HIVE CASE STUDY ASSIGNMENT

Ecommerce Sales Data Analysis

Submitted by:

-Onkar Suryawanshi

-Manish Mishra

Problem Statement:

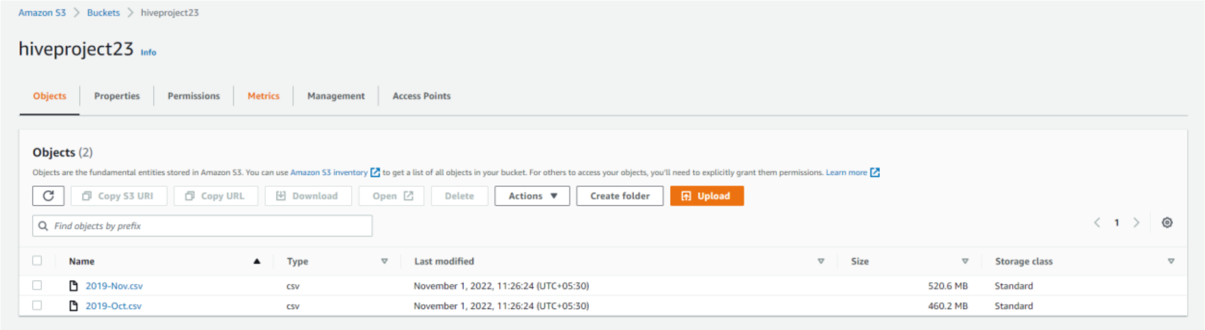
With online sales gaining popularity, tech companies are exploring ways to improve their sales by analysing customer behaviour and gaining insights about product trends. Furthermore, the websites make it easier for customers to find the products they require without much scavenging. Needless to say, 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.**



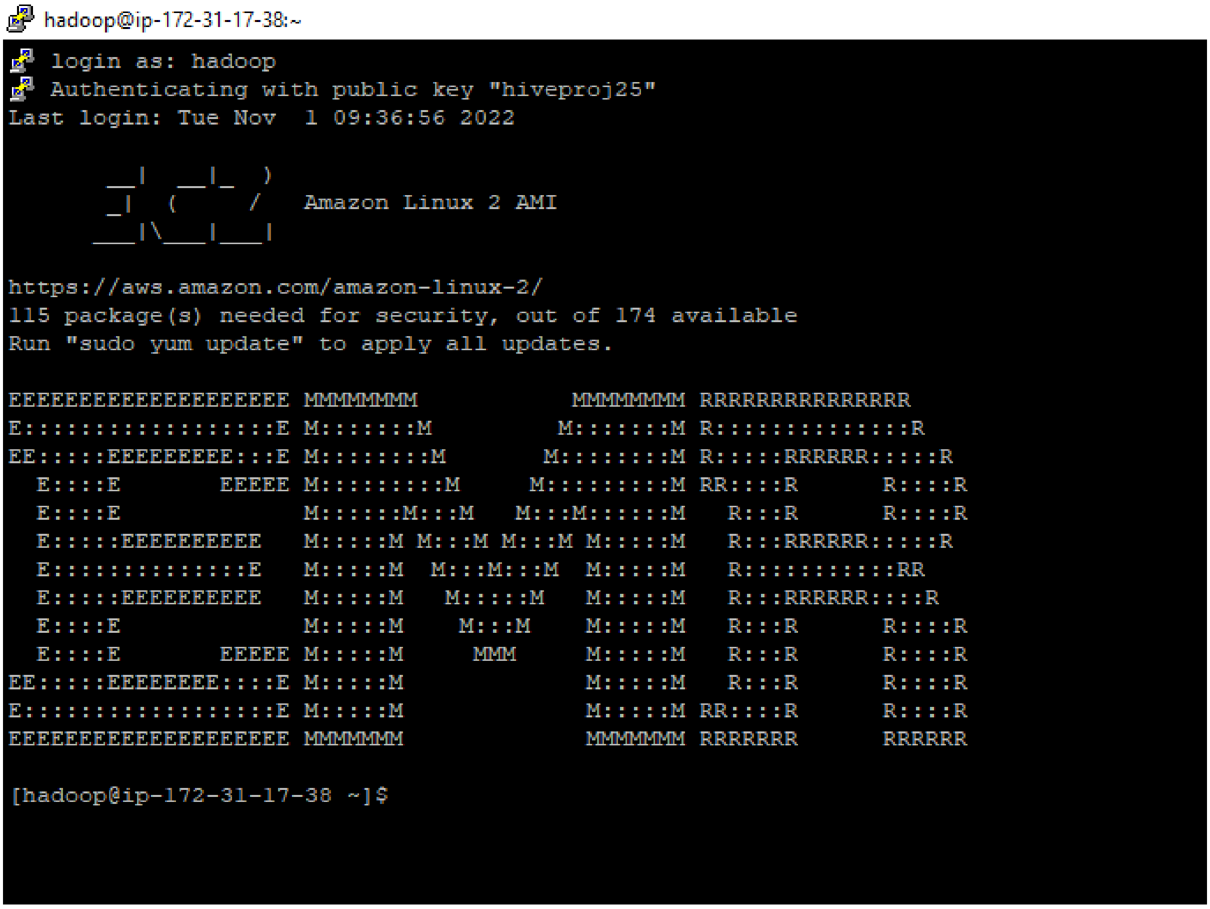
Amazon S3 > Buckets > hiveproject23
hiveproject23 Info
Objects | Properties | Permissions | Metrics | Management | Access Points

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

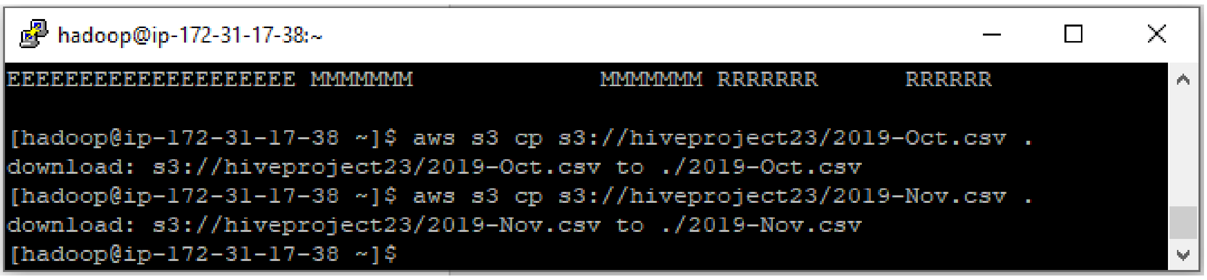
□	Name ▲	Type	Last modified	Size	Storage class
□	2019-Nov.csv	csv	November 1, 2022, 11:26:24 (UTC+05:30)	520.6 MB	Standard
□	2019-Oct.csv	csv	November 1, 2022, 11:26:24 (UTC+05:30)	460.2 MB	Standard

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



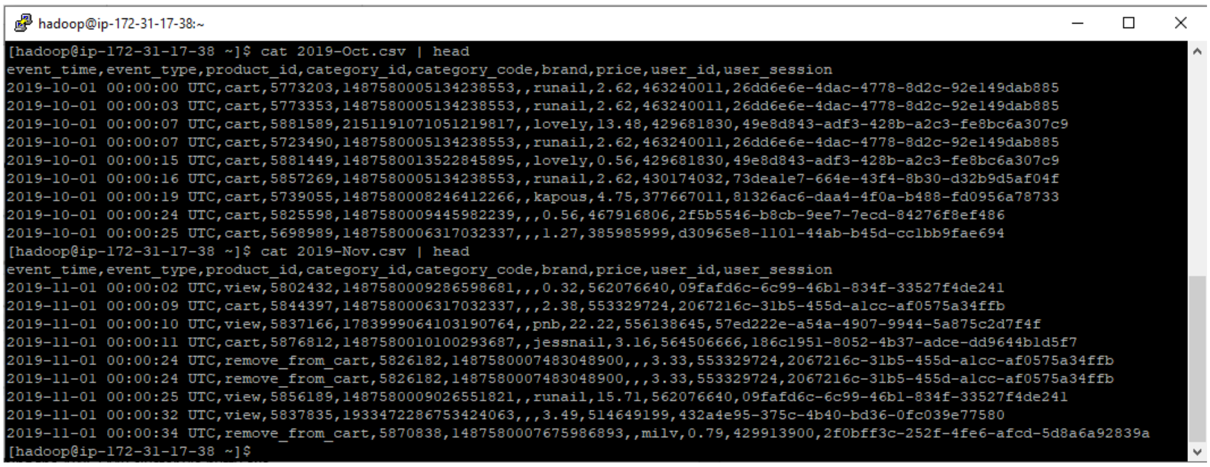
```
login as: hadoop
Authenticating with public key "hiveproj25"
Last login: Tue Nov  1 09:36:56 2022

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

https://aws.amazon.com/amazon-linux-2/
115 package(s) needed for security, out of 174 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:::::EEEEEEEE::::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-172-31-17-38 ~]$
```

3. **Loading both the given datasets in the HDFS.**



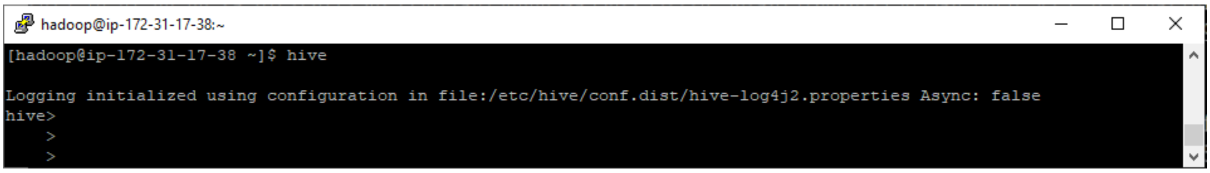
```
EEEEEEEEEEEEEEEEEEEE MMMMMMM             MMMMMMM RRRRRRR      RRRRRR

[hadoop@ip-172-31-17-38 ~]$ aws s3 cp s3://hiveproject23/2019-Oct.csv .
download: s3://hiveproject23/2019-Oct.csv to ./2019-Oct.csv
[hadoop@ip-172-31-17-38 ~]$ aws s3 cp s3://hiveproject23/2019-Nov.csv .
download: s3://hiveproject23/2019-Nov.csv to ./2019-Nov.csv
[hadoop@ip-172-31-17-38 ~]$
```

