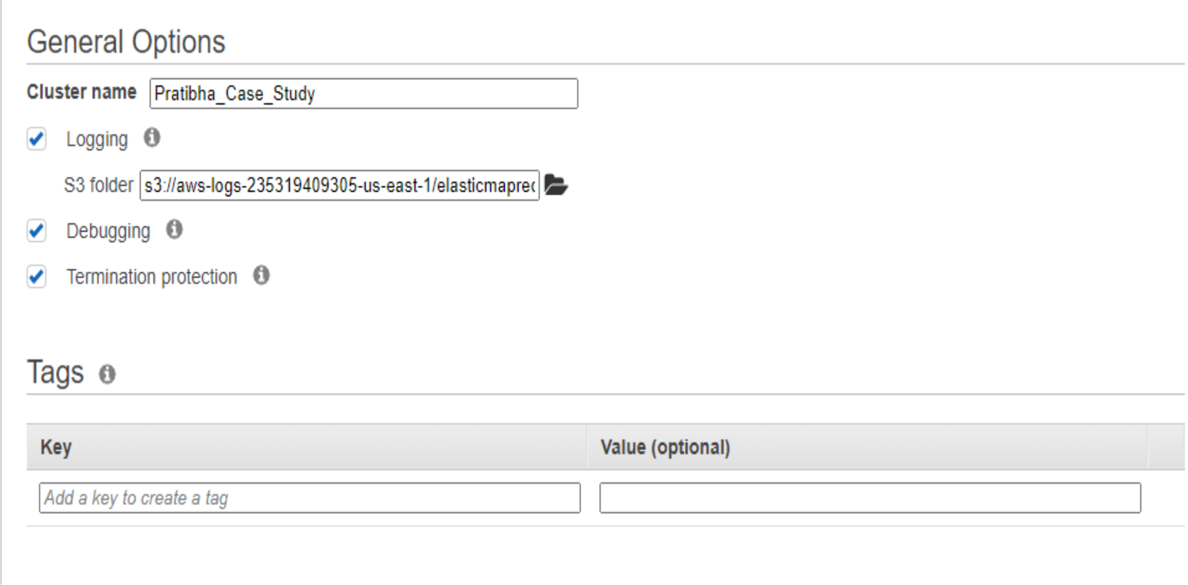
HIVE CASE STUDY

E-COMMERCE EVENT

1. Copying the data set into the HDFS:

A) Launch an EMR cluster that utilizes the Hive services



General Options

Cluster name: Pratibha_Case_Study

☑ Logging

S3 folder: s3://aws-logs-235319409305-us-east-1/elasticmapre(

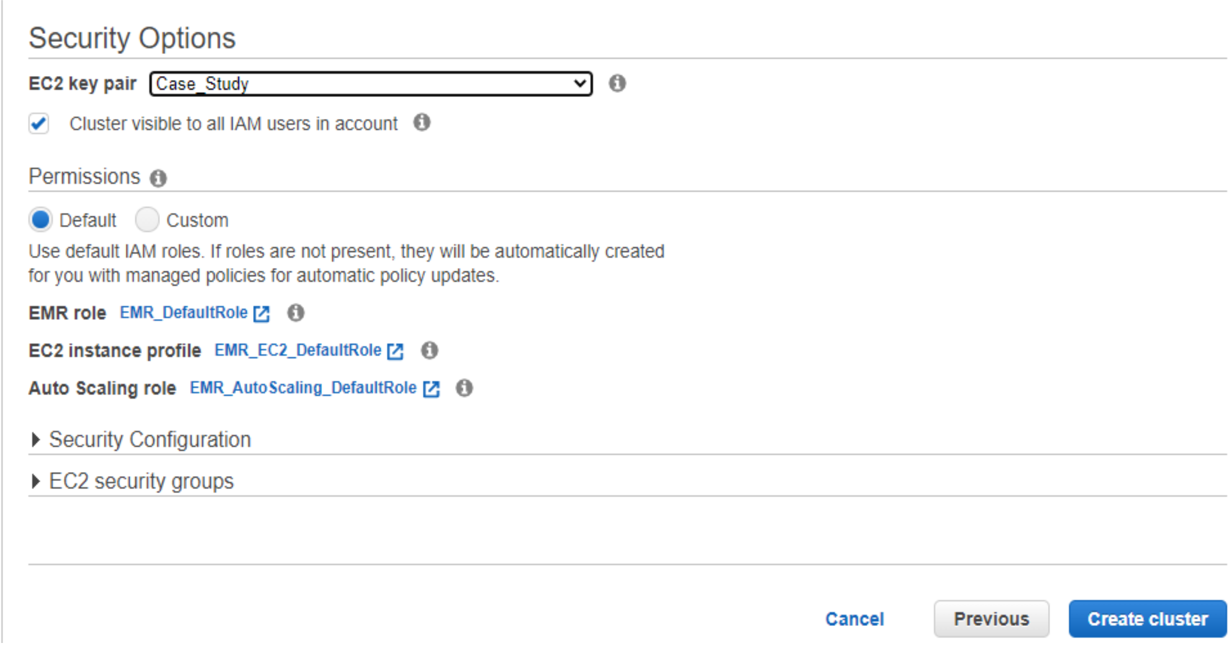
☑ Debugging

☑ Termination protection

Tags

Key	Value (optional)
Add a key to create a tag	

2.



Security Options

EC2 key pair: Case_Study

☑ Cluster visible to all IAM users in account

Permissions

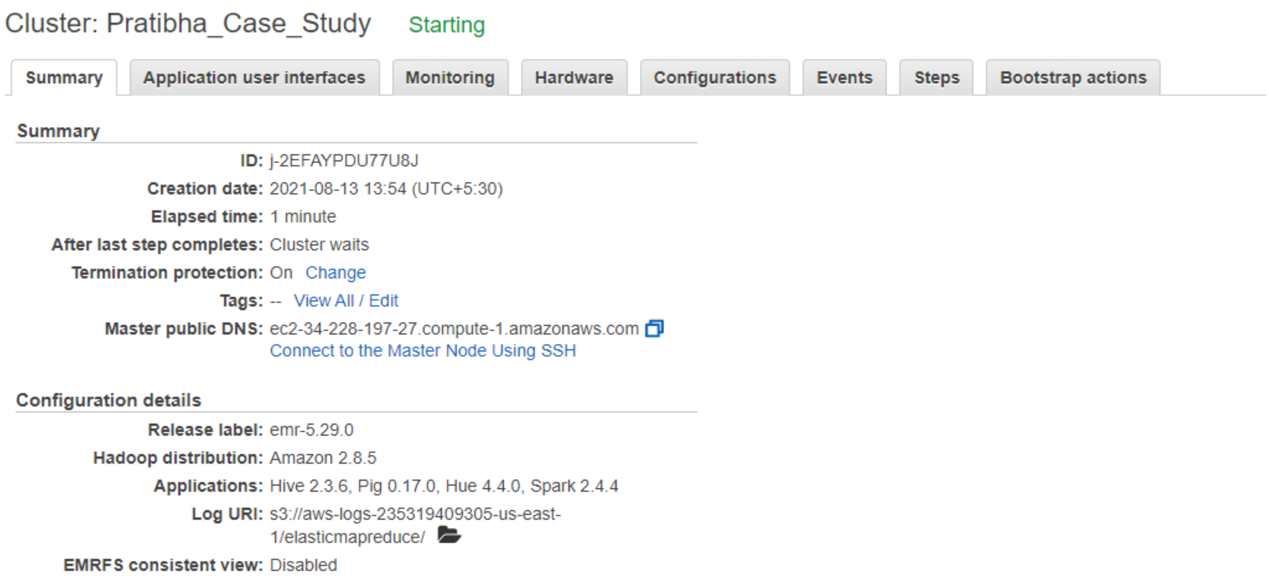
● Default ○ Custom

Use default IAM roles. If roles are not present, they will be automatically created for you with managed policies for automatic policy updates.

