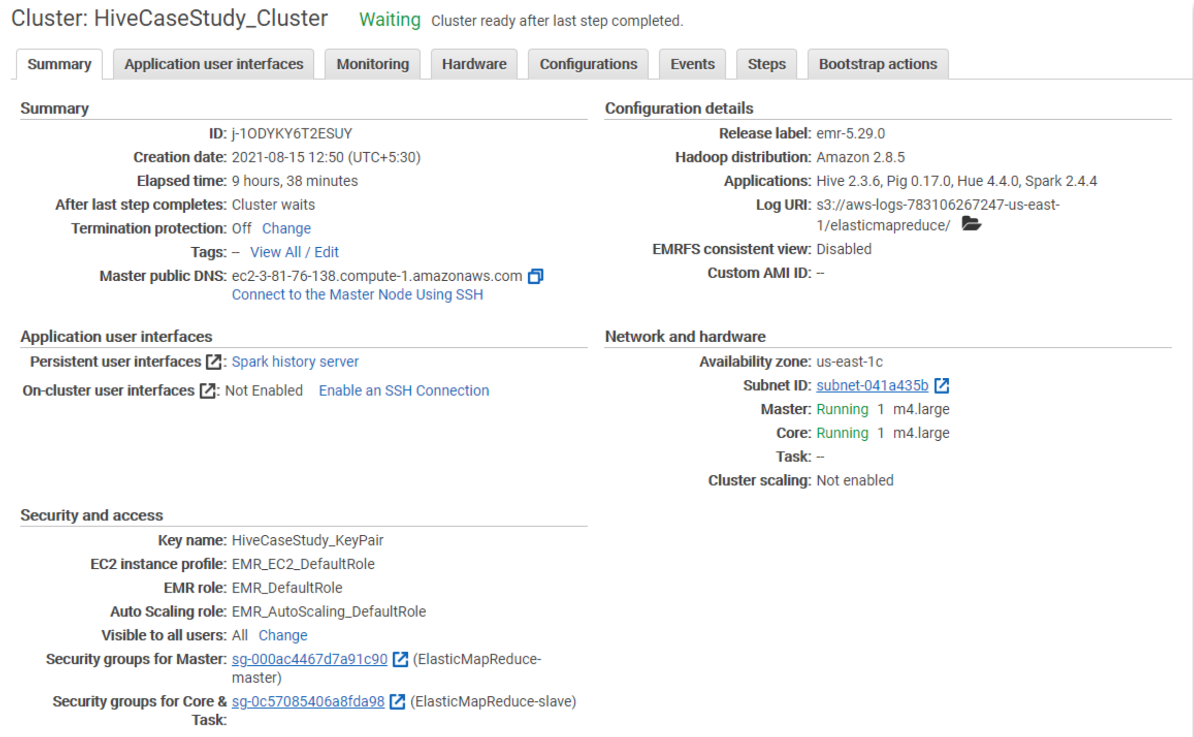
HIVE CASE STUDY

Create and Launch an EMR Cluster

- Select region as N. Virginia(us-east-1).
- Create a key pair.
- Go to AWS account and Create an EMR cluster with 1 master node and 1 core nodes having m4.large instance type.
- Select the key pair as created which was created previously.
- Enable SSH by editing an inbound rule in the master node’s security group and add SSH as port 22 to the rule.
- Open the SSH client (Putty) terminal add the IP address of hostname and put the .pem file in the user private key section.



Cluster: HiveCaseStudy_Cluster Waiting Cluster ready after last step completed.

Summary

ID: j-1ODYKY6T2ESUY
Creation date: 2021-08-15 12:50 (UTC+5:30)
Elapsed time: 9 hours, 38 minutes
After last step completes: Cluster waits
Termination protection: Off Change
Tags: -- View All / Edit
Master public DNS: ec2-3-81-76-138.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-783106267247-us-east-1/elasticmapreduce/
EMRFS consistent view: Disabled
Custom AMI ID: --