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



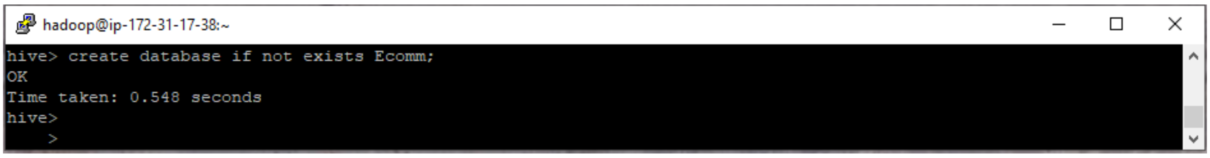
```
[hadoop@ip-172-31-17-38 ~]$ cat 2019-Oct.csv | head
event_time,event_type,product_id,category_id,category_code,brand,price,user_id,user_session
2019-10-01 00:00:00 UTC,cart,5773203,1487580005134238553,,runail,2.62,463240011,26dd6e6e-4dac-4778-8d2c-92e149dab885
2019-10-01 00:00:03 UTC,cart,5773353,1487580005134238553,,runail,2.62,463240011,26dd6e6e-4dac-4778-8d2c-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-172-31-17-38 ~]$ 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-172-31-17-38 ~]$
```

5. **Launching Hive.**



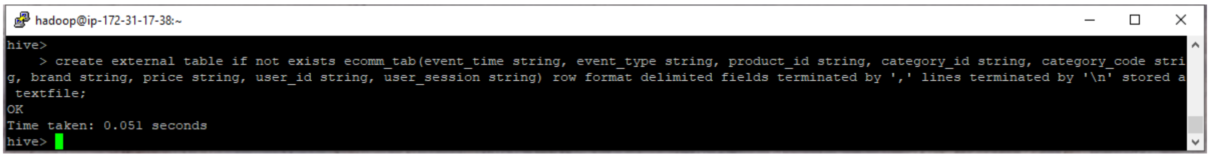
```
[hadoop@ip-172-31-17-38 ~]$ hive

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

6. **Creating the database 'Ecomm' and using it in Hive.**



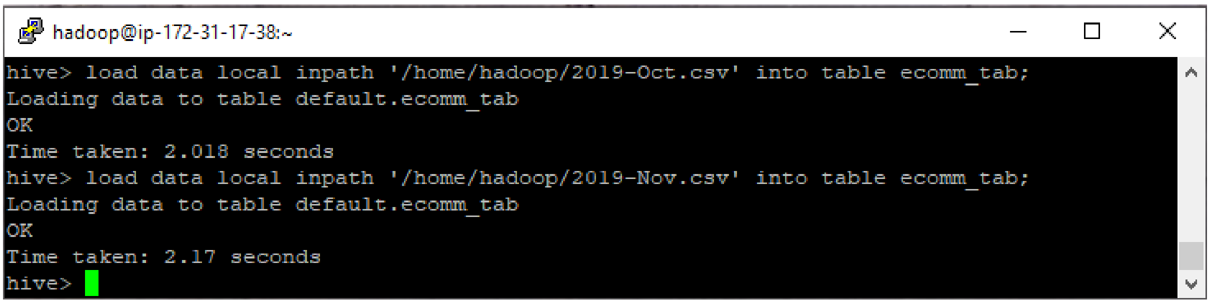
```
hive> create database if not exists Ecomm;
OK
Time taken: 0.548 seconds
hive>
    >
```

7. **Creating an External table 'ecomm_tab'**



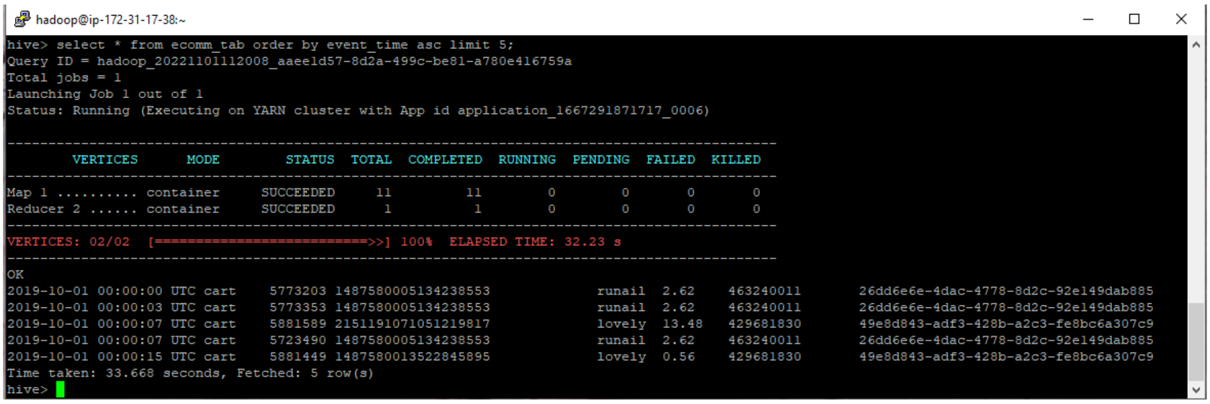
```
hive>
    > create external table if not exists ecomm_tab(event_time string, event_type string, product_id string, category_id string, category_code stri
g, brand string, price string, user_id string, user_session string) row format delimited fields terminated by ',' lines terminated by '\n' stored a
 textfile;
OK
Time taken: 0.051 seconds
hive>
```

8. **Loading and inserting the data 2019-Nov.csv & 2019-Oct.csv in the 'ecomm_tab' table.**



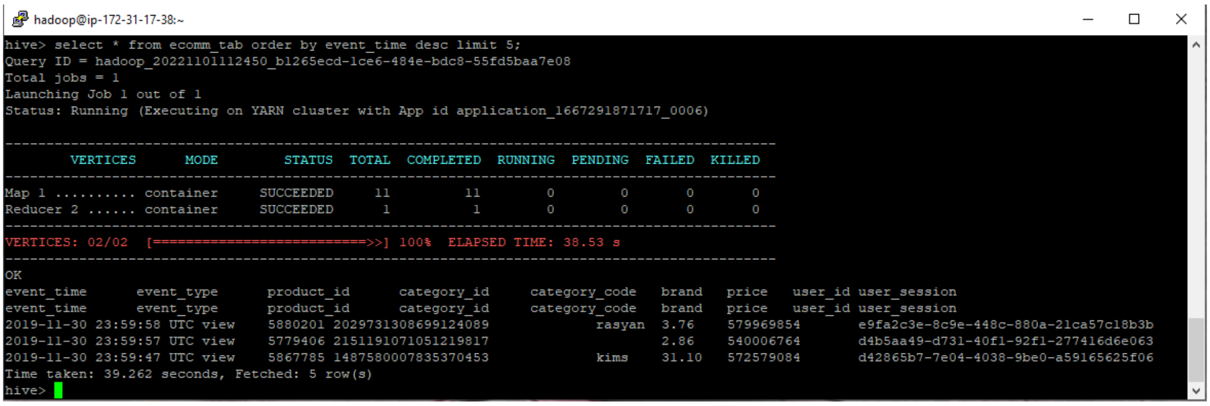
```
hive> load data local inpath '/home/hadoop/2019-Oct.csv' into table ecomm_tab;
Loading data to table default.ecomm_tab
OK
Time taken: 2.018 seconds
hive> load data local inpath '/home/hadoop/2019-Nov.csv' into table ecomm_tab;
Loading data to table default.ecomm_tab
OK
Time taken: 2.17 seconds
hive>
```

9. **Viewing the table records in month – wise manner.**
-Oct-2019



```
hive> select * from ecomm_tab order by event_time asc limit 5;
Query ID = hadoop_20221101112008_aaee1d57-8d2a-499c-be81-a780e416759a
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1667291871717_0006)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED     11         11        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 32.23 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: 33.668 seconds, Fetched: 5 row(s)
hive>
```

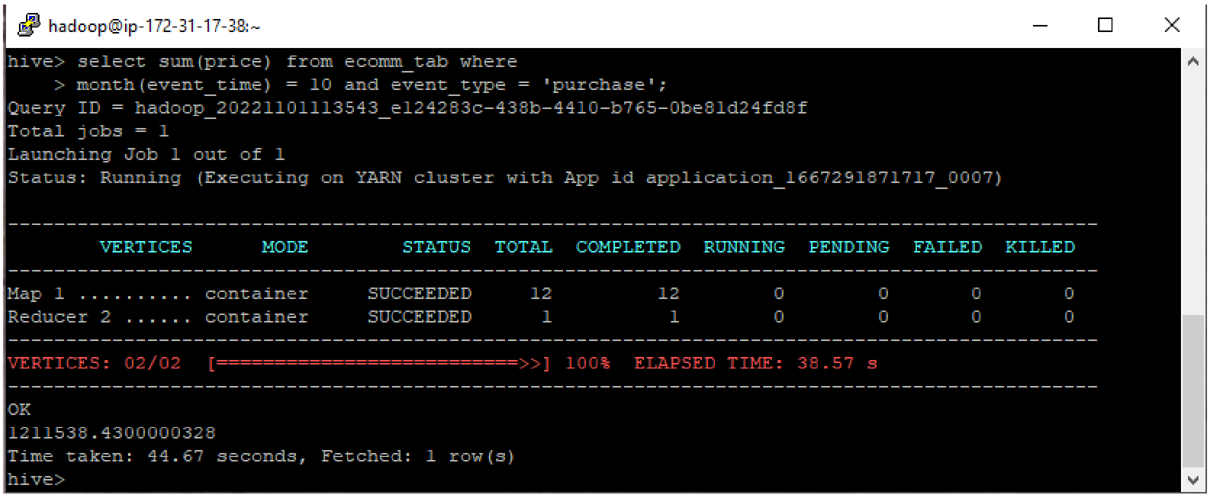
-Nov-2019



