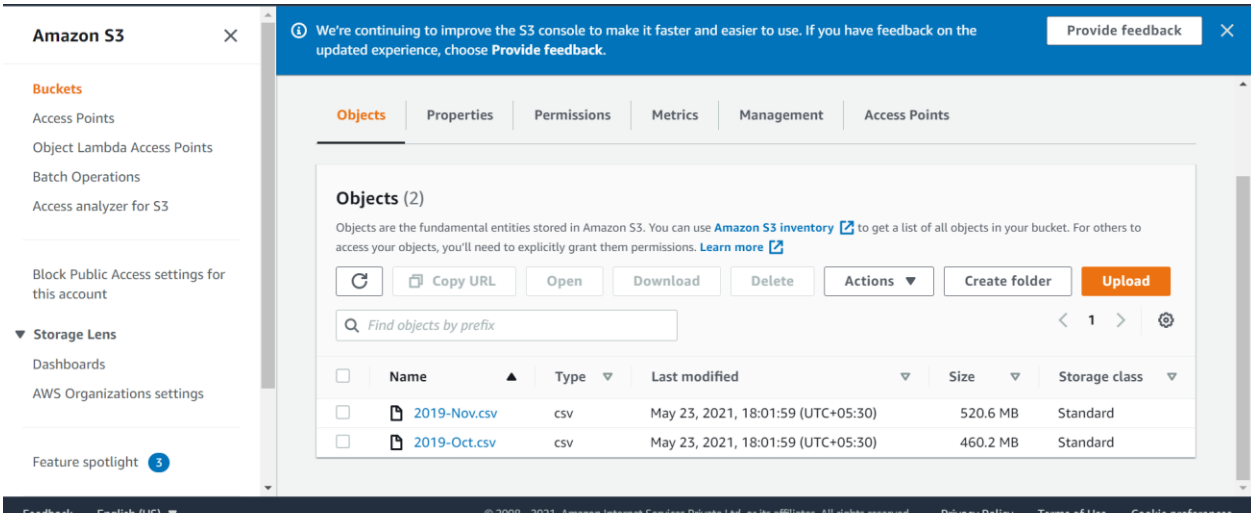
Ecommerce Sales Data Analysis

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.

Collection and Processing of Data:

1. Upload **2019-Nov.csv** & **2019-Oct.csv** data in S3.



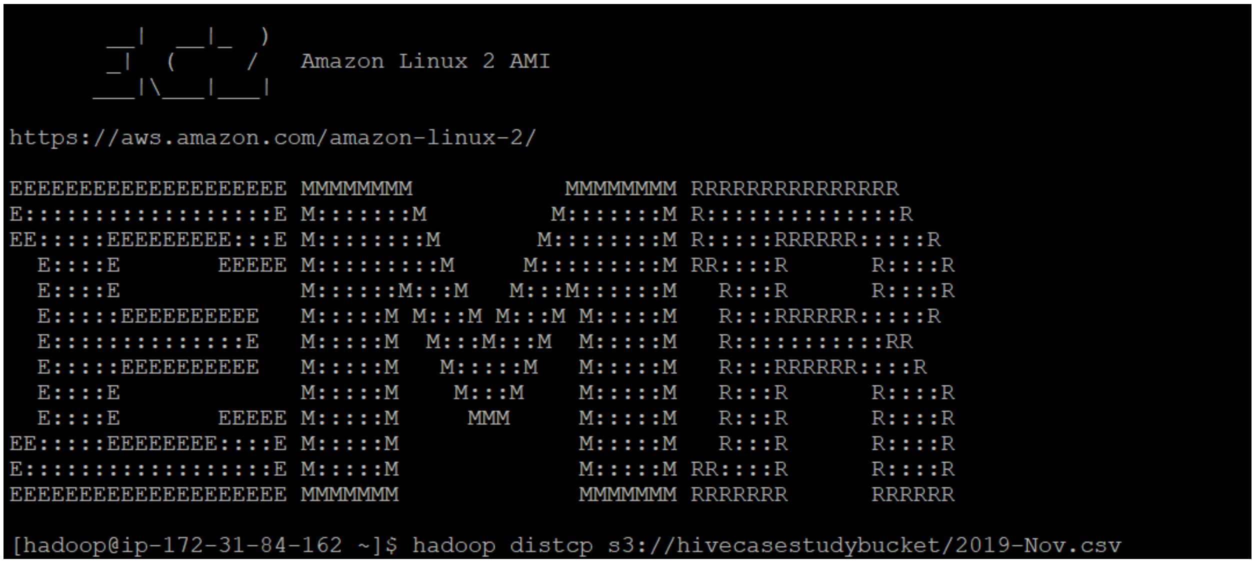
We're continuing to improve the S3 console to make it faster and easier to use. If you have feedback on the updated experience, choose **Provide feedback**.

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**

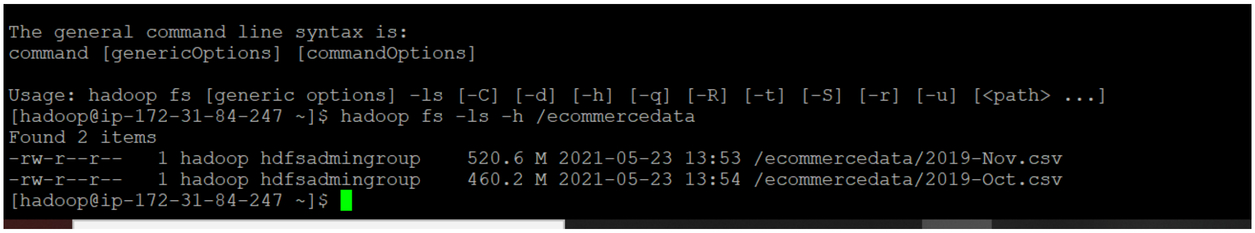
Name	Type	Last modified	Size	Storage class
2019-Nov.csv	csv	May 23, 2021, 18:01:59 (UTC+05:30)	520.6 MB	Standard
2019-Oct.csv	csv	May 23, 2021, 18:01:59 (UTC+05:30)	460.2 MB	Standard

2. Launch the EMR Cluster



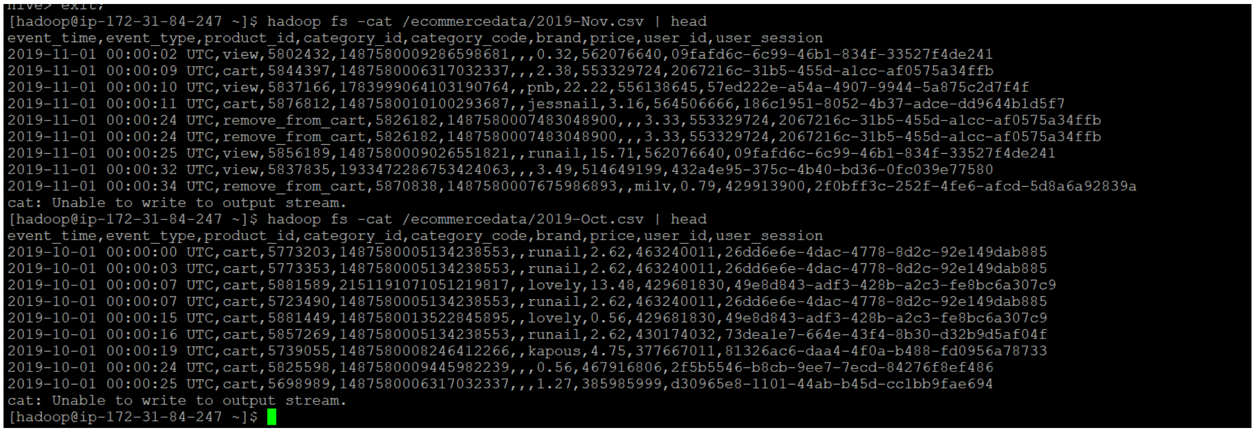
```
       __|  __|_  )
       _|  (     /   Amazon Linux 2 AMI
      ___|\___|___|

https://aws.amazon.com/amazon-linux-2/

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-172-31-84-162 ~]$ hadoop distcp s3://hivecasestudybucket/2019-Nov.csv
```

3. Load both the datasets in HDFS



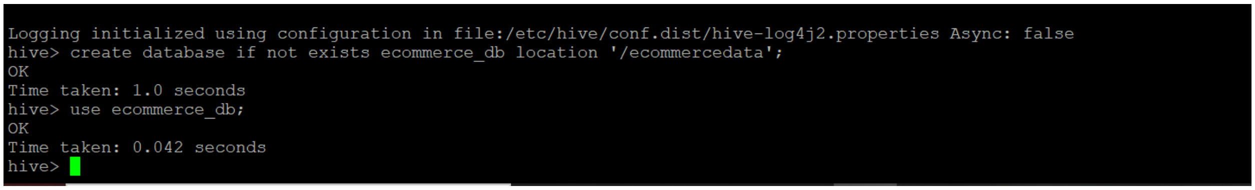
```
The general command line syntax is:
command [genericOptions] [commandOptions]

Usage: hadoop fs [generic options] -ls [-C] [-d] [-h] [-q] [-R] [-t] [-S] [-r] [-u] [<path> ...]
[hadoop@ip-172-31-84-247 ~]$ hadoop fs -ls -h /ecommercedata
Found 2 items
-rw-r--r--   1 hadoop hdfsadmingroup    520.6 M 2021-05-23 13:53 /ecommercedata/2019-Nov.csv
-rw-r--r--   1 hadoop hdfsadmingroup    460.2 M 2021-05-23 13:54 /ecommercedata/2019-Oct.csv
[hadoop@ip-172-31-84-247 ~]$
```