EMR role: EMR_DefaultRole

EC2 instance profile: EMR_EC2_DefaultRole

Auto Scaling role: EMR_AutoScaling_DefaultRole

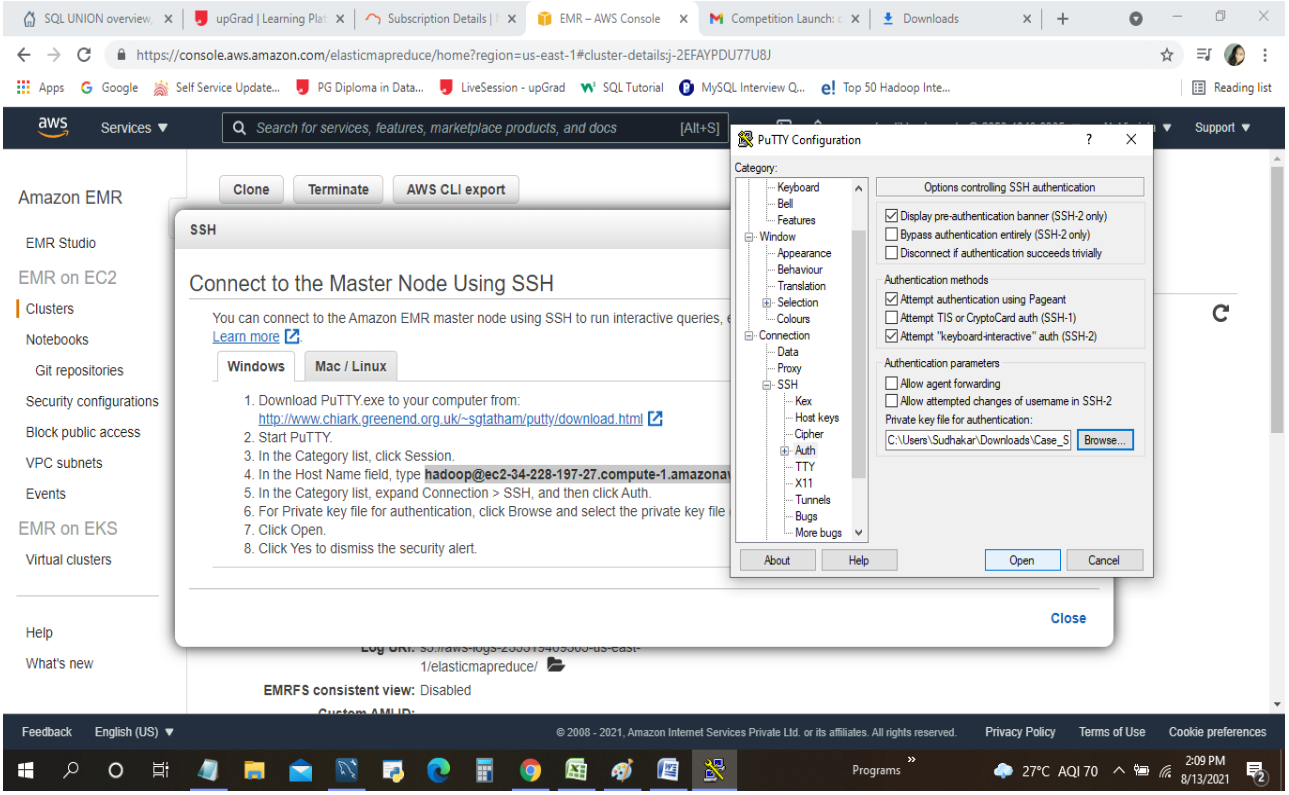

Cluster: Pratibha_Case_Study Starting

Summary

ID: j-2EFAYPDU77U8J
Creation date: 2021-08-13 13:54 (UTC+5:30)
Elapsed time: 1 minute
After last step completes: Cluster waits
Termination protection: On Change
Tags: -- View All / Edit
Master public DNS: ec2-34-228-197-27.compute-1.amazonaws.com
Connect to the Master Node Using SSH

Configuration details

Release label: emr-5.29.0
Hadoop distribution: Amazon 2.8.5
Applications: Hive 2.3.6, Pig 0.17.0, Hue 4.4.0, Spark 2.4.4
Log URI: s3://aws-logs-235319409305-us-east-1/elasticmapreduce/
EMRFS consistent view: Disabled



SSH

Connect to the Master Node Using SSH

You can connect to the Amazon EMR master node using SSH to run interactive queries,
Learn more

1. Download PuTTY.exe to your computer from: http://www.chiark.greenend.org.uk/~sgtatham/putty/download.html
2. Start PuTTY.
3. In the Category list, click Session.
4. In the Host Name field, type **hadoop@ec2-34-228-197-27.compute-1.amazona**
5. In the Category list, expand Connection > SSH, and then click Auth.
6. For Private key file for authentication, click Browse and select the private key file
7. Click Open.
8. Click Yes to dismiss the security alert.

PuTTY Configuration

Options controlling SSH authentication

☑ Display pre-authentication banner (SSH-2 only)
☐ Bypass authentication entirely (SSH-2 only)
☐ Disconnect if authentication succeeds trivially

Authentication methods

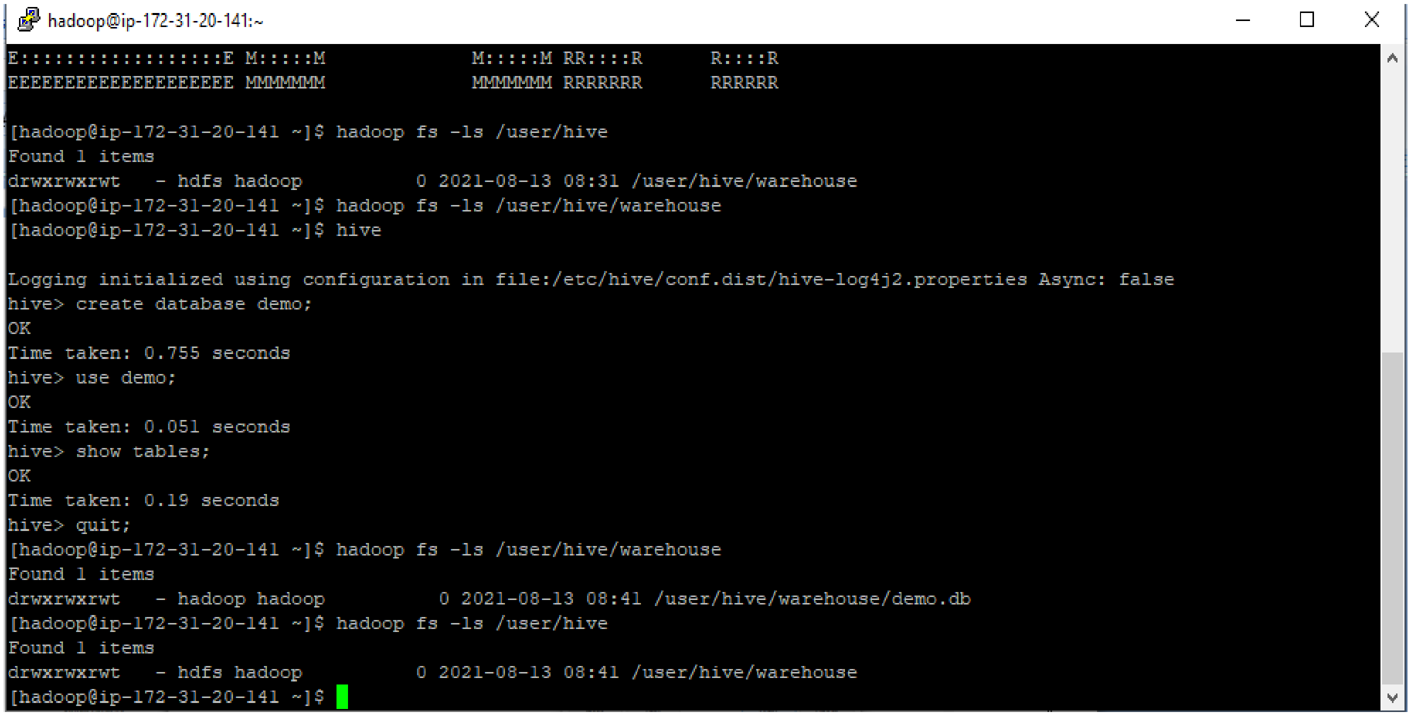
☑ Attempt authentication using Pageant
☐ Attempt TIS or CryptoCard auth (SSH-1)
☑ Attempt "keyboard-interactive" auth (SSH-2)

Authentication parameters

☐ Allow agent forwarding
☐ Allow attempted changes of username in SSH-2

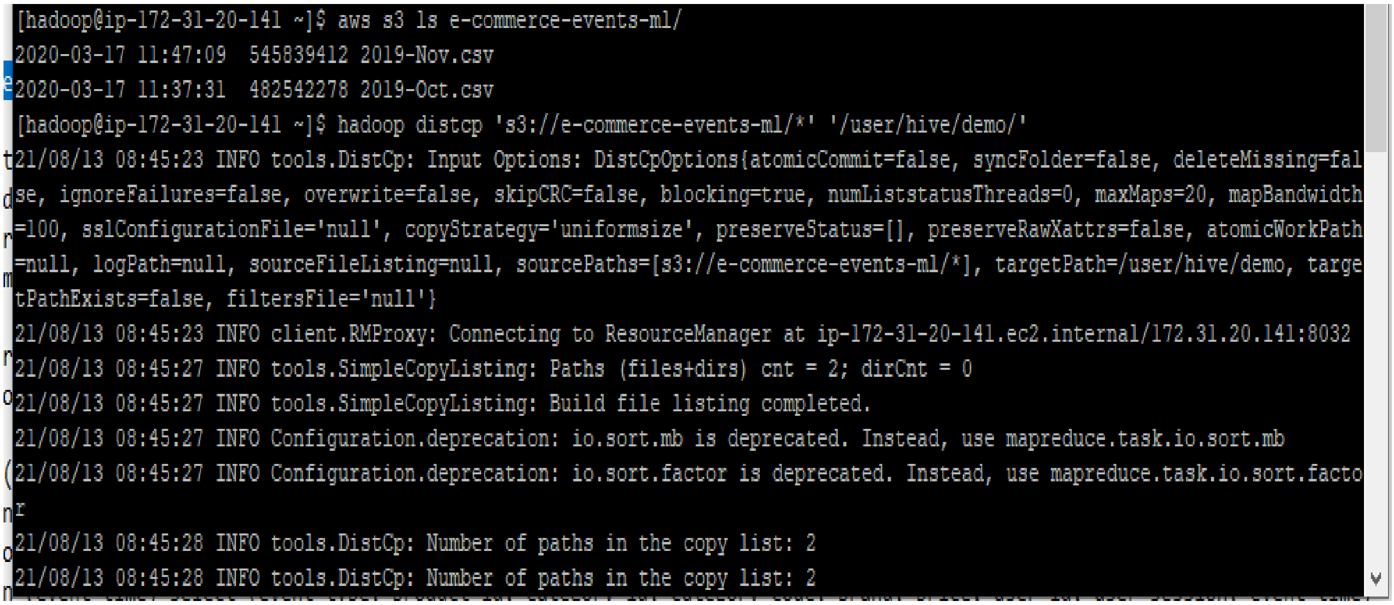
Private key file for authentication: C:\Users\Sudhakar\Downloads\Case_S

