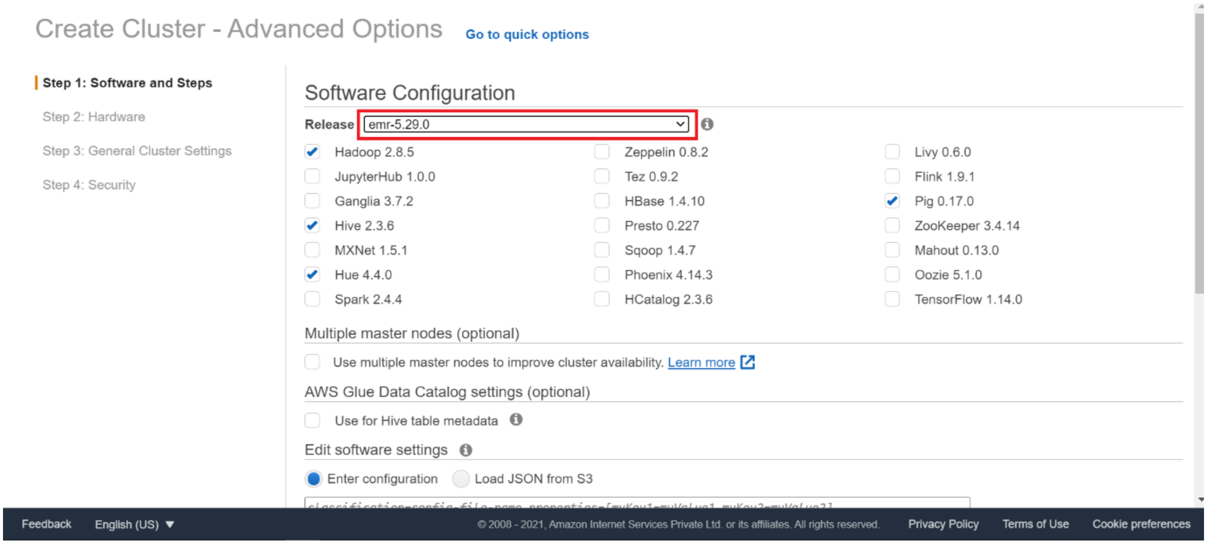
HIVE CASE STUDY

EMR CLUSTER CREATION

STEP 1: Login to your AWS account & search EMR services. After the EMR home page appears click on **Create cluster** & follow the steps as mentioned. We have chosen cluster release version **5.29.0** in our case study.



Create Cluster - Advanced Options Go to quick options

Software Configuration

Release emr-5.29.0

☑ Hadoop 2.8.5	□ Zeppelin 0.8.2	□ Livy 0.6.0
□ JupyterHub 1.0.0	□ Tez 0.9.2	□ Flink 1.9.1
□ Ganglia 3.7.2	□ HBase 1.4.10	☑ Pig 0.17.0
☑ Hive 2.3.6	□ Presto 0.227	□ ZooKeeper 3.4.14
□ MXNet 1.5.1	□ Sqoop 1.4.7	□ Mahout 0.13.0
☑ Hue 4.4.0	□ Phoenix 4.14.3	□ Oozie 5.1.0
□ Spark 2.4.4	□ HCatalog 2.3.6	□ TensorFlow 1.14.0

Multiple master nodes (optional)

□ Use multiple master nodes to improve cluster availability. Learn more

AWS Glue Data Catalog settings (optional)

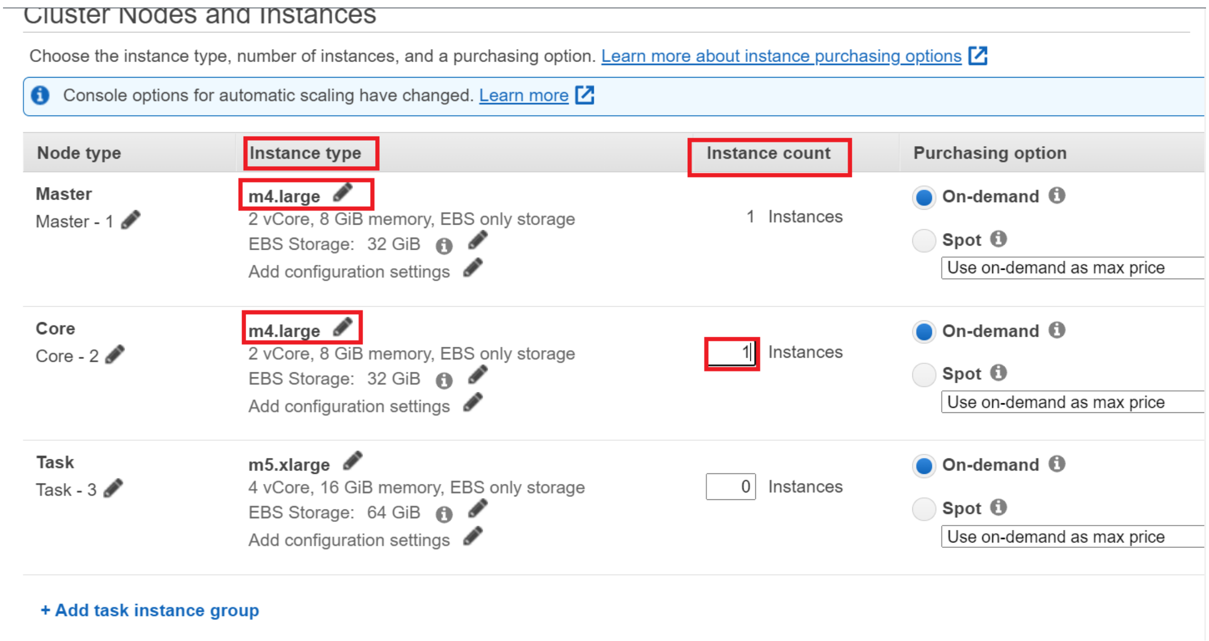
□ Use for Hive table metadata

Edit software settings

● Enter configuration ○ Load JSON from S3

Feedback
English (US)
© 2008 - 2021, Amazon Internet Services Private Ltd. or its affiliates. All rights reserved.
Privacy Policy
Terms of Use
Cookie preferences

STEP 2: We will be going for a **2-node cluster** for our analysis & we will select **m4.large** instance type each for both master & core node.



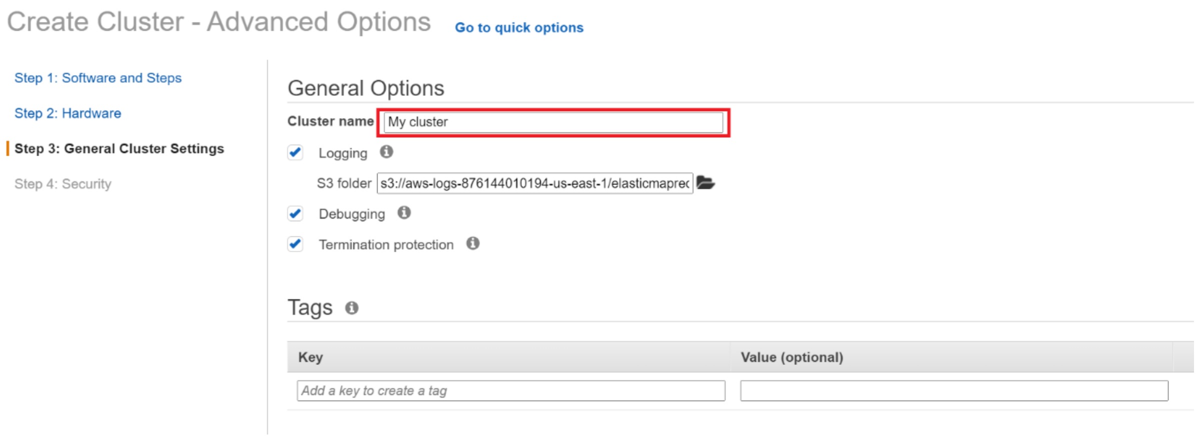
Cluster Nodes and Instances

Choose the instance type, number of instances, and a purchasing option. Learn more about instance purchasing options

Console options for automatic scaling have changed. Learn more

Node type	Instance type	Instance count	Purchasing option
Master Master - 1	m4.large 2 vCore, 8 GiB memory, EBS only storage EBS Storage: 32 GiB Add configuration settings	1 Instances	● On-demand ○ Spot Use on-demand as max price
Core Core - 2	m4.large 2 vCore, 8 GiB memory, EBS only storage EBS Storage: 32 GiB Add configuration settings	1 Instances	● On-demand ○ Spot Use on-demand as max price
Task Task - 3	m5.xlarge 4 vCore, 16 GiB memory, EBS only storage EBS Storage: 64 GiB Add configuration settings	0 Instances	● On-demand ○ Spot Use on-demand as max price

Step 3: Select a cluster name. Here we have taken the cluster name as **My cluster**.



Create Cluster - Advanced Options
Go to quick options
Step 1: Software and Steps
Step 2: Hardware
Step 3: General Cluster Settings
Step 4: Security

General Options

Cluster name: My cluster

☑ Logging

S3 folder: s3://aws-logs-876144010194-us-east-1/elasticmaprec

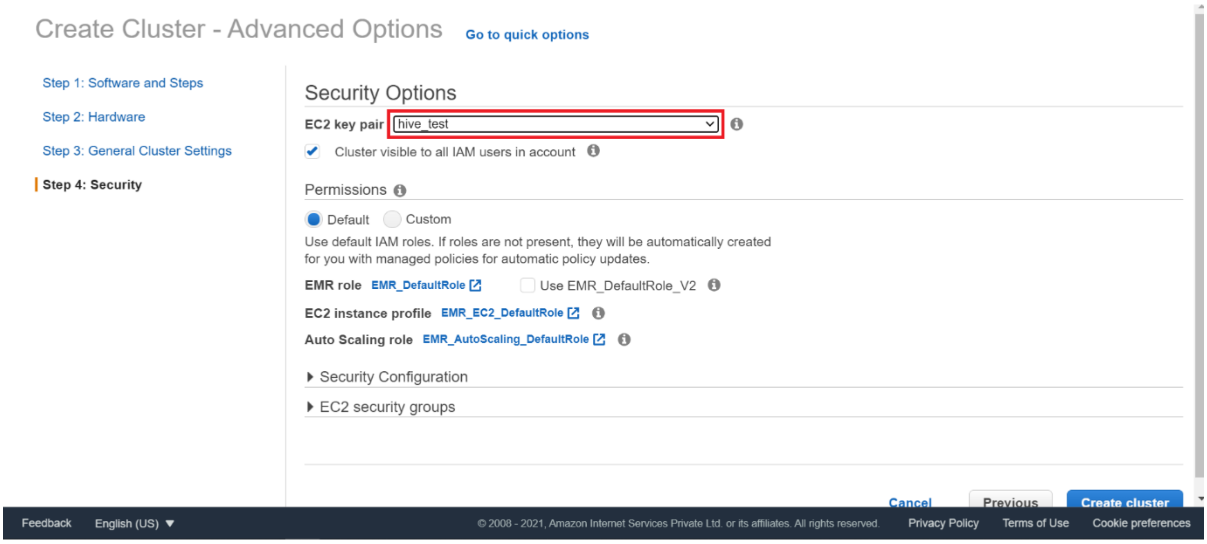
☑ Debugging

☑ Termination protection

Tags

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

Step 4: Select an already created key-pair which will be used while connecting to master node and **hive_test** is the name of the key that we have used.



Create Cluster - Advanced Options
Go to quick options
Step 1: Software and Steps
Step 2: Hardware
Step 3: General Cluster Settings
Step 4: Security

Security Options

EC2 key pair: hive_test

☑ 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 □ Use EMR_DefaultRole_V2

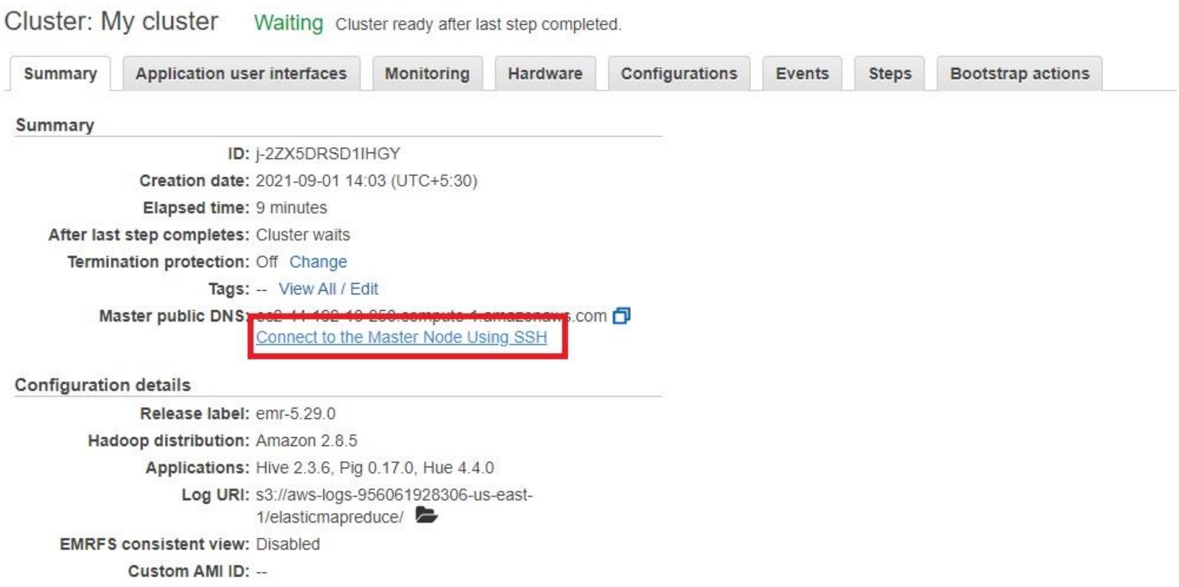
EC2 instance profile: EMR_EC2_DefaultRole

Auto Scaling role: EMR_AutoScaling_DefaultRole

▸ Security Configuration

▸ EC2 security groups

Step 5: Our cluster has been created successfully and is in **waiting state** which indicates its ready to be connected from the local system.



Cluster: My cluster Waiting Cluster ready after last step completed.