4. View datasets **2019-Nov.csv** and **2019-Oct.csv**



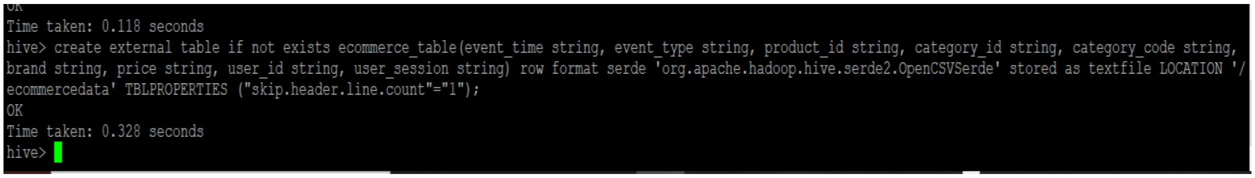
```
hive> exit;
[hadoop@ip-172-31-84-247 ~]$ hadoop fs -cat /ecommercedata/2019-Nov.csv | head
event_time,event_type,product_id,category_id,category_code,brand,price,user_id,user_session
2019-11-01 00:00:02 UTC,view,5802432,1487580009286598681,,,0.32,562076640,09fafd6c-6c99-46b1-834f-33527f4de241
2019-11-01 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
cat: Unable to write to output stream.
[hadoop@ip-172-31-84-247 ~]$ hadoop fs -cat /ecommercedata/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,2.62,430174032,73dea1e7-664e-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
cat: Unable to write to output stream.
[hadoop@ip-172-31-84-247 ~]$
```

5. Creation and Use of Database **‘ecommerce_db’** in Hive



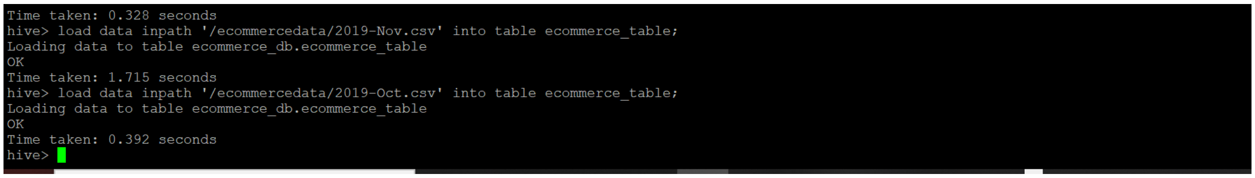
```
Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j2.properties Async: false
hive> create database if not exists ecommerce_db location '/ecommercedata';
OK
Time taken: 1.0 seconds
hive> use ecommerce_db;
OK
Time taken: 0.042 seconds
hive>
```

6. Create External table **'ecommerce_table'**



```
OK
Time taken: 0.118 seconds
hive> create external table if not exists ecommerce_table(event_time string, event_type string, product_id string, category_id string, category_code string,
brand string, price string, user_id string, user_session string) row format serde 'org.apache.hadoop.hive.serde2.OpenCSVSerde' stored as textfile LOCATION '/
ecommercedata' TBLPROPERTIES ("skip.header.line.count"="1");
OK
Time taken: 0.328 seconds
hive>
```

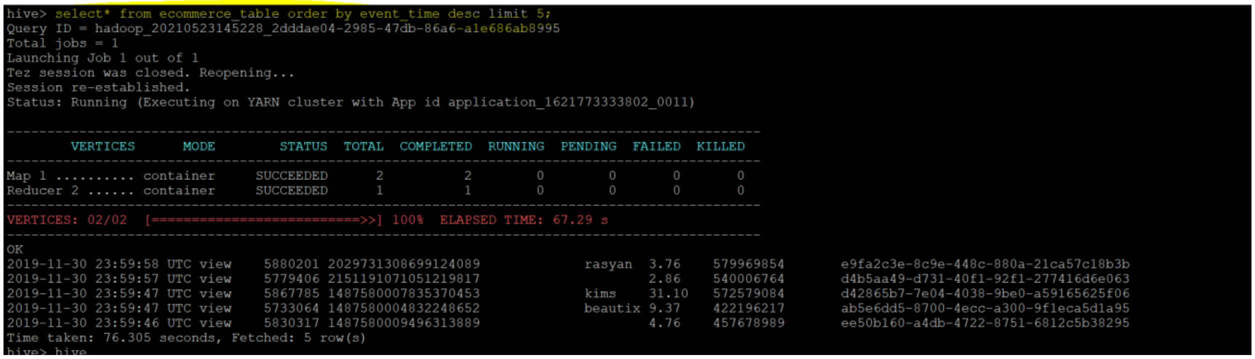
7. Load data **2019-Nov.csv** and **2019-Oct.csv** in the '**ecommerce_table'** table



```
Time taken: 0.328 seconds
hive> load data inpath '/ecommercedata/2019-Nov.csv' into table ecommerce_table;
Loading data to table ecommerce_db.ecommerce_table
OK
Time taken: 1.715 seconds
hive> load data inpath '/ecommercedata/2019-Oct.csv' into table ecommerce_table;
Loading data to table ecommerce_db.ecommerce_table
OK
Time taken: 0.392 seconds
hive>
```

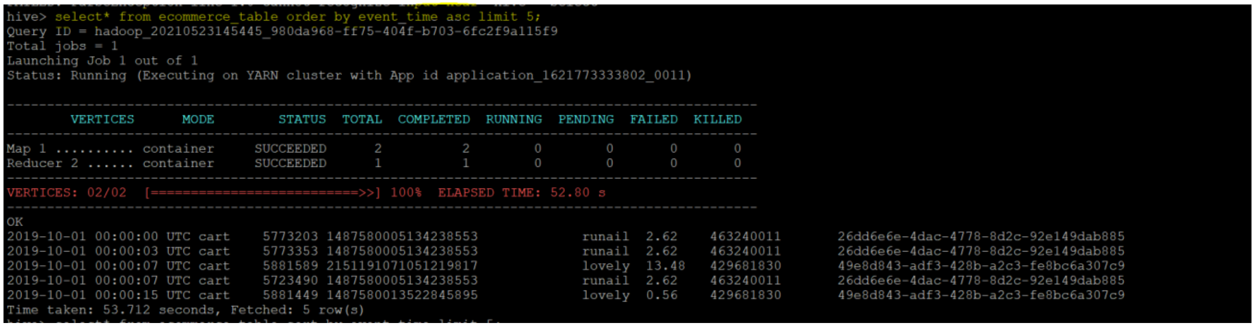
8. View table records month-wise

Nov records



```
hive> select* from ecommerce_table order by event_time desc limit 5;
Query ID = hadoop_20210523145228_2dddae04-2985-47db-86a6-a1e686ab8995
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_1621773333802_0011)

----------------------------------------------------------------------------------------------
        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: 67.29 s
----------------------------------------------------------------------------------------------
OK
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
2019-11-30 23:59:47 UTC view    5733064 1487580004832248652             beautix 9.37    422196217       ab5e6dd5-8700-4ecc-a300-9f1eca5d1a95
2019-11-30 23:59:46 UTC view    5830317 1487580009496313889                     4.76    457678989       ee50b160-a4db-4722-8751-6812c5b38295
Time taken: 76.305 seconds, Fetched: 5 row(s)
hive> hive
```

Oct records



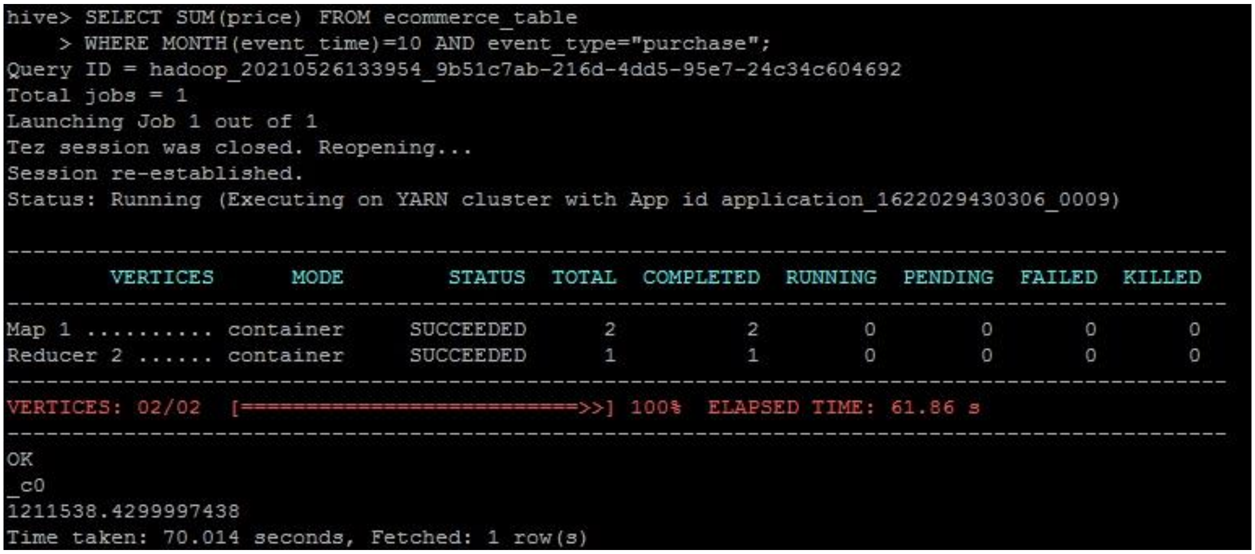
```
hive> select* from ecommerce_table order by event_time asc limit 5;
Query ID = hadoop_20210523145445_980da968-ff75-404f-b703-6fc2f9a115f9
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1621773333802_0011)

----------------------------------------------------------------------------------------------
        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: 52.80 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: 53.712 seconds, Fetched: 5 row(s)
```

DATA Analysis:

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

Sol:

```
SELECT SUM(price)
FROM   ecommerce_table
WHERE  Month(event_time) = 10
      AND event_type = 'purchase';
```