```
hive> select * from ecomm_tab order by event_time desc limit 5;
Query ID = hadoop_20221101112450_b1265ecd-1ce6-484e-bdc8-55fd5baa7e08
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1667291871717_0006)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED     11         11        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 38.53 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: 39.262 seconds, Fetched: 5 row(s)
hive>
```

Solved Questions

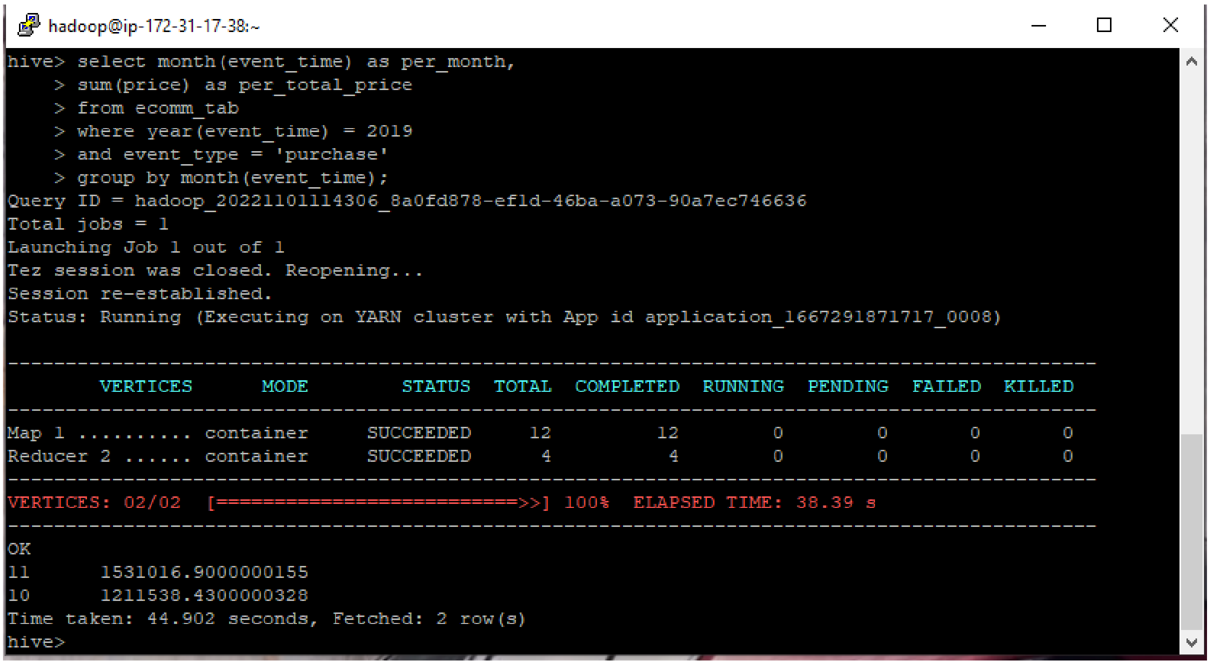
1. **Find the total revenue generated due to purchases made in October.**



hadoop@ip-172-31-17-38:~

```
hive> select sum(price) from ecomm_tab where
    > month(event_time) = 10 and event_type = 'purchase';
Query ID = hadoop_20221101113543_e124283c-438b-4410-b765-0be81d24fd8f
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1667291871717_0007)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED     12         12        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 38.57 s
----------------------------------------------------------------------------------------------
OK
1211538.4300000328
Time taken: 44.67 seconds, Fetched: 1 row(s)
hive>
```

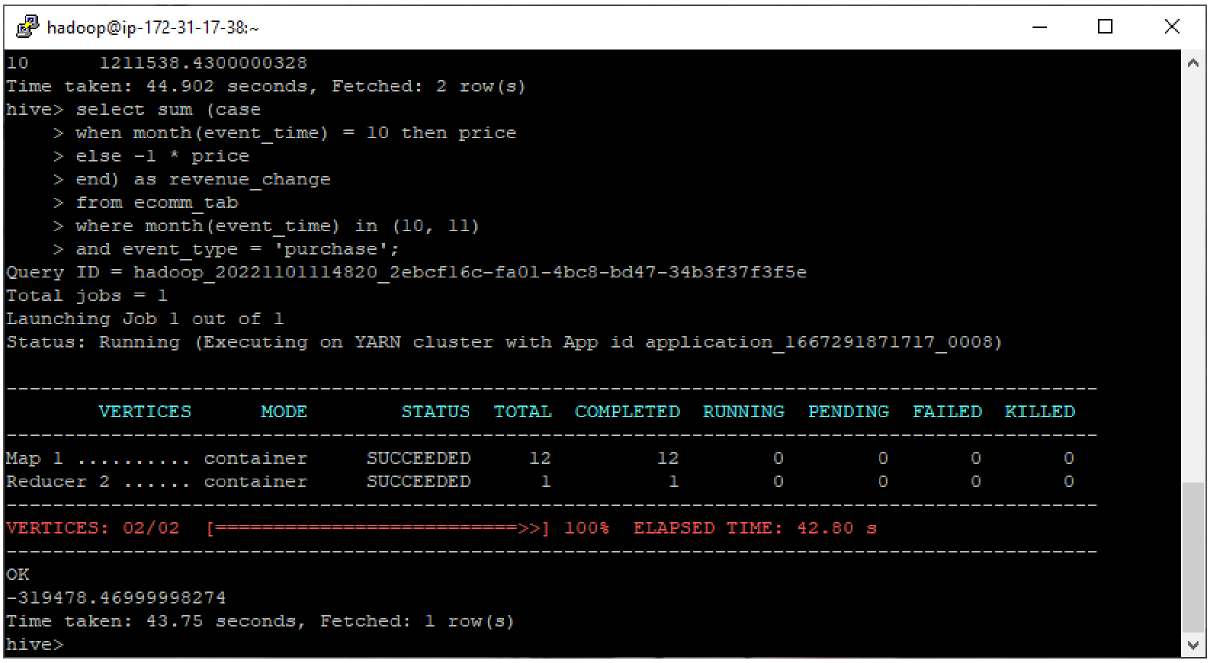
2. **Write a query to yield the total sum of purchases per month in a single output.**



hadoop@ip-172-31-17-38:~

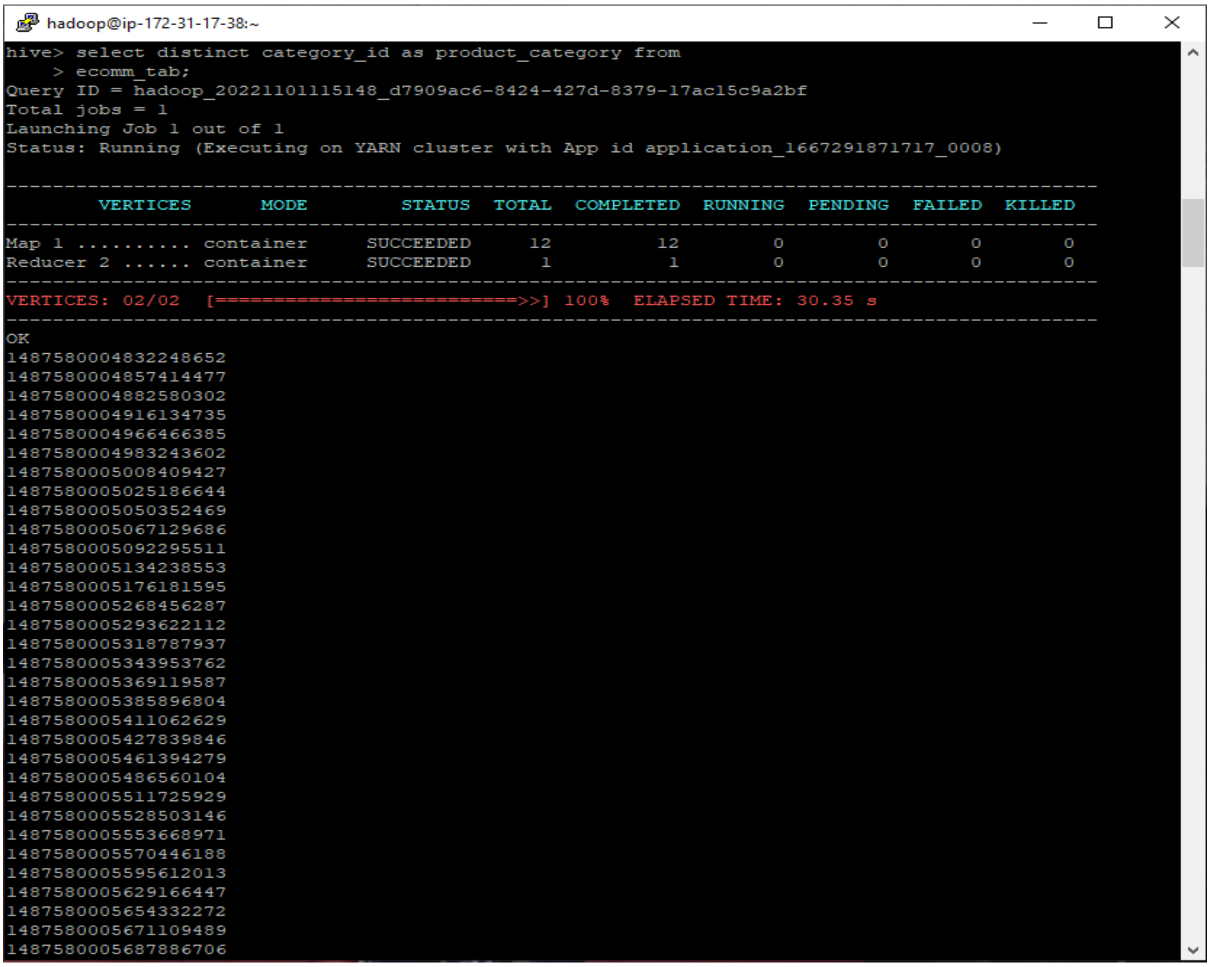
```
hive> select month(event_time) as per_month,
    > sum(price) as per_total_price
    > from ecomm_tab
    > where year(event_time) = 2019
    > and event_type = 'purchase'
    > group by month(event_time);
Query ID = hadoop_20221101114306_8a0fd878-ef1d-46ba-a073-90a7ec746636
Total jobs = 1
Launching Job 1 out of 1
Tez session was closed. Reopening...
Session re-established.
Status: Running (Executing on YARN cluster with App id application_1667291871717_0008)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED     12         12        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      4          4        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 38.39 s
----------------------------------------------------------------------------------------------
OK
11      1531016.9000000155
10      1211538.4300000328
Time taken: 44.902 seconds, Fetched: 2 row(s)
hive>
```

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

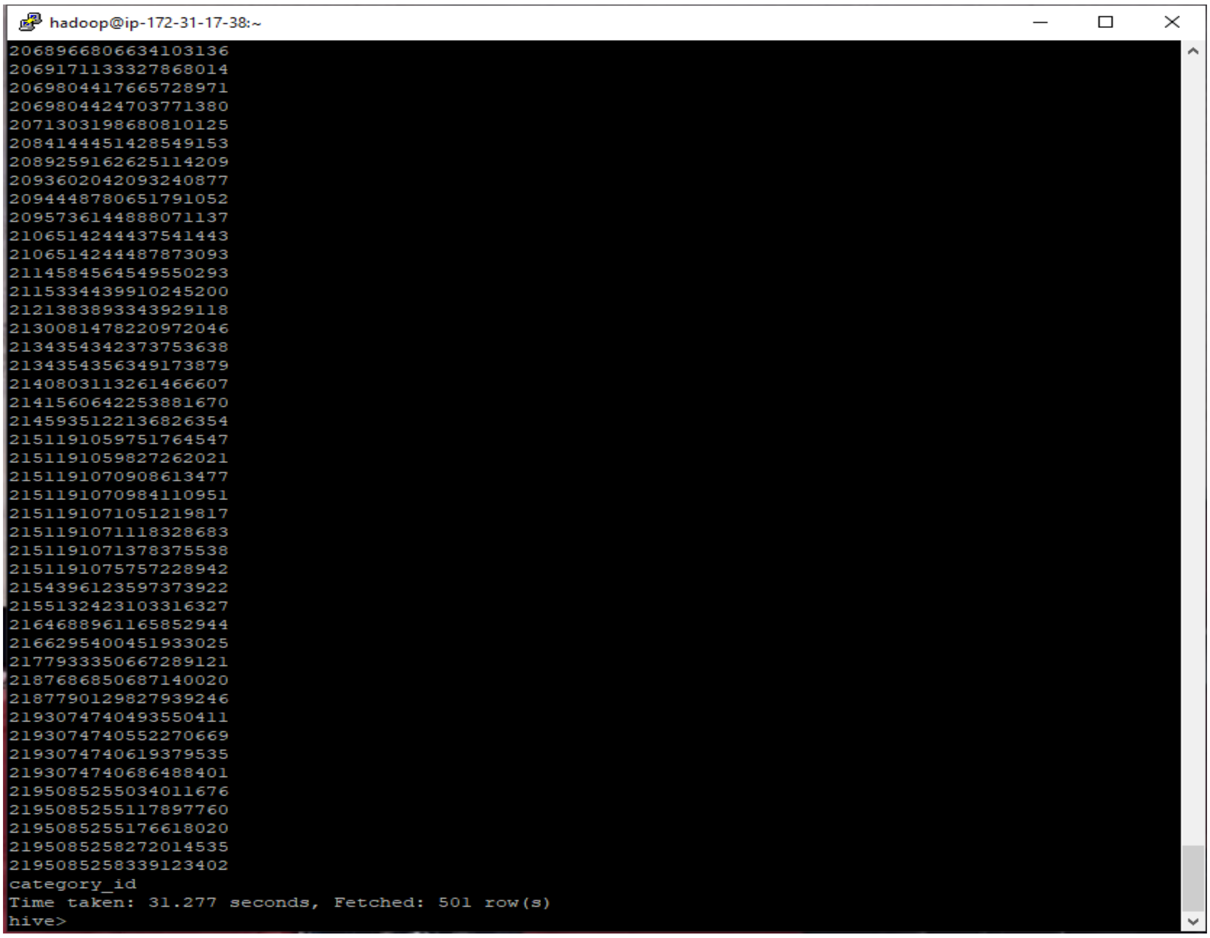