Log URI: s3://aws-logs-235319409305-us-east-1/elasticmapreduce/
EMRFS consistent view: Disabled

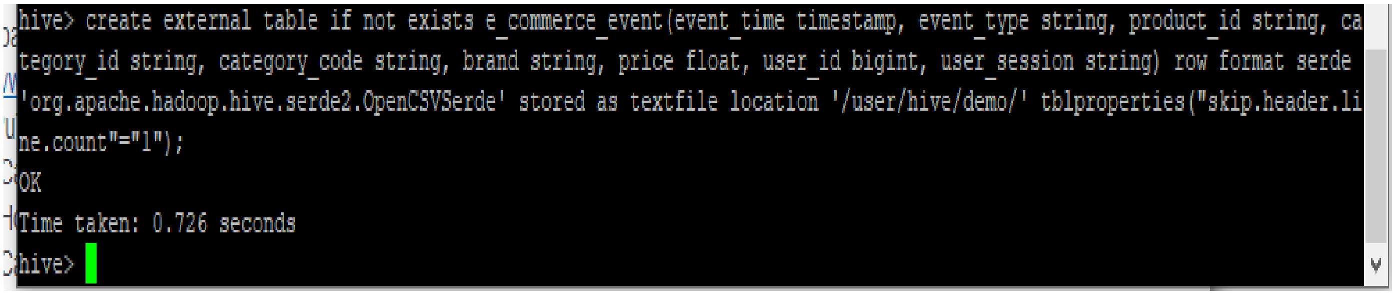


```
E::::::::::::::::::E M:::::M             M:::::M RR::::R      R::::R
EEEEEEEEEEEEEEEEEEEEE MMMMMMM             MMMMMMM RRRRRRR      RRRRRR

[hadoop@ip-172-31-20-141 ~]$ hadoop fs -ls /user/hive
Found 1 items
drwxrwxrwt   - hdfs hadoop          0 2021-08-13 08:31 /user/hive/warehouse
[hadoop@ip-172-31-20-141 ~]$ hadoop fs -ls /user/hive/warehouse
[hadoop@ip-172-31-20-141 ~]$ hive

Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j2.properties Async: false
hive> create database demo;
OK
Time taken: 0.755 seconds
hive> use demo;
OK
Time taken: 0.051 seconds
hive> show tables;
OK
Time taken: 0.19 seconds
hive> quit;
[hadoop@ip-172-31-20-141 ~]$ hadoop fs -ls /user/hive/warehouse
Found 1 items
drwxrwxrwt   - hadoop hadoop          0 2021-08-13 08:41 /user/hive/warehouse/demo.db
[hadoop@ip-172-31-20-141 ~]$ hadoop fs -ls /user/hive
Found 1 items
drwxrwxrwt   - hdfs hadoop          0 2021-08-13 08:41 /user/hive/warehouse
[hadoop@ip-172-31-20-141 ~]$
```

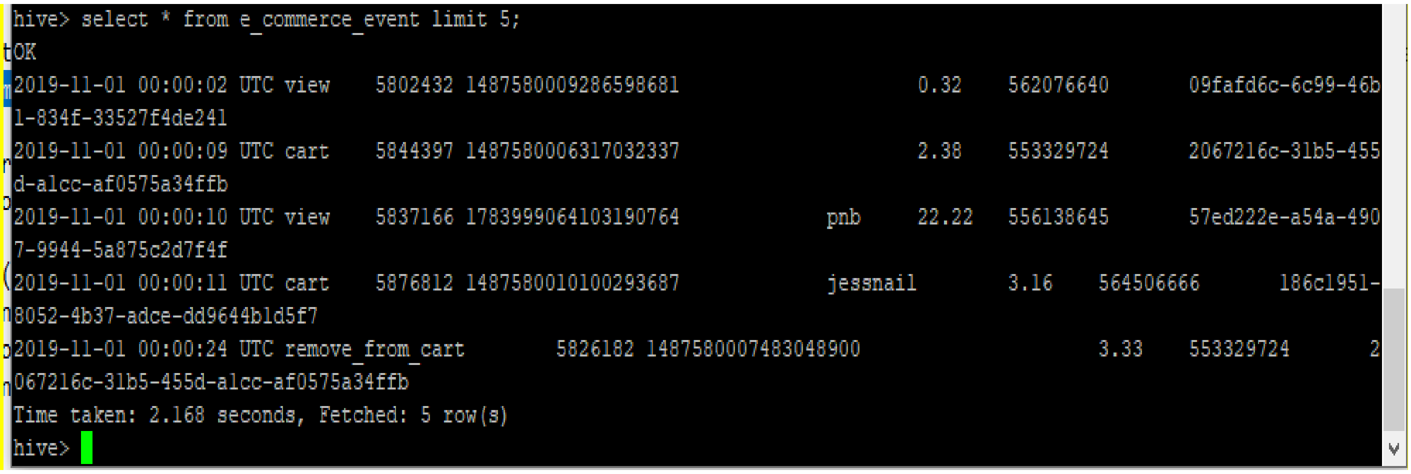
B) Move the data from the S3 bucket into the HDFS



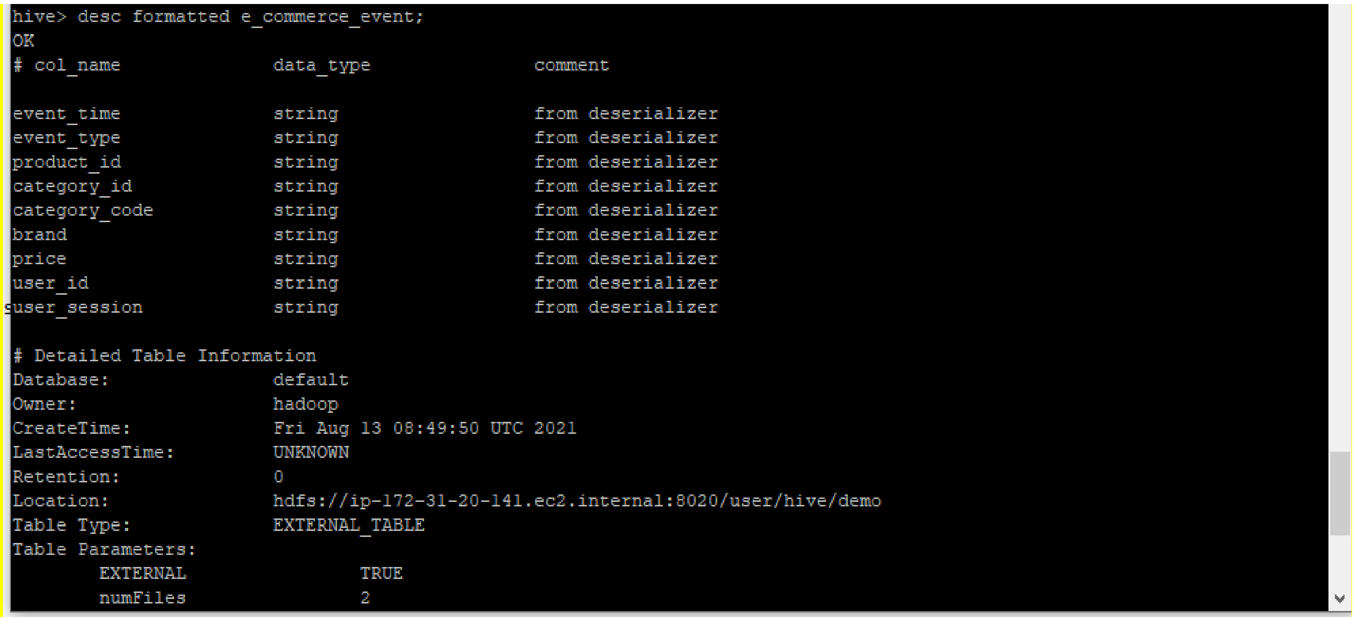
```
[hadoop@ip-172-31-20-141 ~]$ aws s3 ls e-commerce-events-ml/
2020-03-17 11:47:09  545839412 2019-Nov.csv
2020-03-17 11:37:31  482542278 2019-Oct.csv
[hadoop@ip-172-31-20-141 ~]$ hadoop distcp 's3://e-commerce-events-ml/*' '/user/hive/demo/'
21/08/13 08:45:23 INFO tools.DistCp: Input Options: DistCpOptions{atomicCommit=false, syncFolder=false, deleteMissing=fal
se, ignoreFailures=false, overwrite=false, skipCRC=false, blocking=true, numListstatusThreads=0, maxMaps=20, mapBandwidth
=100, sslConfigurationFile='null', copyStrategy='uniformsize', preserveStatus=[], preserveRawXattrs=false, atomicWorkPath
=null, logPath=null, sourceFileListing=null, sourcePaths=[s3://e-commerce-events-ml/*], targetPath=/user/hive/demo, targe
tPathExists=false, filtersFile='null'}
21/08/13 08:45:23 INFO client.RMProxy: Connecting to ResourceManager at ip-172-31-20-141.ec2.internal/172.31.20.141:8032
21/08/13 08:45:27 INFO tools.SimpleCopyListing: Paths (files+dirs) cnt = 2; dirCnt = 0
21/08/13 08:45:27 INFO tools.SimpleCopyListing: Build file listing completed.
21/08/13 08:45:27 INFO Configuration.deprecation: io.sort.mb is deprecated. Instead, use mapreduce.task.io.sort.mb
21/08/13 08:45:27 INFO Configuration.deprecation: io.sort.factor is deprecated. Instead, use mapreduce.task.io.sort.facto
r
21/08/13 08:45:28 INFO tools.DistCp: Number of paths in the copy list: 2
21/08/13 08:45:28 INFO tools.DistCp: Number of paths in the copy list: 2
```