Application user interfaces

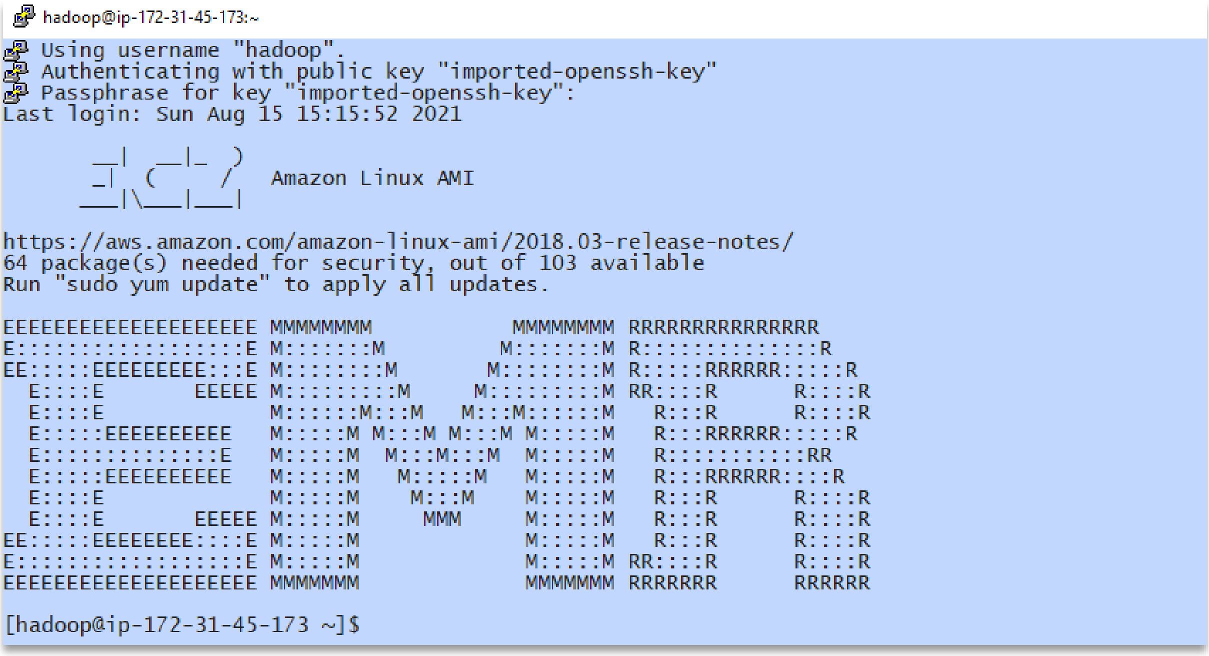
Persistent user interfaces: Spark history server
On-cluster user interfaces: Not Enabled Enable an SSH Connection

Network and hardware

Availability zone: us-east-1c
Subnet ID: subnet-041a435b
Master: Running 1 m4.large
Core: Running 1 m4.large
Task: --
Cluster scaling: Not enabled

Security and access

Key name: HiveCaseStudy_KeyPair
EC2 instance profile: EMR_EC2_DefaultRole
EMR role: EMR_DefaultRole
Auto Scaling role: EMR_AutoScaling_DefaultRole
Visible to all users: All Change
Security groups for Master: sg-000ac4467d7a91c90 (ElasticMapReduce-master)
Security groups for Core & Task: sg-0c57085406a8fda98 (ElasticMapReduce-slave)

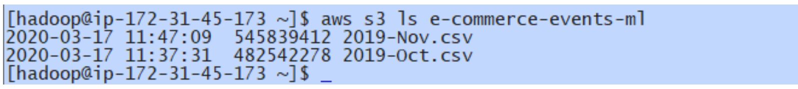


hadoop@ip-172-31-45-173:~

```
Using username "hadoop".
Authenticating with public key "imported-openssh-key"
Passphrase for key "imported-openssh-key":
Last login: Sun Aug 15 15:15:52 2021

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

https://aws.amazon.com/amazon-linux-ami/2018.03-release-notes/
64 package(s) needed for security, out of 103 available
Run "sudo yum update" to apply all updates.

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

[hadoop@ip-172-31-45-173 ~]$
```

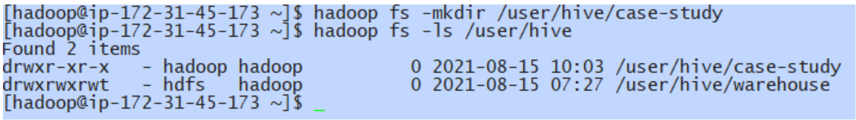
Move the data from the S3 bucket into the HDFS