```
10      1211538.4300000328
Time taken: 44.902 seconds, Fetched: 2 row(s)
hive> select sum (case
    > when month(event_time) = 10 then price
    > else -1 * price
    > end) as revenue_change
    > from ecomm_tab
    > where month(event_time) in (10, 11)
    > and event_type = 'purchase';
Query ID = hadoop_20221101114820_2ebcf16c-fa01-4bc8-bd47-34b3f37f3f5e
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1667291871717_0008)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED     12         12        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 42.80 s
----------------------------------------------------------------------------------------------
OK
-319478.46999998274
Time taken: 43.75 seconds, Fetched: 1 row(s)
hive>
```

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



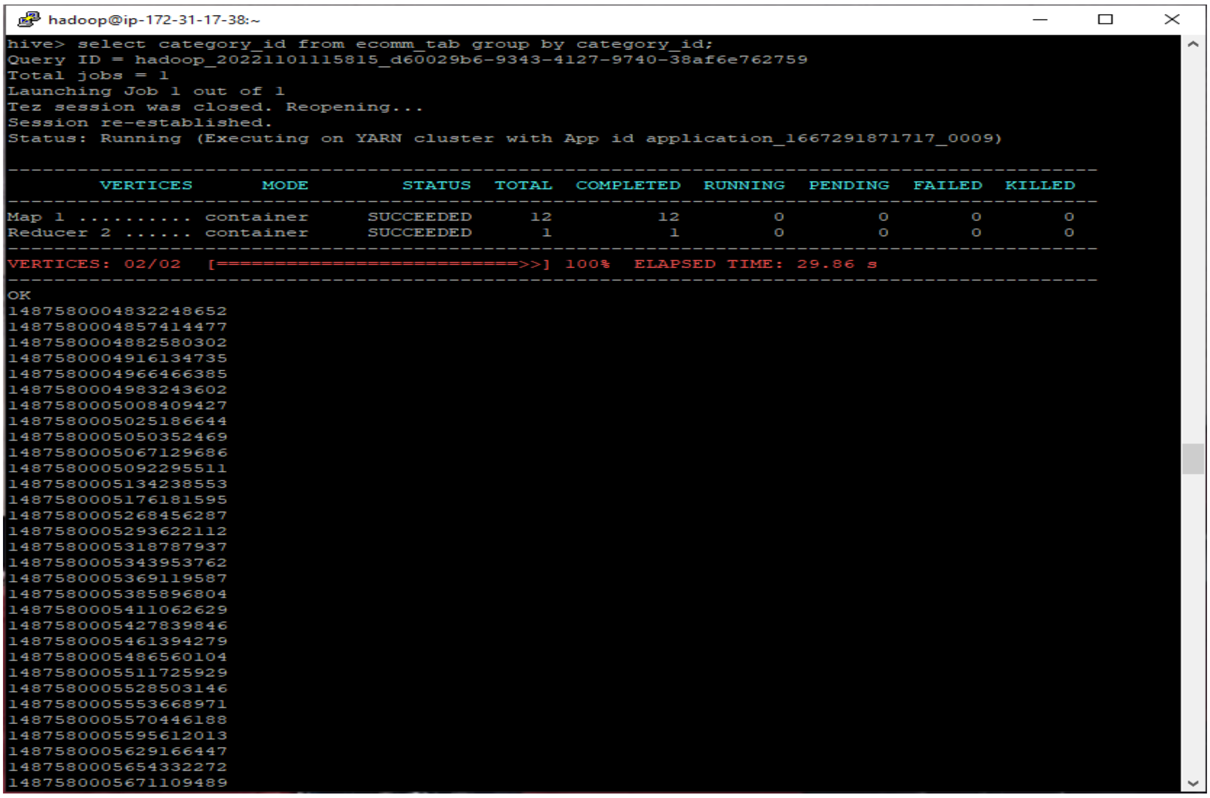
```
hive> select distinct category_id as product_category from
    > ecomm_tab;
Query ID = hadoop_20221101115148_d7909ac6-8424-427d-8379-17ac15c9a2bf
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1667291871717_0008)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED     12         12        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 30.35 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
```



hadoop@ip-172-31-17-38:~

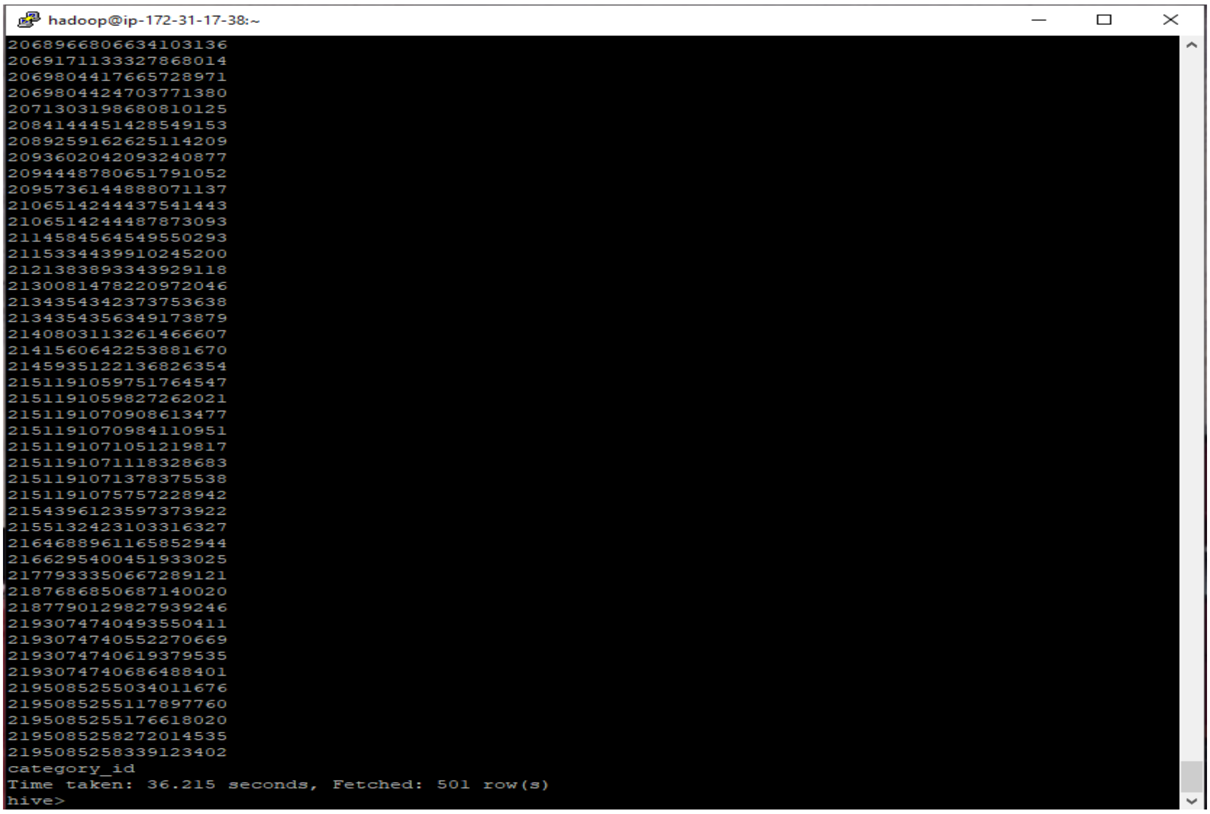
```
2068966806634103136
2069171133327868014
2069804417665728971
2069804424703771380
2071303198680810125
2084144451428549153
2089259162625114209
2093602042093240877
2094448780651791052
2095736144888071137
2106514244437541443
2106514244487873093
2114584564549550293
2115334439910245200
2121383893343929118
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: 31.277 seconds, Fetched: 501 row(s)
hive>
```

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



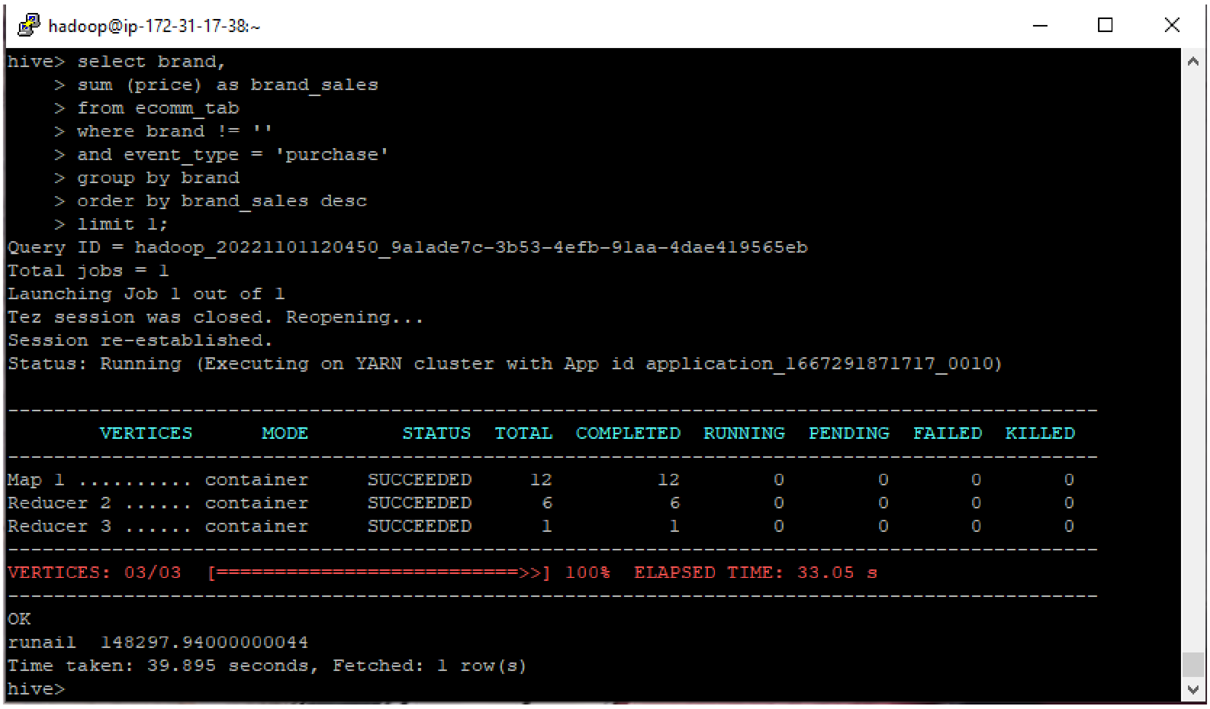
hadoop@ip-172-31-17-38:~

