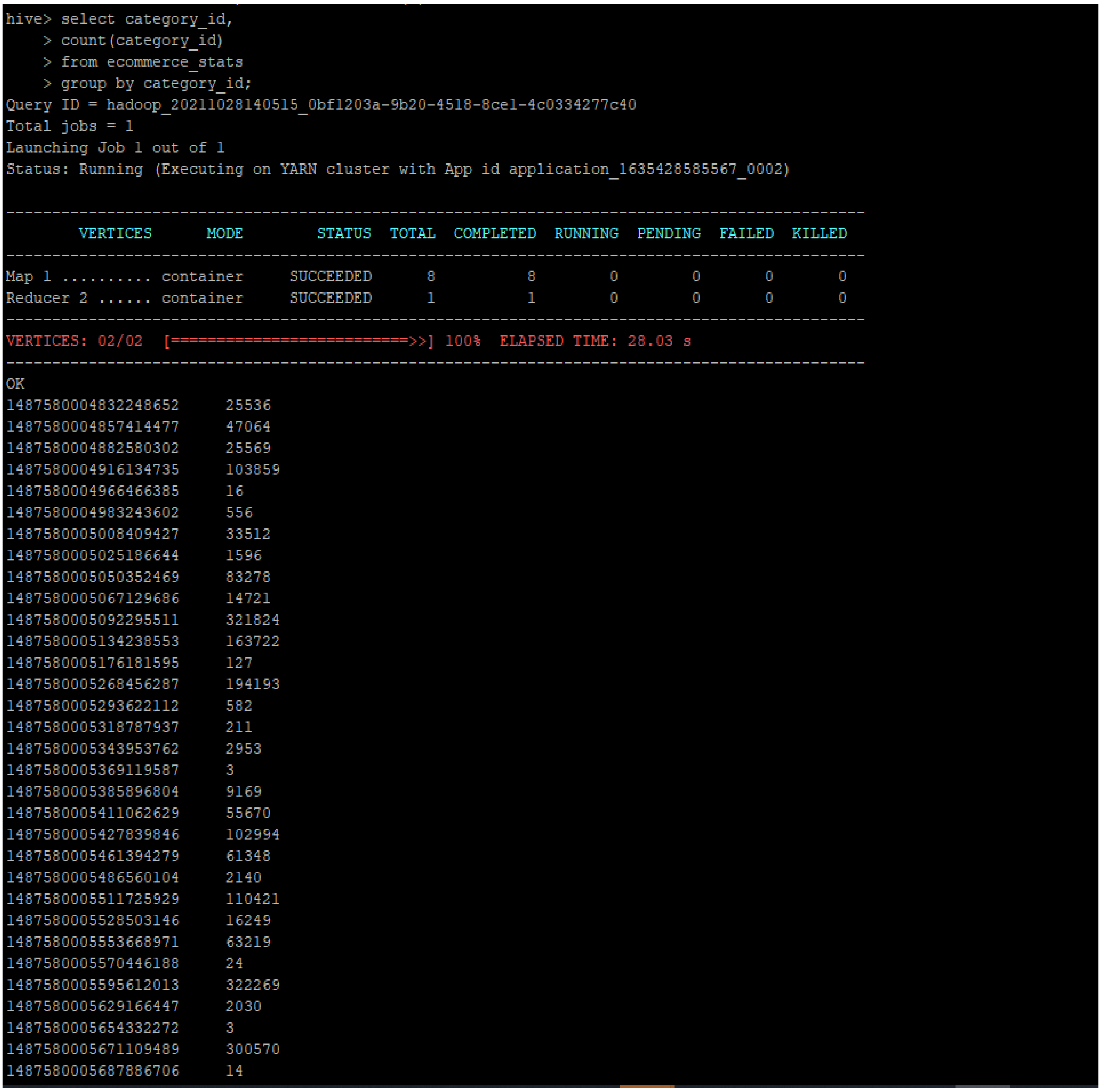
Q.5> Find the total number of products available under each category.

Ans> SELECT category_id,

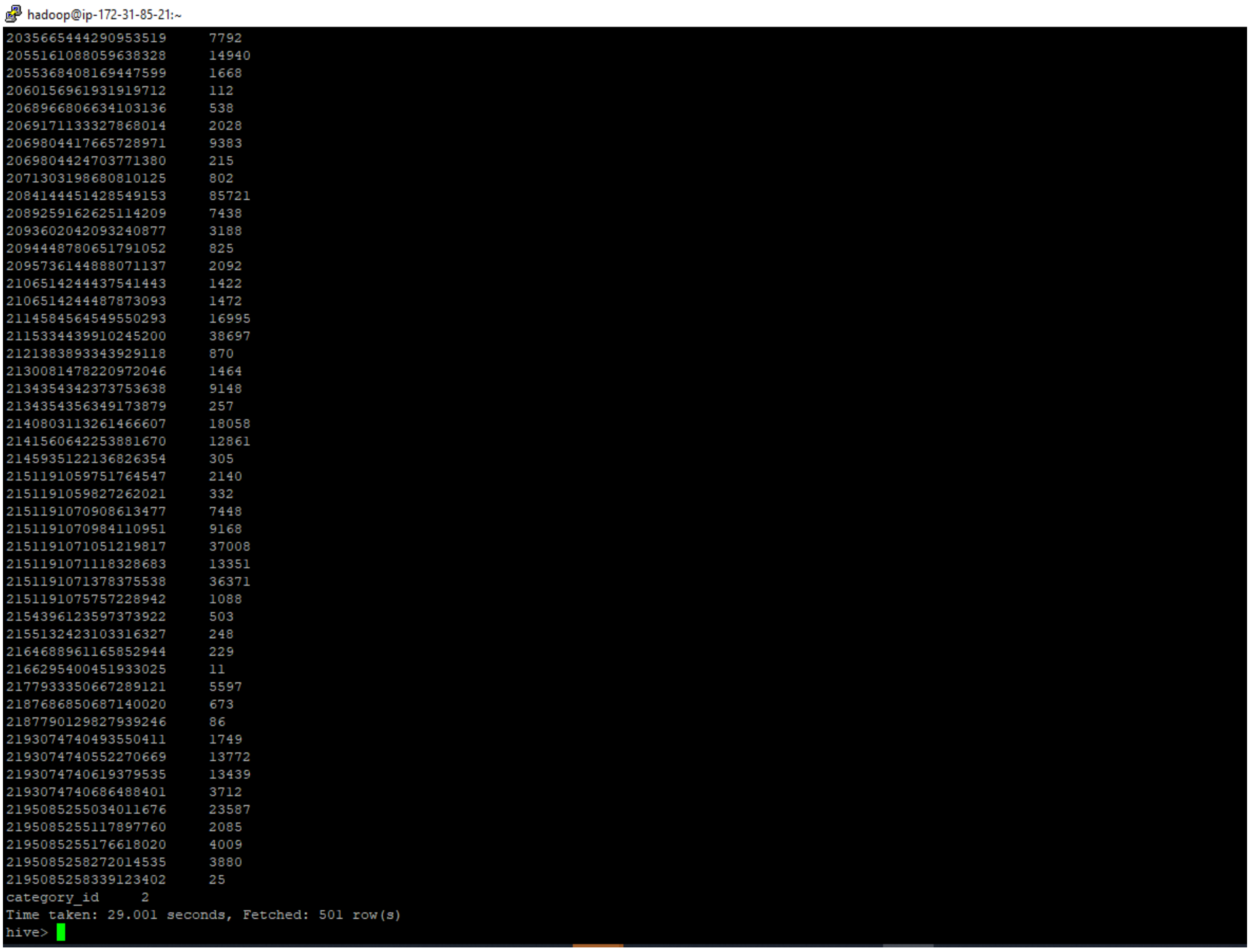
COUNT(category_id)