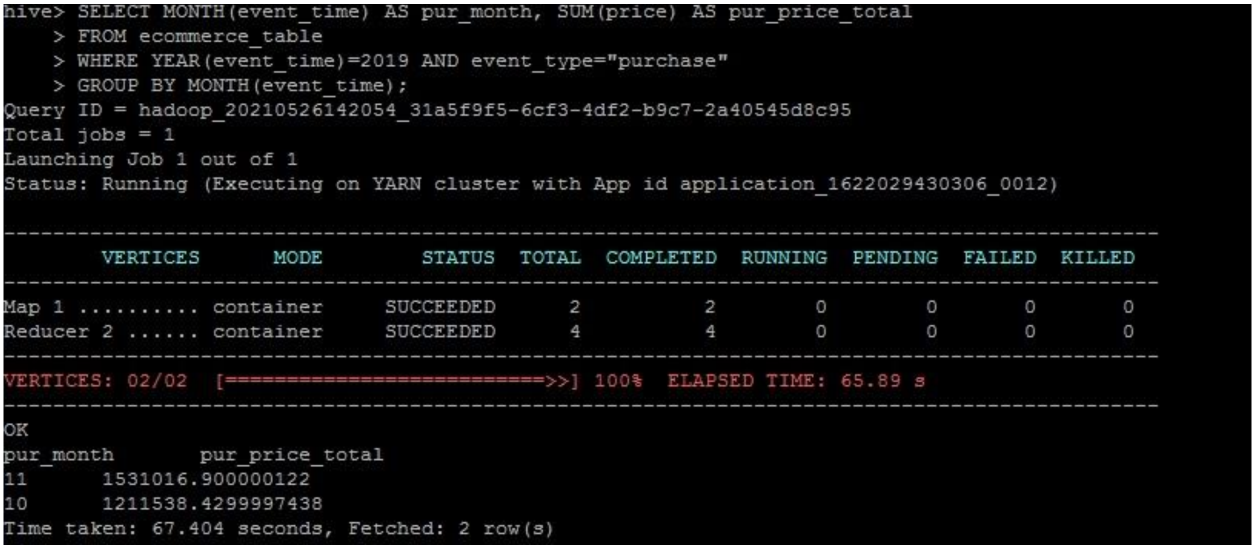


```
hive> SELECT SUM(price) FROM ecommerce_table
    > WHERE MONTH(event_time)=10 AND event_type="purchase";
Query ID = hadoop_20210526133954_9b51c7ab-216d-4dd5-95e7-24c34c604692
Total jobs = 1
Launching Job 1 out of 1
Tez session was closed. Reopening...
Session re-established.
Status: Running (Executing on YARN cluster with App id application_1622029430306_0009)

----------------------------------------------------------------------------------------------
        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: 61.86 s
----------------------------------------------------------------------------------------------
OK
_c0
1211538.4299997438
Time taken: 70.014 seconds, Fetched: 1 row(s)
```

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

Sol:

```
SELECT Month (event_time) AS pur_month,
      SUM(price) AS pur_price_total
FROM   ecommerce_table
WHERE  Year (event_time) = 2019
      AND event_type = 'purchase'
GROUP  BY Month(event_time);
```

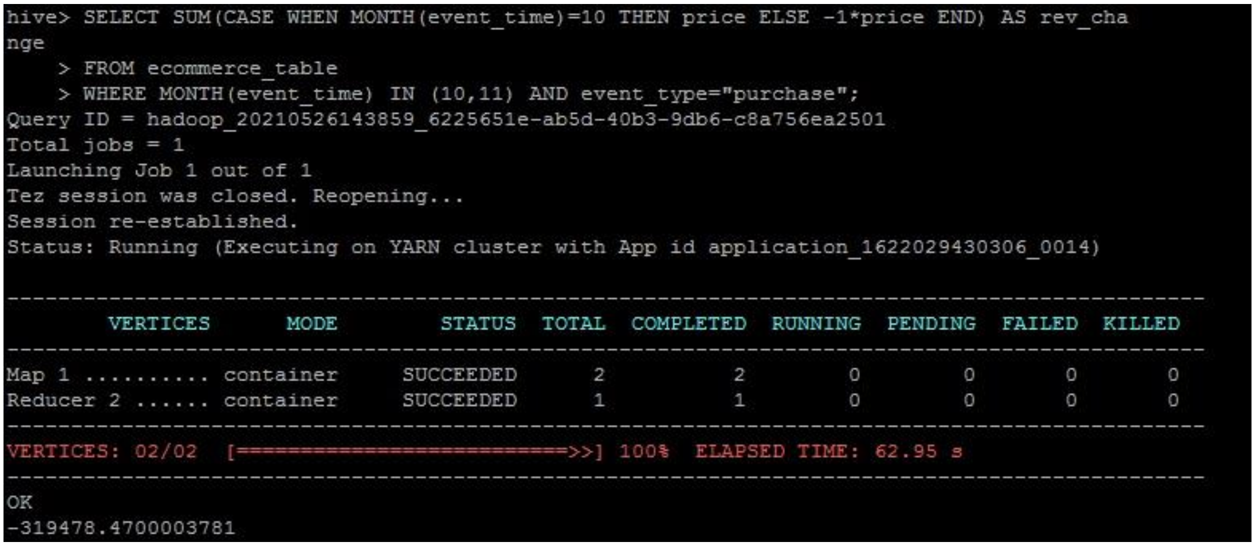


```
hive> SELECT MONTH(event_time) AS pur_month, SUM(price) AS pur_price_total
    > FROM ecommerce_table
    > WHERE YEAR(event_time)=2019 AND event_type="purchase"
    > GROUP BY MONTH(event_time);
Query ID = hadoop_20210526142054_31a5f9f5-6cf3-4df2-b9c7-2a40545d8c95
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1622029430306_0012)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      2          2        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      4          4        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 65.89 s
----------------------------------------------------------------------------------------------
OK
pur_month       pur_price_total
11      1531016.900000122
10      1211538.4299997438
Time taken: 67.404 seconds, Fetched: 2 row(s)
```

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

Sol:

```
SELECT SUM (CASE
            WHEN Month(event_time) = 10 THEN price
            ELSE -1 * price
          END) AS rev_change
FROM   ecommerce_table
WHERE  Month(event_time) IN ( 10, 11 )
      AND event_type = 'purchase';
```

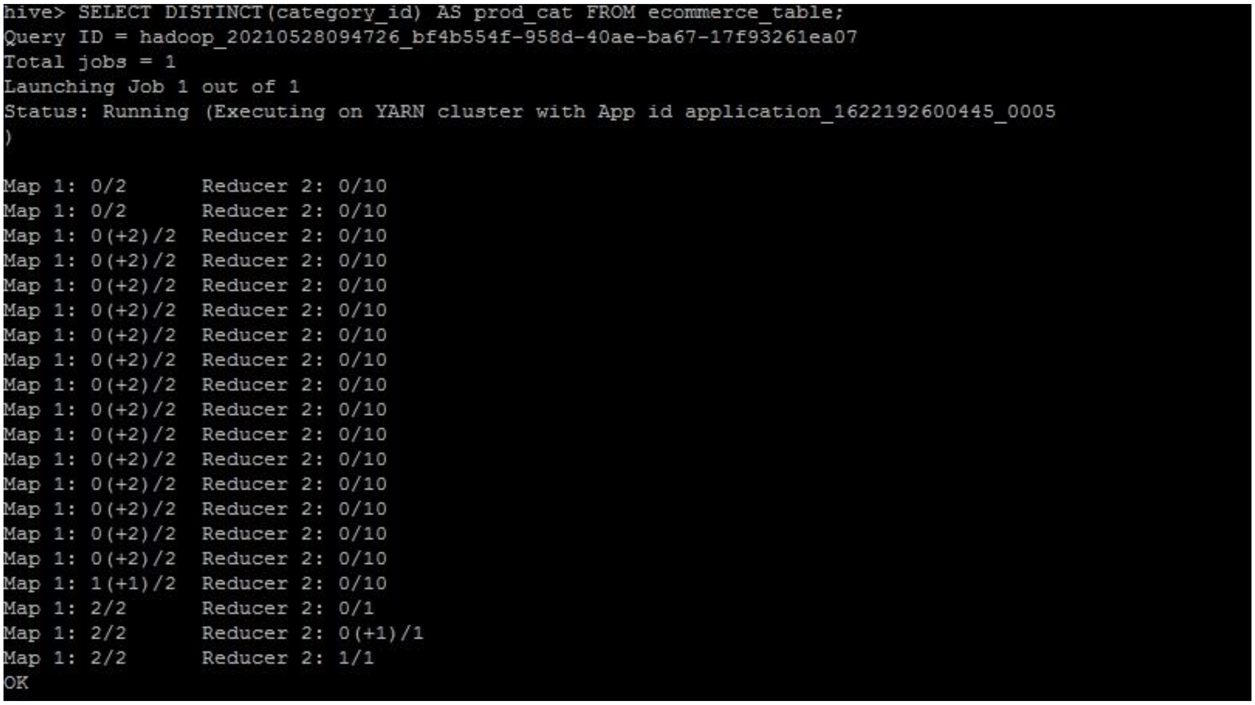


```
hive> SELECT SUM(CASE WHEN MONTH(event_time)=10 THEN price ELSE -1*price END) AS rev_cha
nge
    > FROM ecommerce_table
    > WHERE MONTH(event_time) IN (10,11) AND event_type="purchase";
Query ID = hadoop_20210526143859_6225651e-ab5d-40b3-9db6-c8a756ea2501
Total jobs = 1
Launching Job 1 out of 1
Tez session was closed. Reopening...
Session re-established.
Status: Running (Executing on YARN cluster with App id application_1622029430306_0014)

----------------------------------------------------------------------------------------------
        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: 62.95 s
----------------------------------------------------------------------------------------------
OK
-319478.4700003781
```