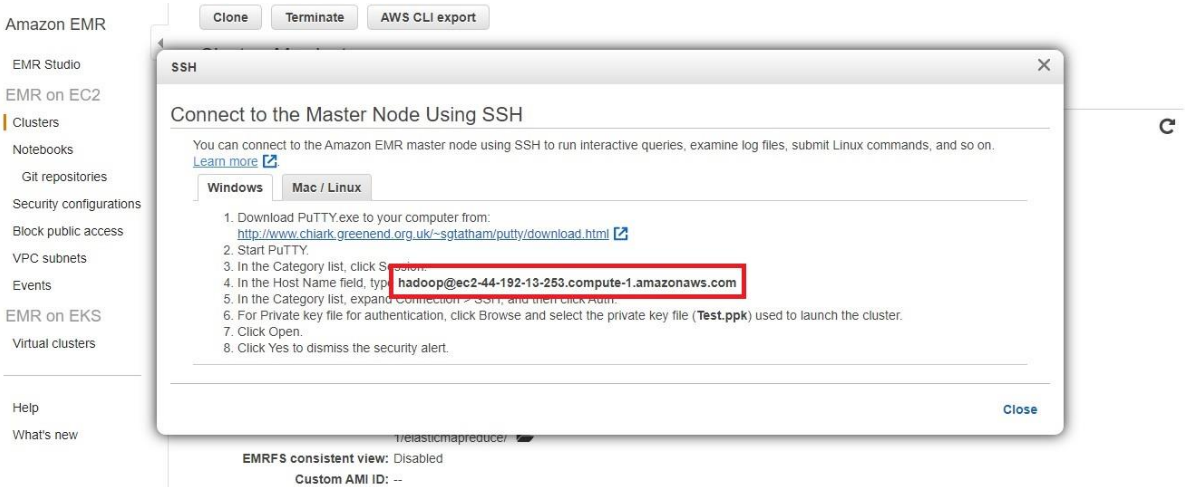
Summary

ID: j-2ZX5DRSD1IHGY
Creation date: 2021-09-01 14:03 (UTC+5:30)
Elapsed time: 9 minutes
After last step completes: Cluster waits
Termination protection: Off Change
Tags: -- View All / Edit
Master public DNS: ec2-44-192-13-253.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
Log URI: s3://aws-logs-956061928306-us-east-1/elasticmapreduce/
EMRFS consistent view: Disabled
Custom AMI ID: --

Step 6: Copy the highlighted link i.e., the master public DNS.



Connect to the Master Node Using SSH

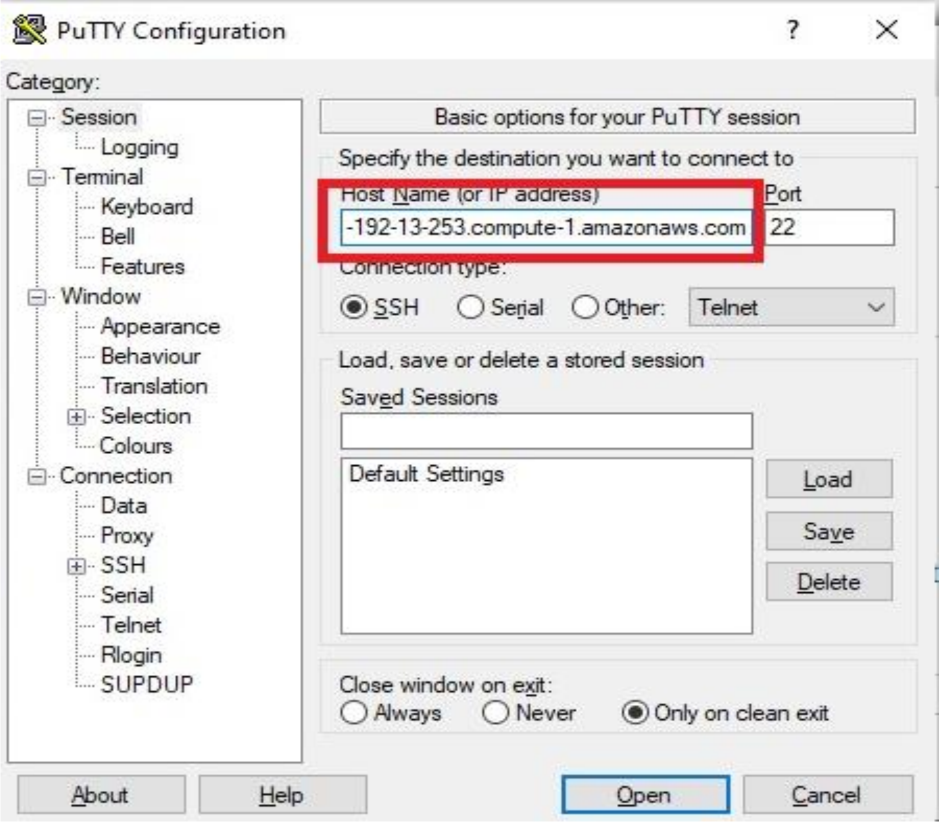
You can connect to the Amazon EMR master node using SSH to run interactive queries, examine log files, submit Linux commands, and so on. Learn more.

Windows | Mac / Linux

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@ec2-44-192-13-253.compute-1.amazonaws.com**
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 (**Test.ppk**) used to launch the cluster.
7. Click Open.
8. Click Yes to dismiss the security alert.

1/elasticmapreduce/
EMRFS consistent view: Disabled
Custom AMI ID: --

Step 7: Next, paste the address in the **Host Name** field.



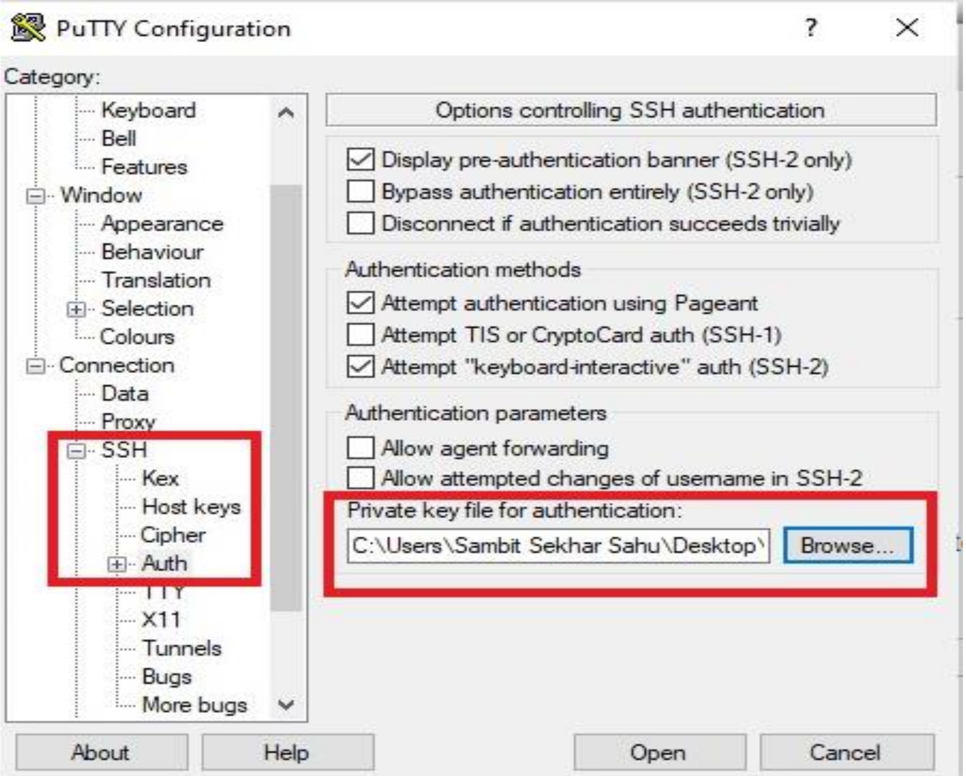
Host Name (or IP address): -192-13-253.compute-1.amazonaws.com

Port: 22

Connection type: ● SSH ○ Serial ○ Other: Telnet

Close window on exit: ○ Always ○ Never ● Only on clean exit

Step 7: Click on **SSH** & then **Auth**. Give the location where the key pair is stored in the local system and click on **Open.**



☑ 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)

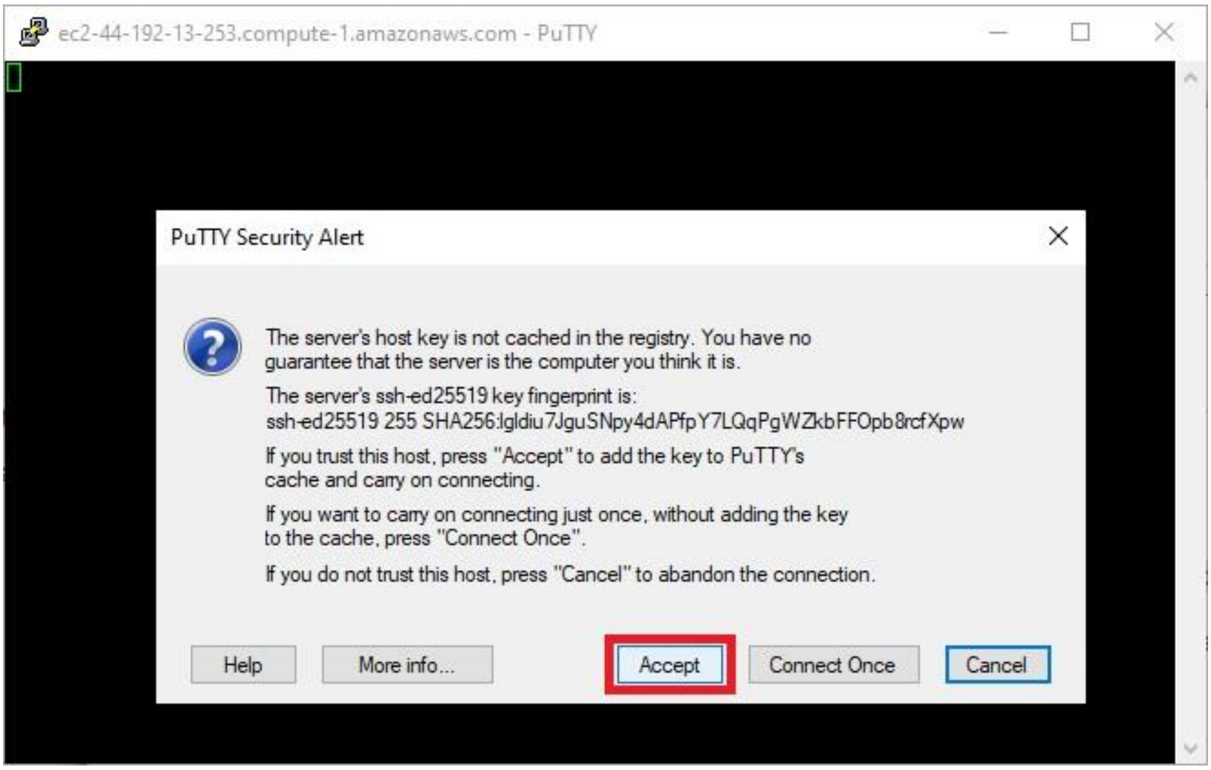
Authentication parameters

□ Allow agent forwarding

□ Allow attempted changes of username in SSH-2

Private key file for authentication: C:\Users\Sambit Sekhar Sahu\Desktop\

Step 8: Next click on Accept which will open the SSH terminal.



The server's host key is not cached in the registry. You have no guarantee that the server is the computer you think it is.

The server's ssh-ed25519 key fingerprint is:
ssh-ed25519 255 SHA256:lgldiu7JguSNpy4dAPfpY7LQqPgWZkbFFOpb8rcfXpw

If you trust this host, press "Accept" to add the key to PuTTY's cache and carry on connecting.

If you want to carry on connecting just once, without adding the key to the cache, press "Connect Once".

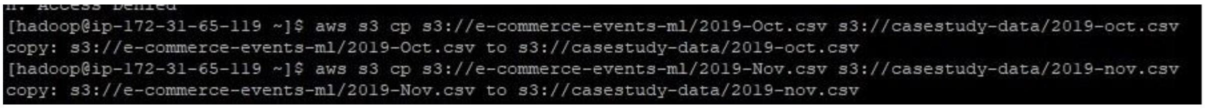
If you do not trust this host, press "Cancel" to abandon the connection.

After we have created an EMR cluster & successfully connected to it via putty we can begin to code in the SSH Terminal.

DATA LOADING INTO S3 BUCKET

aws s3 cp s3://e-commerce-events-ml/2019-Oct.csv s3://casestudy-data/2019-oct.csv

aws s3 cp s3://e-commerce-events-ml/2019-Nov.csv s3://casestudy-data/2019-nov.csv



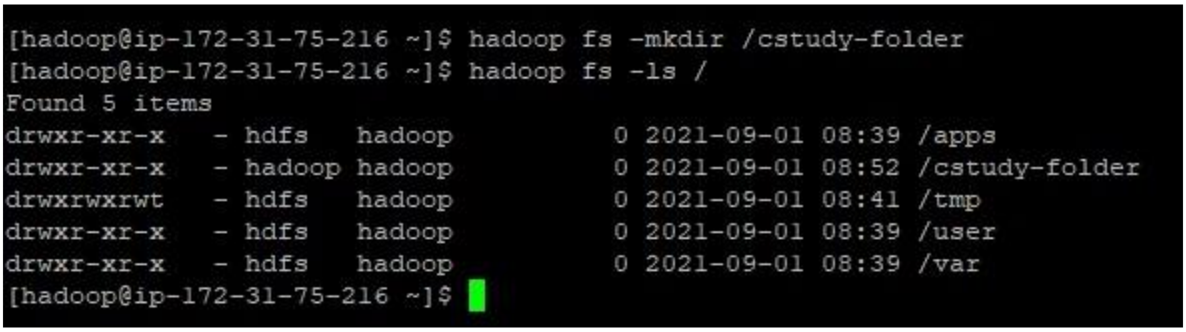
```
n: Access Denied
[hadoop@ip-172-31-65-119 ~]$ aws s3 cp s3://e-commerce-events-ml/2019-Oct.csv s3://casestudy-data/2019-oct.csv
copy: s3://e-commerce-events-ml/2019-Oct.csv to s3://casestudy-data/2019-oct.csv
[hadoop@ip-172-31-65-119 ~]$ aws s3 cp s3://e-commerce-events-ml/2019-Nov.csv s3://casestudy-data/2019-nov.csv
copy: s3://e-commerce-events-ml/2019-Nov.csv to s3://casestudy-data/2019-nov.csv
```

using the above code, we could directly copy the data files into Hadoop without actually downloading the data files into our system, and the copying completes instantly.