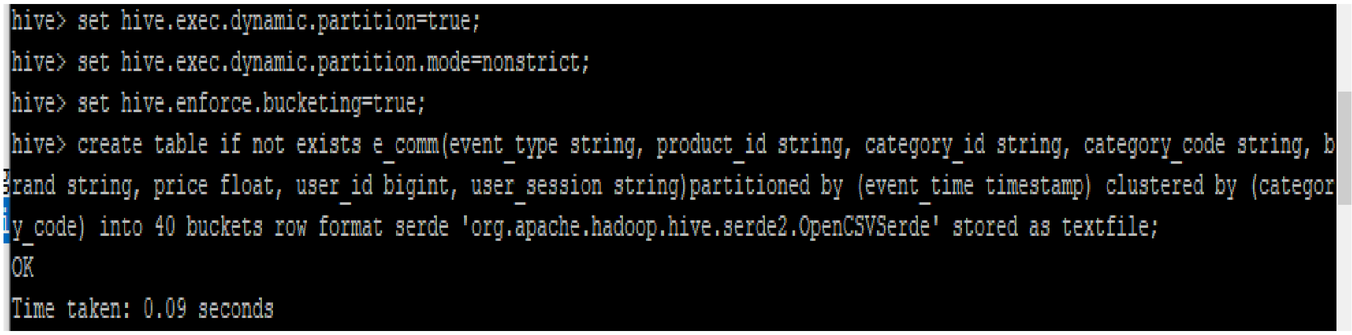
```
hive> create external table if not exists e_commerce_event(event_time timestamp, event_type string, product_id string, ca
tegory_id string, category_code string, brand string, price float, user_id bigint, user_session string) row format serde
'org.apache.hadoop.hive.serde2.OpenCSVSerde' stored as textfile location '/user/hive/demo/' tblproperties("skip.header.li
ne.count"="1");
OK
Time taken: 0.726 seconds
hive>
```



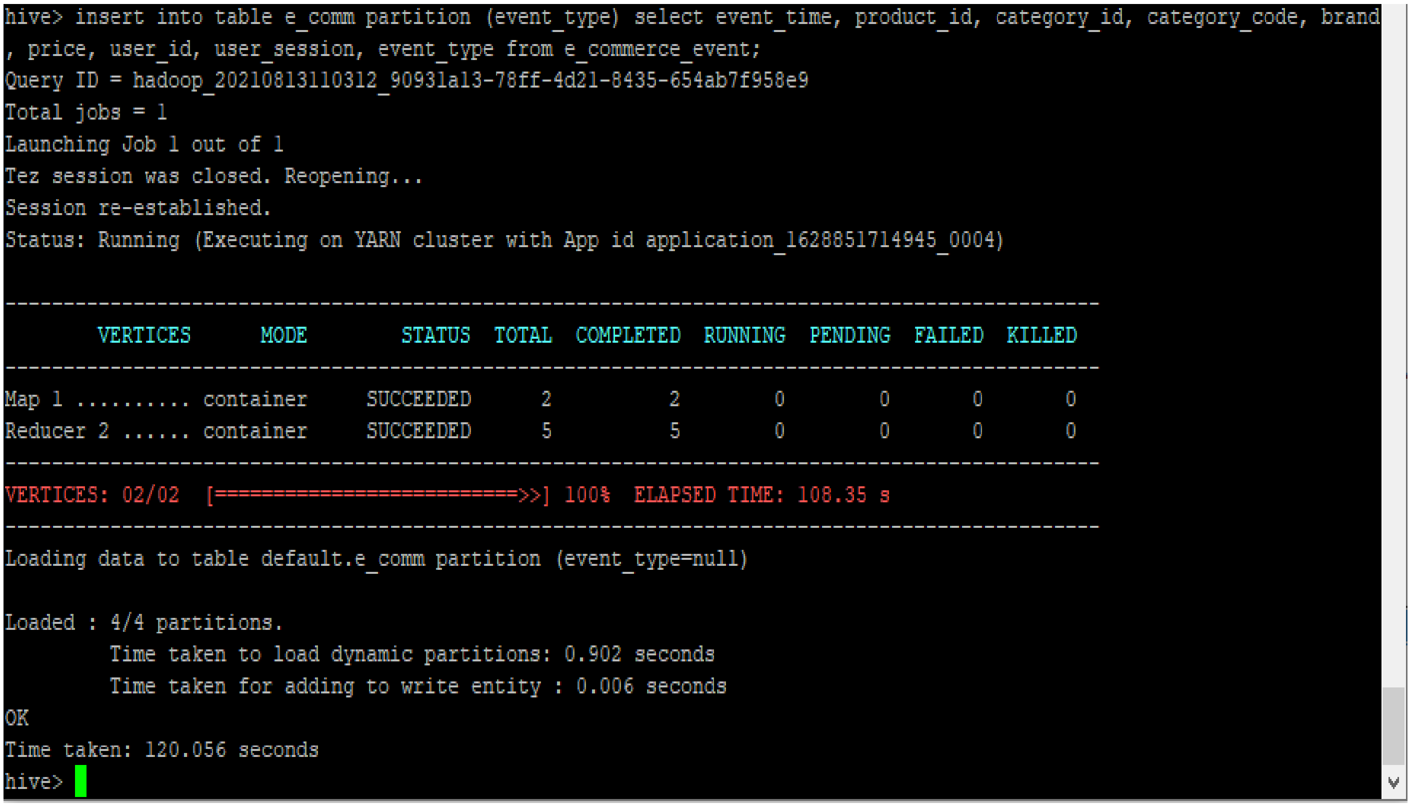
```
hive> select * from e_commerce_event limit 5;
OK
2019-11-01 00:00:02 UTC view    5802432 1487580009286598681                     0.32    562076640       09fafd6c-6c99-46b
1-834f-33527f4de241
2019-11-01 00:00:09 UTC cart    5844397 1487580006317032337                     2.38    553329724       2067216c-31b5-455
d-alcc-af0575a34ffb
2019-11-01 00:00:10 UTC view    5837166 1783999064103190764             pnb     22.22   556138645       57ed222e-a54a-490
7-9944-5a875c2d7f4f
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       2
067216c-31b5-455d-alcc-af0575a34ffb
Time taken: 2.168 seconds, Fetched: 5 row(s)
hive>
```



```
hive> desc formatted e_commerce_event;
OK
# col_name              data_type               comment

event_time              string                  from deserializer
event_type              string                  from deserializer
product_id              string                  from deserializer
category_id             string                  from deserializer
category_code           string                  from deserializer
brand                   string                  from deserializer
price                   string                  from deserializer
user_id                 string                  from deserializer
user_session            string                  from deserializer

# Detailed Table Information
Database:               default
Owner:                  hadoop
CreateTime:             Fri Aug 13 08:49:50 UTC 2021
LastAccessTime:         UNKNOWN
Retention:              0
Location:               hdfs://ip-172-31-20-141.ec2.internal:8020/user/hive/demo
Table Type:             EXTERNAL_TABLE
Table Parameters:
        EXTERNAL                TRUE
        numFiles                2
```