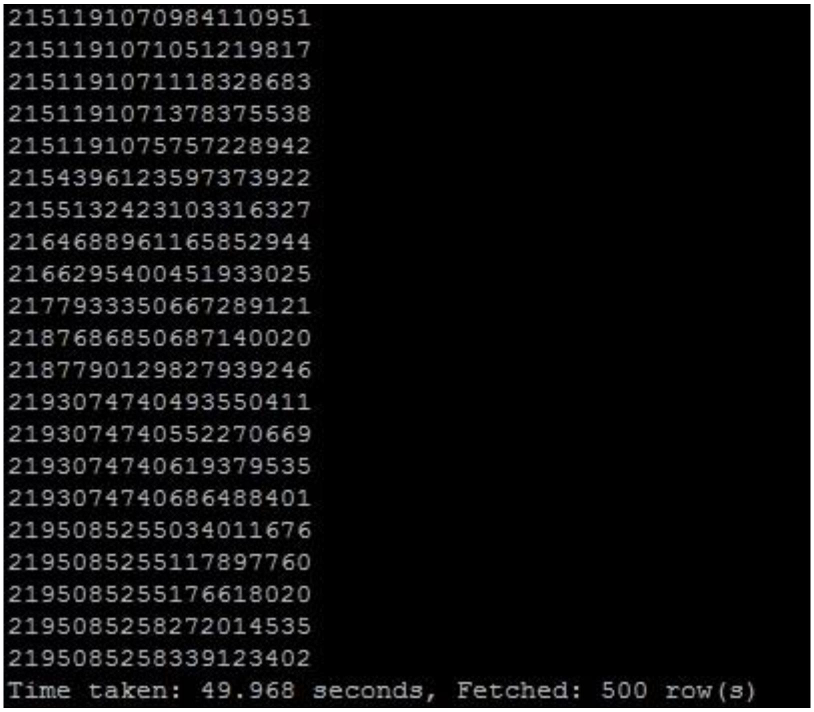
Q-4: Find distinct categories of products. Categories with null category code can be ignored.

Sol:

```
SELECT DISTINCT category_id AS prod_cat
FROM   ecommerce_table;
```



```
hive> SELECT DISTINCT(category_id) AS prod_cat FROM ecommerce_table;
Query ID = hadoop_20210528094726_bf4b554f-958d-40ae-ba67-17f93261ea07
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1622192600445_0005
)

Map 1: 0/2      Reducer 2: 0/10
Map 1: 0/2      Reducer 2: 0/10
Map 1: 0(+2)/2  Reducer 2: 0/10
Map 1: 0(+2)/2  Reducer 2: 0/10
Map 1: 0(+2)/2  Reducer 2: 0/10
Map 1: 0(+2)/2  Reducer 2: 0/10
Map 1: 0(+2)/2  Reducer 2: 0/10
Map 1: 0(+2)/2  Reducer 2: 0/10
Map 1: 0(+2)/2  Reducer 2: 0/10
Map 1: 0(+2)/2  Reducer 2: 0/10
Map 1: 0(+2)/2  Reducer 2: 0/10
Map 1: 0(+2)/2  Reducer 2: 0/10
Map 1: 0(+2)/2  Reducer 2: 0/10
Map 1: 0(+2)/2  Reducer 2: 0/10
Map 1: 0(+2)/2  Reducer 2: 0/10
Map 1: 0(+2)/2  Reducer 2: 0/10
Map 1: 1(+1)/2  Reducer 2: 0/10
Map 1: 2/2      Reducer 2: 0/1
Map 1: 2/2      Reducer 2: 0(+1)/1
Map 1: 2/2      Reducer 2: 1/1
OK
```

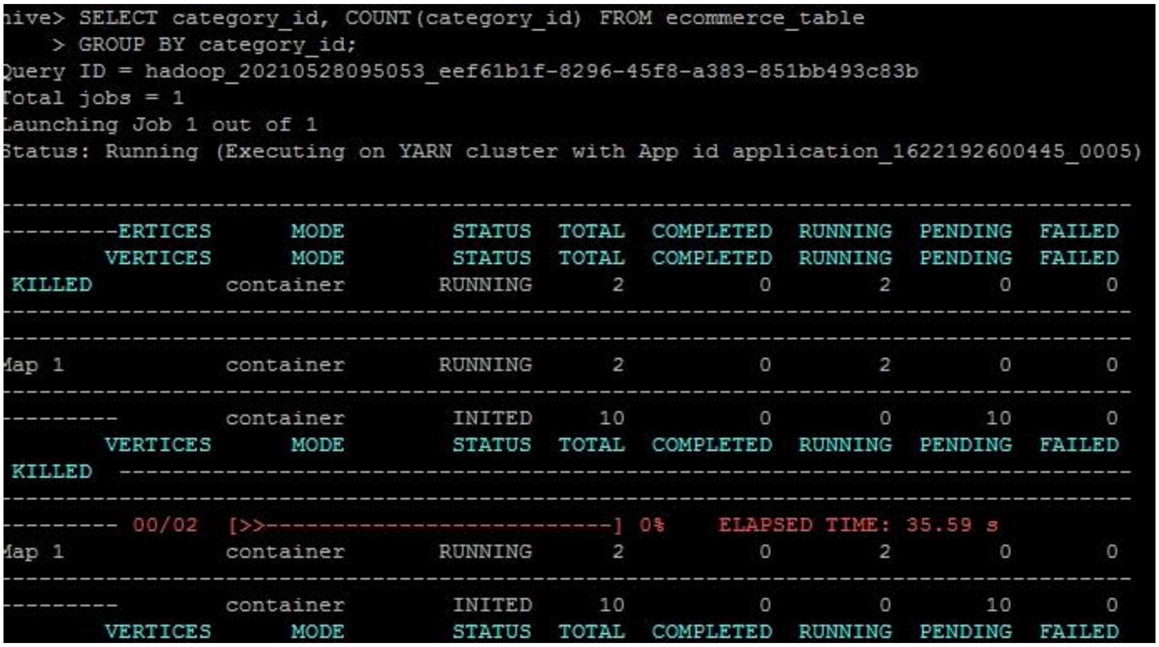


```
2151191070984110951
2151191071051219817
2151191071118328683
2151191071378375538
2151191075757228942
2154396123597373922
2155132423103316327
2164688961165852944
2166295400451933025
2177933350667289121
2187686850687140020
2187790129827939246
2193074740493550411
2193074740552270669
2193074740619379535
2193074740686488401
2195085255034011676
2195085255117897760
2195085255176618020
2195085258272014535
2195085258339123402
Time taken: 49.968 seconds, Fetched: 500 row(s)
```

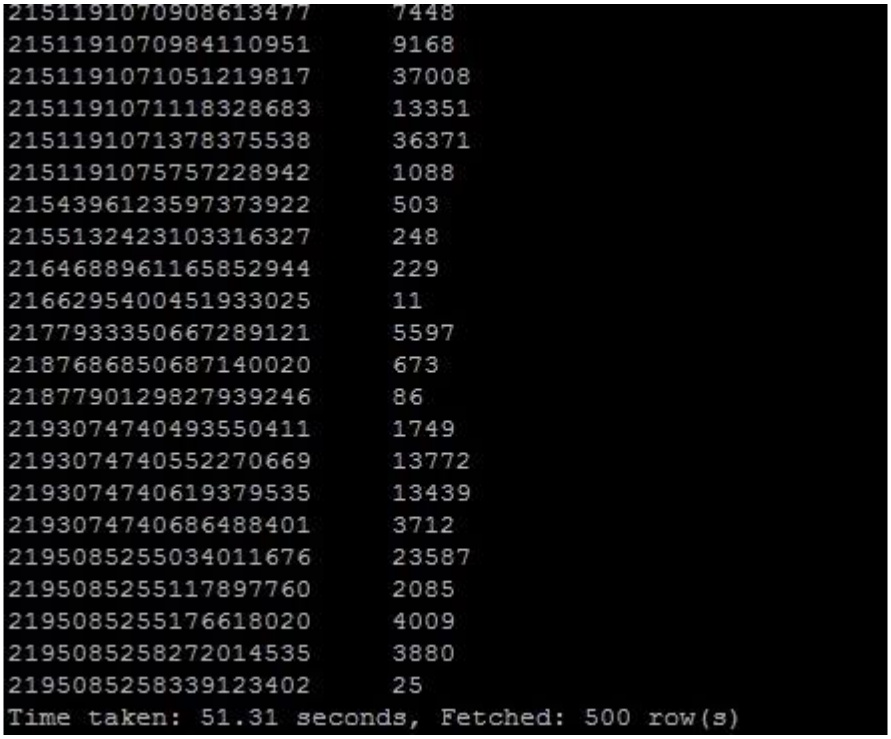
Q-5: Find the total number of products available under each category.