WORKING WITH HDFS

- **Creating a folder in Hadoop**

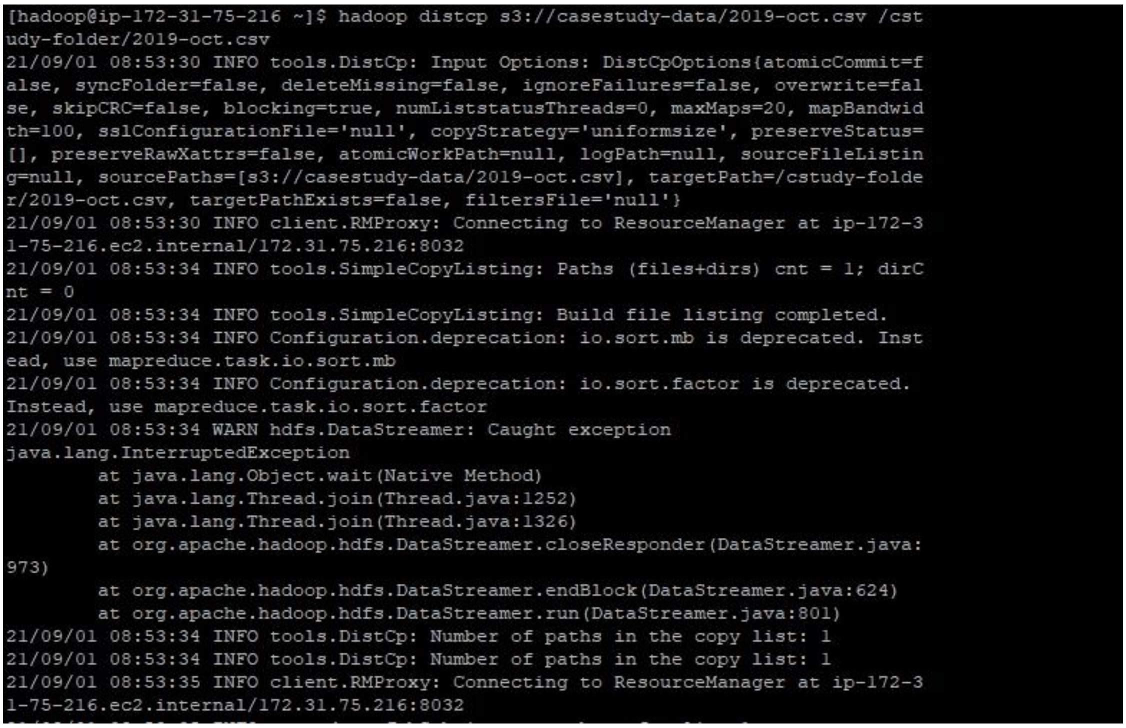
  hadoop fs -mkdir /cstudy-folder
  hadoop fs -ls /



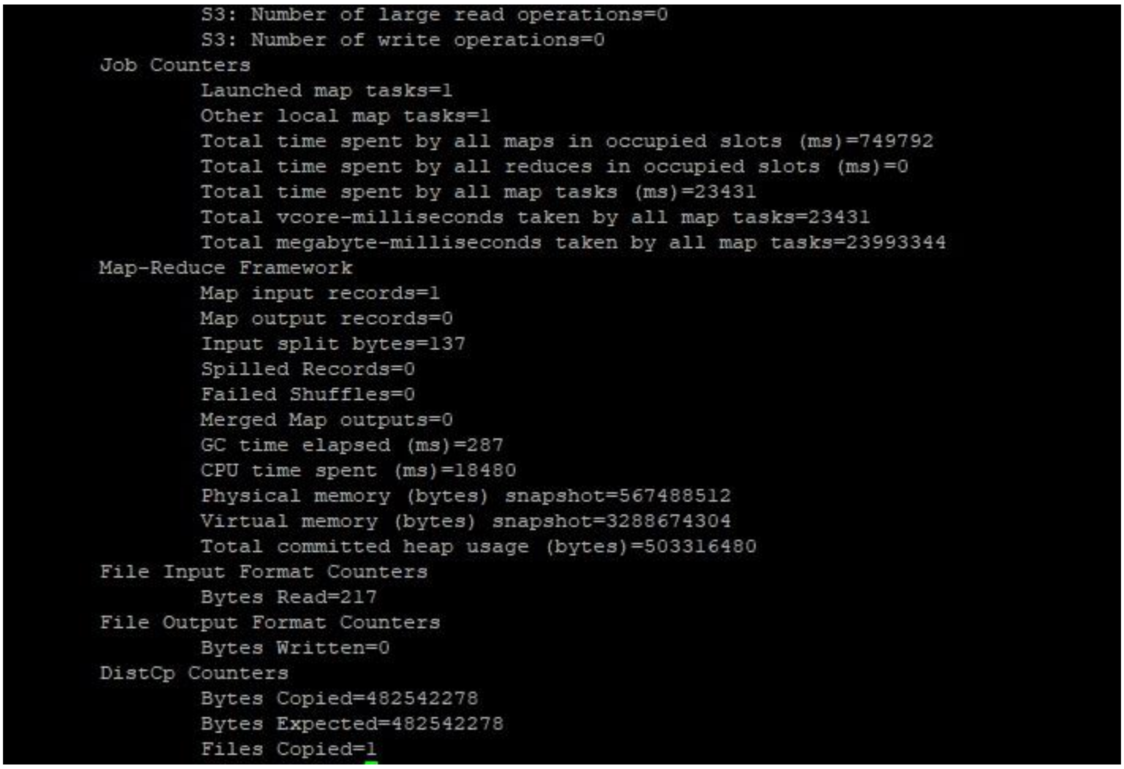
```
[hadoop@ip-172-31-75-216 ~]$ hadoop fs -mkdir /cstudy-folder
[hadoop@ip-172-31-75-216 ~]$ hadoop fs -ls /
Found 5 items
drwxr-xr-x   - hdfs   hadoop          0 2021-09-01 08:39 /apps
drwxr-xr-x   - hadoop hadoop          0 2021-09-01 08:52 /cstudy-folder
drwxrwxrwt   - hdfs   hadoop          0 2021-09-01 08:41 /tmp
drwxr-xr-x   - hdfs   hadoop          0 2021-09-01 08:39 /user
drwxr-xr-x   - hdfs   hadoop          0 2021-09-01 08:39 /var
[hadoop@ip-172-31-75-216 ~]$
```

- **Copying October & November data from S3 bucket into HDFS**

  hadoop distcp s3://casestudy-data/2019-oct.csv /cstudy-folder/2019-oct.csv
  hadoop distcp s3://casestudy-data/2019-nov.csv /cstudy-folder/2019-nov.csv



```
[hadoop@ip-172-31-75-216 ~]$ hadoop distcp s3://casestudy-data/2019-oct.csv /cst
udy-folder/2019-oct.csv
21/09/01 08:53:30 INFO tools.DistCp: Input Options: DistCpOptions{atomicCommit=f
alse, syncFolder=false, deleteMissing=false, ignoreFailures=false, overwrite=fal
se, skipCRC=false, blocking=true, numListstatusThreads=0, maxMaps=20, mapBandwid
th=100, sslConfigurationFile='null', copyStrategy='uniformsize', preserveStatus=
[], preserveRawXattrs=false, atomicWorkPath=null, logPath=null, sourceFileListin
g=null, sourcePaths=[s3://casestudy-data/2019-oct.csv], targetPath=/cstudy-folde
r/2019-oct.csv, targetPathExists=false, filtersFile='null'}
21/09/01 08:53:30 INFO client.RMProxy: Connecting to ResourceManager at ip-172-3
1-75-216.ec2.internal/172.31.75.216:8032
21/09/01 08:53:34 INFO tools.SimpleCopyListing: Paths (files+dirs) cnt = 1; dirC
nt = 0
21/09/01 08:53:34 INFO tools.SimpleCopyListing: Build file listing completed.
21/09/01 08:53:34 INFO Configuration.deprecation: io.sort.mb is deprecated. Inst
ead, use mapreduce.task.io.sort.mb
21/09/01 08:53:34 INFO Configuration.deprecation: io.sort.factor is deprecated.
Instead, use mapreduce.task.io.sort.factor
21/09/01 08:53:34 WARN hdfs.DataStreamer: Caught exception
java.lang.InterruptedException
        at java.lang.Object.wait(Native Method)
        at java.lang.Thread.join(Thread.java:1252)
        at java.lang.Thread.join(Thread.java:1326)
        at org.apache.hadoop.hdfs.DataStreamer.closeResponder(DataStreamer.java:
973)
        at org.apache.hadoop.hdfs.DataStreamer.endBlock(DataStreamer.java:624)
        at org.apache.hadoop.hdfs.DataStreamer.run(DataStreamer.java:801)
21/09/01 08:53:34 INFO tools.DistCp: Number of paths in the copy list: 1
21/09/01 08:53:34 INFO tools.DistCp: Number of paths in the copy list: 1
21/09/01 08:53:35 INFO client.RMProxy: Connecting to ResourceManager at ip-172-3
1-75-216.ec2.internal/172.31.75.216:8032
```

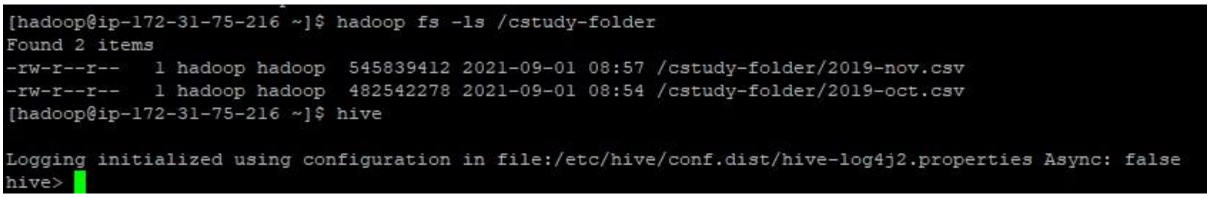


```
                S3: Number of large read operations=0
                S3: Number of write operations=0
        Job Counters
                Launched map tasks=1
                Other local map tasks=1
                Total time spent by all maps in occupied slots (ms)=749792
                Total time spent by all reduces in occupied slots (ms)=0
                Total time spent by all map tasks (ms)=23431
                Total vcore-milliseconds taken by all map tasks=23431
                Total megabyte-milliseconds taken by all map tasks=23993344
        Map-Reduce Framework
                Map input records=1
                Map output records=0
                Input split bytes=137
                Spilled Records=0
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=287
                CPU time spent (ms)=18480
                Physical memory (bytes) snapshot=567488512
                Virtual memory (bytes) snapshot=3288674304
                Total committed heap usage (bytes)=503316480
        File Input Format Counters
                Bytes Read=217
        File Output Format Counters
                Bytes Written=0
        DistCp Counters
                Bytes Copied=482542278
                Bytes Expected=482542278
                Files Copied=1
```

Similarly, November data file was also copied, as shown in above image.

- **Verifying if data has been copied successfully**

  hadoop fs -ls /cstudy-folder

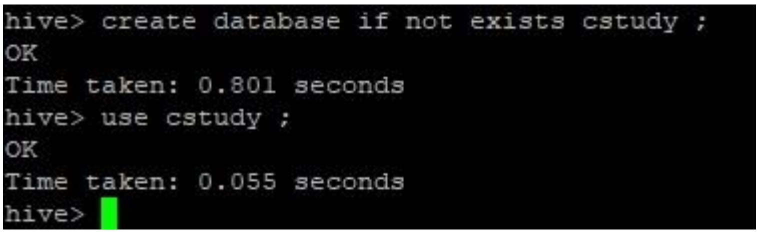


```
[hadoop@ip-172-31-75-216 ~]$ hadoop fs -ls /cstudy-folder
Found 2 items
-rw-r--r--   1 hadoop hadoop  545839412 2021-09-01 08:57 /cstudy-folder/2019-nov.csv
-rw-r--r--   1 hadoop hadoop  482542278 2021-09-01 08:54 /cstudy-folder/2019-oct.csv
[hadoop@ip-172-31-75-216 ~]$ hive

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

WORKING ON HIVE

- **Getting into Hive CLI – making a new database named *cstudy***

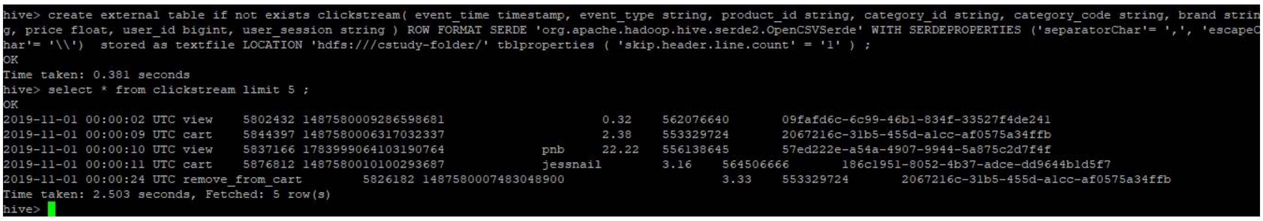
  create database if not exists cstudy ;
  use cstudy ;



```
hive> create database if not exists cstudy ;
OK
Time taken: 0.801 seconds
hive> use cstudy ;
OK
Time taken: 0.055 seconds
hive>
```

- **Creating a common table named clickstream and storing both October & November data in it**

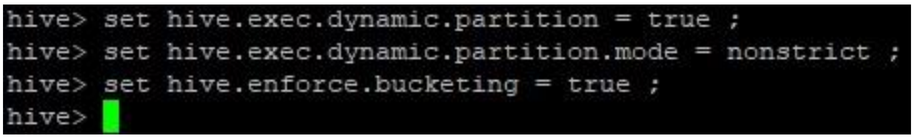
  create external table if not exists clickstream(event_time timestamp, event_type string, product_id string, category_id string, category_code string, brand string, price float, user_id bigint, user_session string) ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.OpenCSVSerde' WITH SERDEPROPERTIES ('separatorChar'= ',', 'escapeChar'= '\\') stored as textfile LOCATION 'hdfs:///cstudy-folder/' tblproperties ('skip.header.line.count' = '1') ;