FROM ecommerce_stats

GROUP BY category_id;



```
hive> select category_id,
    > count(category_id)
    > from ecommerce_stats
    > group by category_id;
Query ID = hadoop_20211028140515_0bf1203a-9b20-4518-8ce1-4c0334277c40
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1635428585567_0002)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      8          8        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 28.03 s
----------------------------------------------------------------------------------------------
OK
1487580004832248652     25536
1487580004857414477     47064
1487580004882580302     25569
1487580004916134735     103859
1487580004966466385     16
1487580004983243602     556
1487580005008409427     33512
1487580005025186644     1596
1487580005050352469     83278
1487580005067129686     14721
1487580005092295511     321824
1487580005134238553     163722
1487580005176181595     127
1487580005268456287     194193
1487580005293622112     582
1487580005318787937     211
1487580005343953762     2953
1487580005369119587     3
1487580005385896804     9169
1487580005411062629     55670
1487580005427839846     102994
1487580005461394279     61348
1487580005486560104     2140
1487580005511725929     110421
1487580005528503146     16249
1487580005553668971     63219
1487580005570446188     24
1487580005595612013     322269
1487580005629166447     2030
1487580005654332272     3
1487580005671109489     300570
1487580005687886706     14
```



hadoop@ip-172-31-85-21:~

```
2035665444290953519     7792
2055161088059638328     14940
2055368408169447599     1668
2060156961931919712     112
2068966806634103136     538
2069171133327868014     2028
2069804417665728971     9383
2069804424703771380     215
2071303198680810125     802
2084144451428549153     85721
2089259162625114209     7438
2093602042093240877     3188
2094448780651791052     825
2095736144888071137     2092
2106514244437541443     1422
2106514244487873093     1472
2114584564549550293     16995
2115334439910245200     38697
2121383893343929118     870
2130081478220972046     1464
2134354342373753638     9148
2134354356349173879     257
2140803113261466607     18058
2141560642253881670     12861
2145935122136826354     305
2151191059751764547     2140
2151191059827262021     332
2151191070908613477     7448
2151191070984110951     9168
2151191071051219817     37008
2151191071118328683     13351
2151191071378375538     36371
2151191075757228942     1088
2154396123597373922     503
2155132423103316327     248
2164688961165852944     229
2166295400451933025     11
2177933350667289121     5597
2187686850687140020     673
2187790129827939246     86
2193074740493550411     1749
2193074740552270669     13772
2193074740619379535     13439
2193074740686488401     3712
2195085255034011676     23587
2195085255117897760     2085
2195085255176618020     4009
2195085258272014535     3880
2195085258339123402     25
category_id     2
Time taken: 29.001 seconds, Fetched: 501 row(s)
hive> 
```

Q.6> Which brand had the maximum sales in October and November combined?

Ans> SELECT brand,

SUM (price) AS brand_sales

FROM ecommerce_stats

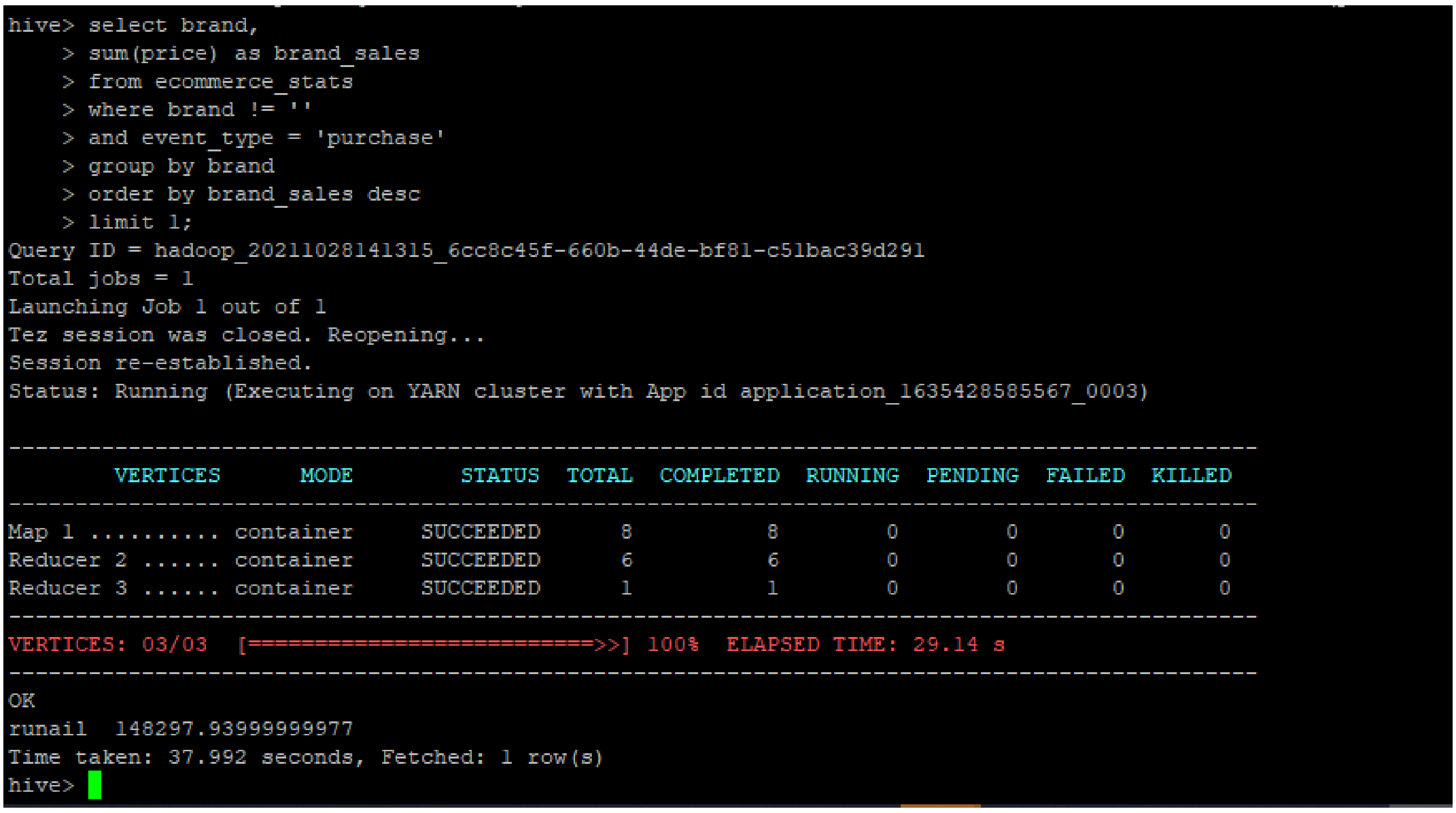
WHERE brand != ''

AND event_type = 'purchase'

GROUP BY brand

ORDER BY brand_sales DESC

LIMIT 1;



```
hive> select brand,
    > sum(price) as brand_sales
    > from ecommerce_stats
    > where brand != ''
    > and event_type = 'purchase'
    > group by brand
    > order by brand_sales desc
    > limit 1;
Query ID = hadoop_20211028141315_6cc8c45f-660b-44de-bf81-c51bac39d291
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_1635428585567_0003)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      8          8        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      6          6        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 29.14 s
----------------------------------------------------------------------------------------------
OK
runail  148297.93999999977
Time taken: 37.992 seconds, Fetched: 1 row(s)
hive>
```

Q.7> Which brands increased their sales from October to November?