Sol:

```
SELECT category_id,
      COUNT(category_id)
FROM   ecommerce_table
GROUP  BY category_id;
```



```
hive> SELECT category_id, COUNT(category_id) FROM ecommerce_table
    > GROUP BY category_id;
Query ID = hadoop_20210528095053_eef61b1f-8296-45f8-a383-851bb493c83b
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1622192600445_0005)

----------------------------------------------------------------------------------------------
---------ERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED
 KILLED         container       RUNNING      2          0        2        0       0
----------------------------------------------------------------------------------------------
----------------------------------------------------------------------------------------------
Map 1           container       RUNNING      2          0        2        0       0
----------------------------------------------------------------------------------------------
---------       container        INITED     10          0        0       10       0
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED
 KILLED  -------------------------------------------------------------------------------------
----------------------------------------------------------------------------------------------
--------- 00/02  [>>----------------------------] 0%    ELAPSED TIME: 35.59 s
Map 1           container       RUNNING      2          0        2        0       0
----------------------------------------------------------------------------------------------
---------       container        INITED     10          0        0       10       0
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED
```



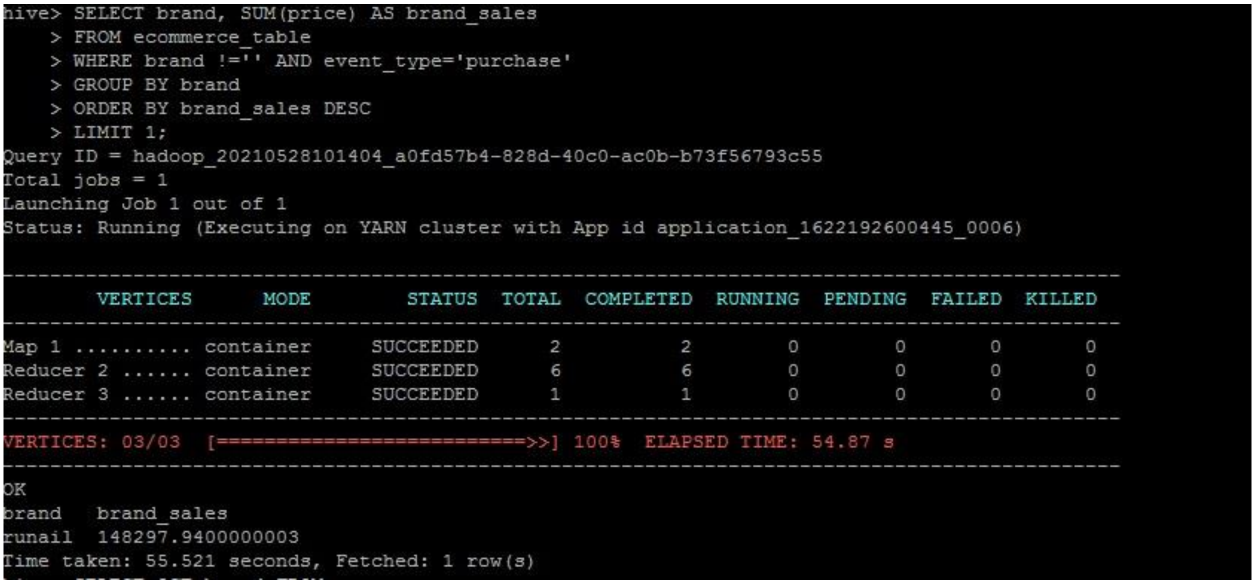
```
2151191070908613477     7448
2151191070984110951     9168
2151191071051219817     37008
2151191071118328683     13351
2151191071378375538     36371
2151191075757228942     1088
2154396123597373922     503
2155132423103316327     248
2164688961165852944     229
2166295400451933025     11
2177933350667289121     5597
2187686850687140020     673
2187790129827939246     86
2193074740493550411     1749
2193074740552270669     13772
2193074740619379535     13439
2193074740686488401     3712
2195085255034011676     23587
2195085255117897760     2085
2195085255176618020     4009
2195085258272014535     3880
2195085258339123402     25
Time taken: 51.31 seconds, Fetched: 500 row(s)
```

Gurjas Singh   Aishwarya Gyanjyoti

Q-6: Which brand had the maximum sales in October and November combined?

Sol:

```
SELECT brand,
       SUM (price) AS brand_sales
FROM   ecommerce_table
WHERE  brand != ''
       AND event_type = 'purchase'
GROUP  BY brand
ORDER  BY sales DESC
LIMIT  1;
```



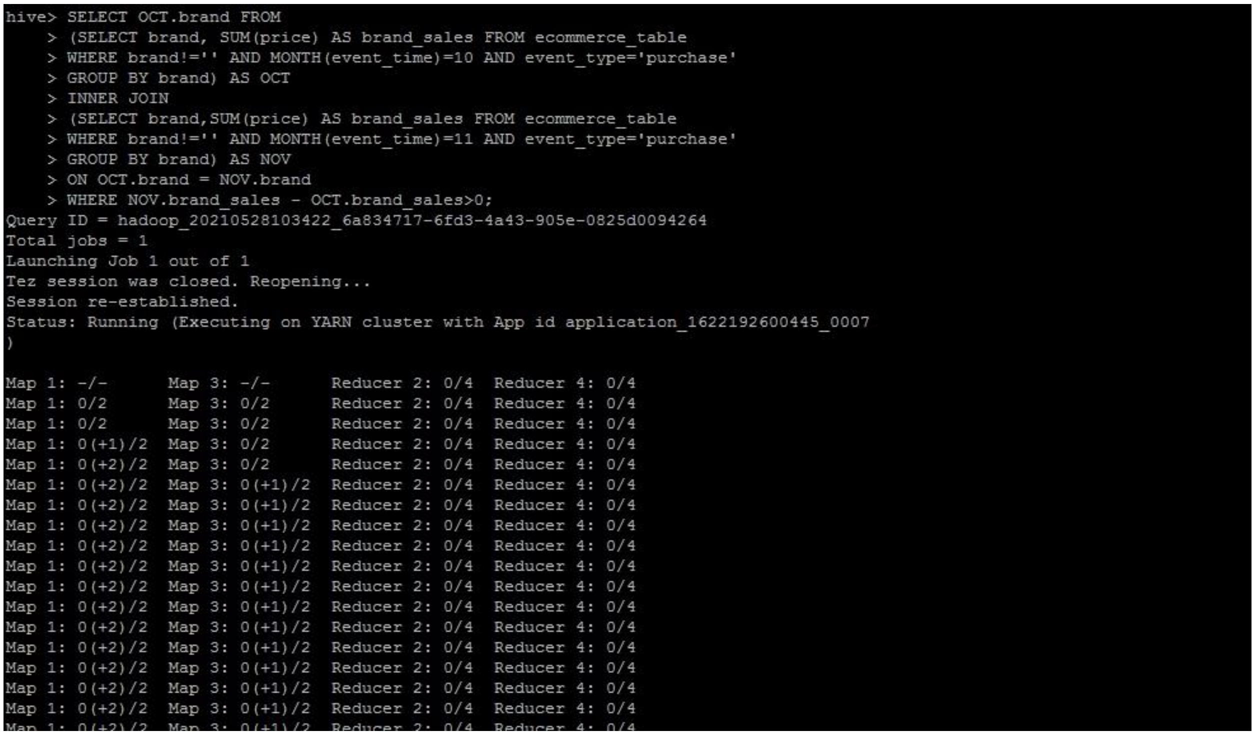
```
hive> SELECT brand, SUM(price) AS brand_sales
    > FROM ecommerce_table
    > WHERE brand !='' AND event_type='purchase'
    > GROUP BY brand
    > ORDER BY brand_sales DESC
    > LIMIT 1;
Query ID = hadoop_20210528101404_a0fd57b4-828d-40c0-ac0b-b73f56793c55
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1622192600445_0006)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      2          2        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: 54.87 s
----------------------------------------------------------------------------------------------
OK
brand   brand_sales
runail  148297.9400000003
Time taken: 55.521 seconds, Fetched: 1 row(s)
```

Q-7: Which brands increased their sales from October to November?

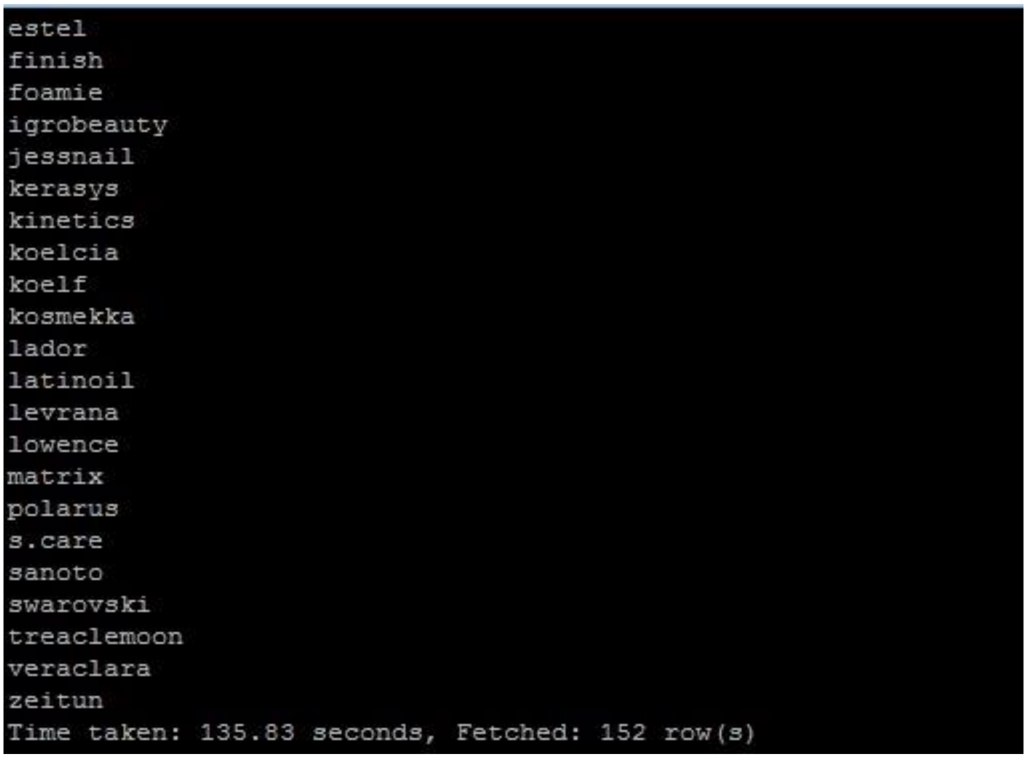
Sol:

```
SELECT OCT. brand
FROM   (SELECT brand,
               SUM(price) AS brand_sales
        FROM   ecommerce_table
        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_table
                   WHERE  brand != ''
                          AND event_type = 'purchase'
                          AND Month(event_time) = 11
```

```
              GROUP  BY brand) AS Nov
           ON Nov.brand = Oct.brand
WHERE  Nov.brand_sales - Oct.brand_sales > 0;
```



```
hive> SELECT OCT.brand FROM
    > (SELECT brand, SUM(price) AS brand_sales FROM ecommerce_table
    > 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_table
    > 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_20210528103422_6a834717-6fd3-4a43-905e-0825d0094264
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_1622192600445_0007
)

Map 1: -/-      Map 3: -/-      Reducer 2: 0/4  Reducer 4: 0/4
Map 1: 0/2      Map 3: 0/2      Reducer 2: 0/4  Reducer 4: 0/4
Map 1: 0/2      Map 3: 0/2      Reducer 2: 0/4  Reducer 4: 0/4
Map 1: 0(+1)/2  Map 3: 0/2      Reducer 2: 0/4  Reducer 4: 0/4
Map 1: 0(+2)/2  Map 3: 0/2      Reducer 2: 0/4  Reducer 4: 0/4
Map 1: 0(+2)/2  Map 3: 0(+1)/2  Reducer 2: 0/4  Reducer 4: 0/4
Map 1: 0(+2)/2  Map 3: 0(+1)/2  Reducer 2: 0/4  Reducer 4: 0/4
Map 1: 0(+2)/2  Map 3: 0(+1)/2  Reducer 2: 0/4  Reducer 4: 0/4
Map 1: 0(+2)/2  Map 3: 0(+1)/2  Reducer 2: 0/4  Reducer 4: 0/4
Map 1: 0(+2)/2  Map 3: 0(+1)/2  Reducer 2: 0/4  Reducer 4: 0/4
Map 1: 0(+2)/2  Map 3: 0(+1)/2  Reducer 2: 0/4  Reducer 4: 0/4
Map 1: 0(+2)/2  Map 3: 0(+1)/2  Reducer 2: 0/4  Reducer 4: 0/4
Map 1: 0(+2)/2  Map 3: 0(+1)/2  Reducer 2: 0/4  Reducer 4: 0/4
Map 1: 0(+2)/2  Map 3: 0(+1)/2  Reducer 2: 0/4  Reducer 4: 0/4
Map 1: 0(+2)/2  Map 3: 0(+1)/2  Reducer 2: 0/4  Reducer 4: 0/4
Map 1: 0(+2)/2  Map 3: 0(+1)/2  Reducer 2: 0/4  Reducer 4: 0/4
Map 1: 0(+2)/2  Map 3: 0(+1)/2  Reducer 2: 0/4  Reducer 4: 0/4
Map 1: 0(+2)/2  Map 3: 0(+1)/2  Reducer 2: 0/4  Reducer 4: 0/4
```