- To access the public s3 bucket.
  aws s3 ls e-commerce-events-ml



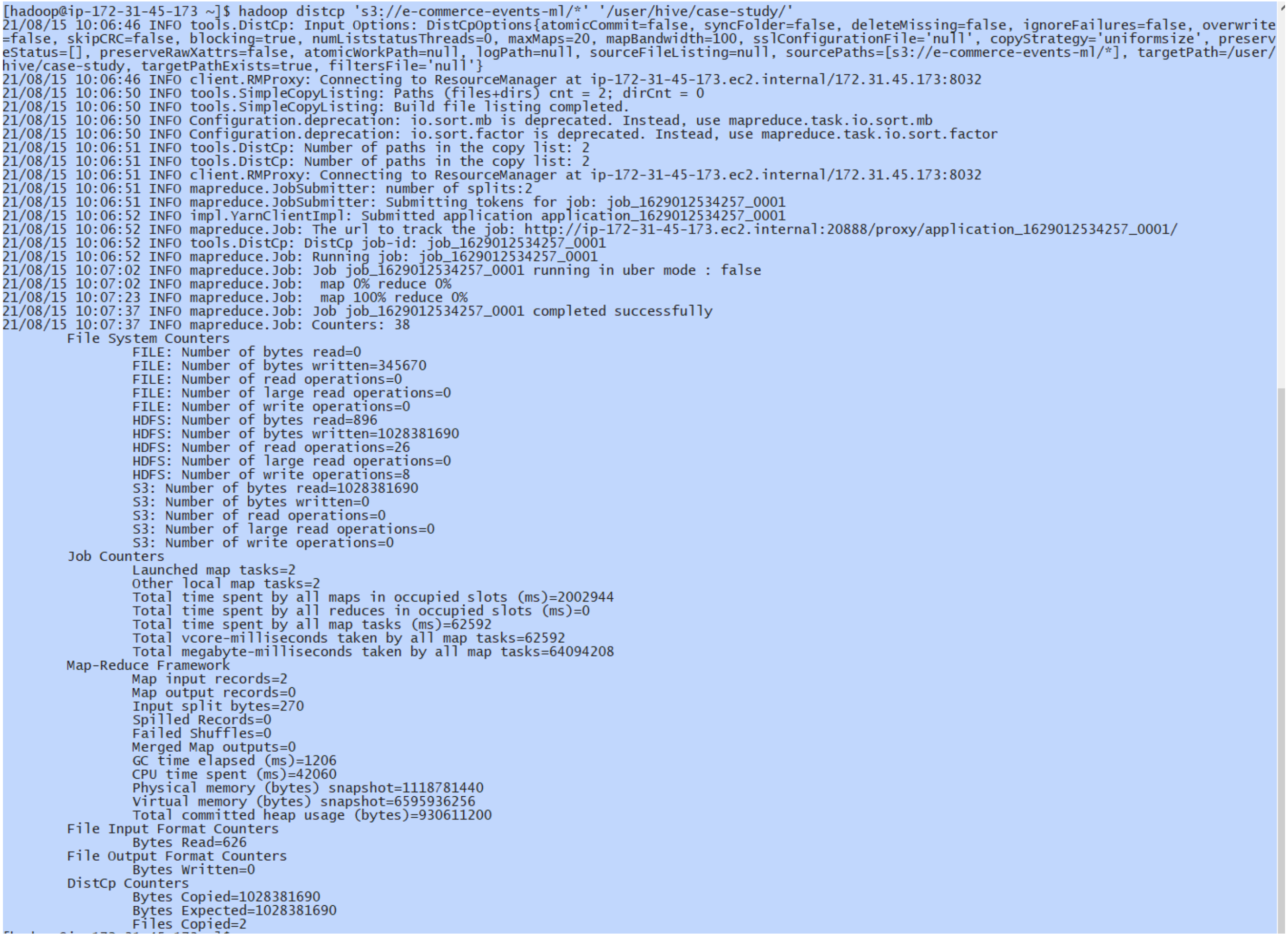
```
[hadoop@ip-172-31-45-173 ~]$ aws s3 ls e-commerce-events-ml
2020-03-17 11:47:09  545839412 2019-Nov.csv
2020-03-17 11:37:31  482542278 2019-Oct.csv
[hadoop@ip-172-31-45-173 ~]$ _
```