```
hive> select category_id from ecomm_tab group by category_id;
Query ID = hadoop_20221101115815_d60029b6-9343-4127-9740-38af6e762759
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_1667291871717_0009)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED     12         12        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 29.86 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
```



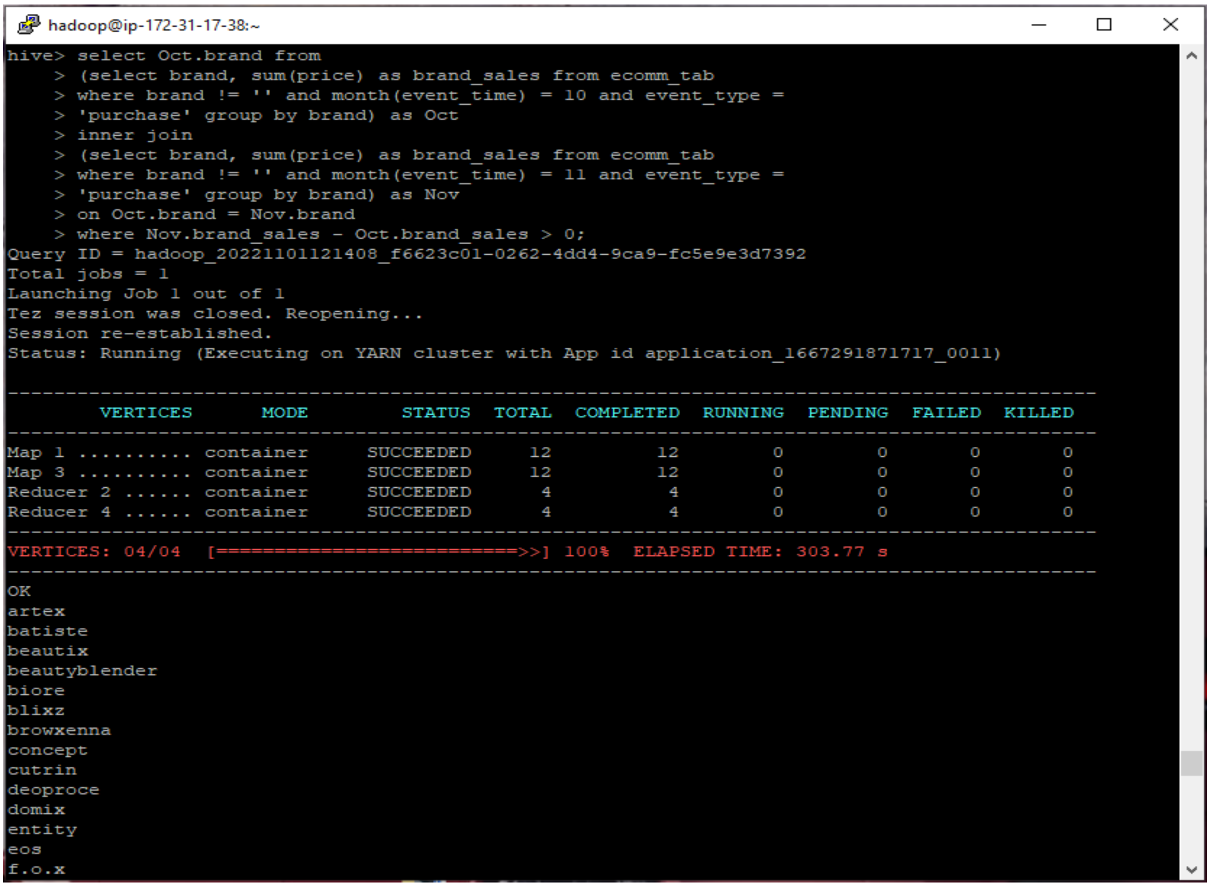
```
2068966806634103136
2069171133327868014
2069804417665728971
2069804424703771380
2071303198680810125
2084144451428549153
2089259162625114209
2093602042093240877
2094448780651791052
2095736144888071137
2106514244437541443
2106514244487873093
2114584564549550293
2115334439910245200
2121383893343929118
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: 36.215 seconds, Fetched: 501 row(s)
hive>
```

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

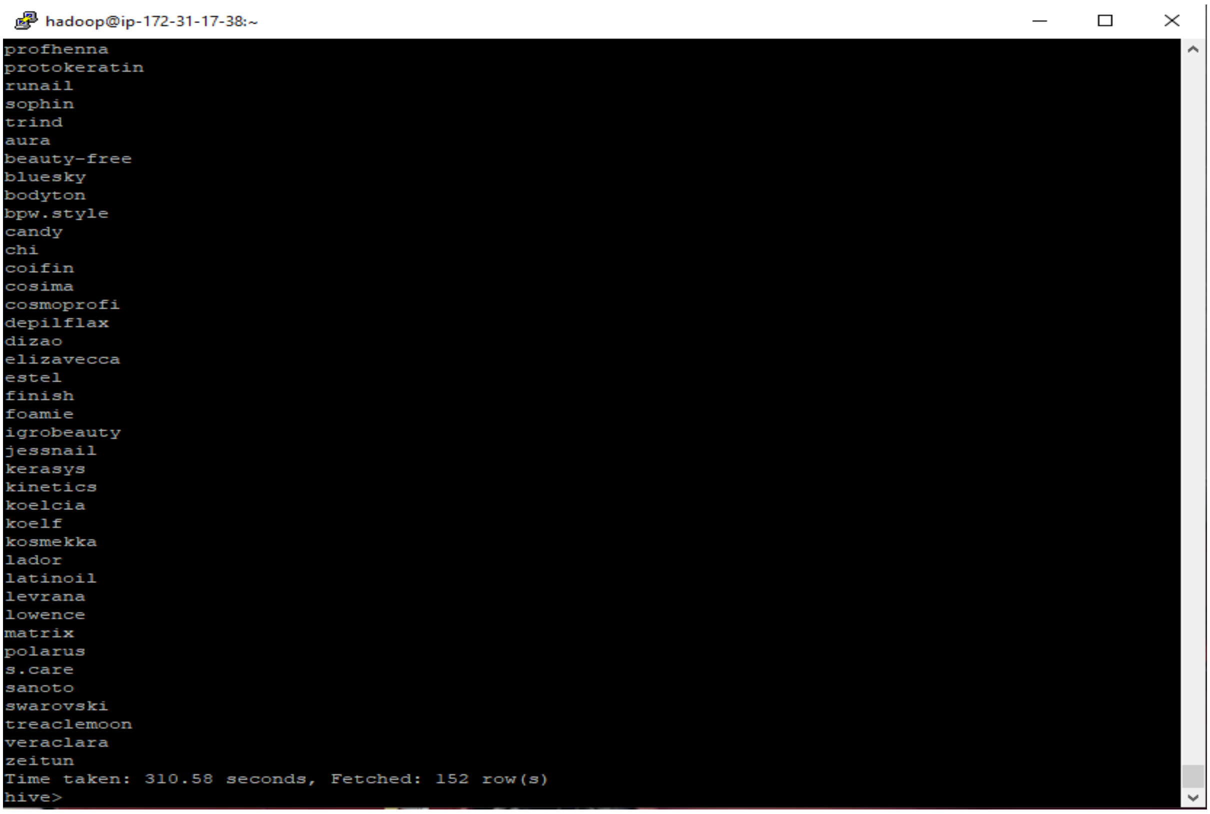