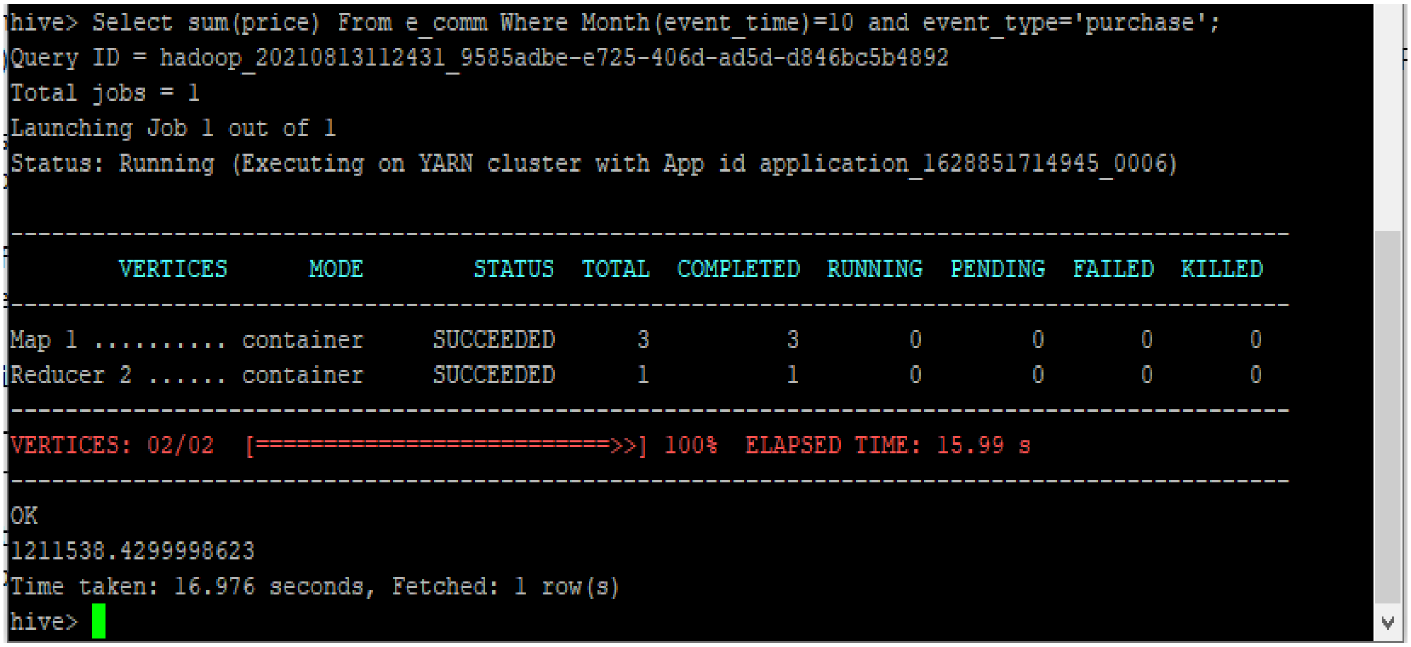


```
hive> set hive.exec.dynamic.partition=true;
hive> set hive.exec.dynamic.partition.mode=nonstrict;
hive> set hive.enforce.bucketing=true;
hive> create table if not exists e_comm(event_type string, product_id string, category_id string, category_code string, b
rand string, price float, user_id bigint, user_session string)partitioned by (event_time timestamp) clustered by (categor
y_code) into 40 buckets row format serde 'org.apache.hadoop.hive.serde2.OpenCSVSerde' stored as textfile;
OK
Time taken: 0.09 seconds
```



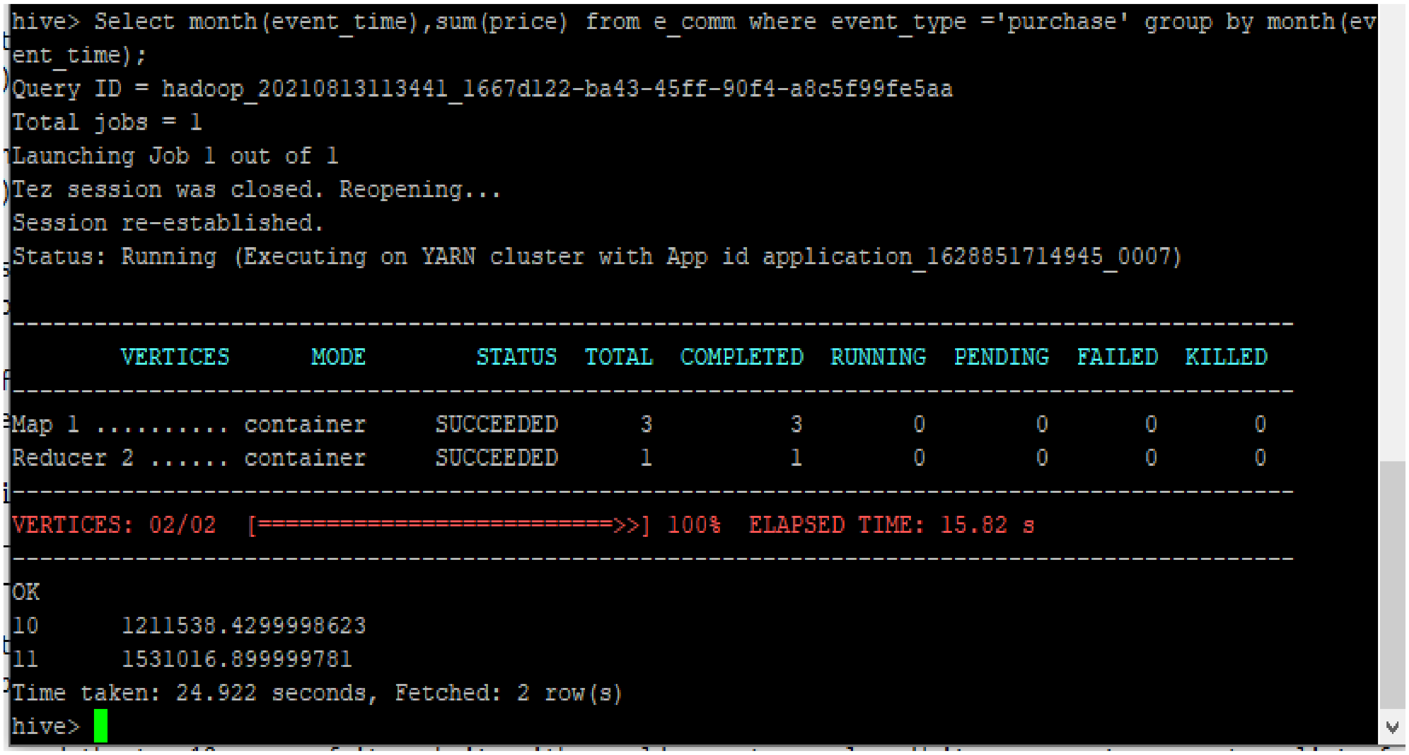
```
hive> insert into table e_comm partition (event_type) select event_time, product_id, category_id, category_code, brand
, price, user_id, user_session, event_type from e_commerce_event;
Query ID = hadoop_20210813110312_90931a13-78ff-4d21-8435-654ab7f958e9
Total jobs = 1
Launching Job 1 out of 1
Tez session was closed. Reopening...
Session re-established.
Status: Running (Executing on YARN cluster with App id application_1628851714945_0004)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      2          2        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      5          5        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 108.35 s
----------------------------------------------------------------------------------------------
Loading data to table default.e_comm partition (event_type=null)

Loaded : 4/4 partitions.
         Time taken to load dynamic partitions: 0.902 seconds
         Time taken for adding to write entity : 0.006 seconds
OK
Time taken: 120.056 seconds
hive>
```

Q1. Find the total revenue generated due to purchases made in October.