- Creating a directory in HDFS.
  hadoop fs -mkdir /user/hive/case-study



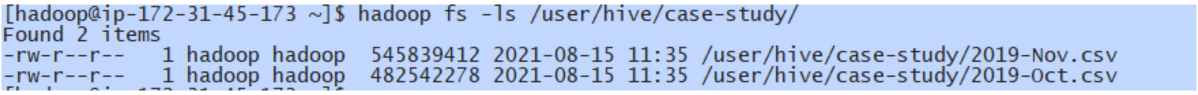
```
[hadoop@ip-172-31-45-173 ~]$ hadoop fs -mkdir /user/hive/case-study
[hadoop@ip-172-31-45-173 ~]$ hadoop fs -ls /user/hive
Found 2 items
drwxr-xr-x   - hadoop hadoop          0 2021-08-15 10:03 /user/hive/case-study
drwxrwxrwt   - hdfs   hadoop          0 2021-08-15 07:27 /user/hive/warehouse
[hadoop@ip-172-31-45-173 ~]$ _
```

- Loading the s3 public dataset to created directory "*case-study*" in hadoop.
  hadoop distcp 's3://e-commerce-events-ml/' '/user/hive/case-study/'*



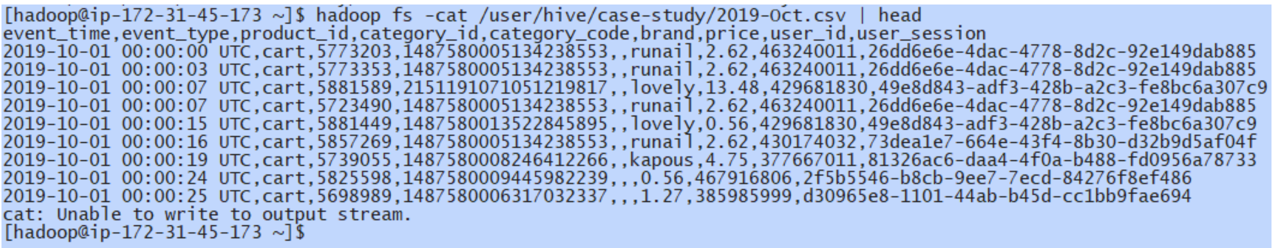
```
[hadoop@ip-172-31-45-173 ~]$ hadoop distcp 's3://e-commerce-events-ml/*' '/user/hive/case-study/'
21/08/15 10:06:46 INFO tools.DistCp: Input Options: DistCpOptions{atomicCommit=false, syncFolder=false, deleteMissing=false, ignoreFailures=false, overwrite
=false, skipCRC=false, blocking=true, numListstatusThreads=0, maxMaps=20, mapBandwidth=100, sslConfigurationFile='null', copyStrategy='uniformsize', preserv
eStatus=[], preserveRawXattrs=false, atomicWorkPath=null, logPath=null, sourceFileListing=null, sourcePaths=[s3://e-commerce-events-ml/*], targetPath=/user/
hive/case-study, targetPathExists=true, filtersFile='null'}
21/08/15 10:06:46 INFO client.RMProxy: Connecting to ResourceManager at ip-172-31-45-173.ec2.internal/172.31.45.173:8032
21/08/15 10:06:50 INFO tools.SimpleCopyListing: Paths (files+dirs) cnt = 2; dirCnt = 0
21/08/15 10:06:50 INFO tools.SimpleCopyListing: Build file listing completed.
21/08/15 10:06:50 INFO Configuration.deprecation: io.sort.mb is deprecated. Instead, use mapreduce.task.io.sort.mb
21/08/15 10:06:50 INFO Configuration.deprecation: io.sort.factor is deprecated. Instead, use mapreduce.task.io.sort.factor
21/08/15 10:06:51 INFO tools.DistCp: Number of paths in the copy list: 2
21/08/15 10:06:51 INFO tools.DistCp: Number of paths in the copy list: 2
21/08/15 10:06:51 INFO client.RMProxy: Connecting to ResourceManager at ip-172-31-45-173.ec2.internal/172.31.45.173:8032
21/08/15 10:06:51 INFO mapreduce.JobSubmitter: number of splits:2
21/08/15 10:06:51 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1629012534257_0001
21/08/15 10:06:52 INFO impl.YarnClientImpl: Submitted application application_1629012534257_0001
21/08/15 10:06:52 INFO mapreduce.Job: The url to track the job: http://ip-172-31-45-173.ec2.internal:20888/proxy/application_1629012534257_0001/
21/08/15 10:06:52 INFO tools.DistCp: DistCp job-id: job_1629012534257_0001
21/08/15 10:06:52 INFO mapreduce.Job: Running job: job_1629012534257_0001
21/08/15 10:07:02 INFO mapreduce.Job: Job job_1629012534257_0001 running in uber mode : false
21/08/15 10:07:02 INFO mapreduce.Job:  map 0% reduce 0%
21/08/15 10:07:23 INFO mapreduce.Job:  map 100% reduce 0%
21/08/15 10:07:37 INFO mapreduce.Job: Job job_1629012534257_0001 completed successfully
21/08/15 10:07:37 INFO mapreduce.Job: Counters: 38
        File System Counters
                FILE: Number of bytes read=0
                FILE: Number of bytes written=345670
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=896
                HDFS: Number of bytes written=1028381690
                HDFS: Number of read operations=26
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=8
                S3: Number of bytes read=1028381690
                S3: Number of bytes written=0
                S3: Number of read operations=0
                S3: Number of large read operations=0
                S3: Number of write operations=0
        Job Counters
                Launched map tasks=2
                Other local map tasks=2
                Total time spent by all maps in occupied slots (ms)=2002944
                Total time spent by all reduces in occupied slots (ms)=0
                Total time spent by all map tasks (ms)=62592
                Total vcore-milliseconds taken by all map tasks=62592
                Total megabyte-milliseconds taken by all map tasks=64094208
        Map-Reduce Framework
                Map input records=2
                Map output records=0
                Input split bytes=270
                Spilled Records=0
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=1206
                CPU time spent (ms)=42060
                Physical memory (bytes) snapshot=1118781440
                Virtual memory (bytes) snapshot=6595936256
                Total committed heap usage (bytes)=930611200
        File Input Format Counters
                Bytes Read=626
        File Output Format Counters
                Bytes Written=0
        DistCp Counters
                Bytes Copied=1028381690
                Bytes Expected=1028381690
                Files Copied=2
```