```
hive> create external table if not exists clickstream( event_time timestamp, event_type string, product_id string, category_id string, category_code string, brand strin
g, price float, user_id bigint, user_session string ) ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.OpenCSVSerde' WITH SERDEPROPERTIES ('separatorChar'= ',', 'escapeC
har'= '\\')  stored as textfile LOCATION 'hdfs:///cstudy-folder/' tblproperties ( 'skip.header.line.count' = '1' ) ;
OK
Time taken: 0.381 seconds
hive> select * from clickstream limit 5 ;
OK
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
Time taken: 2.503 seconds, Fetched: 5 row(s)
hive> 
```

- **To create optimised table having partitions & buckets we need to enable some settings**

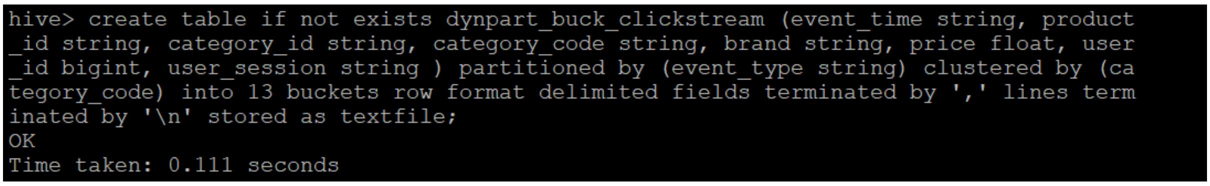
  set hive.exec.dynamic.partition = true ;
  set hive.exec.dynamic.partition.mode = nonstrict ;
  set hive.enforce.bucketing = true ;



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

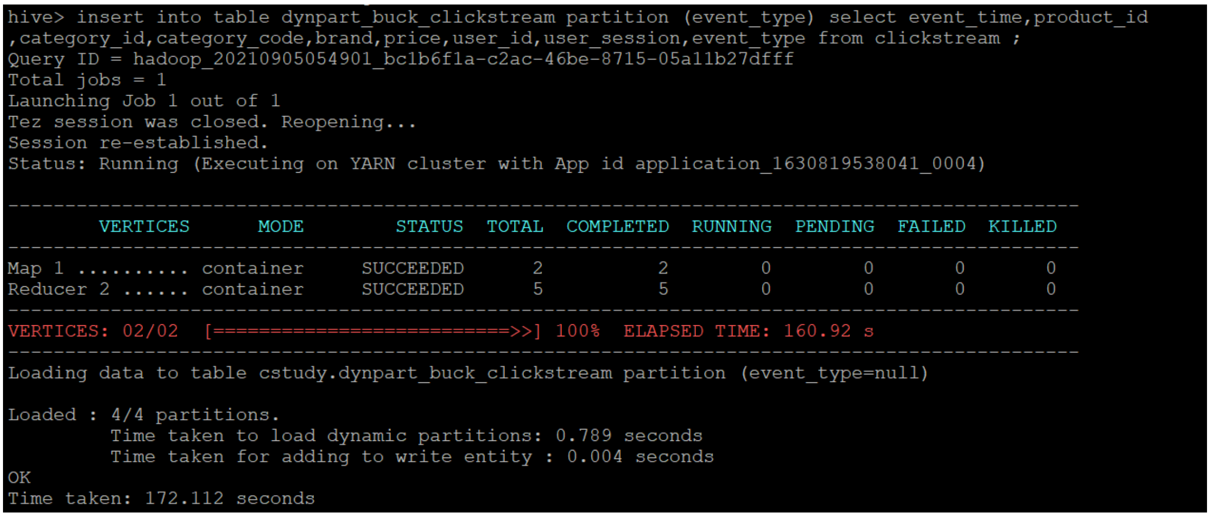
- **Creating table with dynamic partitions and buckets and inserting data into it**

  create table if not exists dynpart_buck_clickstream (event_time string, product_id string, category_id string, category_code string, brand string, price float, user_id bigint, user_session string) partitioned by (event_type string) clustered by (category_code) into 13 buckets row format delimited fields terminated by ',' lines terminated by '\n' stored as textfile;



```
hive> create table if not exists dynpart_buck_clickstream (event_time string, product
_id string, category_id string, category_code string, brand string, price float, user
_id bigint, user_session string ) partitioned by (event_type string) clustered by (ca
tegory_code) into 13 buckets row format delimited fields terminated by ',' lines term
inated by '\n' stored as textfile;
OK
Time taken: 0.111 seconds
```

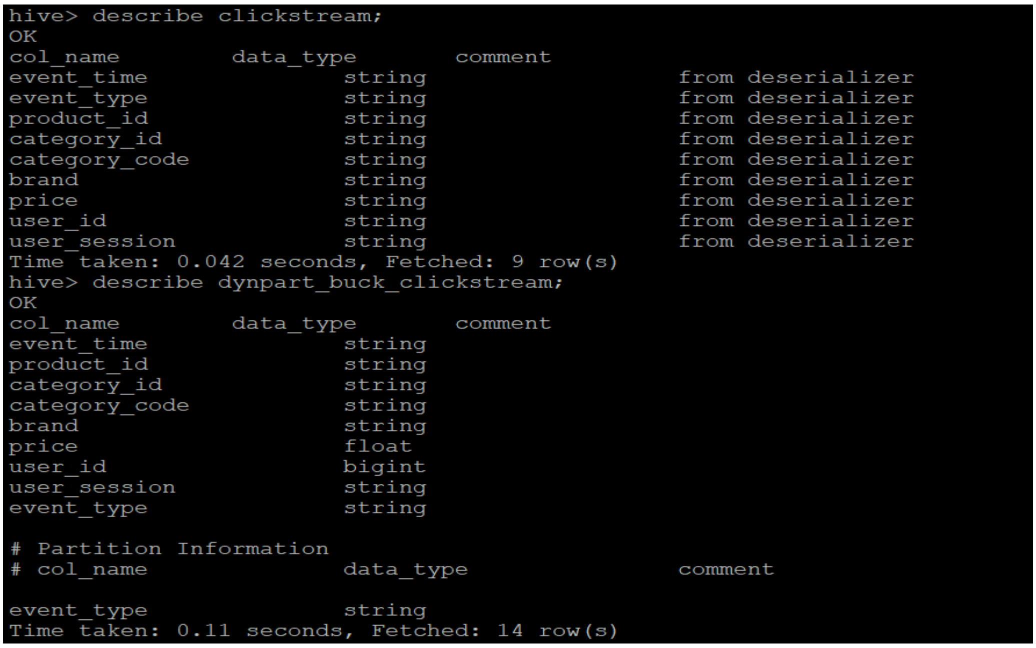
insert into table dynpart_buck_clickstream partition (event_type) select event_time, product_id,category_id,category_code,brand,price,user_id,user_session,event_type from clickstream ;



```
hive> insert into table dynpart_buck_clickstream partition (event_type) select event_time,product_id
,category_id,category_code,brand,price,user_id,user_session,event_type from clickstream ;
Query ID = hadoop_20210905054901_bc1b6f1a-c2ac-46be-8715-05a11b27dfff
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_1630819538041_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: 160.92 s
----------------------------------------------------------------------------------------------
Loading data to table cstudy.dynpart_buck_clickstream partition (event_type=null)

Loaded : 4/4 partitions.
         Time taken to load dynamic partitions: 0.789 seconds
         Time taken for adding to write entity : 0.004 seconds
OK
Time taken: 172.112 seconds
```

- **Describing both tables**

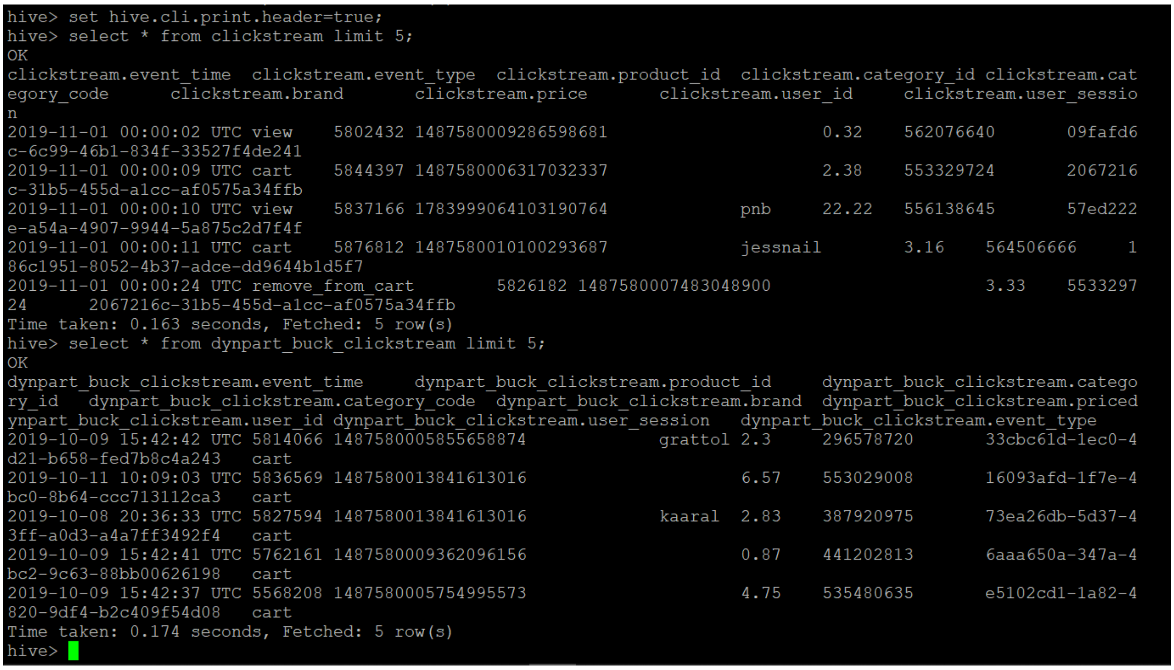
describe clickstream;
describe dynpart_buck_clickstream;



```
hive> describe clickstream;
OK
col_name        data_type       comment
event_time              string                  from deserializer
event_type              string                  from deserializer
product_id              string                  from deserializer
category_id             string                  from deserializer
category_code           string                  from deserializer
brand                   string                  from deserializer
price                   string                  from deserializer
user_id                 string                  from deserializer
user_session            string                  from deserializer
Time taken: 0.042 seconds, Fetched: 9 row(s)
hive> describe dynpart_buck_clickstream;
OK
col_name        data_type       comment
event_time              string
product_id              string
category_id             string
category_code           string
brand                   string
price                   float
user_id                 bigint
user_session            string
event_type              string

# Partition Information
# col_name              data_type               comment

event_type              string
Time taken: 0.11 seconds, Fetched: 14 row(s)
```

- **Checking data in both tables**

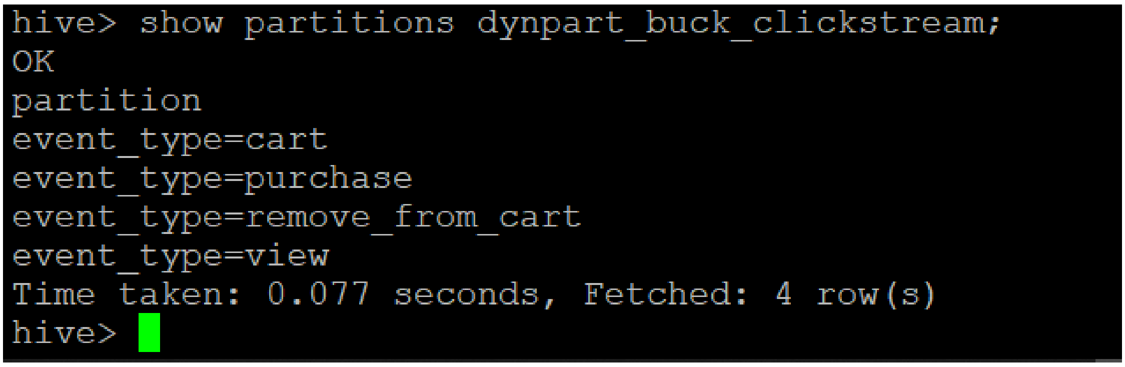
  set hive.cli.print.header=true;
  *select * from clickstream limit 5 ;*
  *select * from dynpart_buck_clickstream limit 5 ;*



```
hive> set hive.cli.print.header=true;
hive> select * from clickstream limit 5;
OK
clickstream.event_time  clickstream.event_type  clickstream.product_id  clickstream.category_id clickstream.cat
egory_code      clickstream.brand       clickstream.price       clickstream.user_id     clickstream.user_sessio
n
2019-11-01 00:00:02 UTC view    5802432 1487580009286598681                     0.32    562076640       09fafd6
c-6c99-46b1-834f-33527f4de241
2019-11-01 00:00:09 UTC cart    5844397 1487580006317032337                     2.38    553329724       2067216
c-31b5-455d-a1cc-af0575a34ffb
2019-11-01 00:00:10 UTC view    5837166 1783999064103190764             pnb     22.22   556138645       57ed222
e-a54a-4907-9944-5a875c2d7f4f
2019-11-01 00:00:11 UTC cart    5876812 1487580010100293687             jessnail        3.16    564506666     1
86c1951-8052-4b37-adce-dd9644b1d5f7
2019-11-01 00:00:24 UTC remove_from_cart        5826182 1487580007483048900                     3.33    5533297
24      2067216c-31b5-455d-a1cc-af0575a34ffb
Time taken: 0.163 seconds, Fetched: 5 row(s)
hive> select * from dynpart_buck_clickstream limit 5;
OK
dynpart_buck_clickstream.event_time     dynpart_buck_clickstream.product_id     dynpart_buck_clickstream.catego
ry_id   dynpart_buck_clickstream.category_code  dynpart_buck_clickstream.brand  dynpart_buck_clickstream.priced
ynpart_buck_clickstream.user_id dynpart_buck_clickstream.user_session   dynpart_buck_clickstream.event_type
2019-10-09 15:42:42 UTC 5814066 1487580005855658874             grattol 2.3     296578720       33cbc61d-1ec0-4
d21-b658-fed7b8c4a243   cart
2019-10-11 10:09:03 UTC 5836569 1487580013841613016                     6.57    553029008       16093afd-1f7e-4
bc0-8b64-ccc713112ca3   cart
2019-10-08 20:36:33 UTC 5827594 1487580013841613016             kaaral  2.83    387920975       73ea26db-5d37-4
3ff-a0d3-a4a7ff3492f4   cart
2019-10-09 15:42:41 UTC 5762161 1487580009362096156                     0.87    441202813       6aaa650a-347a-4
bc2-9c63-88bb00626198   cart
2019-10-09 15:42:37 UTC 5568208 1487580005754995573                     4.75    535480635       e5102cd1-1a82-4
820-9df4-b2c409f54d08   cart
Time taken: 0.174 seconds, Fetched: 5 row(s)
hive>
```

- **Checking if partitions were created successfully**

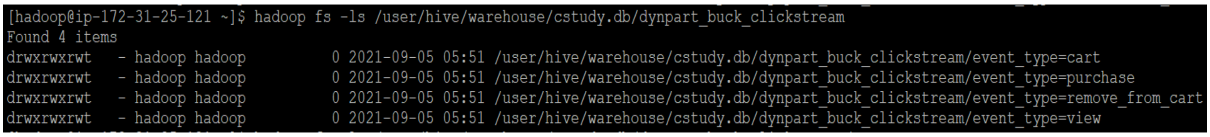
  show partitions dynpart_buck_clickstream;



```
hive> show partitions dynpart_buck_clickstream;
OK
partition
event_type=cart
event_type=purchase
event_type=remove_from_cart
event_type=view
Time taken: 0.077 seconds, Fetched: 4 row(s)
hive>
```

- **Checking in Hadoop if partitions were created successfully**

  hadoop fs -ls /user/hive/warehouse/cstudy.db/dynpart_buck_clickstream