Ans> SELECT Oct.Brand FROM

(SELECT brand, SUM(price) AS brand_sales FROM ecommerce_stats

WHERE brand != '' AND Month(event_time) = 10 AND event_type = 'purchase' GROUP BY brand) AS Oct

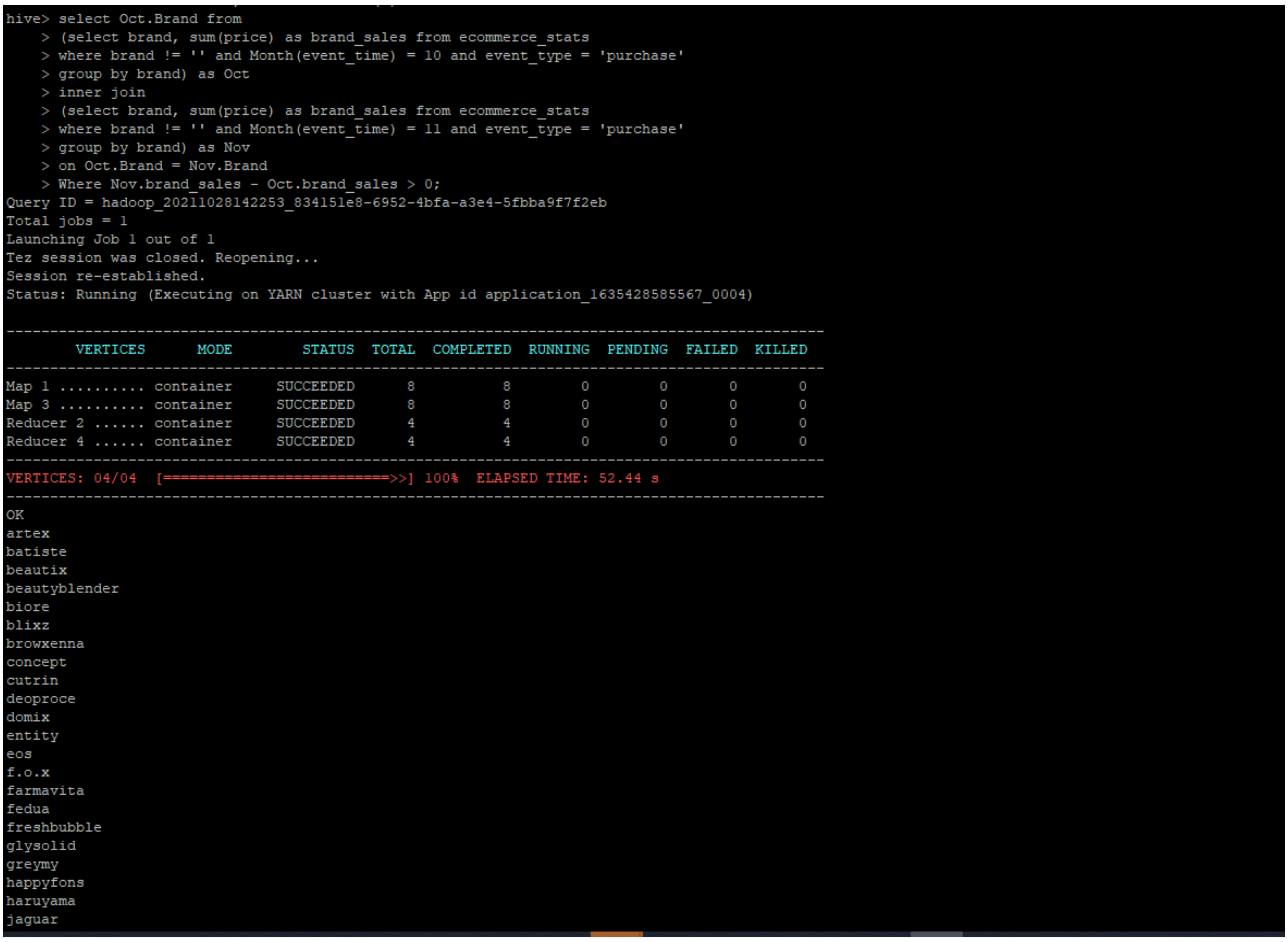
INNER JOIN

(SELECT brand, SUM(price) AS brand_Sales FROM ecommerce_stats

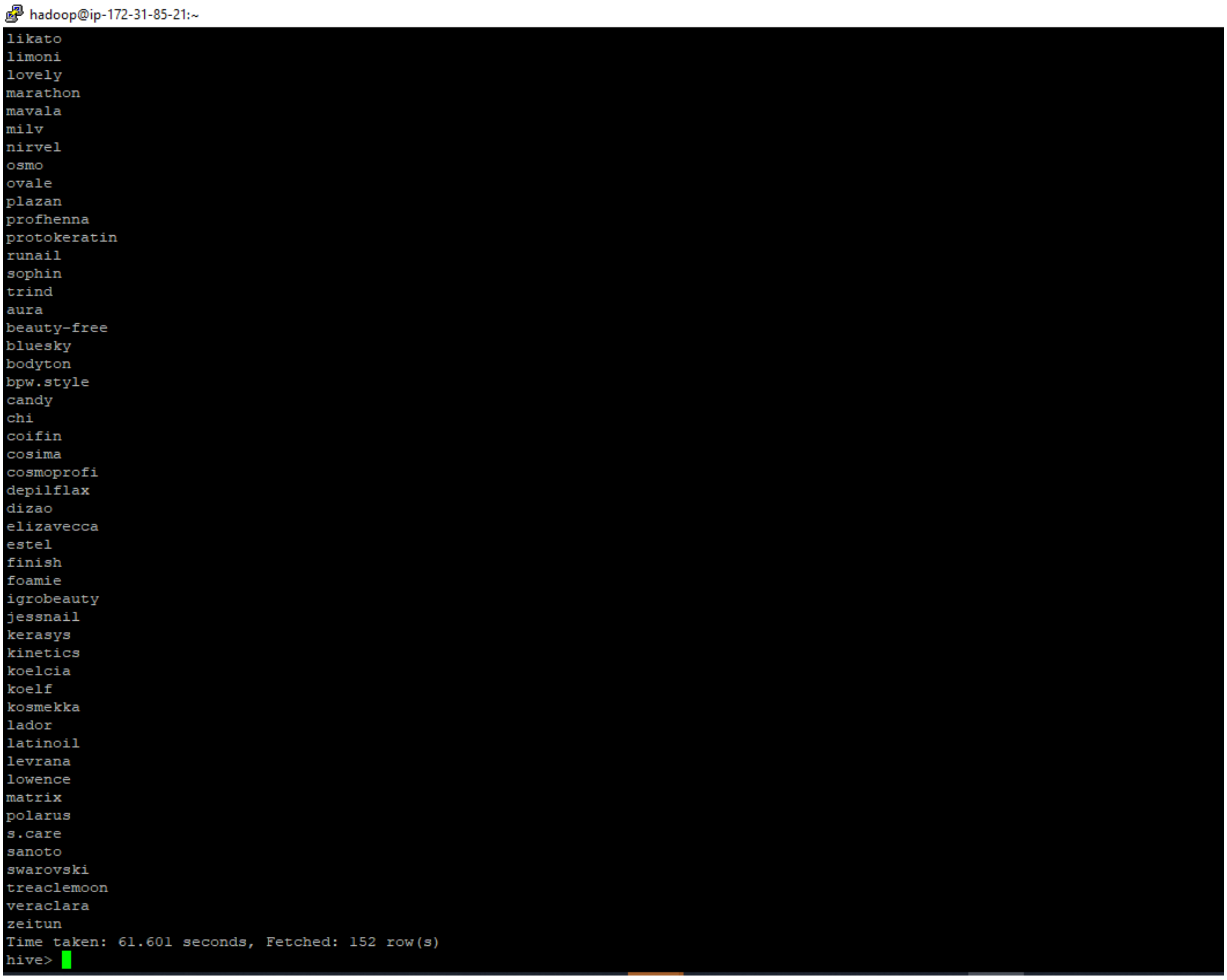
WHERE brand != '' AND Month(event_time) = 11 AND event_type = 'purchase' GROUP BY brand) AS Nov