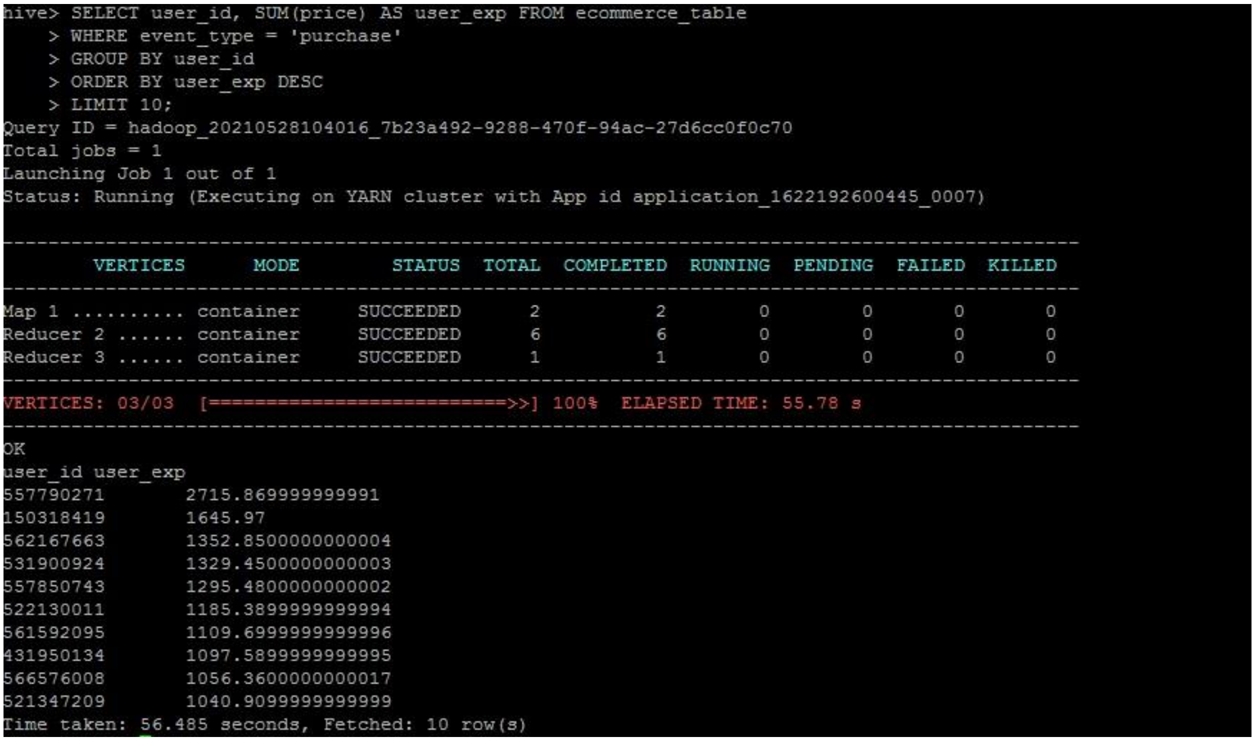


```
estel
finish
foamie
igrobeauty
jessnail
kerasys
kinetics
koelcia
koelf
kosmekka
lador
latinoil
levrana
lowence
matrix
polarus
s.care
sanoto
swarovski
treaclemoon
veraclara
zeitun
Time taken: 135.83 seconds, Fetched: 152 row(s)
```

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.

Sol:

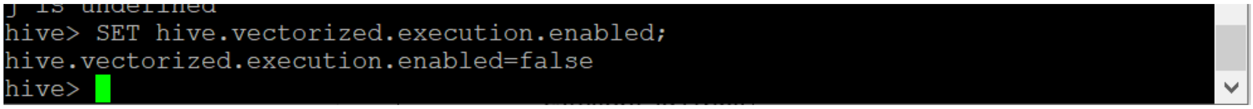
```
SELECT user_id,
      SUM(price) AS user_exp
FROM   ecommerce_table
WHERE  event_type = 'purchase'
GROUP  BY user_id
ORDER  BY user_exp DESC
LIMIT  10;
```



```
hive> SELECT user_id, SUM(price) AS user_exp FROM ecommerce_table
    > WHERE event_type = 'purchase'
    > GROUP BY user_id
    > ORDER BY user_exp DESC
    > LIMIT 10;
Query ID = hadoop_20210528104016_7b23a492-9288-470f-94ac-27d6cc0f0c70
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1622192600445_0007)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      2          2        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: 55.78 s
----------------------------------------------------------------------------------------------
OK
user_id user_exp
557790271       2715.869999999991
150318419       1645.97
562167663       1352.8500000000004
531900924       1329.4500000000003
557850743       1295.4800000000002
522130011       1185.3899999999994
561592095       1109.6999999999996
431950134       1097.5899999999995
566576008       1056.3600000000017
521347209       1040.9099999999999
Time taken: 56.485 seconds, Fetched: 10 row(s)
```

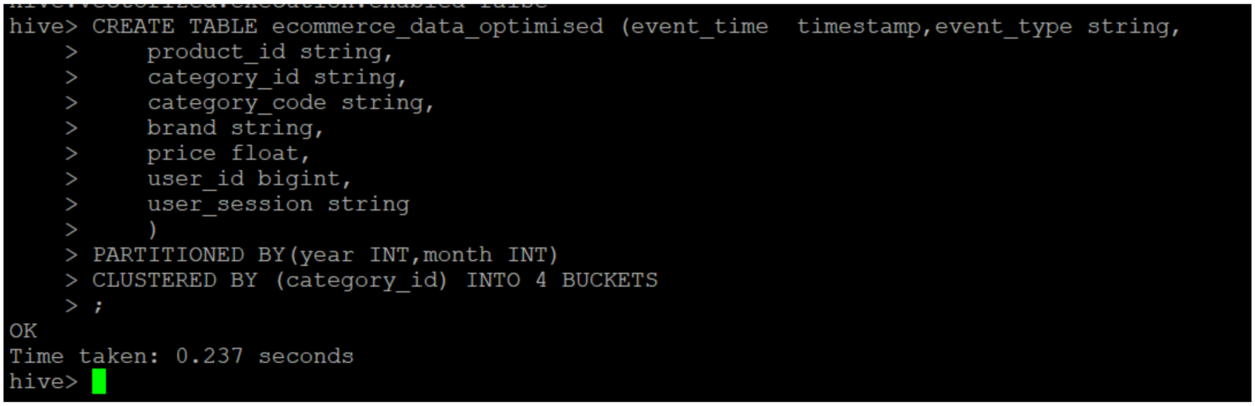
Query Optimization:

1. **SET** hive.vectorised.execution.enabled;



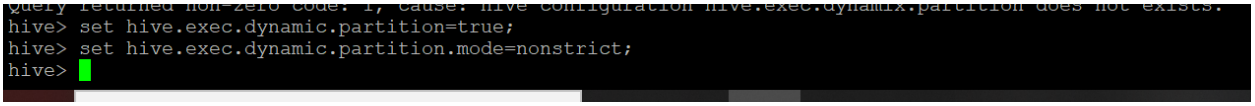
```
j is undefined
hive> SET hive.vectorized.execution.enabled;
hive.vectorized.execution.enabled=false
hive> 
```