```
[hadoop@ip-172-31-25-121 ~]$ hadoop fs -ls /user/hive/warehouse/cstudy.db/dynpart_buck_clickstream
Found 4 items
drwxrwxrwt   - hadoop hadoop          0 2021-09-05 05:51 /user/hive/warehouse/cstudy.db/dynpart_buck_clickstream/event_type=cart
drwxrwxrwt   - hadoop hadoop          0 2021-09-05 05:51 /user/hive/warehouse/cstudy.db/dynpart_buck_clickstream/event_type=purchase
drwxrwxrwt   - hadoop hadoop          0 2021-09-05 05:51 /user/hive/warehouse/cstudy.db/dynpart_buck_clickstream/event_type=remove_from_cart
drwxrwxrwt   - hadoop hadoop          0 2021-09-05 05:51 /user/hive/warehouse/cstudy.db/dynpart_buck_clickstream/event_type=view
```

Overall, we have made two tables,

- One common table named **clickstream** which contains data of both October & November.
- One table with partitions & buckets named **dynpart_buck_clickstream** for optimised querying which also contains data of both October & November.

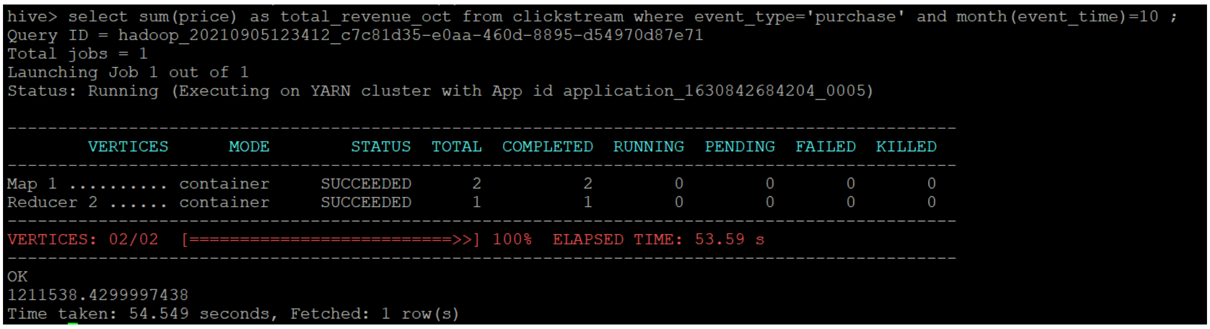
So, all the preparations are done & now we can move to query analysis-

QUERY ANALYSIS

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

a) *Unoptimized query:*

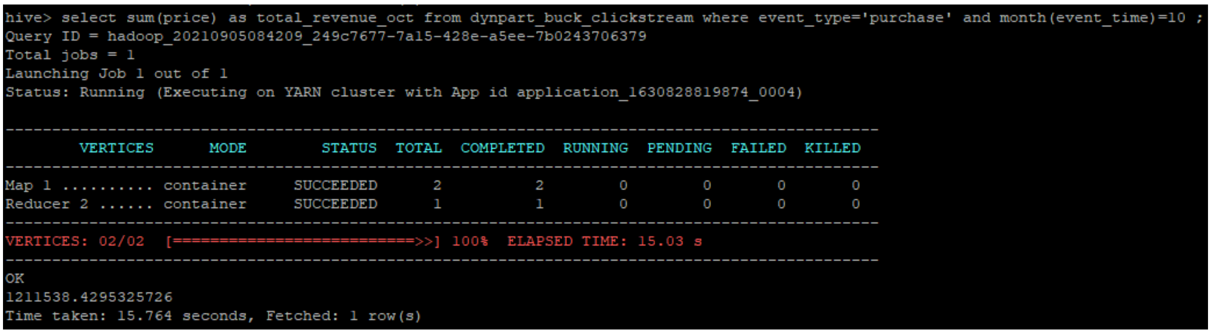
select sum(price) as total_revenue_oct from clickstream where
event_type='purchase' and month(event_time)=10 ;
and month(event_time)=10 ;



```
hive> select sum(price) as total_revenue_oct from clickstream where event_type='purchase' and month(event_time)=10 ;
Query ID = hadoop_20210905123412_c7c81d35-e0aa-460d-8895-d54970d87e71
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1630842684204_0005)

----------------------------------------------------------------------------------------------
        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: 53.59 s
----------------------------------------------------------------------------------------------
OK
1211538.4299997438
Time taken: 54.549 seconds, Fetched: 1 row(s)
```

b) *Optimized query:*

select sum(price) as total_revenue_oct from dynpart_buck_clickstream where
event_type='purchase' and month(event_time)=10 ;



```
hive> select sum(price) as total_revenue_oct from dynpart_buck_clickstream where event_type='purchase' and month(event_time)=10 ;
Query ID = hadoop_20210905084209_249c7677-7a15-428e-a5ee-7b0243706379
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1630828819874_0004)

----------------------------------------------------------------------------------------------
        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: 15.03 s
----------------------------------------------------------------------------------------------
OK
1211538.4295325726
Time taken: 15.764 seconds, Fetched: 1 row(s)
```

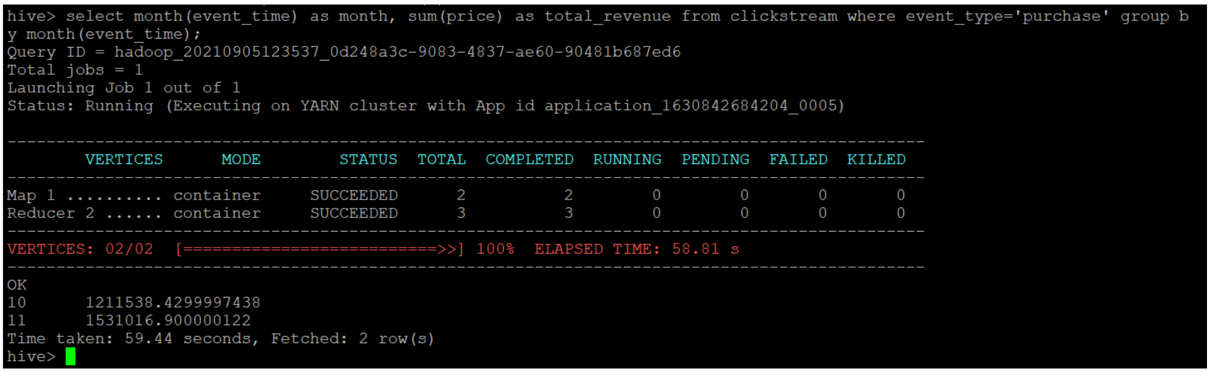
Optimized query using dynpart_buck_clickstream table

*The total revenue generated in October is **1211538.429.** Optimized query took **15.764 secs** while unoptimized query took **54.549 secs** to fetch the same result.*

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

a) *Unoptimized query:*

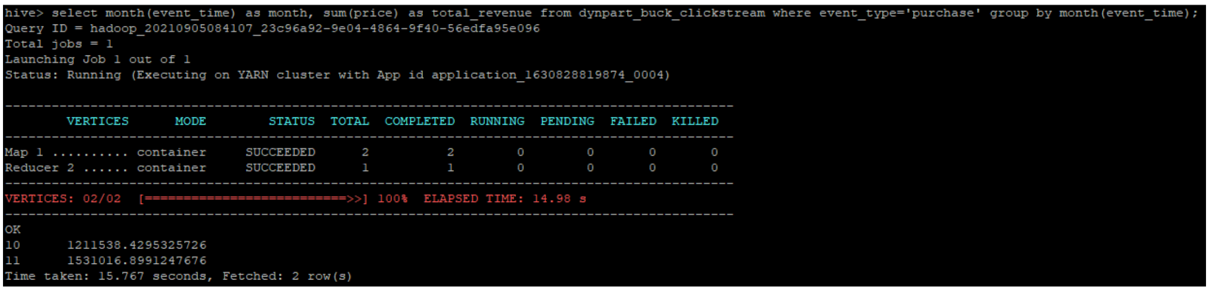
select month(event_time) as month, sum(price) as total_revenue from clickstream where event_type='purchase' group by month(event_time);



```
hive> select month(event_time) as month, sum(price) as total_revenue from clickstream where event_type='purchase' group b
y month(event_time);
Query ID = hadoop_20210905123537_0d248a3c-9083-4837-ae60-90481b687ed6
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1630842684204_0005)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      2          2        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      3          3        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 58.81 s
----------------------------------------------------------------------------------------------
OK
10      1211538.4299997438
11      1531016.900000122
Time taken: 59.44 seconds, Fetched: 2 row(s)
hive>
```

b) *Optimized query:*

select month(event_time) as month, sum(price) as total_revenue from dynpart_buck_clickstream where event_type='purchase' group by month(event_time);



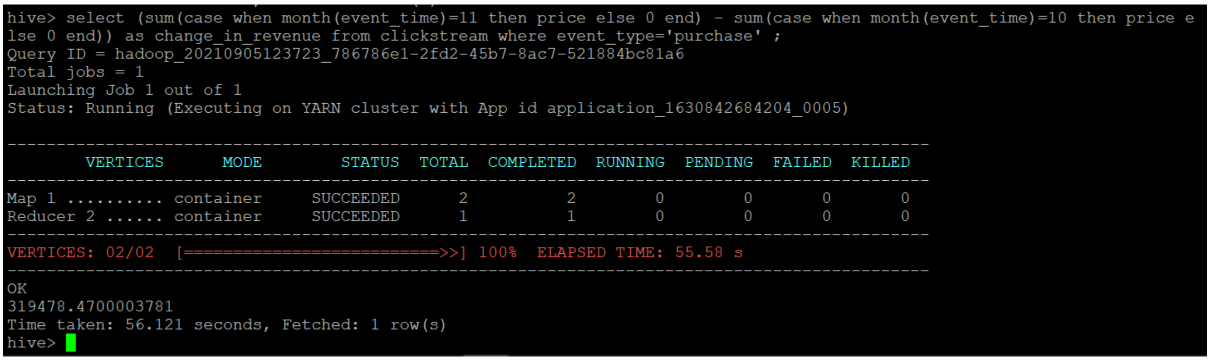
```
hive> select month(event_time) as month, sum(price) as total_revenue from dynpart_buck_clickstream where event_type='purchase' group by month(event_time);
Query ID = hadoop_20210905084107_23c96a92-9e04-4864-9f40-56edfa95e096
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1630828819874_0004)

----------------------------------------------------------------------------------------------
        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: 14.98 s
----------------------------------------------------------------------------------------------
OK
10      1211538.4295325726
11      1531016.8991247676
Time taken: 15.767 seconds, Fetched: 2 row(s)
```

Optimized query using dynpart_buck_clickstream table

*Total sum of purchases for **October** is **1211538.429** while for the **November** it's **1531016.899.** Optimized query took **15.767 secs** while unoptimized query took **59.44 secs**.*

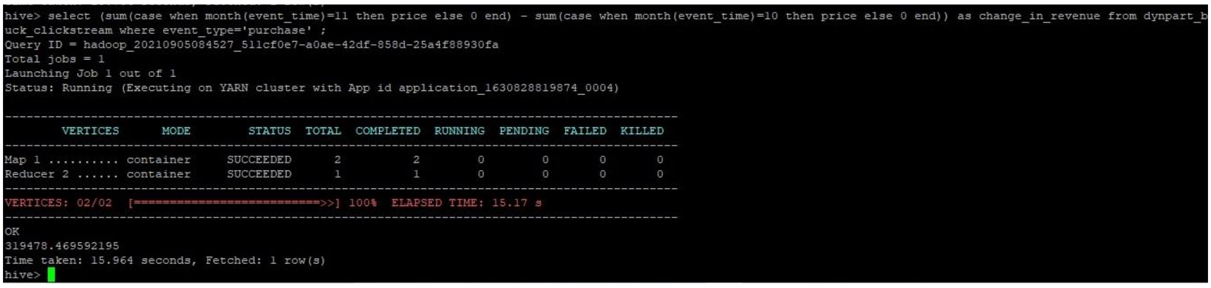
3. Write a query to find the change in revenue due to purchases from October to November.

a) *<u>Unoptimized query:</u>*
select (sum(case when month(event_time)=11 then price else 0 end) - sum(case when month(event_time)=10 then price else 0 end)) as change_in_revenue from clickstream where event_type='purchase' ;



```
hive> select (sum(case when month(event_time)=11 then price else 0 end) - sum(case when month(event_time)=10 then price e
lse 0 end)) as change_in_revenue from clickstream where event_type='purchase' ;
Query ID = hadoop_20210905123723_786786e1-2fd2-45b7-8ac7-521884bc81a6
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1630842684204_0005)

----------------------------------------------------------------------------------------------
        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: 55.58 s
----------------------------------------------------------------------------------------------
OK
319478.4700003781
Time taken: 56.121 seconds, Fetched: 1 row(s)
hive>
```

b) *<u>Optimized query:</u>*
select (sum(case when month(event_time)=11 then price else 0 end) - sum(case when month(event_time)=10 then price else 0 end)) as change_in_revenue from dynpart_buck_clickstream where event_type='purchase' ;



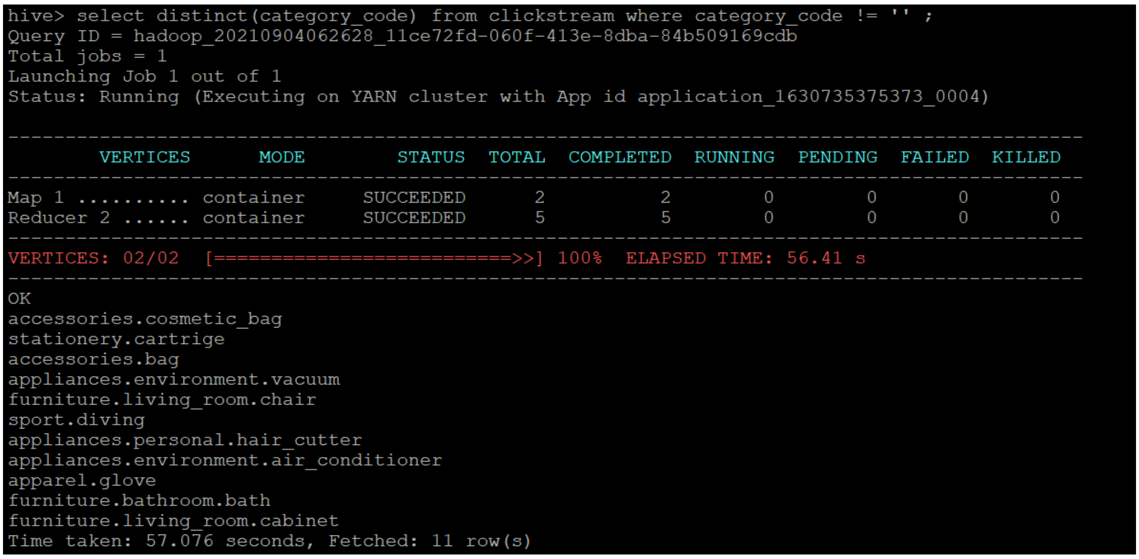
```
hive> select (sum(case when month(event_time)=11 then price else 0 end) - sum(case when month(event_time)=10 then price else 0 end)) as change_in_revenue from dynpart_b
uck_clickstream where event_type='purchase' ;
Query ID = hadoop_20210905084527_511cf0e7-a0ae-42df-858d-25a4f88930fa
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1630828819874_0004)