ON Oct.Brand = Nov.Brand

WHERE Nov.brand_sales - Oct.brand_sales > 0;



```
hive> select Oct.Brand from
    > (select brand, sum(price) as brand_sales from ecommerce_stats
    > where brand != '' and Month(event_time) = 10 and event_type = 'purchase'
    > group by brand) as Oct
    > inner join
    > (select brand, sum(price) as brand_sales from ecommerce_stats
    > where brand != '' and Month(event_time) = 11 and event_type = 'purchase'
    > group by brand) as Nov
    > on Oct.Brand = Nov.Brand
    > Where Nov.brand_sales - Oct.brand_sales > 0;
Query ID = hadoop_20211028142253_834151e8-6952-4bfa-a3e4-5fbba9f7f2eb
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_1635428585567_0004)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      8          8        0        0       0       0
Map 3 .......... container     SUCCEEDED      8          8        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      4          4        0        0       0       0
Reducer 4 ...... container     SUCCEEDED      4          4        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 04/04  [==========================>>] 100%  ELAPSED TIME: 52.44 s
----------------------------------------------------------------------------------------------
OK
artex
batiste
beautix
beautyblender
biore
blixz
browxenna
concept
cutrin
deoproce
domix
entity
eos
f.o.x
farmavita
fedua
freshbubble
glysolid
greymy
happyfons
haruyama
jaguar
```



hadoop@ip-172-31-85-21:~

```
likato
limoni
lovely
marathon
mavala
milv
nirvel
osmo
ovale
plazan
profhenna
protokeratin
runail
sophin
trind
aura
beauty-free
bluesky
bodyton
bpw.style
candy
chi
coifin
cosima
cosmoprofi
depilflax
dizao
elizavecca
estel
finish
foamie
igrobeauty
jessnail
kerasys
kinetics
koelcia
koelf
kosmekka
lador
latinoil
levrana
lowence
matrix
polarus
s.care
sanoto
swarovski
treaclemoon
veraclara
zeitun
Time taken: 61.601 seconds, Fetched: 152 row(s)
hive>
```

Q.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.

Ans> SELECT user_id,

SUM(price) AS User_expense

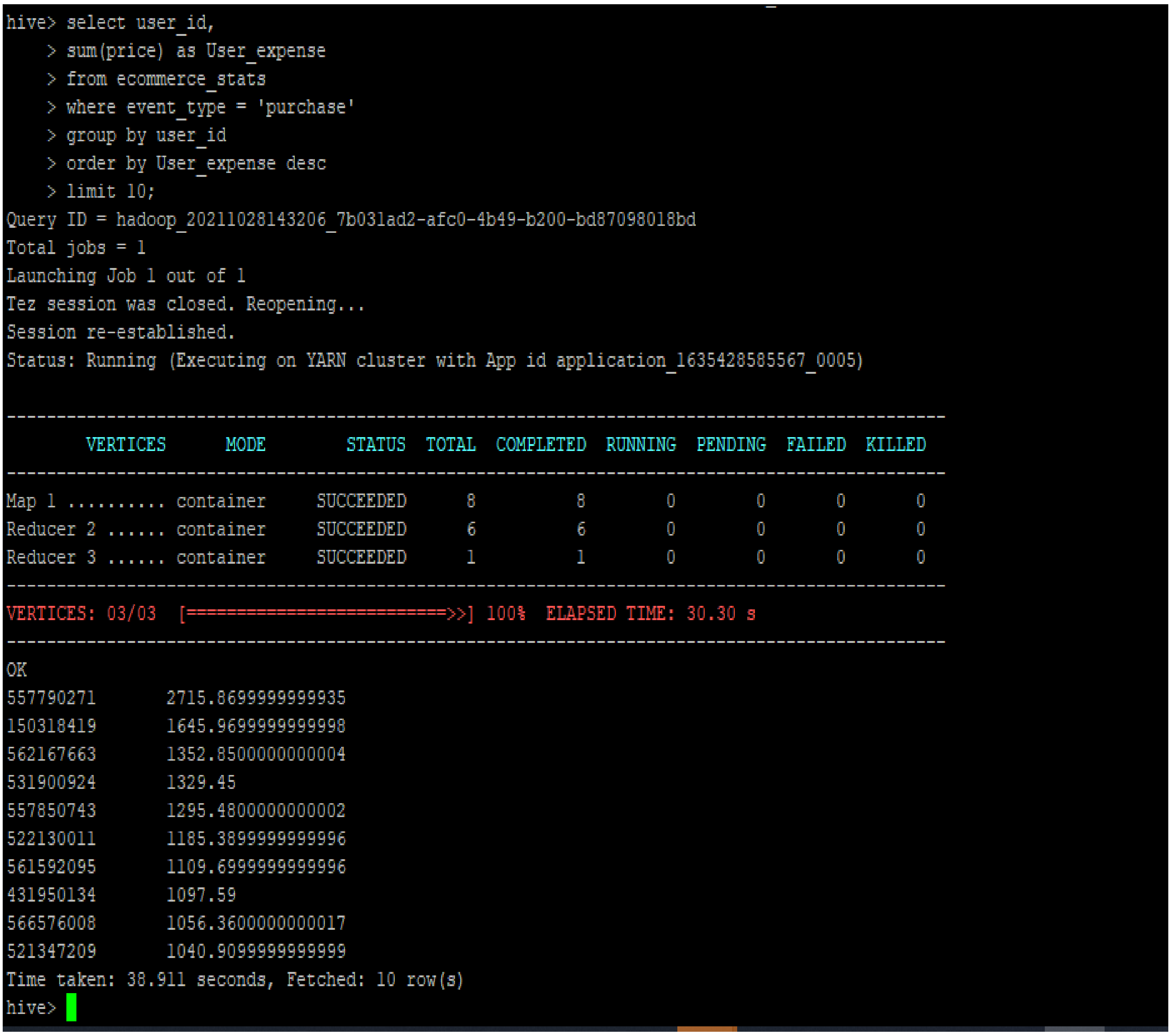
FROM ecommerce_stats

WHERE event_type = 'purchase'