```
hive> select brand,
    > sum (price) as brand_sales
    > from ecomm_tab
    > where brand != ''
    > and event_type = 'purchase'
    > group by brand
    > order by brand_sales desc
    > limit 1;
Query ID = hadoop_20221101120450_9alade7c-3b53-4efb-91aa-4dae419565eb
Total jobs = 1
Launching Job 1 out of 1
Tez session was closed. Reopening...
Session re-established.
Status: Running (Executing on YARN cluster with App id application_1667291871717_0010)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED     12         12        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: 33.05 s
----------------------------------------------------------------------------------------------
OK
runail  148297.94000000044
Time taken: 39.895 seconds, Fetched: 1 row(s)
hive>
```

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

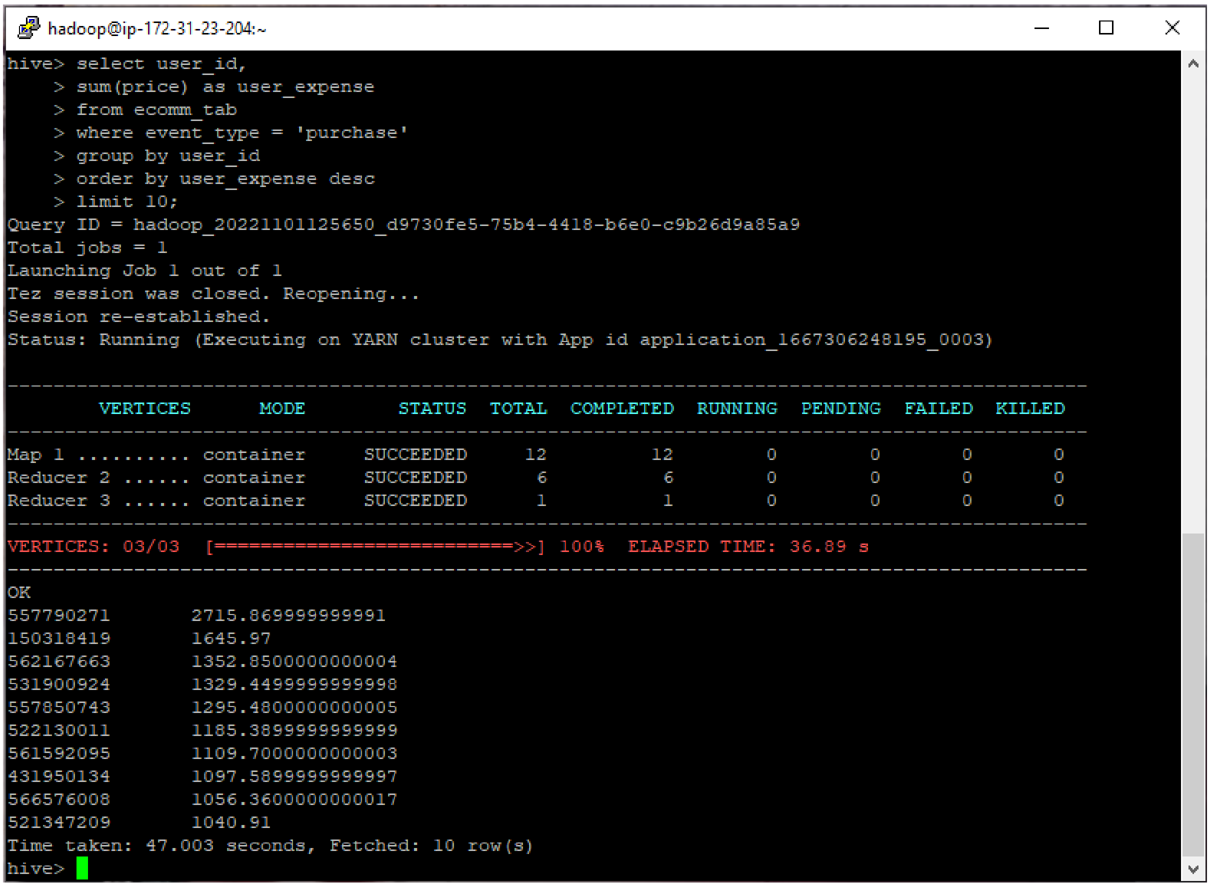


```
hive> select Oct.brand from
    > (select brand, sum(price) as brand_sales from ecomm_tab
    > where brand != '' and month(event_time) = 10 and event_type =
    > 'purchase' group by brand) as Oct
    > inner join
    > (select brand, sum(price) as brand_sales from ecomm_tab
    > where brand != '' 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_20221101121408_f6623c01-0262-4dd4-9ca9-fc5e9e3d7392
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_1667291871717_0011)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED     12         12        0        0       0       0
Map 3 .......... container     SUCCEEDED     12         12        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: 303.77 s
----------------------------------------------------------------------------------------------
OK
artex
batiste
beautix
beautyblender
biore
blixz
browxenna
concept
cutrin
deoproce
domix
entity
eos
f.o.x
```



```
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: 310.58 seconds, Fetched: 152 row(s)
hive>
```

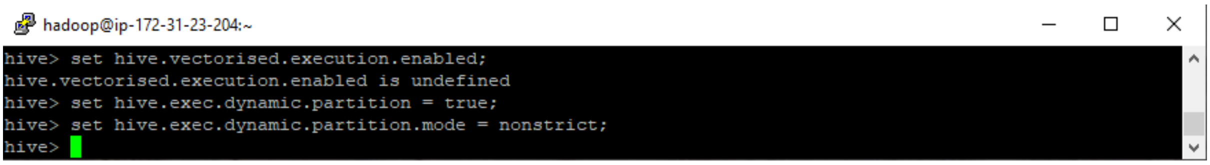
8. Your company wants to reward the top 10 users of its website with a Golden Customer plan. Write a query to generate a list of top 10 users who spend the most.