----------------------------------------------------------------------------------------------
        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: 15.17 s
----------------------------------------------------------------------------------------------
OK
319478.469592195
Time taken: 15.964 seconds, Fetched: 1 row(s)
hive>
```

Optimized query using dynpart_buck_clickstream table

*Change in revenue is **319478.469**. Optimized query took **15.964 secs** while unoptimized query took **56.121 secs**.*

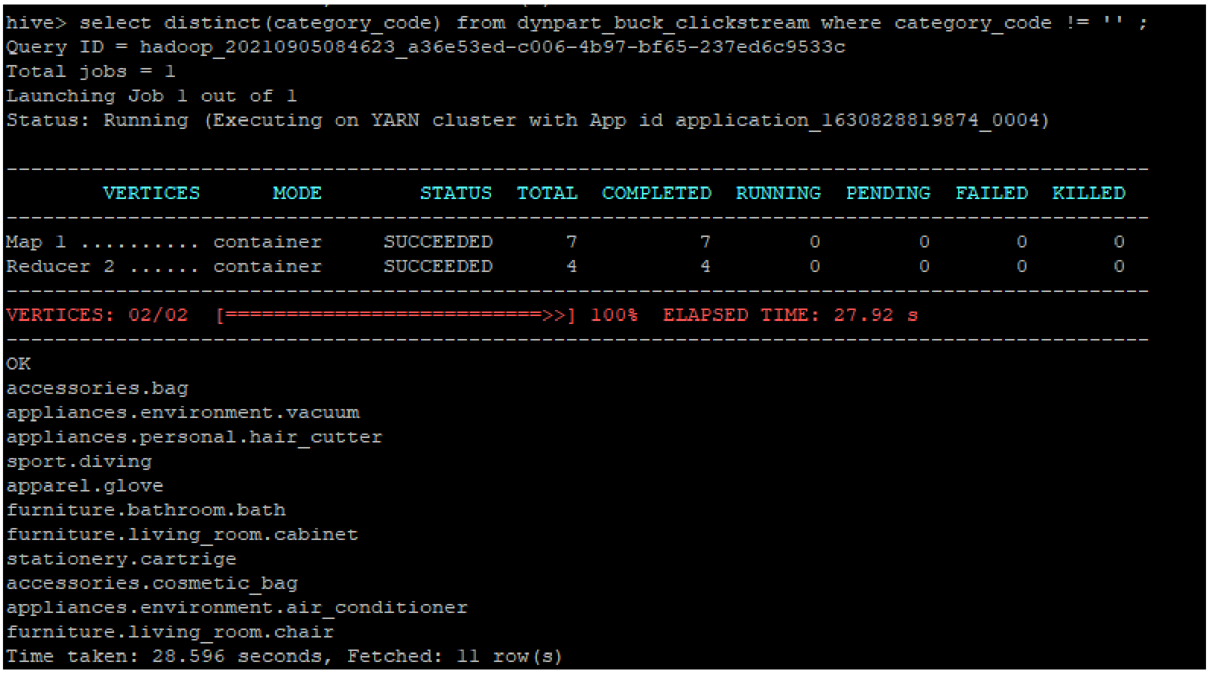
4. **Find distinct categories of products. Categories with null category code can be ignored.**

a) *<u>Unoptimized query:</u>*
select distinct(category_code) from clickstream where category_code != '' ;



```
hive> select distinct(category_code) from clickstream where category_code != '' ;
Query ID = hadoop_20210904062628_11ce72fd-060f-413e-8dba-84b509169cdb
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1630735375373_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: 56.41 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: 57.076 seconds, Fetched: 11 row(s)
```

b) *<u>Optimized query:</u>*
select distinct(category_code) from dynpart_buck_clickstream where category_code != '' ;



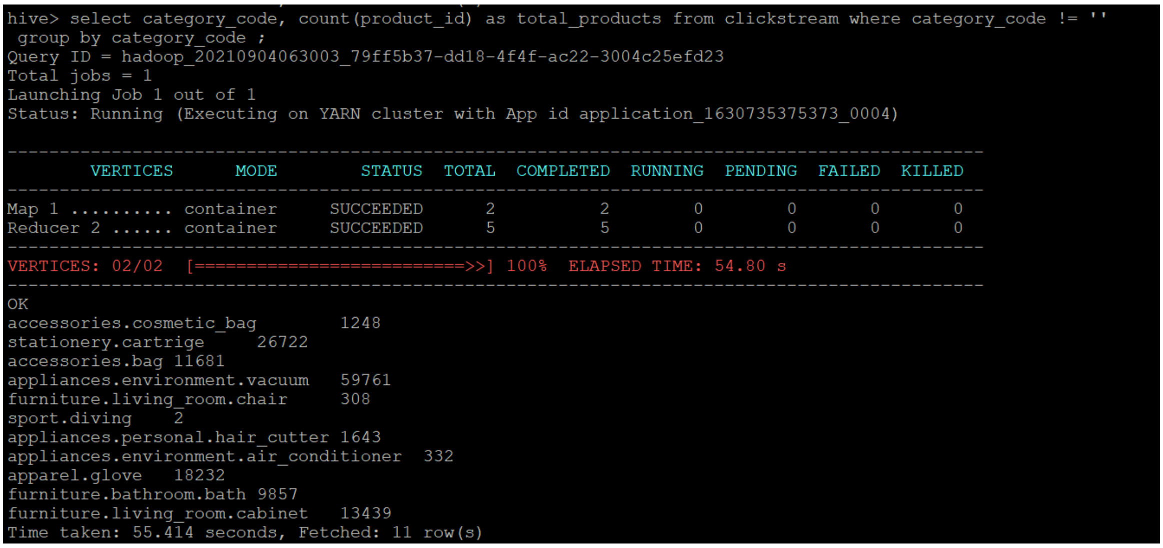
```
hive> select distinct(category_code) from dynpart_buck_clickstream where category_code != '' ;
Query ID = hadoop_20210905084623_a36e53ed-c006-4b97-bf65-237ed6c9533c
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1630828819874_0004)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      7          7        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      4          4        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 27.92 s
----------------------------------------------------------------------------------------------
OK
accessories.bag
appliances.environment.vacuum
appliances.personal.hair_cutter
sport.diving
apparel.glove
furniture.bathroom.bath
furniture.living_room.cabinet
stationery.cartrige
accessories.cosmetic_bag
appliances.environment.air_conditioner
furniture.living_room.chair
Time taken: 28.596 seconds, Fetched: 11 row(s)
```

Optimized query using dynpart_buck_clickstream table

*There are **11 distinct categories**. Optimized query took **28.596 secs** while unoptimized query took **57.076 secs.***

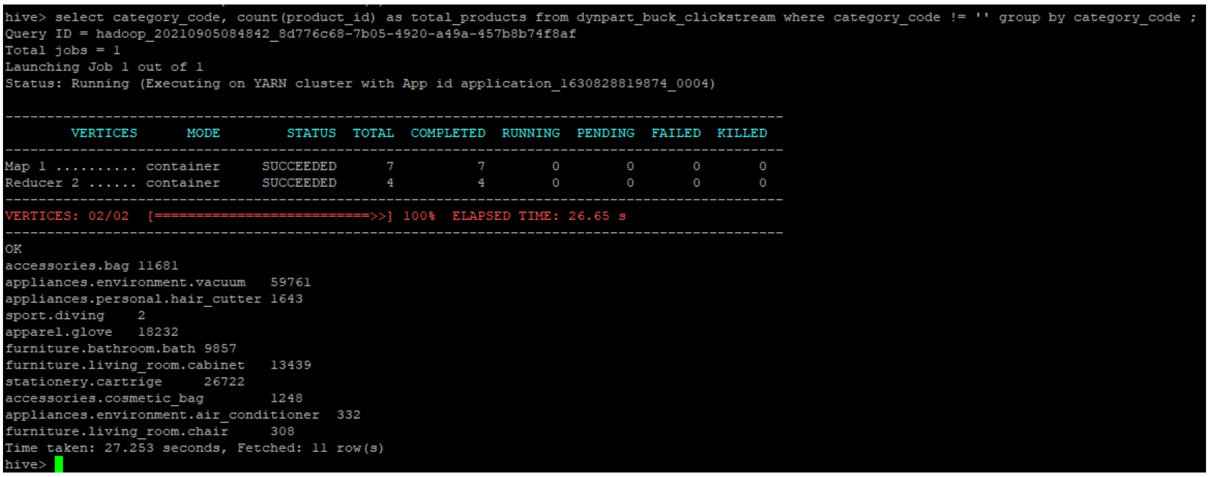
5. **Find the total number of products available under each category.**

a) *<u>Unoptimized query:</u>*
select category_code, count(product_id) as total_products from clickstream where category_code != '' group by category_code ;



```
hive> select category_code, count(product_id) as total_products from clickstream where category_code != ''
 group by category_code ;
Query ID = hadoop_20210904063003_79ff5b37-dd18-4f4f-ac22-3004c25efd23
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1630735375373_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: 54.80 s
----------------------------------------------------------------------------------------------
OK
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: 55.414 seconds, Fetched: 11 row(s)
```

b) *<u>Optimized query:</u>*
select category_code, count(product_id) as total_products from dynpart_buck_clickstream where category_code != '' group by category_code ;



```
hive> select category_code, count(product_id) as total_products from dynpart_buck_clickstream where category_code != '' group by category_code ;
Query ID = hadoop_20210905084842_8d776c68-7b05-4920-a49a-457b8b74f8af
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1630828819874_0004)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      7          7        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      4          4        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 26.65 s
----------------------------------------------------------------------------------------------
OK
accessories.bag 11681
appliances.environment.vacuum   59761
appliances.personal.hair_cutter 1643
sport.diving    2
apparel.glove   18232
furniture.bathroom.bath 9857
furniture.living_room.cabinet   13439
stationery.cartrige     26722
accessories.cosmetic_bag        1248
appliances.environment.air_conditioner  332
furniture.living_room.chair     308
Time taken: 27.253 seconds, Fetched: 11 row(s)
hive>
```

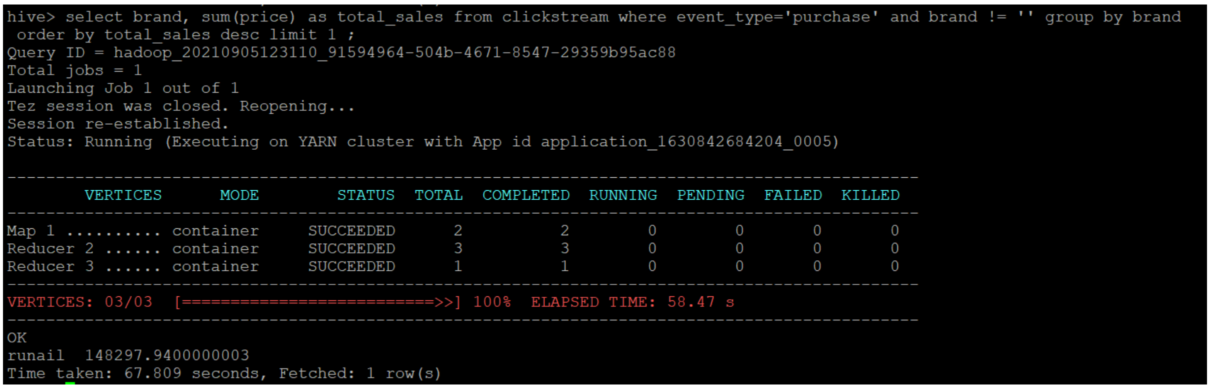
Optimized query using dynpart_buck_clickstream table

*Optimized query took **27.253 secs** while unoptimized query took **55.414 secs**.*

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

a) *<u>Unoptimized query:</u>*

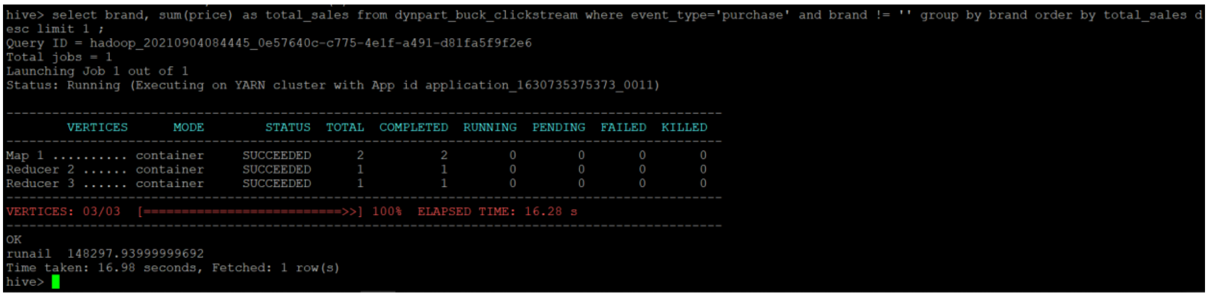
select brand, sum(price) as total_sales from clickstream where event_type='purchase' and brand != '' group by brand order by total_sales desc limit 1 ;



```
hive> select brand, sum(price) as total_sales from clickstream where event_type='purchase' and brand != '' group by brand
 order by total_sales desc limit 1 ;
Query ID = hadoop_20210905123110_91594964-504b-4671-8547-29359b95ac88
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_1630842684204_0005)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      2          2        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      3          3        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 58.47 s
----------------------------------------------------------------------------------------------
OK
runail  148297.9400000003
Time taken: 67.809 seconds, Fetched: 1 row(s)
```

b) *<u>Optimized query:</u>*

select brand, sum(price) as total_sales from dynpart_buck_clickstream where event_type='purchase' and brand != '' group by brand order by total_sales desc limit 1 ;