GROUP BY user_id

ORDER BY User_expense DESC

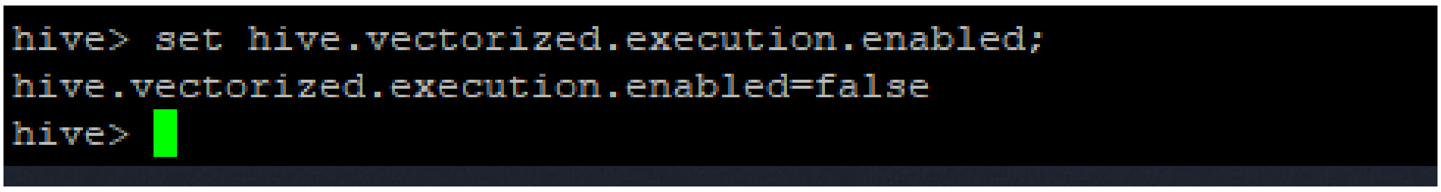
LIMIT 10;



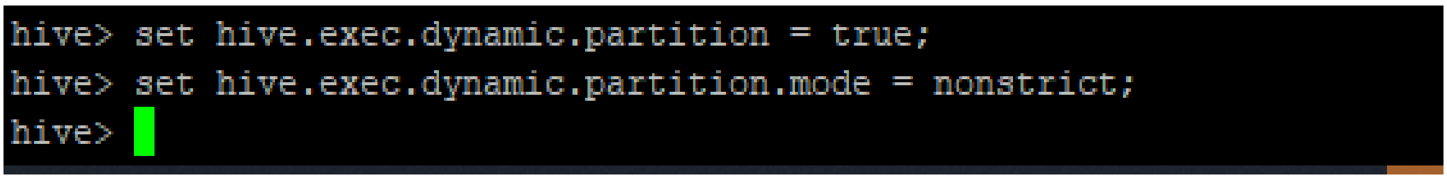
```
hive> select user_id,
    > sum(price) as User_expense
    > from ecommerce_stats
    > where event_type = 'purchase'
    > group by user_id
    > order by User_expense desc
    > limit 10;
Query ID = hadoop_20211028143206_7b031ad2-afc0-4b49-b200-bd87098018bd
Total jobs = 1
Launching Job 1 out of 1
Tez session was closed. Reopening...
Session re-established.
Status: Running (Executing on YARN cluster with App id application_1635428585567_0005)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      8          8        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      6          6        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 30.30 s
----------------------------------------------------------------------------------------------
OK
557790271       2715.8699999999935
150318419       1645.9699999999998
562167663       1352.8500000000004
531900924       1329.45
557850743       1295.4800000000002
522130011       1185.3899999999996
561592095       1109.6999999999996
431950134       1097.59
566576008       1056.3600000000017
521347209       1040.9099999999999
Time taken: 38.911 seconds, Fetched: 10 row(s)
hive>
```

Query Optimization and its Efficiency

1. SET hive.vectorised.execution.enabled;
   SET hive.exec.dynamic.partition = true;
   SET hive.exec.dynamic.partition.mode=nonstrict;

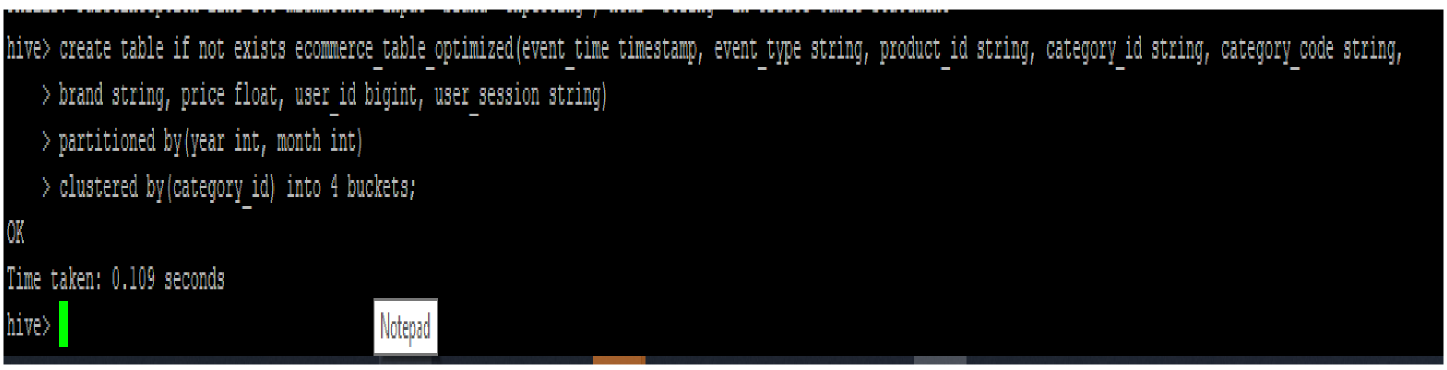


```
hive> set hive.vectorized.execution.enabled;
hive.vectorized.execution.enabled=false
hive>
```



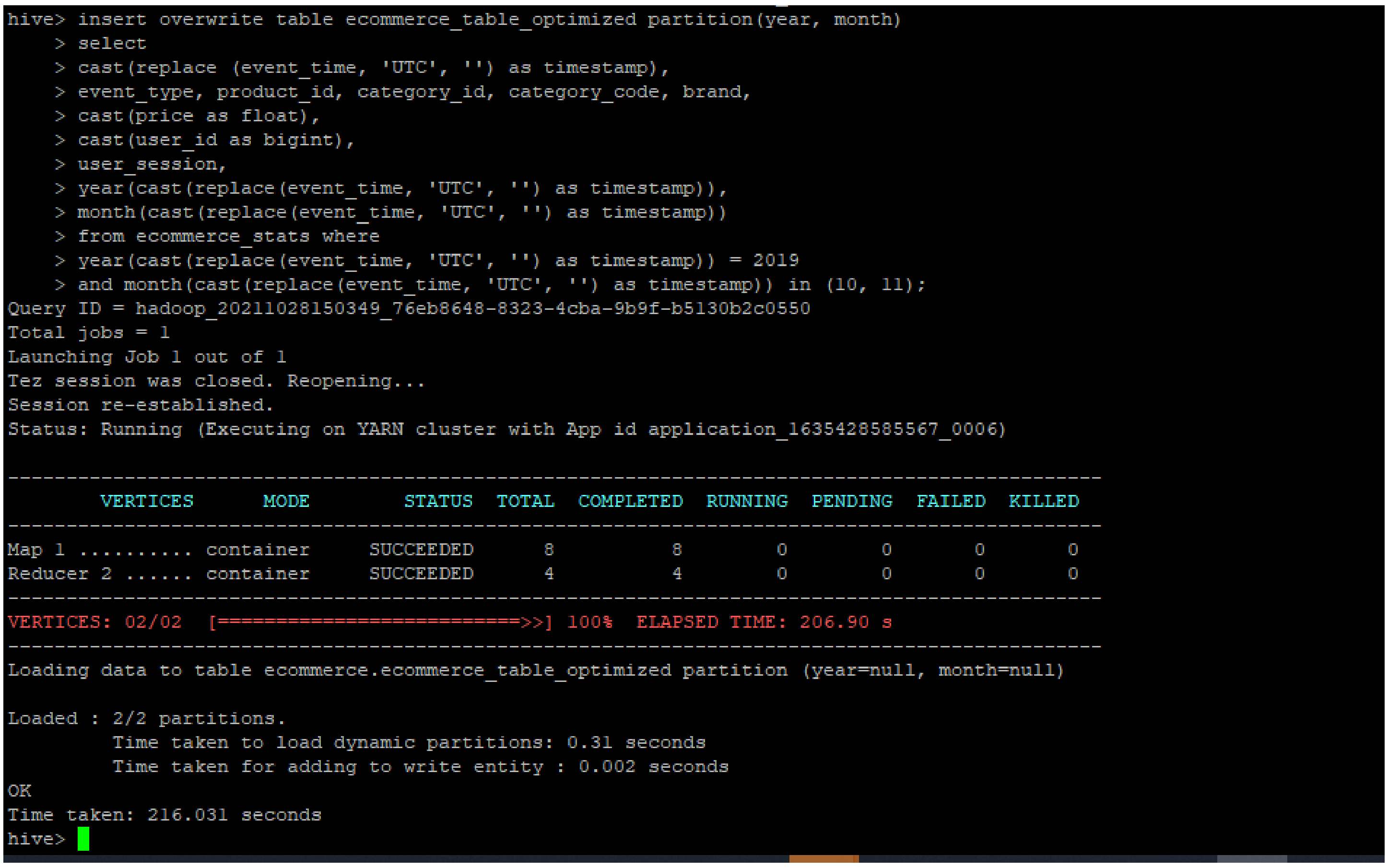
```
hive> set hive.exec.dynamic.partition = true;
hive> set hive.exec.dynamic.partition.mode = nonstrict;
hive>
```