```
hive> select user_id,
    > sum(price) as user_expense
    > from ecomm_tab
    > where event_type = 'purchase'
    > group by user_id
    > order by user_expense desc
    > limit 10;
Query ID = hadoop_20221101125650_d9730fe5-75b4-4418-b6e0-c9b26d9a85a9
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_1667306248195_0003)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED     12         12        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: 36.89 s
----------------------------------------------------------------------------------------------
OK
557790271       2715.869999999991
150318419       1645.97
562167663       1352.8500000000004
531900924       1329.4499999999998
557850743       1295.4800000000005
522130011       1185.3899999999999
561592095       1109.7000000000003
431950134       1097.5899999999997
566576008       1056.3600000000017
521347209       1040.91
Time taken: 47.003 seconds, Fetched: 10 row(s)
hive>
```

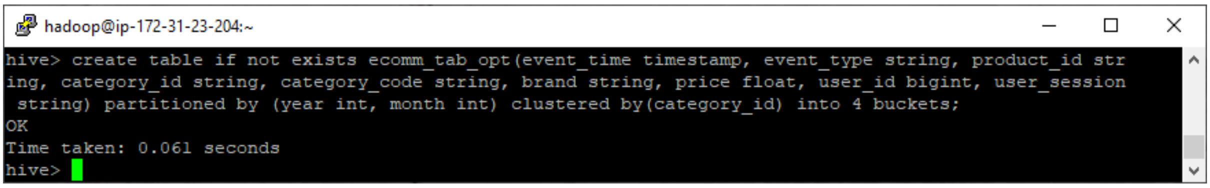
Optimising query and overall efficiency

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



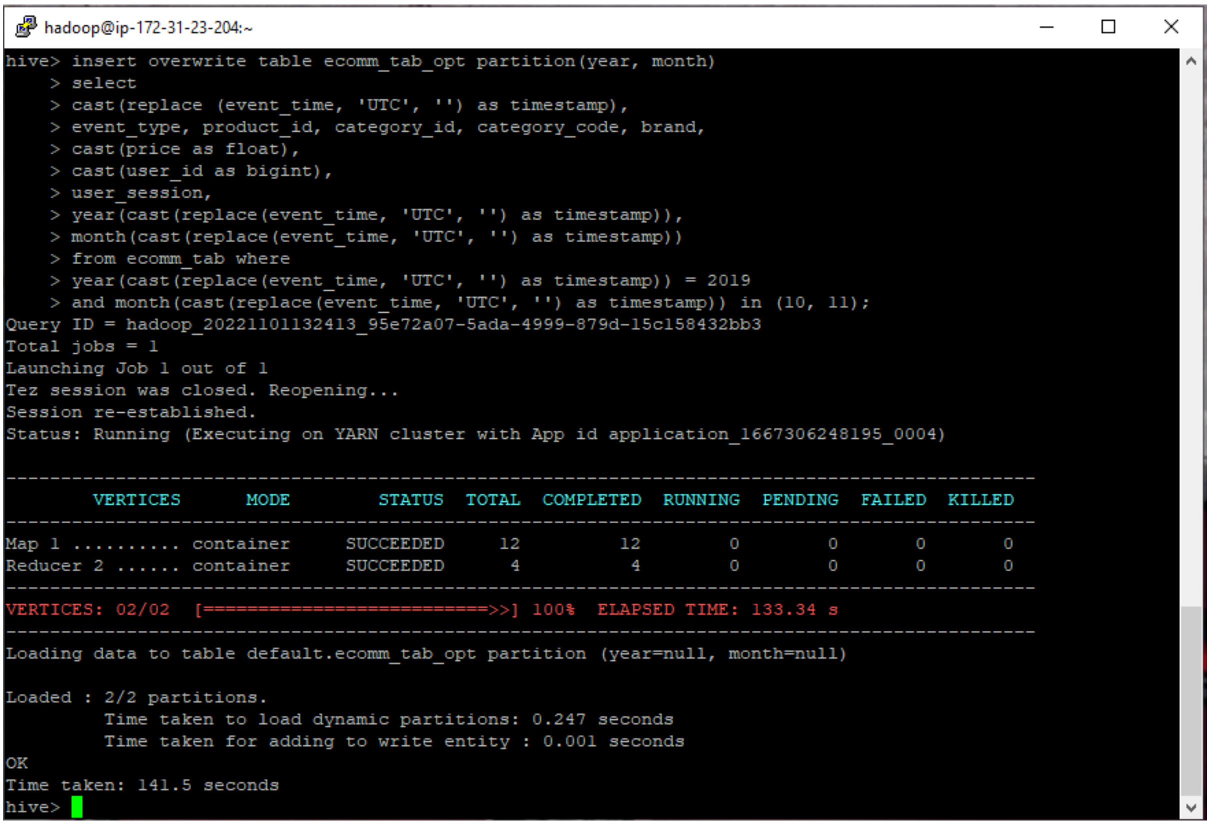
```
hive> set hive.vectorised.execution.enabled;
hive.vectorised.execution.enabled is undefined
hive> set hive.exec.dynamic.partition = true;
hive> set hive.exec.dynamic.partition.mode = nonstrict;
hive> 
```

2. **Creating an optimized table 'ecomm_tab_opt' with partitioning and dividing it into 4 buckets.**



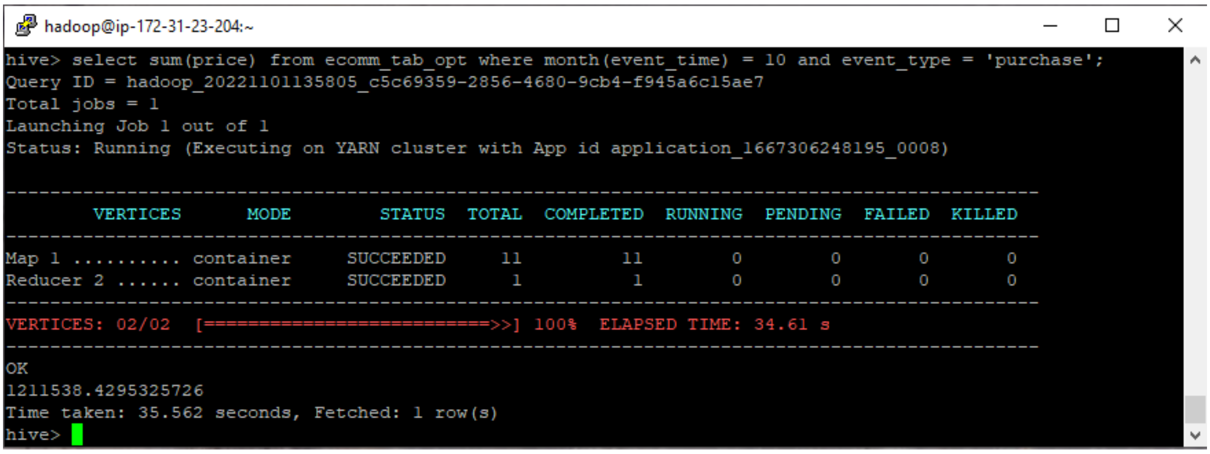
```
hive> create table if not exists ecomm_tab_opt(event_time timestamp, event_type string, product_id str
ing, category_id string, category_code string, brand string, price float, user_id bigint, user_session
 string) partitioned by (year int, month int) clustered by(category_id) into 4 buckets;
OK
Time taken: 0.061 seconds
hive> 
```

3. **Loading and inserting data into optimized table 'ecomm_tab_opt**



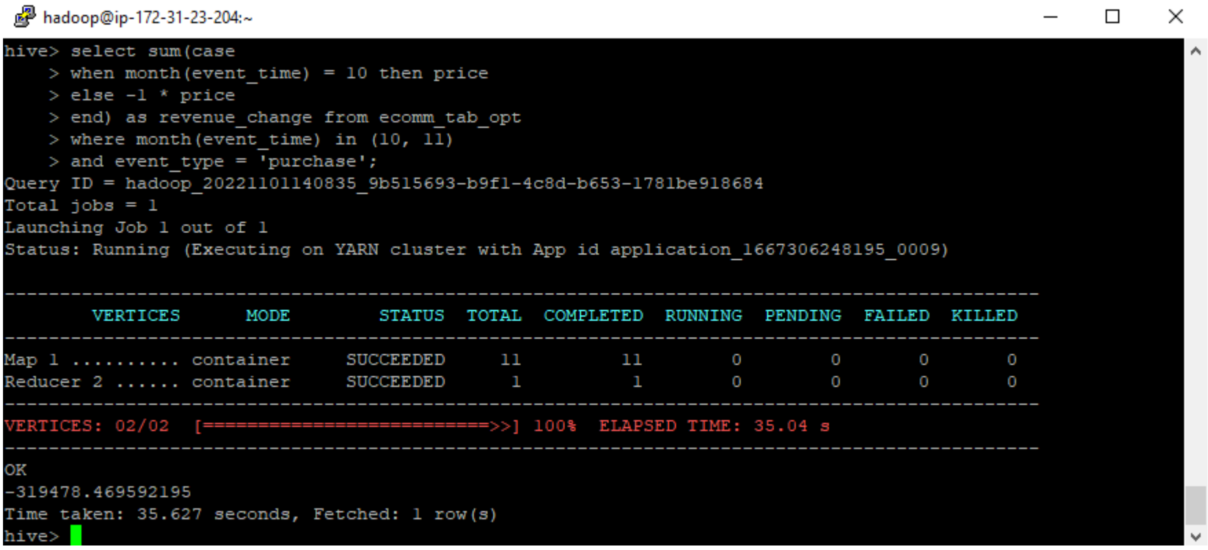
```
hive> insert overwrite table ecomm_tab_opt partition(year, month)
    > select
    > cast(replace (event_time, 'UTC', '') as timestamp),
    > event_type, product_id, category_id, category_code, brand,
    > cast(price as float),
    > cast(user_id as bigint),
    > user_session,
    > year(cast(replace(event_time, 'UTC', '') as timestamp)),
    > month(cast(replace(event_time, 'UTC', '') as timestamp))
    > from ecomm_tab 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_20221101132413_95e72a07-5ada-4999-879d-15c158432bb3
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_1667306248195_0004)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED     12         12        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      4          4        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 133.34 s
----------------------------------------------------------------------------------------------
Loading data to table default.ecomm_tab_opt partition (year=null, month=null)