- After loading the dataset, Checking the dataset files and dataset in the hadoop directory.
  hadoop fs -ls /user/hive/case-study/



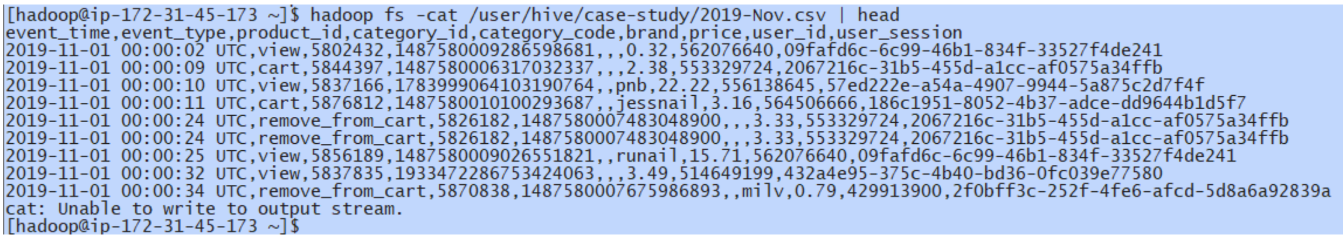
```
[hadoop@ip-172-31-45-173 ~]$ hadoop fs -ls /user/hive/case-study/
Found 2 items
-rw-r--r--   1 hadoop hadoop  545839412 2021-08-15 11:35 /user/hive/case-study/2019-Nov.csv
-rw-r--r--   1 hadoop hadoop  482542278 2021-08-15 11:35 /user/hive/case-study/2019-Oct.csv
```

hadoop fs -cat /user/hive/case-study/2019-Oct.csv | head



```
[hadoop@ip-172-31-45-173 ~]$ hadoop fs -cat /user/hive/case-study/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-172-31-45-173 ~]$
```