2. Creating an optimized table ‘ecommerce_table_optimized’ with partitioning and dividing it into 4 buckets.



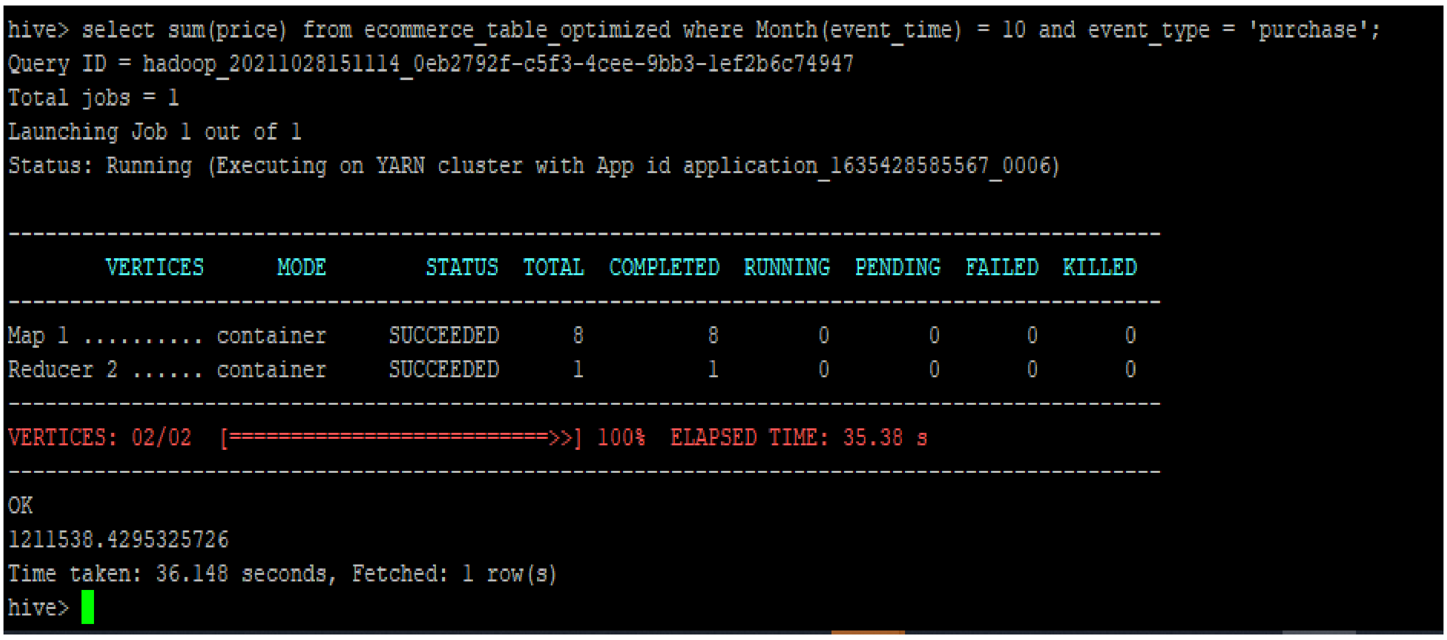
```
hive> create table if not exists ecommerce_table_optimized(event_time timestamp, event_type string, product_id string, category_id string, category_code string,
    > brand string, price float, user_id bigint, user_session string)
    > partitioned by(year int, month int)
    > clustered by(category_id) into 4 buckets;
OK
Time taken: 0.109 seconds
hive>
```

3. Loading and inserting data into optimized table ‘ecommerce_table_optimized’



```
hive> insert overwrite table ecommerce_table_optimized partition(year, month)
    > select
    > cast(replace (event_time, 'UTC', '') as timestamp),
    > event_type, product_id, category_id, category_code, brand,
    > cast(price as float),
    > cast(user_id as bigint),
    > user_session,
    > year(cast(replace(event_time, 'UTC', '') as timestamp)),
    > month(cast(replace(event_time, 'UTC', '') as timestamp))
    > from ecommerce_stats where
    > year(cast(replace(event_time, 'UTC', '') as timestamp)) = 2019
    > and month(cast(replace(event_time, 'UTC', '') as timestamp)) in (10, 11);
Query ID = hadoop_20211028150349_76eb8648-8323-4cba-9b9f-b5130b2c0550
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_1635428585567_0006)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      8          8        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      4          4        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 206.90 s
----------------------------------------------------------------------------------------------
Loading data to table ecommerce.ecommerce_table_optimized partition (year=null, month=null)

Loaded : 2/2 partitions.
         Time taken to load dynamic partitions: 0.31 seconds
         Time taken for adding to write entity : 0.002 seconds
OK
Time taken: 216.031 seconds
hive>
```

4. After optimizing the table running query from Q.1
   Before Optimization – Time taken 37.919 seconds
   After Optimization – Time taken 36.148 seconds