```
hive> select brand, sum(price) as total_sales from dynpart_buck_clickstream where event_type='purchase' and brand != '' group by brand order by total_sales d
esc limit 1 ;
Query ID = hadoop_20210904084445_0e57640c-c775-4e1f-a491-d81fa5f9f2e6
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1630735375373_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
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 16.28 s
----------------------------------------------------------------------------------------------
OK
runail  148297.93999999692
Time taken: 16.98 seconds, Fetched: 1 row(s)
hive>
```

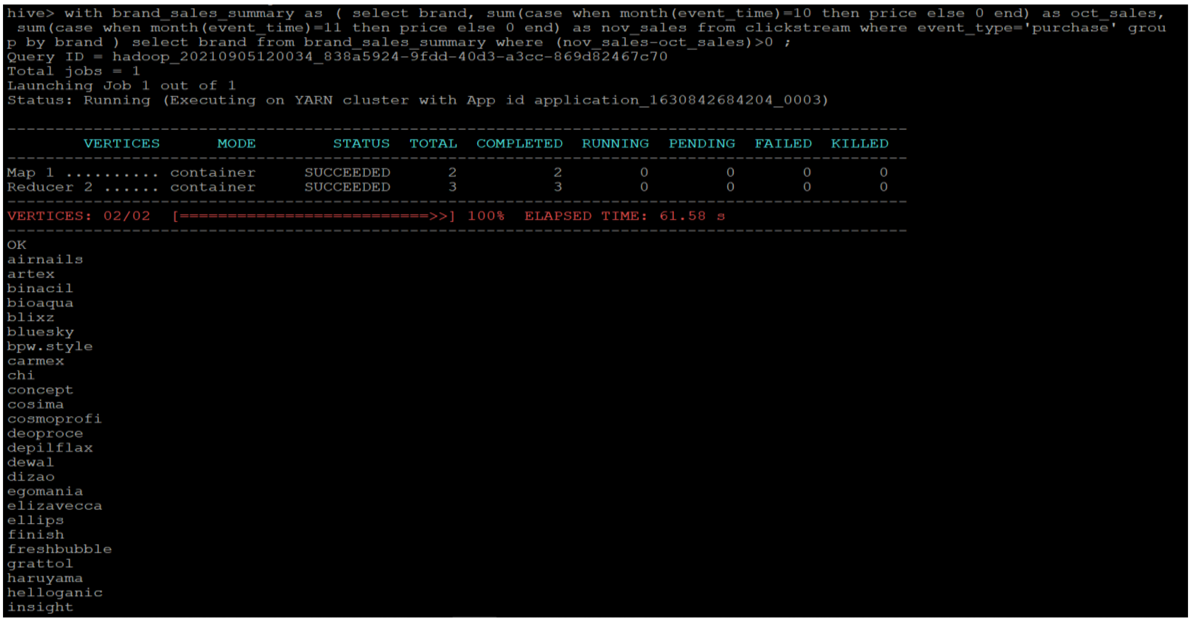
Optimized query using dynpart_buck_clickstream table

*The top brand is **runail** with **total_sales 148297.939** & to show this optimized query took **16.98 secs** while unoptimized query took **67.809 secs.***

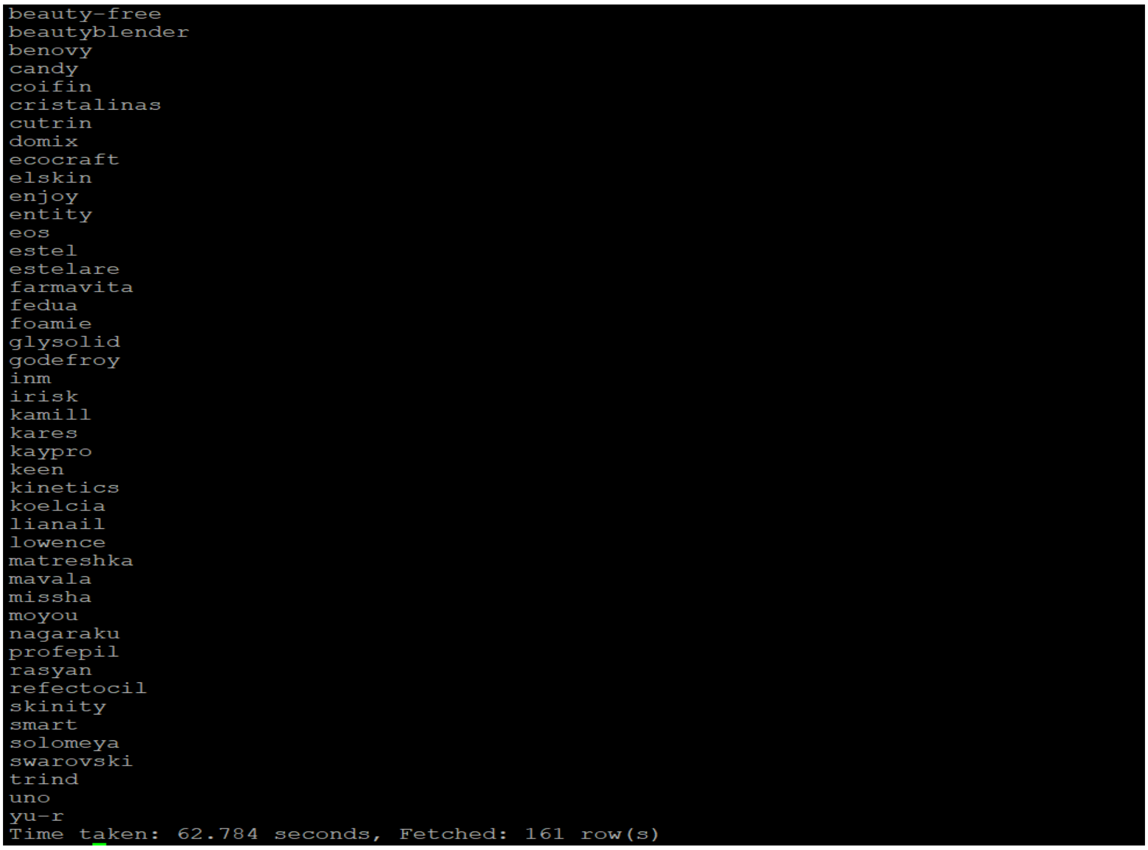
7. Which brands increases their sales from October to November?

a) *<u>Unoptimized query:</u>*

with brand_sales_summary as (select brand, sum(case when month(event_time)=10 then price else 0 end) as oct_sales, sum(case when month(event_time)=11 then price else 0 end) as nov_sales from clickstream where event_type='purchase' group by brand) select brand from brand_sales_summary where (nov_sales-oct_sales)>0 ;



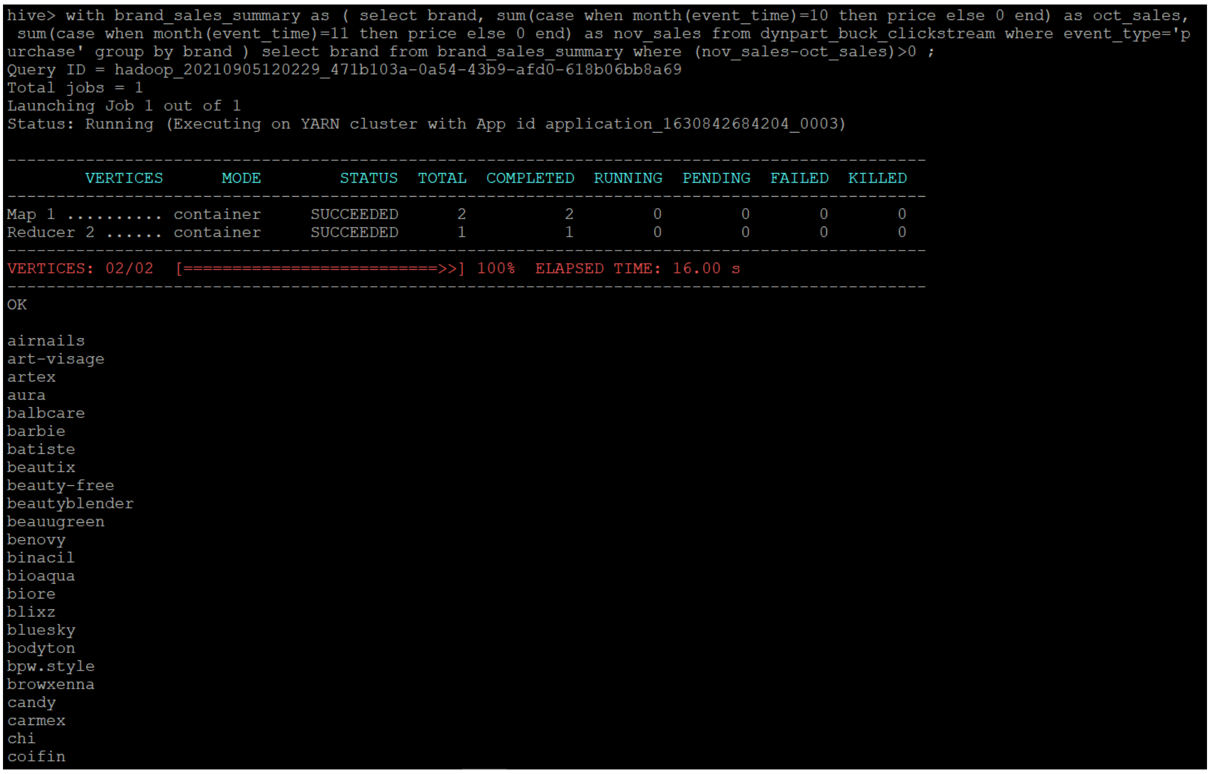
```
hive> with brand_sales_summary as ( select brand, sum(case when month(event_time)=10 then price else 0 end) as oct_sales,
 sum(case when month(event_time)=11 then price else 0 end) as nov_sales from clickstream where event_type='purchase' grou
p by brand ) select brand from brand_sales_summary where (nov_sales-oct_sales)>0 ;
Query ID = hadoop_20210905120034_838a5924-9fdd-40d3-a3cc-869d82467c70
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1630842684204_0003)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      2          2        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      3          3        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 61.58 s
----------------------------------------------------------------------------------------------
OK
airnails
artex
binacil
bioaqua
blixz
bluesky
bpw.style
carmex
chi
concept
cosima
cosmoprofi
deoproce
depilflax
dewal
dizao
egomania
elizavecca
ellips
finish
freshbubble
grattol
haruyama
helloganic
insight
```



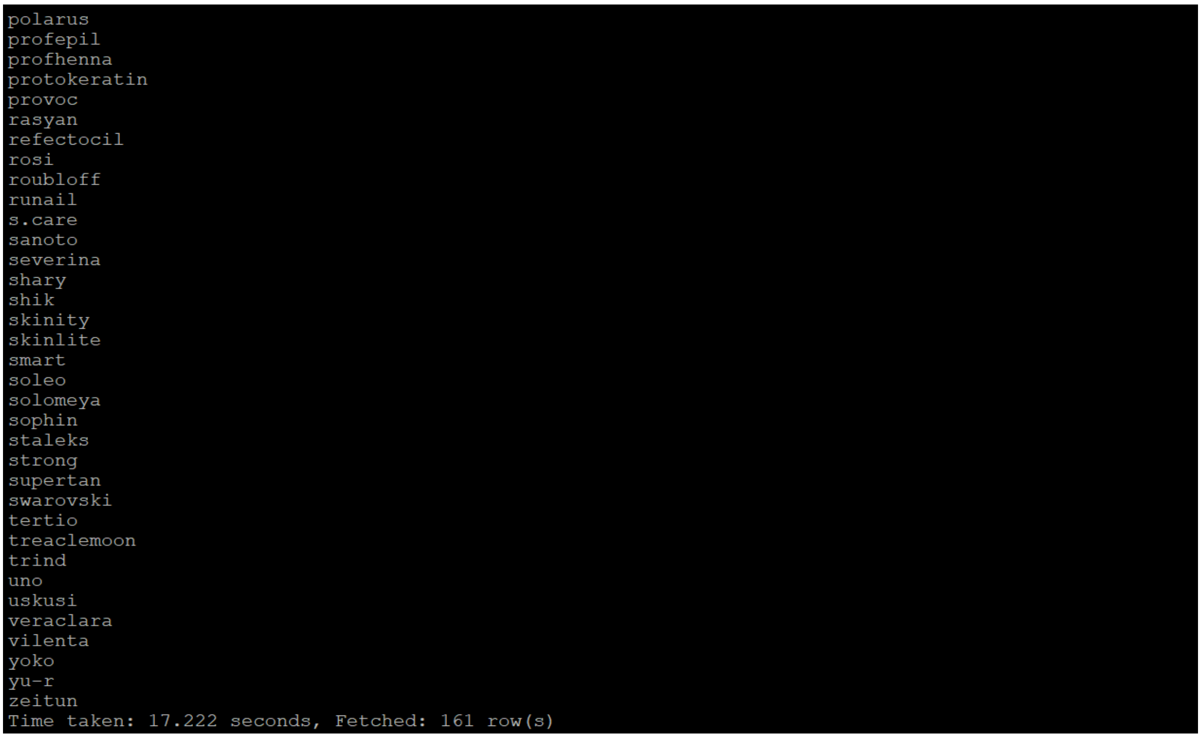
```
beauty-free
beautyblender
benovy
candy
coifin
cristalinas
cutrin
domix
ecocraft
elskin
enjoy
entity
eos
estel
estelare
farmavita
fedua
foamie
glysolid
godefroy
inm
irisk
kamill
kares
kaypro
keen
kinetics
koelcia
lianail
lowence
matreshka
mavala
missha
moyou
nagaraku
profepil
rasyan
refectocil
skinity
smart
solomeya
swarovski
trind
uno
yu-r
Time taken: 62.784 seconds, Fetched: 161 row(s)
```

b) *<u>Optimized query:</u>*

with brand_sales_summary as (select brand, sum(case when month(event_time)=10 then price else 0 end) as oct_sales, sum(case when month(event_time)=11 then price else 0 end) as nov_sales from dynpart_buck_clickstream where event_type='purchase' group by brand) select brand from brand_sales_summary where (nov_sales-oct_sales)>0 ;



```
hive> with brand_sales_summary as ( select brand, sum(case when month(event_time)=10 then price else 0 end) as oct_sales,
 sum(case when month(event_time)=11 then price else 0 end) as nov_sales from dynpart_buck_clickstream where event_type='p
urchase' group by brand ) select brand from brand_sales_summary where (nov_sales-oct_sales)>0 ;
Query ID = hadoop_20210905120229_471b103a-0a54-43b9-afd0-618b06bb8a69
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1630842684204_0003)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      2          2        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 16.00 s
----------------------------------------------------------------------------------------------
OK

airnails
art-visage
artex
aura
balbcare
barbie
batiste
beautix
beauty-free
beautyblender
beauugreen
benovy
binacil
bioaqua
biore
blixz
bluesky
bodyton
bpw.style
browxenna
candy
carmex
chi
coifin
```

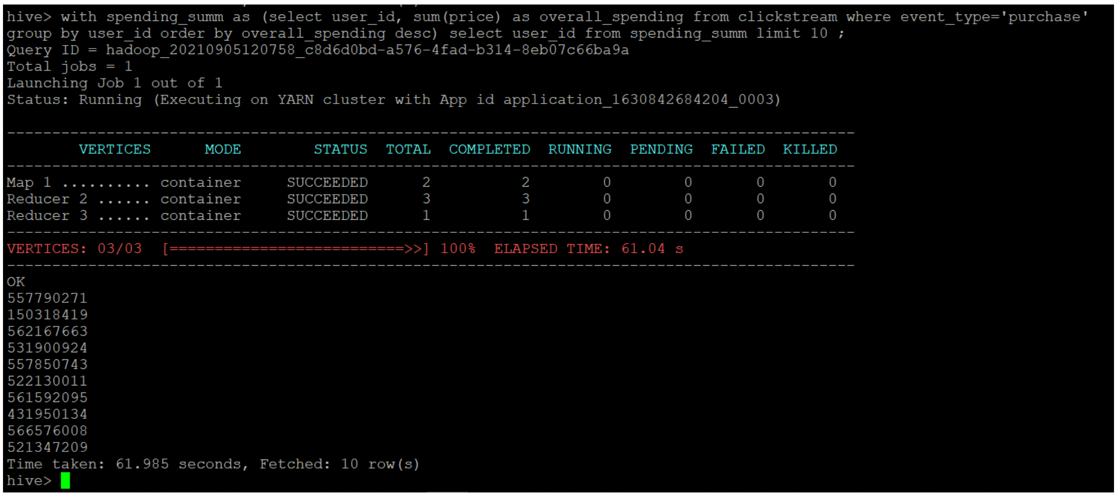