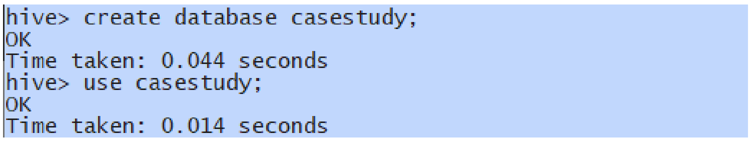
hadoop fs -cat /user/hive/case-study/2019-Nov.csv | head



```
[hadoop@ip-172-31-45-173 ~]$ hadoop fs -cat /user/hive/case-study/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-172-31-45-173 ~]$
```

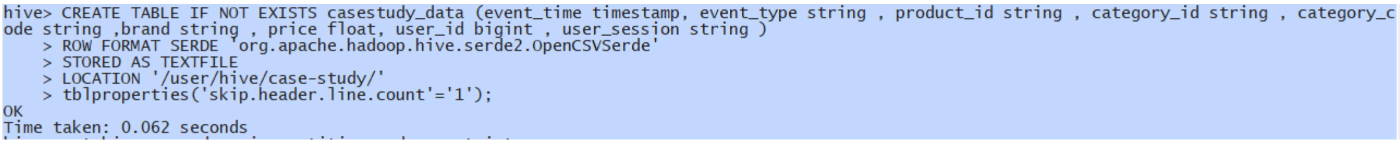
Creating the database and tables to launch Hive queries on EMR cluster

- Create the Database



```
hive> create database casestudy;
OK
Time taken: 0.044 seconds
hive> use casestudy;
OK
Time taken: 0.014 seconds
```

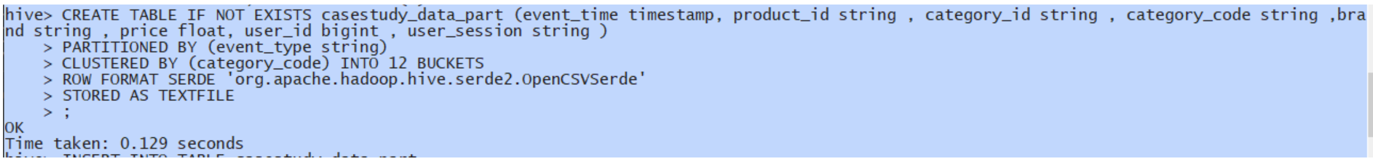
- Created the base table(*casestudy_data*) and check for the data in the table.
  CREATE TABLE IF NOT EXISTS casestudy_data (event_time timestamp, event_type string , product_id string , category_id string , category_code string ,brand string , price float, user_id bigint , user_session string)
  ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.OpenCSVSerde'
  STORED AS TEXTFILE
  LOCATION '/user/hive/case-study/'
  tblproperties('skip.header.line.count'='1');



```
hive> CREATE TABLE IF NOT EXISTS casestudy_data (event_time timestamp, event_type string , product_id string , category_id string , category_c
ode 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/case-study/'
    > tblproperties('skip.header.line.count'='1');
OK
Time taken: 0.062 seconds
```

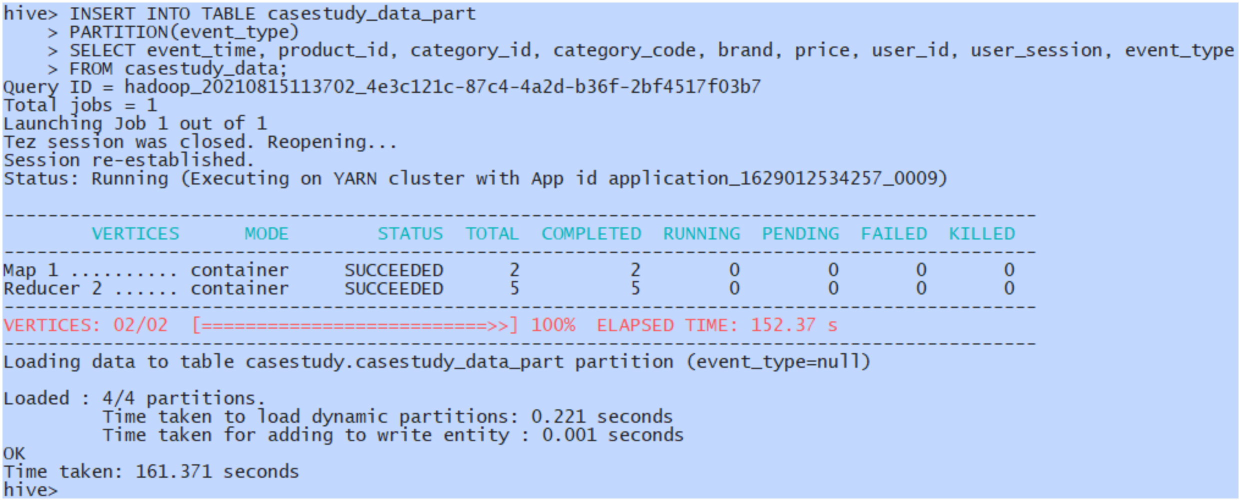
- Once the base table is created, Optimize the table for quick query result through partitioning and bucketing. Our optimized table name is *casestudy_data_part*.
  set hive.exec.dynamic.partition.mode=nonstrict;
  set hive.exec.dynamic.partition.mode=true;
  set hive.enforce.bucketing=true;