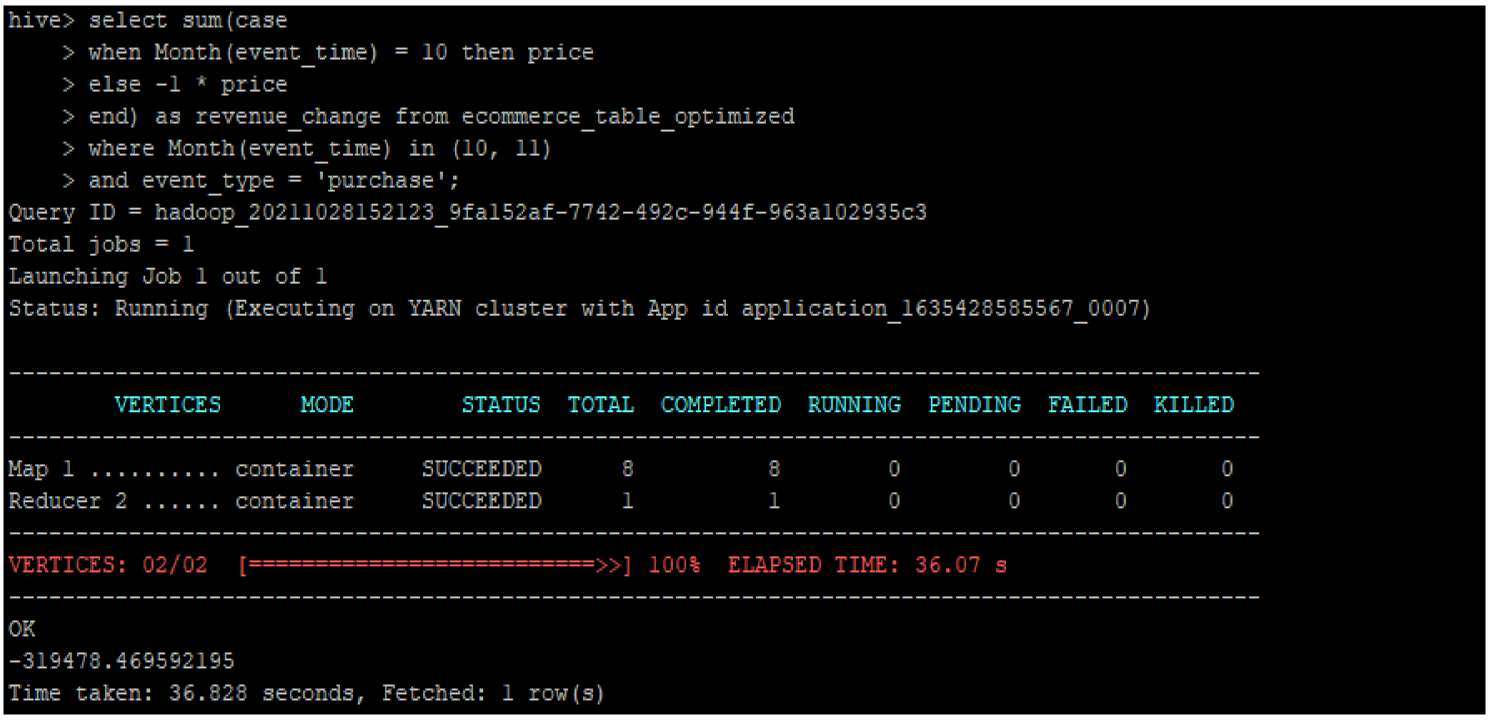


```
hive> select sum(price) from ecommerce_table_optimized where Month(event_time) = 10 and event_type = 'purchase';
Query ID = hadoop_20211028151114_0eb2792f-c5f3-4cee-9bb3-1ef2b6c74947
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1635428585567_0006)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      8          8        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 35.38 s
----------------------------------------------------------------------------------------------
OK
1211538.4295325726
Time taken: 36.148 seconds, Fetched: 1 row(s)
hive>
```

5. After optimizing the table running query from Q.3

   Before Optimization – Time taken 38.482 seconds

   After Optimization – Time taken 36.828 seconds

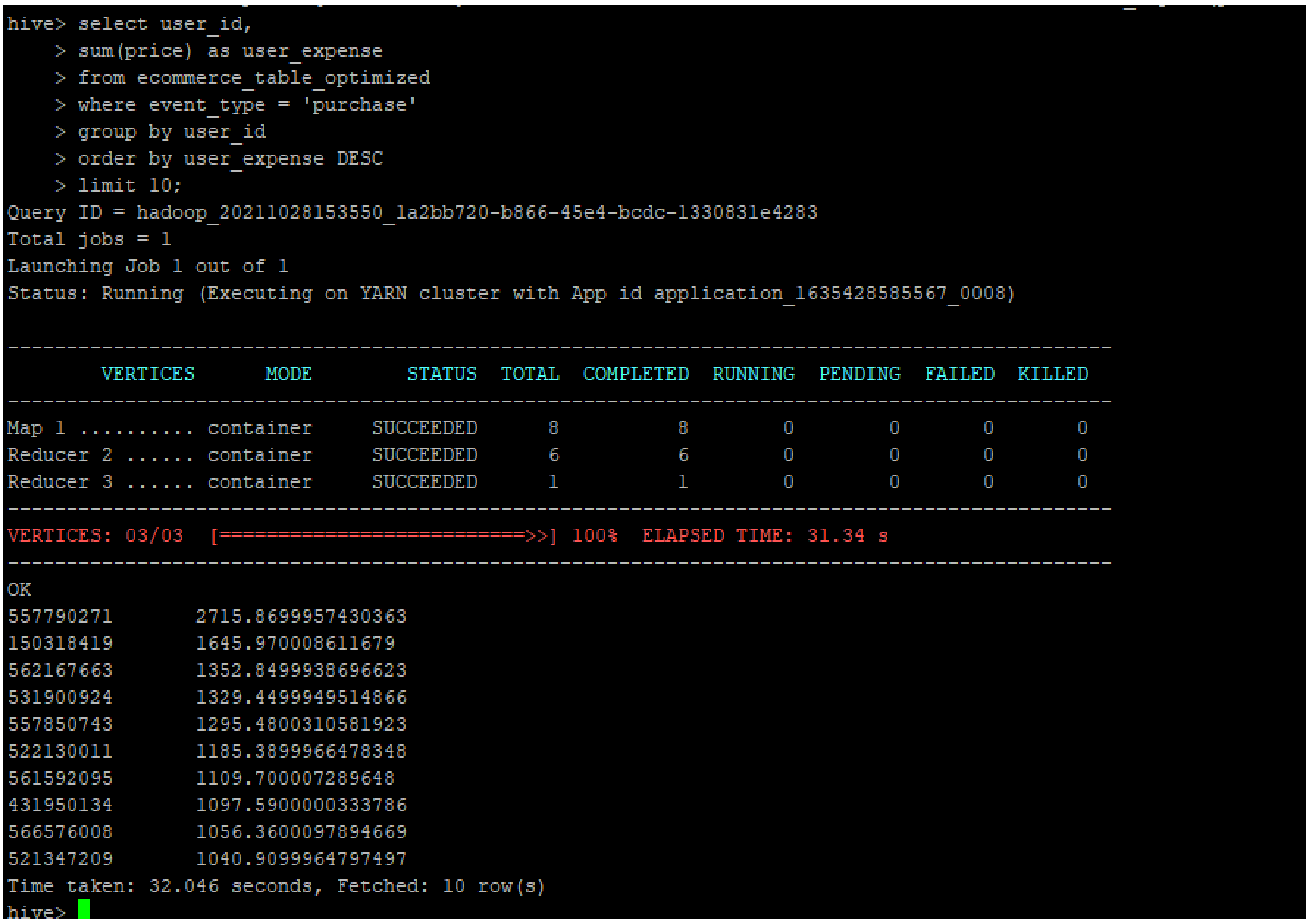


```
hive> select sum(case
    > when Month(event_time) = 10 then price
    > else -1 * price
    > end) as revenue_change from ecommerce_table_optimized
    > where Month(event_time) in (10, 11)
    > and event_type = 'purchase';
Query ID = hadoop_20211028152123_9fa152af-7742-492c-944f-963a102935c3
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1635428585567_0007)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      8          8        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 36.07 s
----------------------------------------------------------------------------------------------
OK
-319478.469592195
Time taken: 36.828 seconds, Fetched: 1 row(s)
```