2. Create table **'ecommerce_data_optimised'** with **partioning** & **buckets** into **4 buckets**



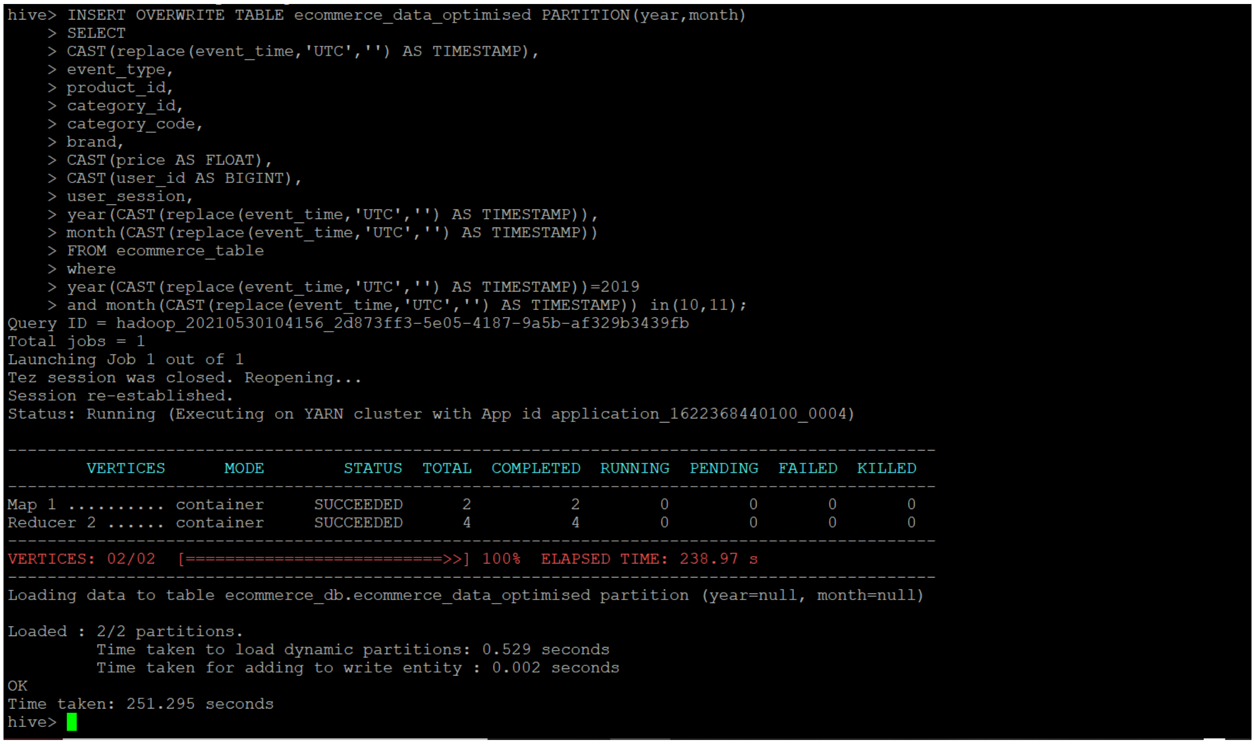
```
hive> CREATE TABLE ecommerce_data_optimised (event_time  timestamp,event_type string,
    >     product_id string,
    >     category_id string,
    >     category_code string,
    >     brand string,
    >     price float,
    >     user_id bigint,
    >     user_session string
    >     )
    > PARTITIONED BY(year INT,month INT)
    > CLUSTERED BY (category_id) INTO 4 BUCKETS
    > ;
OK
Time taken: 0.237 seconds
hive>
```

3. SET hive.exec.dynamic.partition=true;
   SET hive.exec.dynamic.partition.mode=nonstrict;



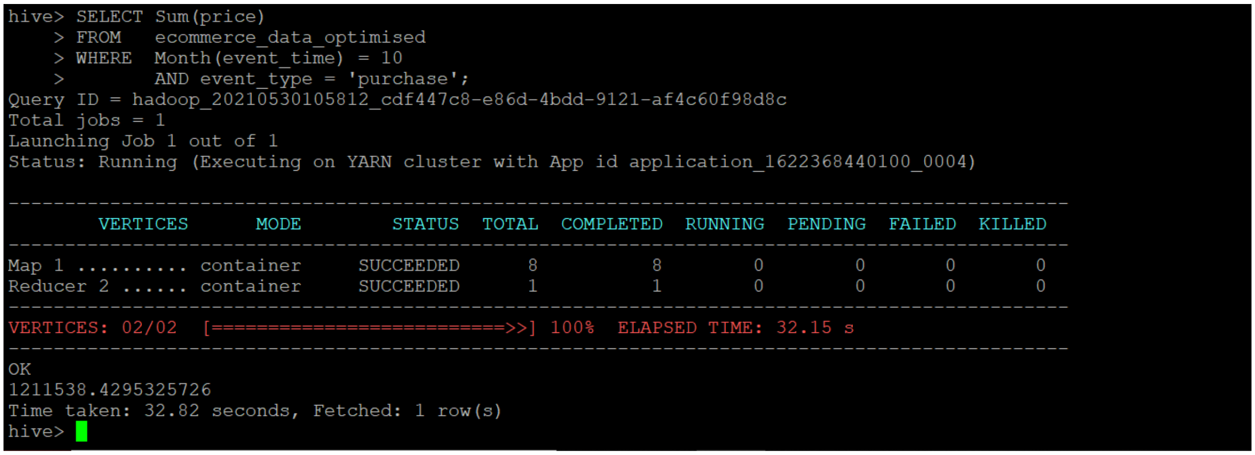
```
Query returned non-zero code: 1, cause: hive configuration hive.exec.dynamix.partition does not exists.
hive> set hive.exec.dynamic.partition=true;
hive> set hive.exec.dynamic.partition.mode=nonstrict;
hive>
```

4. Insert data into **'ecommerce_data_optimised'** table (**Optimized table**)



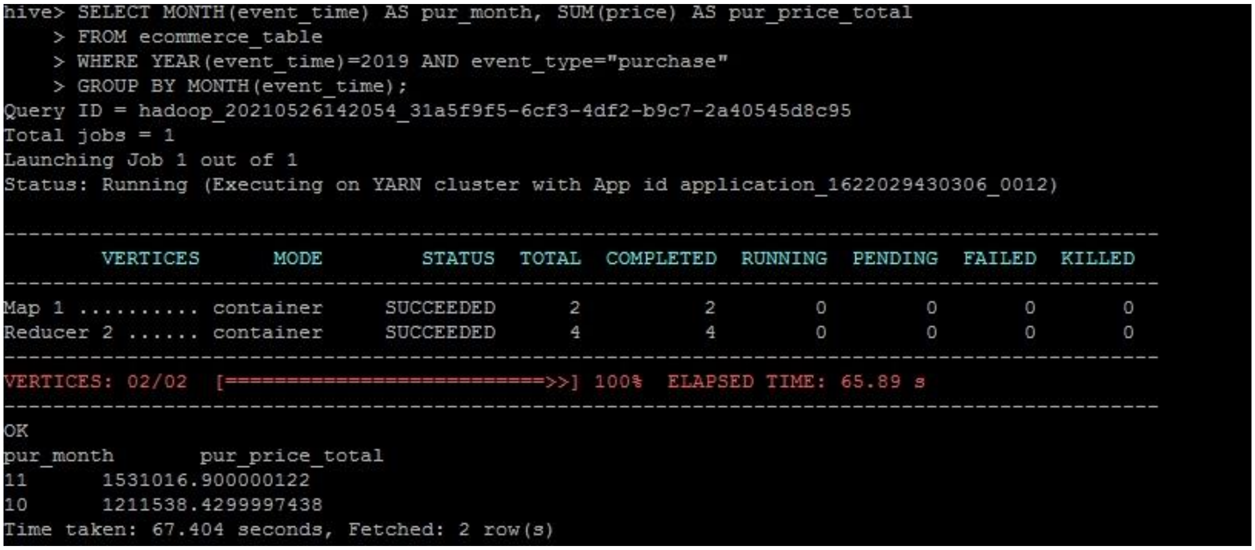
```
hive> INSERT OVERWRITE TABLE ecommerce_data_optimised PARTITION(year,month)
    > SELECT
    > CAST(replace(event_time,'UTC','') AS TIMESTAMP),
    > event_type,
    > product_id,
    > category_id,
    > category_code,
    > brand,
    > CAST(price AS FLOAT),
    > CAST(user_id AS BIGINT),
    > user_session,
    > year(CAST(replace(event_time,'UTC','') AS TIMESTAMP)),
    > month(CAST(replace(event_time,'UTC','') AS TIMESTAMP))
    > FROM ecommerce_table
    > 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_20210530104156_2d873ff3-5e05-4187-9a5b-af329b3439fb
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_1622368440100_0004)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      2          2        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      4          4        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 238.97 s
----------------------------------------------------------------------------------------------
Loading data to table ecommerce_db.ecommerce_data_optimised partition (year=null, month=null)

Loaded : 2/2 partitions.
         Time taken to load dynamic partitions: 0.529 seconds
         Time taken for adding to write entity : 0.002 seconds
OK
Time taken: 251.295 seconds
hive>
```