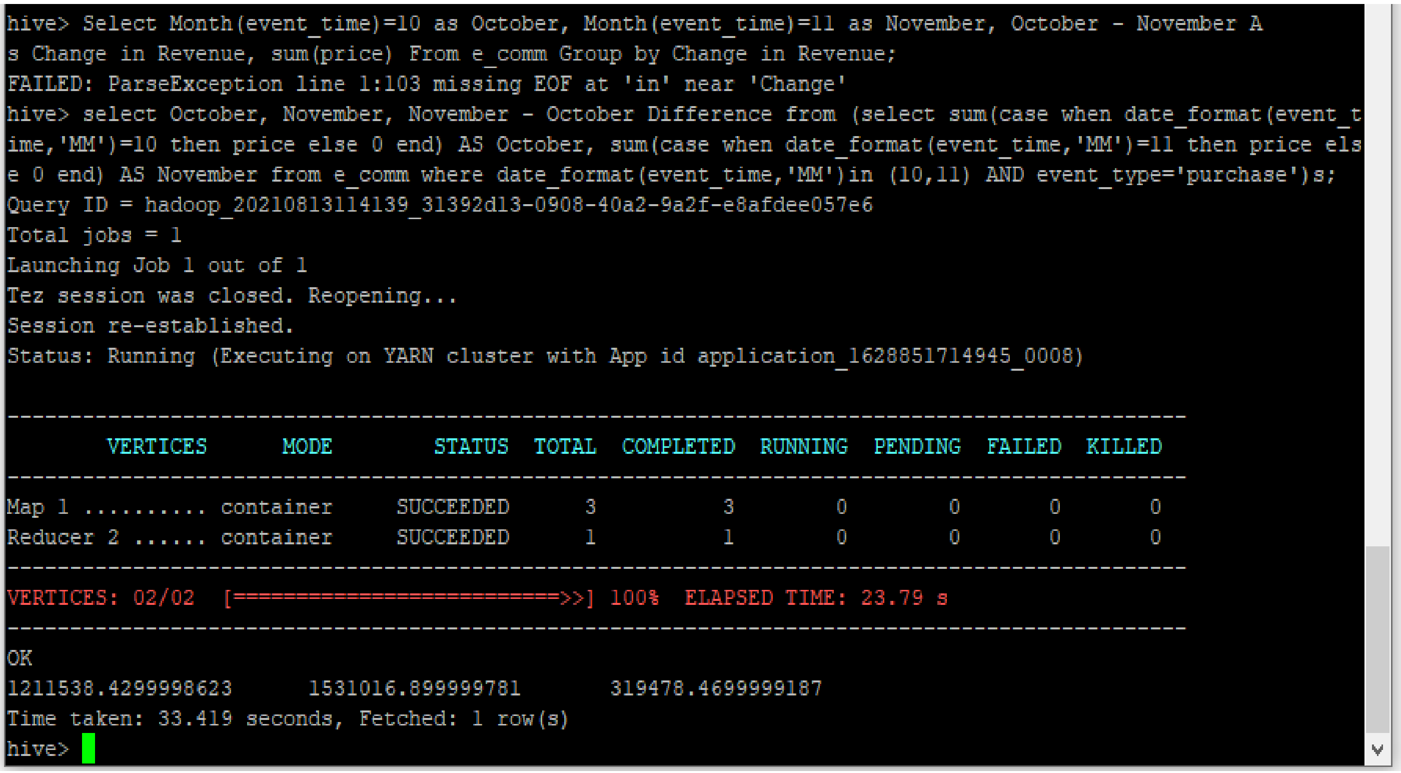
```
hive> Select sum(price) From e_comm Where Month(event_time)=10 and event_type='purchase';
Query ID = hadoop_20210813112431_9585adbe-e725-406d-ad5d-d846bc5b4892
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1628851714945_0006)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      3          3        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 15.99 s
----------------------------------------------------------------------------------------------
OK
1211538.4299998623
Time taken: 16.976 seconds, Fetched: 1 row(s)
hive>
```

Q2. Write a query to yield the total sum of purchases per month in a single output.



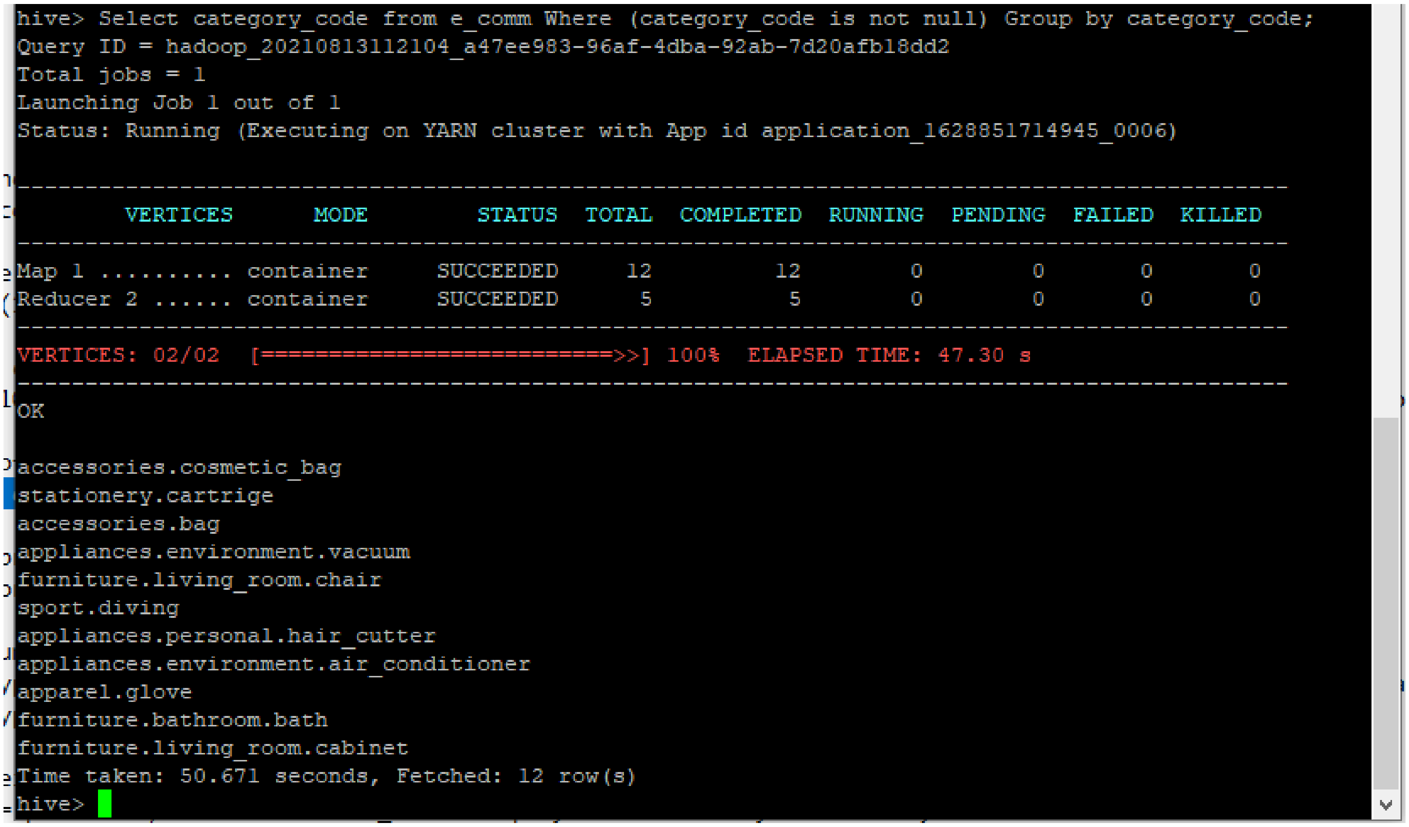
```
hive> Select month(event_time),sum(price) from e_comm where event_type ='purchase' group by month(ev
ent_time);
Query ID = hadoop_20210813113441_1667d122-ba43-45ff-90f4-a8c5f99fe5aa
Total jobs = 1
Launching Job 1 out of 1
Tez session was closed. Reopening...
Session re-established.
Status: Running (Executing on YARN cluster with App id application_1628851714945_0007)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      3          3        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 15.82 s
----------------------------------------------------------------------------------------------
OK
10      1211538.4299998623
11      1531016.899999781
Time taken: 24.922 seconds, Fetched: 2 row(s)
hive>
```

Q3. Write a query to find the change in revenue generate due to purchases from October to November.