6. After optimizing the table running query from Q.8

   Before Optimization – Time taken 38.911 seconds

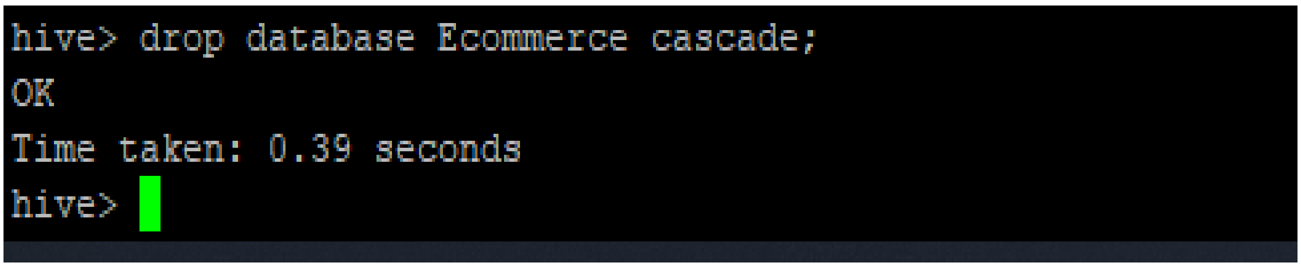
   After Optimization – Time taken 32.046 seconds



```
hive> select user_id,
    > sum(price) as user_expense
    > from ecommerce_table_optimized
    > where event_type = 'purchase'
    > group by user_id
    > order by user_expense DESC
    > limit 10;
Query ID = hadoop_20211028153550_1a2bb720-b866-45e4-bcdc-1330831e4283
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1635428585567_0008)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      8          8        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      6          6        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 31.34 s
----------------------------------------------------------------------------------------------
OK
557790271       2715.8699957430363
150318419       1645.970008611679
562167663       1352.8499938696623
531900924       1329.4499949514866
557850743       1295.4800310581923
522130011       1185.3899966478348
561592095       1109.700007289648
431950134       1097.5900000333786
566576008       1056.3600097894669
521347209       1040.9099964797497
Time taken: 32.046 seconds, Fetched: 10 row(s)
hive>
```

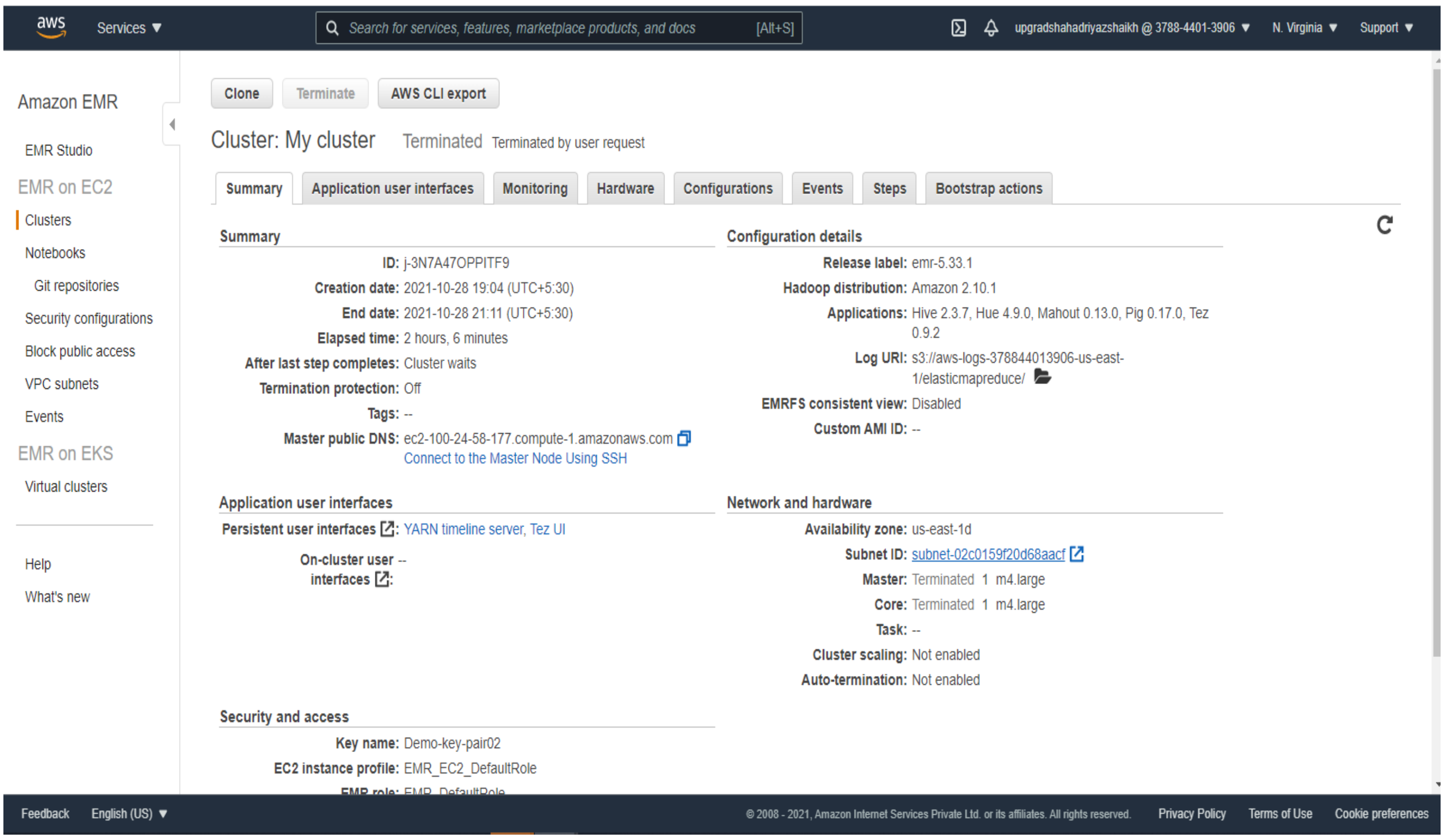
Clean – Up Process

1. Dropping the previously created database ‘Ecommerce’.



```
hive> drop database Ecommerce cascade;
OK
Time taken: 0.39 seconds
hive>
```

2. Terminating the AWS EMR cluster.



Cluster: My cluster Terminated Terminated by user request

Summary

ID: j-3N7A47OPPITF9
Creation date: 2021-10-28 19:04 (UTC+5:30)
End date: 2021-10-28 21:11 (UTC+5:30)
Elapsed time: 2 hours, 6 minutes
After last step completes: Cluster waits
Termination protection: Off
Tags: --
Master public DNS: ec2-100-24-58-177.compute-1.amazonaws.com
Connect to the Master Node Using SSH

Configuration details

Release label: emr-5.33.1
Hadoop distribution: Amazon 2.10.1
Applications: Hive 2.3.7, Hue 4.9.0, Mahout 0.13.0, Pig 0.17.0, Tez 0.9.2
Log URI: s3://aws-logs-378844013906-us-east-1/elasticmapreduce/
EMRFS consistent view: Disabled
Custom AMI ID: --

Application user interfaces

Persistent user interfaces: YARN timeline server, Tez UI
On-cluster user interfaces: --

Network and hardware

Availability zone: us-east-1d
Subnet ID: subnet-02c0159f20d68aacf
Master: Terminated 1 m4.large
Core: Terminated 1 m4.large
Task: --
Cluster scaling: Not enabled
Auto-termination: Not enabled

Security and access

Key name: Demo-key-pair02
EC2 instance profile: EMR_EC2_DefaultRole
EMR role: EMR_DefaultRole