CREATE TABLE IF NOT EXISTS casestudy_data_part2 (event_time timestamp, 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 12 BUCKETS
ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.OpenCSVSerde'
STORED AS TEXTFILE;



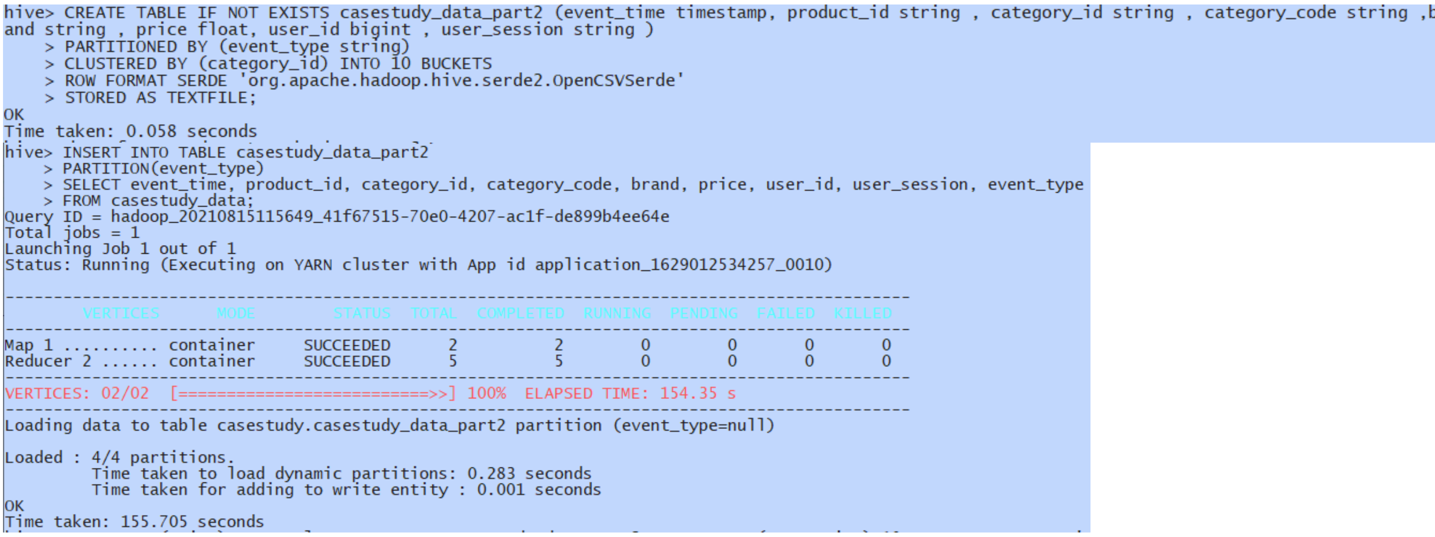
```
hive> CREATE TABLE IF NOT EXISTS casestudy_data_part (event_time timestamp, product_id string , category_id string , category_code string ,bra
nd string , price float, user_id bigint , user_session string )
    > PARTITIONED BY (event_type string)
    > CLUSTERED BY (category_code) INTO 12 BUCKETS
    > ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.OpenCSVSerde'
    > STORED AS TEXTFILE
    > ;
OK
Time taken: 0.129 seconds
```