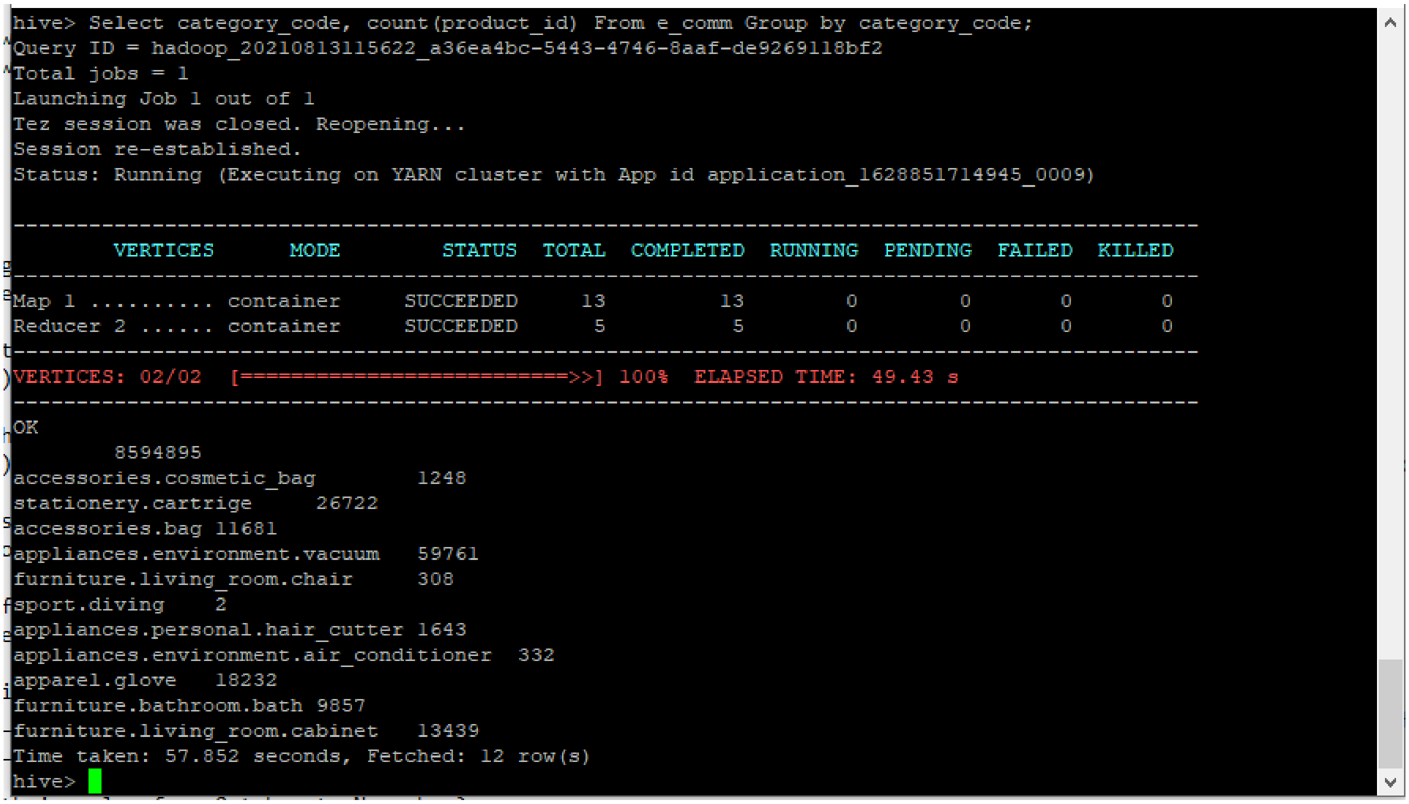
```
hive> Select Month(event_time)=10 as October, Month(event_time)=11 as November, October - November A
s Change in Revenue, sum(price) From e_comm Group by Change in Revenue;
FAILED: ParseException line 1:103 missing EOF at 'in' near 'Change'
hive> select October, November, November - October Difference from (select sum(case when date_format(event_t
ime,'MM')=10 then price else 0 end) AS October, sum(case when date_format(event_time,'MM')=11 then price els
e 0 end) AS November from e_comm where date_format(event_time,'MM')in (10,11) AND event_type='purchase')s;
Query ID = hadoop_20210813114139_31392d13-0908-40a2-9a2f-e8afdee057e6
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_1628851714945_0008)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      3          3        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 23.79 s
----------------------------------------------------------------------------------------------
OK
1211538.4299998623      1531016.899999781       319478.4699999187
Time taken: 33.419 seconds, Fetched: 1 row(s)
hive>
```

Q4. Find distinct categories of products. Categories with null category code can be ignored.



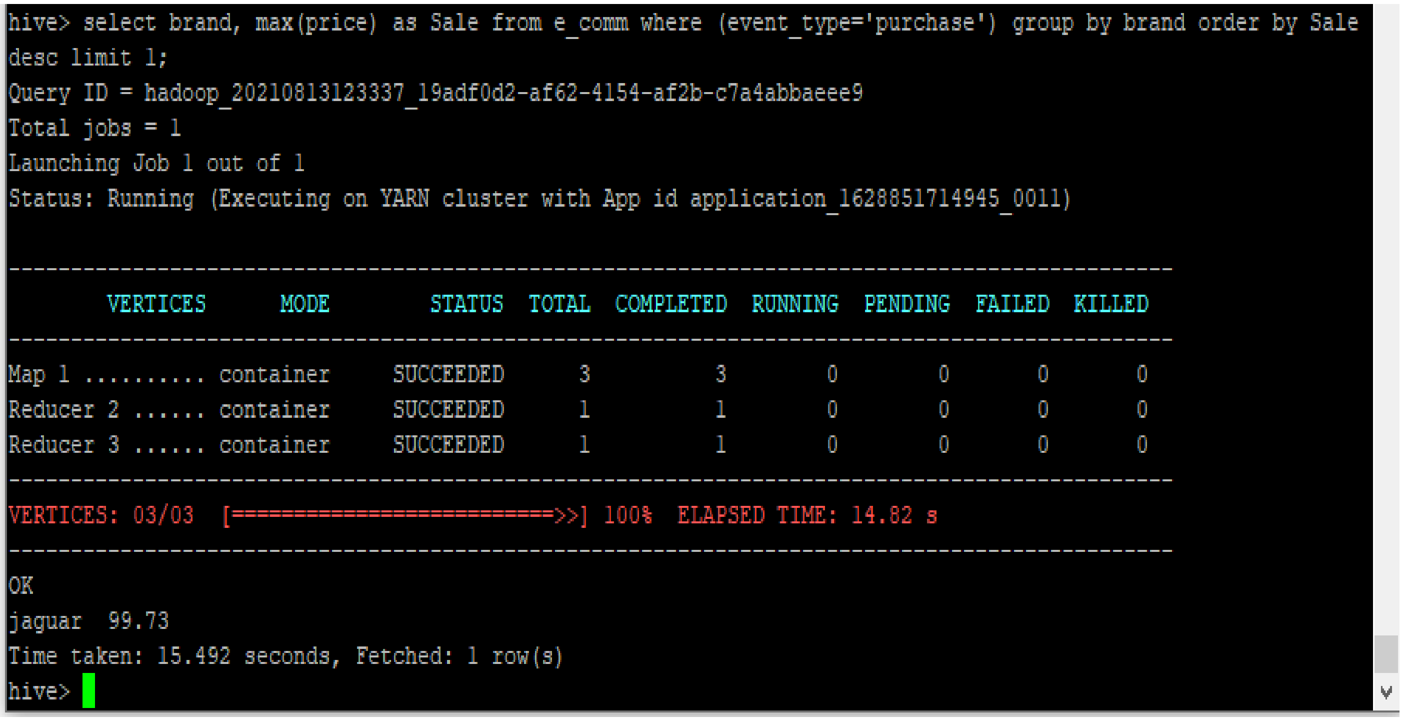
```
hive> Select category_code from e_comm Where (category_code is not null) Group by category_code;
Query ID = hadoop_20210813112104_a47ee983-96af-4dba-92ab-7d20afb18dd2
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1628851714945_0006)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED     12         12        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      5          5        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 47.30 s
----------------------------------------------------------------------------------------------
OK

accessories.cosmetic_bag
stationery.cartrige
accessories.bag
appliances.environment.vacuum
furniture.living_room.chair
sport.diving
appliances.personal.hair_cutter
appliances.environment.air_conditioner
apparel.glove
furniture.bathroom.bath
furniture.living_room.cabinet
Time taken: 50.671 seconds, Fetched: 12 row(s)
hive>
```

Q5. Find the total number of product available under each category.