Loaded : 2/2 partitions.
         Time taken to load dynamic partitions: 0.247 seconds
         Time taken for adding to write entity : 0.001 seconds
OK
Time taken: 141.5 seconds
hive> 
```

4. **After optimizing the table running query from Q.1**
   Before Optimization – Time taken 44.67 seconds
   After Optimization – Time taken 35.562 seconds



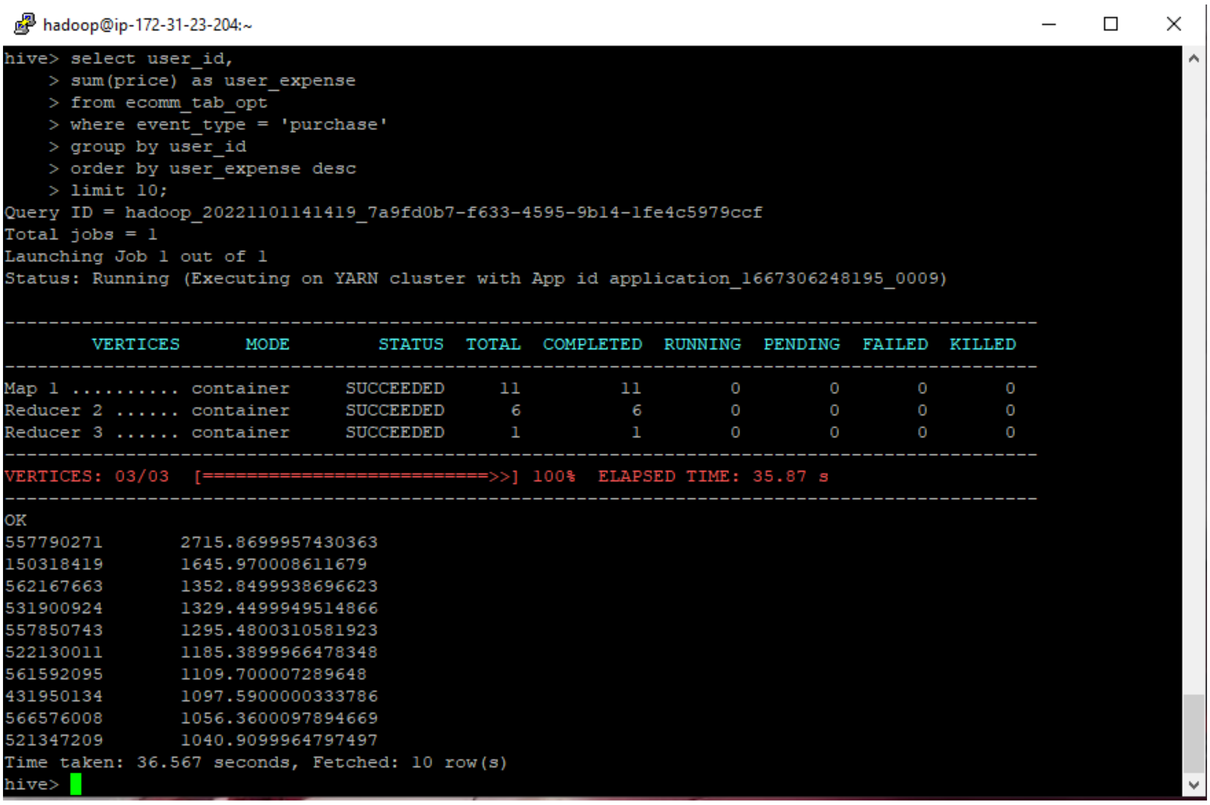
```
hive> select sum(price) from ecomm_tab_opt where month(event_time) = 10 and event_type = 'purchase';
Query ID = hadoop_20221101135805_c5c69359-2856-4680-9cb4-f945a6c15ae7
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1667306248195_0008)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED     11         11        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 34.61 s
----------------------------------------------------------------------------------------------
OK
1211538.4295325726
Time taken: 35.562 seconds, Fetched: 1 row(s)
hive>
```

5. **After optimizing the table running query from Q.3**
   Before Optimization – Time taken 43.75 seconds
   After Optimization – Time taken 35.627 seconds



```
hive> select sum(case
    > when month(event_time) = 10 then price
    > else -1 * price
    > end) as revenue_change from ecomm_tab_opt
    > where month(event_time) in (10, 11)
    > and event_type = 'purchase';
Query ID = hadoop_20221101140835_9b515693-b9f1-4c8d-b653-1781be918684
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1667306248195_0009)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED     11         11        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 35.04 s
----------------------------------------------------------------------------------------------
OK
-319478.469592195
Time taken: 35.627 seconds, Fetched: 1 row(s)
hive>
```

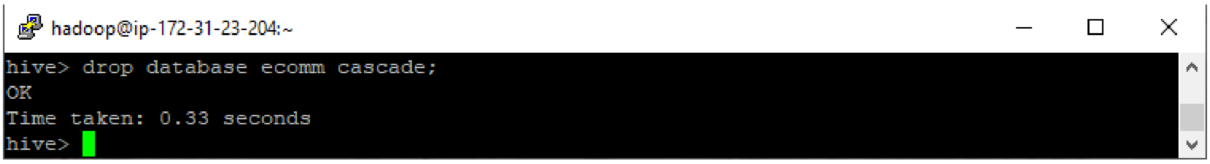
6. **After optimizing the table running query from Q.8**
   Before Optimization – Time taken 47.003 seconds
   After Optimization – Time taken 36.567 seconds



```
hive> select user_id,
    > sum(price) as user_expense
    > from ecomm_tab_opt
    > where event_type = 'purchase'
    > group by user_id
    > order by user_expense desc
    > limit 10;
Query ID = hadoop_20221101141419_7a9fd0b7-f633-4595-9b14-1fe4c5979ccf
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1667306248195_0009)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED     11         11        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: 35.87 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: 36.567 seconds, Fetched: 10 row(s)
hive>
```

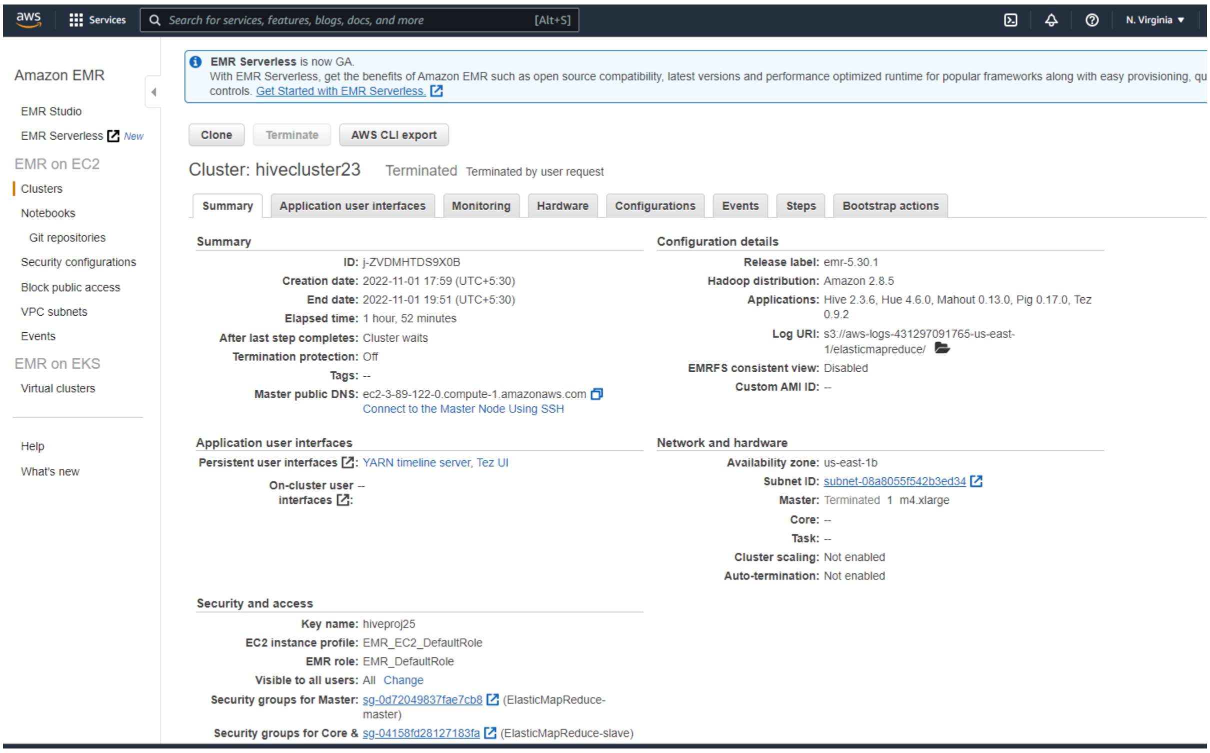
Cleaning

1. Dropping the previously created database ‘Ecommerce’



```
hive> drop database ecomm cascade;
OK
Time taken: 0.33 seconds
hive>
```

2. Terminating the AWS EMR cluster



EMR Serverless is now GA.
With EMR Serverless, get the benefits of Amazon EMR such as open source compatibility, latest versions and performance optimized runtime for popular frameworks along with easy provisioning, qu
controls. Get Started with EMR Serverless.

Cluster: hivecluster23 Terminated Terminated by user request

Summary

ID: j-ZVDMHTDS9X0B
Creation date: 2022-11-01 17:59 (UTC+5:30)
End date: 2022-11-01 19:51 (UTC+5:30)
Elapsed time: 1 hour, 52 minutes
After last step completes: Cluster waits
Termination protection: Off
Tags: --
Master public DNS: ec2-3-89-122-0.compute-1.amazonaws.com
Connect to the Master Node Using SSH

Configuration details

Release label: emr-5.30.1
Hadoop distribution: Amazon 2.8.5
Applications: Hive 2.3.6, Hue 4.6.0, Mahout 0.13.0, Pig 0.17.0, Tez 0.9.2
Log URI: s3://aws-logs-431297091765-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-1b
Subnet ID: subnet-08a8055f542b3ed34
Master: Terminated 1 m4.xlarge
Core: --
Task: --
Cluster scaling: Not enabled
Auto-termination: Not enabled

Security and access

Key name: hiveproj25
EC2 instance profile: EMR_EC2_DefaultRole
EMR role: EMR_DefaultRole
Visible to all users: All Change
Security groups for Master: sg-0d72049837fae7cb8 (ElasticMapReduce-master)
Security groups for Core & sg-04158fd28127183fa (ElasticMapReduce-slave)