```
polarus
profepil
profhenna
protokeratin
provoc
rasyan
refectocil
rosi
roubloff
runail
s.care
sanoto
severina
shary
shik
skinity
skinlite
smart
soleo
solomeya
sophin
staleks
strong
supertan
swarovski
tertio
treaclemoon
trind
uno
uskusi
veraclara
vilenta
yoko
yu-r
zeitun
Time taken: 17.222 seconds, Fetched: 161 row(s)
```

*There is a total of **161 brands** that had increased sales from October to November. To show this optimized query took **17.222 secs** while unoptimized query took **62.784 secs.***

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.**

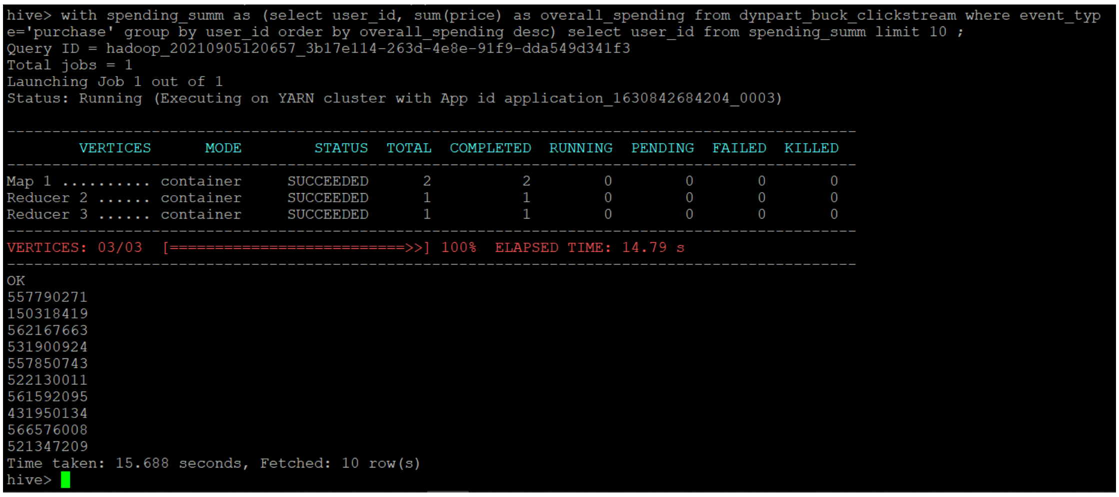
a) *<u>Unoptimized query:</u>*
with spending_summ as (select user_id, sum(price) as overall_spending from clickstream where event_type='purchase' group by user_id order by overall_spending desc) select user_id from spending_summ limit 10 ;



```
hive> with spending_summ as (select user_id, sum(price) as overall_spending from clickstream where event_type='purchase'
group by user_id order by overall_spending desc) select user_id from spending_summ limit 10 ;
Query ID = hadoop_20210905120758_c8d6d0bd-a576-4fad-b314-8eb07c66ba9a
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1630842684204_0003)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      2          2        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      3          3        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 61.04 s
----------------------------------------------------------------------------------------------
OK
557790271
150318419
562167663
531900924
557850743
522130011
561592095
431950134
566576008
521347209
Time taken: 61.985 seconds, Fetched: 10 row(s)
hive>
```

b) <u>*Optimized query:*</u>

with spending_summ as (select user_id, sum(price) as overall_spending from dynpart_buck_clickstream where event_type='purchase' group by user_id order by overall_spending desc) select user_id from spending_summ limit 10 ;



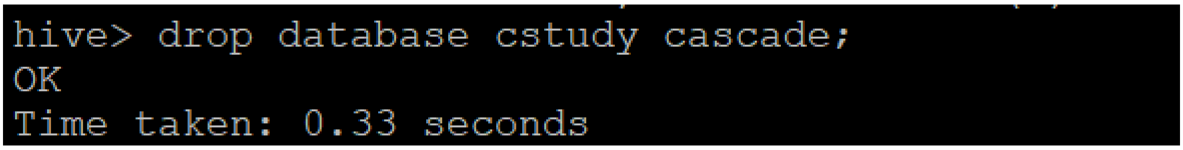
```
hive> with spending_summ as (select user_id, sum(price) as overall_spending from dynpart_buck_clickstream where event_typ
e='purchase' group by user_id order by overall_spending desc) select user_id from spending_summ limit 10 ;
Query ID = hadoop_20210905120657_3b17e114-263d-4e8e-91f9-dda549d341f3
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1630842684204_0003)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      2          2        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 14.79 s
----------------------------------------------------------------------------------------------
OK
557790271
150318419
562167663
531900924
557850743
522130011
561592095
431950134
566576008
521347209
Time taken: 15.688 seconds, Fetched: 10 row(s)
hive>
```

Optimized query using dynpart_buck_clickstream table

So above are the top 10 customers who spend the most and should be awarded with golden customer plan. To show this the optimized query took ***15.688 secs*** *while unoptimized query took* ***61.985 secs****.*

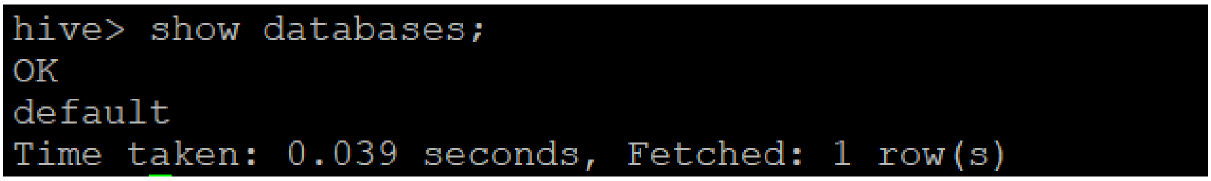
- **Dropping database**

drop database cstudy cascade;



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

show databases;

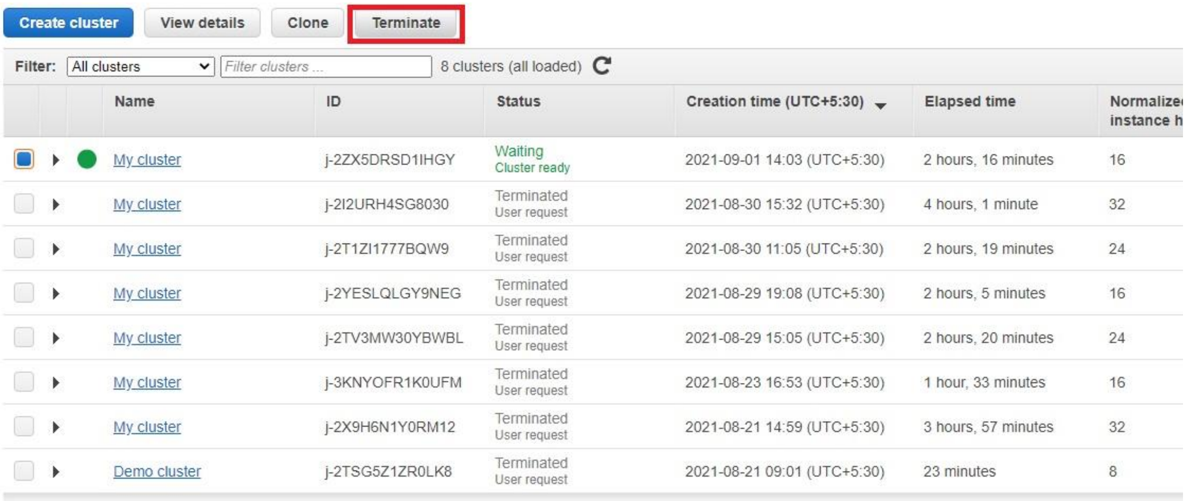


```
hive> show databases;
OK
default
Time taken: 0.039 seconds, Fetched: 1 row(s)
```

We are done with our analysis. Finally, we will move towards terminating our cluster

TERMINATING THE EMR CLUSTER

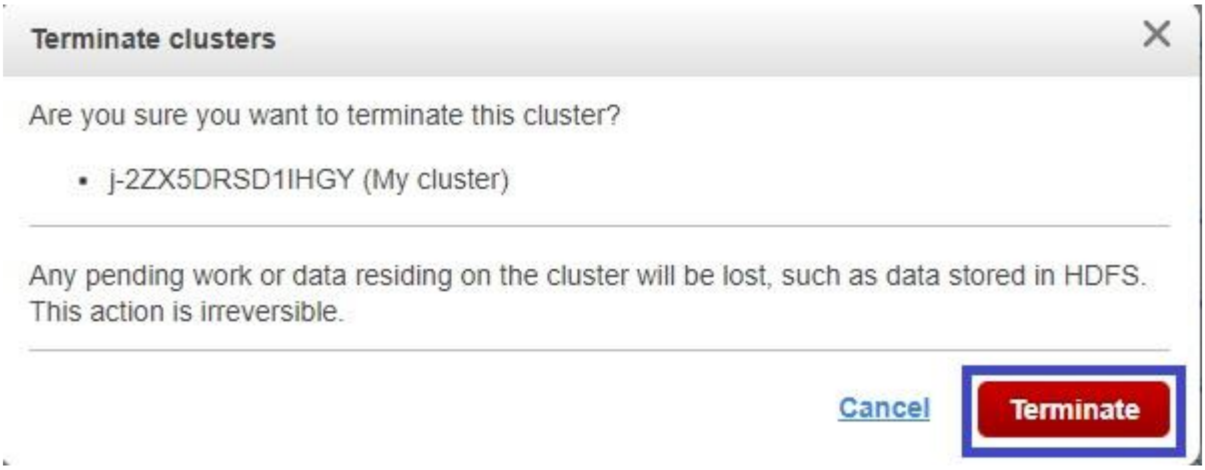
Step 1: Click on Terminate.



Filter: All clusters | Filter clusters ... | 8 clusters (all loaded)

			Name	ID	Status	Creation time (UTC+5:30)	Elapsed time	Normalize instance h
☑	▸	●	My cluster	j-2ZX5DRSD1IHGY	Waiting Cluster ready	2021-09-01 14:03 (UTC+5:30)	2 hours, 16 minutes	16
□	▸		My cluster	j-2I2URH4SG8030	Terminated User request	2021-08-30 15:32 (UTC+5:30)	4 hours, 1 minute	32
□	▸		My cluster	j-2T1ZI1777BQW9	Terminated User request	2021-08-30 11:05 (UTC+5:30)	2 hours, 19 minutes	24
□	▸		My cluster	j-2YESLQLGY9NEG	Terminated User request	2021-08-29 19:08 (UTC+5:30)	2 hours, 5 minutes	16
□	▸		My cluster	j-2TV3MW30YBWBL	Terminated User request	2021-08-29 15:05 (UTC+5:30)	2 hours, 20 minutes	24
□	▸		My cluster	j-3KNYOFR1K0UFM	Terminated User request	2021-08-23 16:53 (UTC+5:30)	1 hour, 33 minutes	16
□	▸		My cluster	j-2X9H6N1Y0RM12	Terminated User request	2021-08-21 14:59 (UTC+5:30)	3 hours, 57 minutes	32
□	▸		Demo cluster	j-2TSG5Z1ZR0LK8	Terminated User request	2021-08-21 09:01 (UTC+5:30)	23 minutes	8

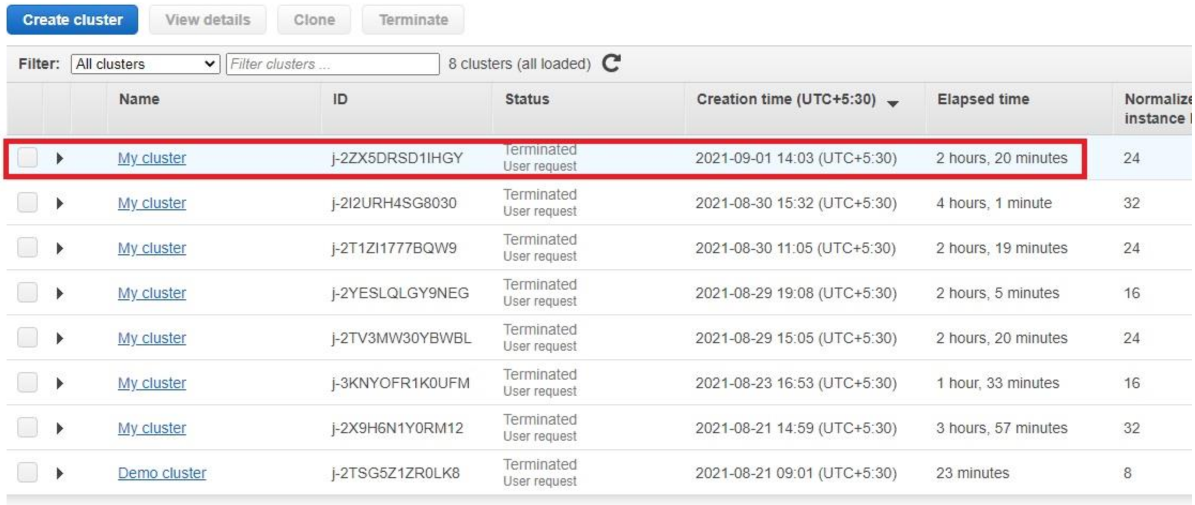
Step 2: A security/confirmation message will pop up. Click on Terminate.



Are you sure you want to terminate this cluster?

- j-2ZX5DRSD1IHGY (My cluster)

Any pending work or data residing on the cluster will be lost, such as data stored in HDFS. This action is irreversible.



		Name	ID	Status	Creation time (UTC+5:30)	Elapsed time	Normalize instance
□	▸	My cluster	j-2ZX5DRSD1IHGY	Terminated User request	2021-09-01 14:03 (UTC+5:30)	2 hours, 20 minutes	24
□	▸	My cluster	j-2I2URH4SG8030	Terminated User request	2021-08-30 15:32 (UTC+5:30)	4 hours, 1 minute	32
□	▸	My cluster	j-2T1ZI1777BQW9	Terminated User request	2021-08-30 11:05 (UTC+5:30)	2 hours, 19 minutes	24
□	▸	My cluster	j-2YESLQLGY9NEG	Terminated User request	2021-08-29 19:08 (UTC+5:30)	2 hours, 5 minutes	16
□	▸	My cluster	j-2TV3MW30YBWBL	Terminated User request	2021-08-29 15:05 (UTC+5:30)	2 hours, 20 minutes	24
□	▸	My cluster	j-3KNYOFR1K0UFM	Terminated User request	2021-08-23 16:53 (UTC+5:30)	1 hour, 33 minutes	16
□	▸	My cluster	j-2X9H6N1Y0RM12	Terminated User request	2021-08-21 14:59 (UTC+5:30)	3 hours, 57 minutes	32
□	▸	Demo cluster	j-2TSG5Z1ZR0LK8	Terminated User request	2021-08-21 09:01 (UTC+5:30)	23 minutes	8

The Cluster has been terminated. Now we can log out of our AWS account.