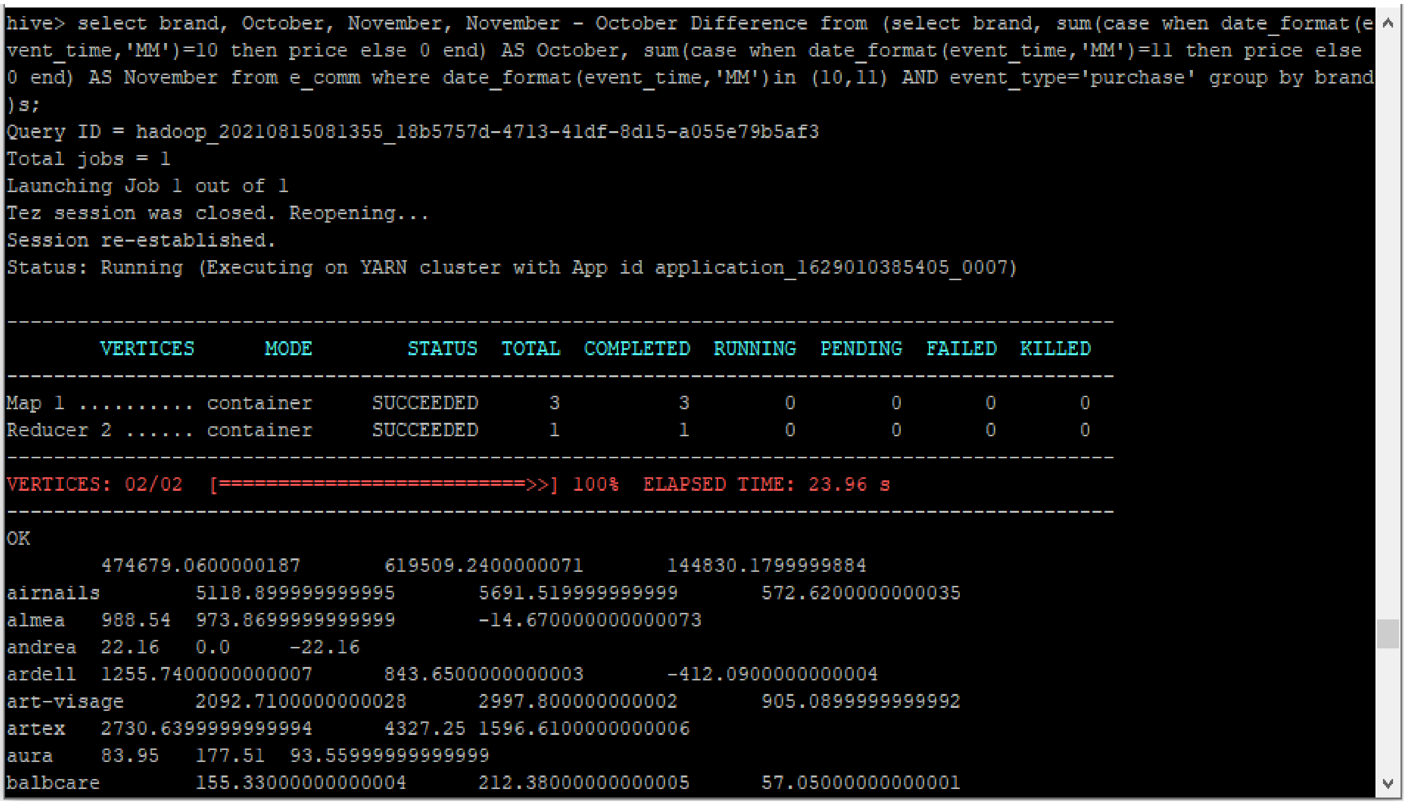
```
hive> Select category_code, count(product_id) From e_comm Group by category_code;
Query ID = hadoop_20210813115622_a36ea4bc-5443-4746-8aaf-de9269118bf2
Total jobs = 1
Launching Job 1 out of 1
Tez session was closed. Reopening...
Session re-established.
Status: Running (Executing on YARN cluster with App id application_1628851714945_0009)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED     13         13        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      5          5        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 49.43 s
----------------------------------------------------------------------------------------------
OK
        8594895
accessories.cosmetic_bag        1248
stationery.cartrige     26722
accessories.bag 11681
appliances.environment.vacuum   59761
furniture.living_room.chair     308
sport.diving    2
appliances.personal.hair_cutter 1643
appliances.environment.air_conditioner  332
apparel.glove   18232
furniture.bathroom.bath 9857
furniture.living_room.cabinet   13439
Time taken: 57.852 seconds, Fetched: 12 row(s)
hive>
```

Q6. Which brand had the maximum sales in October and November combined?



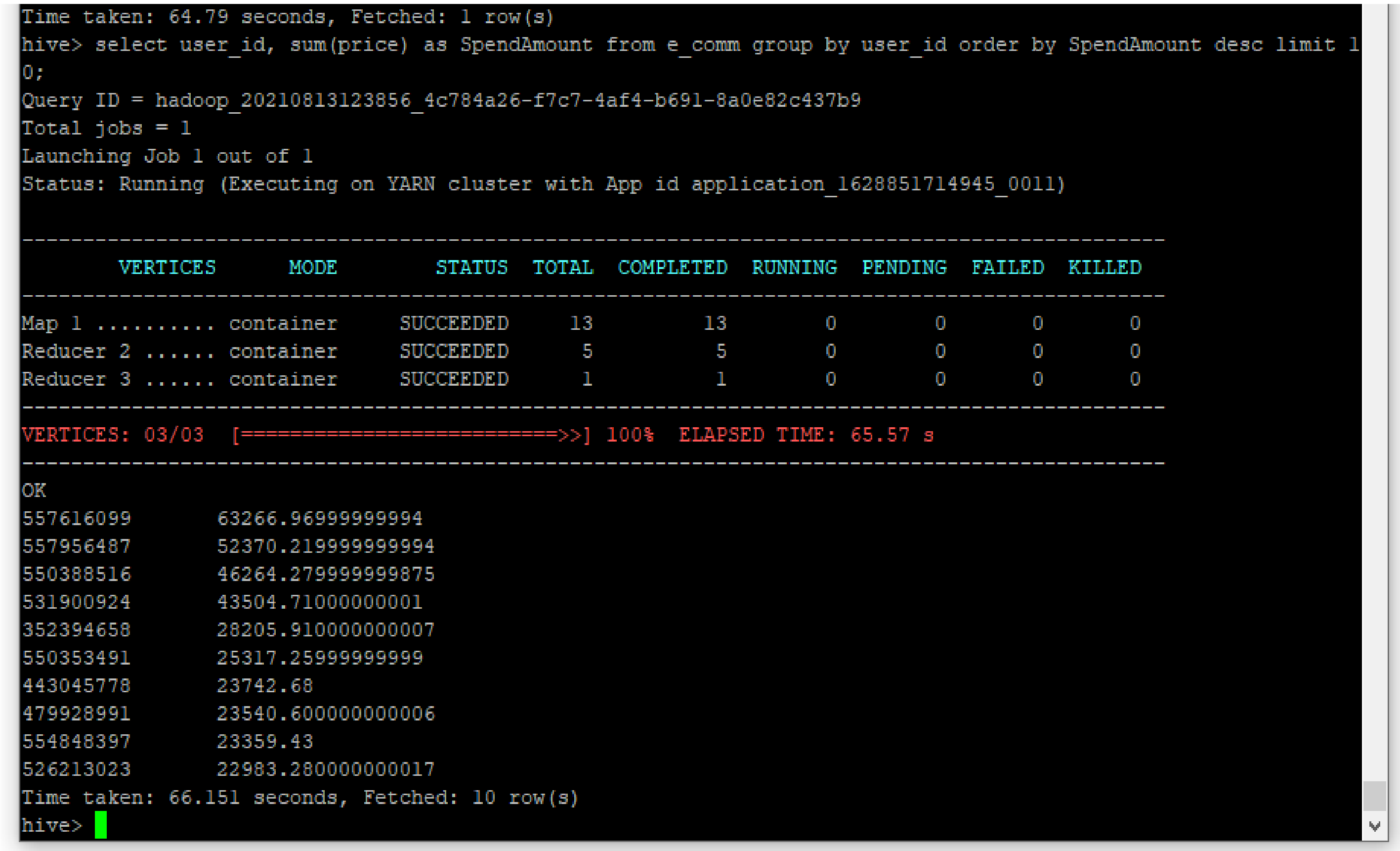
```
hive> select brand, max(price) as Sale from e_comm where (event_type='purchase') group by brand order by Sale
desc limit 1;
Query ID = hadoop_20210813123337_19adf0d2-af62-4154-af2b-c7a4abbaeee9
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1628851714945_0011)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      3          3        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 14.82 s
----------------------------------------------------------------------------------------------
OK
jaguar  99.73
Time taken: 15.492 seconds, Fetched: 1 row(s)
hive>
```

Q7. Which brands increased their sales from October to November?



```
hive> select brand, October, November, November - October Difference from (select brand, sum(case when date_format(e
vent_time,'MM')=10 then price else 0 end) AS October, sum(case when date_format(event_time,'MM')=11 then price else
0 end) AS November from e_comm where date_format(event_time,'MM')in (10,11) AND event_type='purchase' group by brand
)s;
Query ID = hadoop_20210815081355_18b5757d-4713-41df-8d15-a055e79b5af3
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_1629010385405_0007)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      3          3        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 23.96 s
----------------------------------------------------------------------------------------------
OK
        474679.0600000187       619509.2400000071       144830.1799999884
airnails        5118.899999999995       5691.519999999999       572.6200000000035
almea   988.54  973.8699999999999       -14.670000000000073
andrea  22.16   0.0     -22.16
ardell  1255.7400000000007      843.6500000000003       -412.0900000000004
art-visage      2092.7100000000028      2997.800000000002       905.0899999999992
artex   2730.6399999999994      4327.25 1596.6100000000006
aura    83.95   177.51  93.55999999999999
balbcare        155.33000000000004      212.38000000000005      57.05000000000001
```

Q8. 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.



```
Time taken: 64.79 seconds, Fetched: 1 row(s)
hive> select user_id, sum(price) as SpendAmount from e_comm group by user_id order by SpendAmount desc limit 1
0;
Query ID = hadoop_20210813123856_4c784a26-f7c7-4af4-b691-8a0e82c437b9
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1628851714945_0011)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED     13         13        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      5          5        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 65.57 s
----------------------------------------------------------------------------------------------
OK
557616099       63266.96999999994
557956487       52370.219999999994
550388516       46264.279999999875
531900924       43504.71000000001
352394658       28205.910000000007
550353491       25317.25999999999
443045778       23742.68
479928991       23540.600000000006
554848397       23359.43
526213023       22983.280000000017
Time taken: 66.151 seconds, Fetched: 10 row(s)
hive>
```