5. Solved **Q-1** after optimizing in **32.82 seconds**:



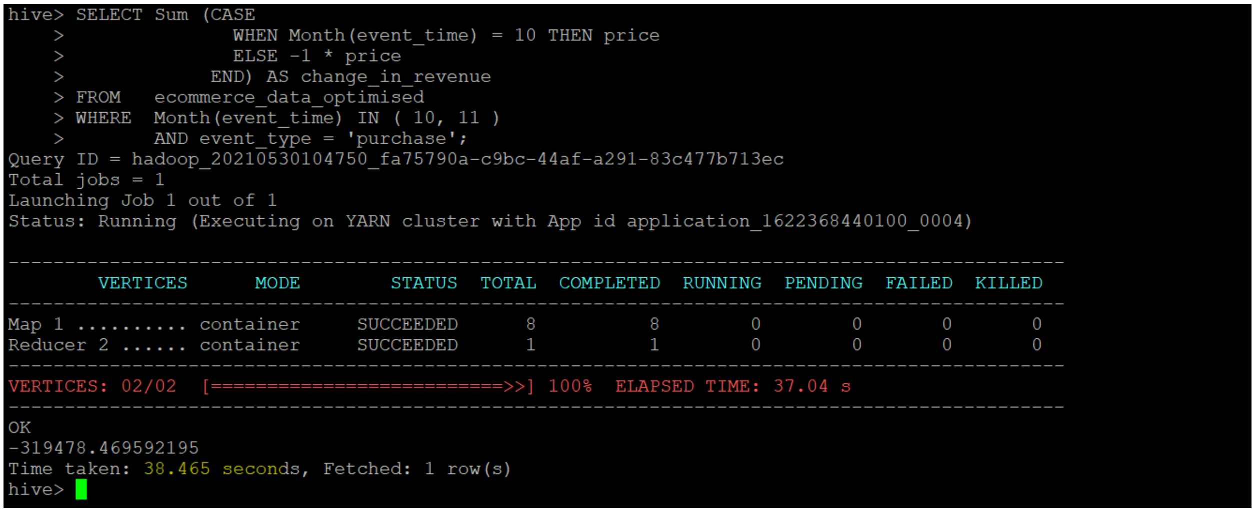
```
hive> SELECT Sum(price)
    > FROM   ecommerce_data_optimised
    > WHERE  Month(event_time) = 10
    >        AND event_type = 'purchase';
Query ID = hadoop_20210530105812_cdf447c8-e86d-4bdd-9121-af4c60f98d8c
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1622368440100_0004)

----------------------------------------------------------------------------------------------
        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: 32.15 s
----------------------------------------------------------------------------------------------
OK
1211538.4295325726
Time taken: 32.82 seconds, Fetched: 1 row(s)
hive> 
```

Solved **Q-2** without optimization in **67.40 seconds**:



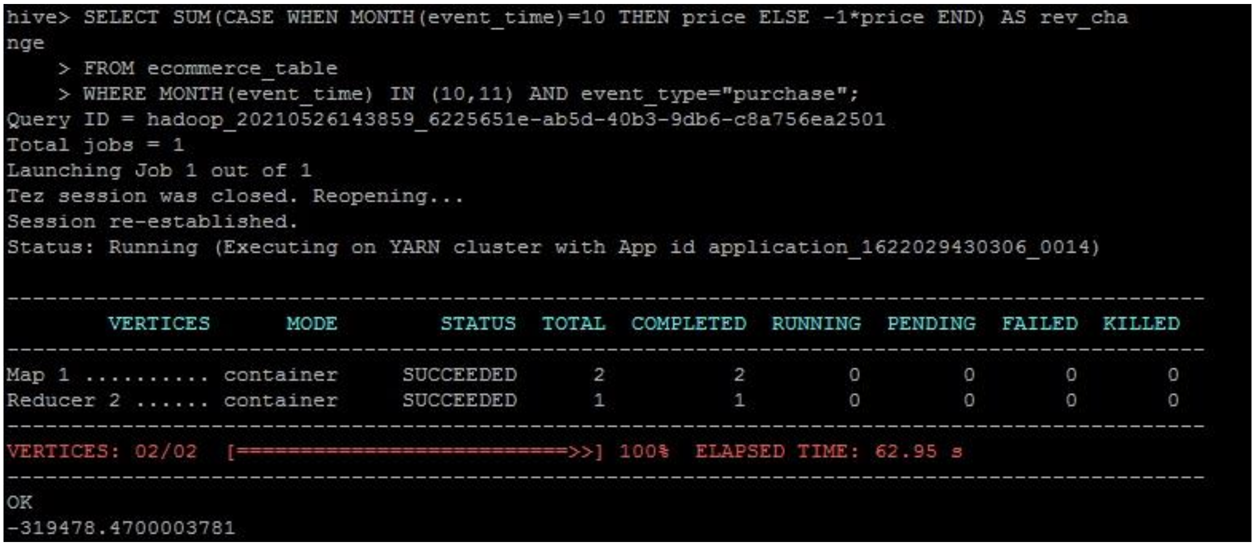
```
hive> SELECT MONTH(event_time) AS pur_month, SUM(price) AS pur_price_total
    > FROM ecommerce_table
    > WHERE YEAR(event_time)=2019 AND event_type="purchase"
    > GROUP BY MONTH(event_time);
Query ID = hadoop_20210526142054_31a5f9f5-6cf3-4df2-b9c7-2a40545d8c95
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1622029430306_0012)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      2          2        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      4          4        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 65.89 s
----------------------------------------------------------------------------------------------
OK
pur_month       pur_price_total
11      1531016.900000122
10      1211538.4299997438
Time taken: 67.404 seconds, Fetched: 2 row(s)
```

6. Solved **Q-3** after optimization in **38.465 seconds**:



```
hive> SELECT Sum (CASE
    >               WHEN Month(event_time) = 10 THEN price
    >               ELSE -1 * price
    >             END) AS change_in_revenue
    > FROM   ecommerce_data_optimised
    > WHERE  Month(event_time) IN ( 10, 11 )
    >        AND event_type = 'purchase';
Query ID = hadoop_20210530104750_fa75790a-c9bc-44af-a291-83c477b713ec
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1622368440100_0004)

----------------------------------------------------------------------------------------------
        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: 37.04 s
----------------------------------------------------------------------------------------------
OK
-319478.469592195
Time taken: 38.465 seconds, Fetched: 1 row(s)
hive> 
```

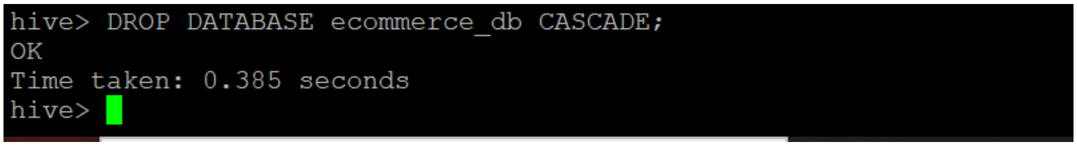
Solved **Q-3** without optimization in **62.95 seconds**:



```
hive> SELECT SUM(CASE WHEN MONTH(event_time)=10 THEN price ELSE -1*price END) AS rev_cha
nge
    > FROM ecommerce_table
    > WHERE MONTH(event_time) IN (10,11) AND event_type="purchase";
Query ID = hadoop_20210526143859_6225651e-ab5d-40b3-9db6-c8a756ea2501
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_1622029430306_0014)

----------------------------------------------------------------------------------------------
        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: 62.95 s
----------------------------------------------------------------------------------------------
OK
-319478.4700003781
```

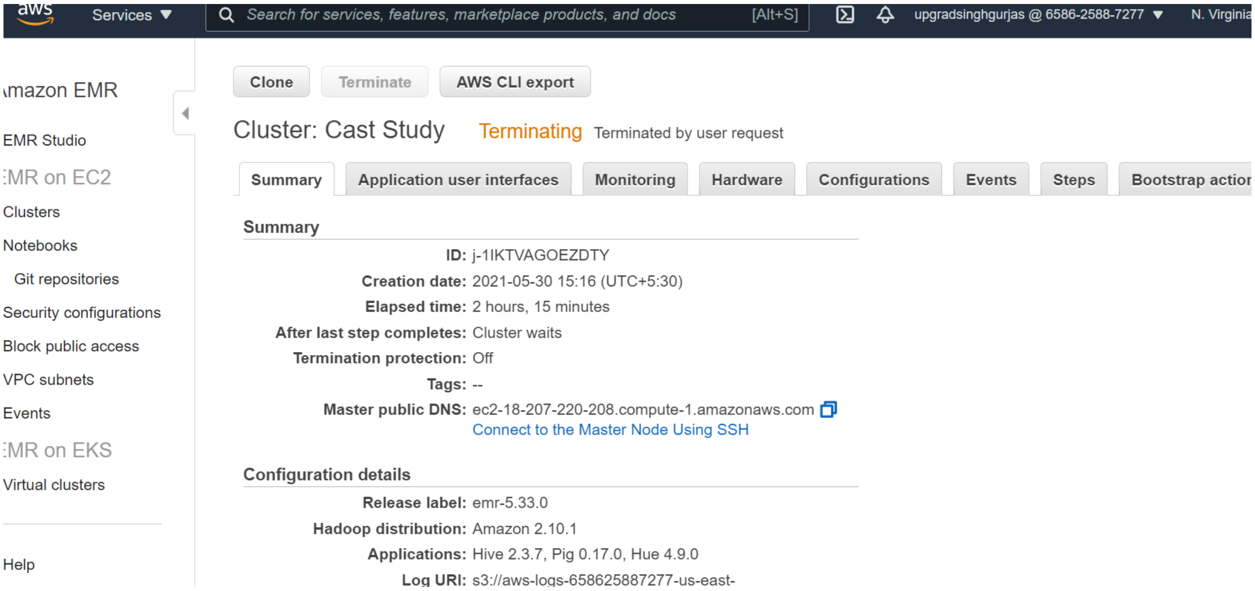
Cleaning Up:

1. Drop the database



```
hive> DROP DATABASE ecommerce_db CASCADE;
OK
Time taken: 0.385 seconds
hive>
```

2. Terminate the cluster



Cluster: Cast Study Terminating Terminated by user request

Summary

ID: j-1IKTVAGOEZDTY
Creation date: 2021-05-30 15:16 (UTC+5:30)
Elapsed time: 2 hours, 15 minutes
After last step completes: Cluster waits
Termination protection: Off
Tags: --
Master public DNS: ec2-18-207-220-208.compute-1.amazonaws.com
Connect to the Master Node Using SSH

Configuration details

Release label: emr-5.33.0
Hadoop distribution: Amazon 2.10.1
Applications: Hive 2.3.7, Pig 0.17.0, Hue 4.9.0
Log URI: s3://aws-logs-658625887277-us-east-