- Loading the data into optimize table from base table.
  INSERT INTO TABLE casestudy_data_part
  PARTITION(event_type)
  SELECT event_time, product_id, category_id, category_code, brand, price, user_id, user_session, event_type
  FROM casestudy_data;



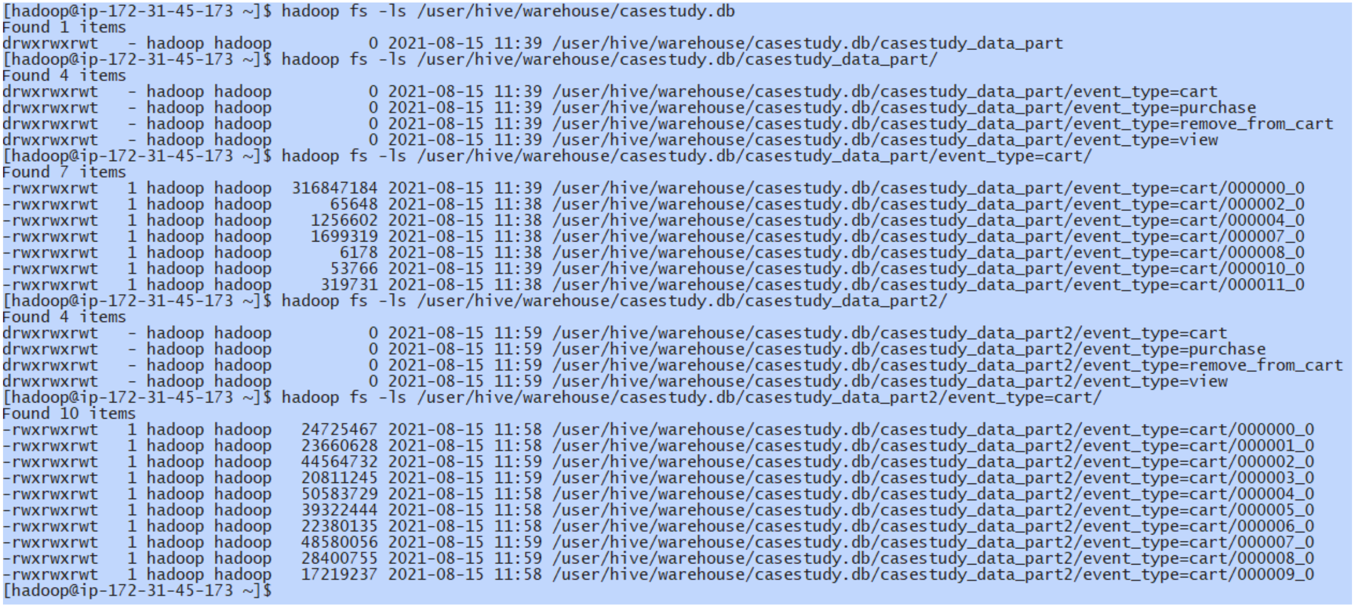
```
hive> INSERT INTO TABLE casestudy_data_part
    > PARTITION(event_type)
    > SELECT event_time, product_id, category_id, category_code, brand, price, user_id, user_session, event_type
    > FROM casestudy_data;
Query ID = hadoop_20210815113702_4e3c121c-87c4-4a2d-b36f-2bf4517f03b7
Total jobs = 1
Launching Job 1 out of 1
Tez session was closed. Reopening...
Session re-established.
Status: Running (Executing on YARN cluster with App id application_1629012534257_0009)

----------------------------------------------------------------------------------------------
        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: 152.37 s
----------------------------------------------------------------------------------------------
Loading data to table casestudy.casestudy_data_part partition (event_type=null)

Loaded : 4/4 partitions.
         Time taken to load dynamic partitions: 0.221 seconds
         Time taken for adding to write entity : 0.001 seconds
OK
Time taken: 161.371 seconds
hive>
```

- We created another optimized table *casestudy_data_part2*, this time clustering it by 'category_id' into 10 buckets.
  CREATE TABLE IF NOT EXISTS casestudy_data_part2 (event_time timestamp, 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_id) INTO 10 BUCKETS
  ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.OpenCSVSerde'
  STORED AS TEXTFILE;



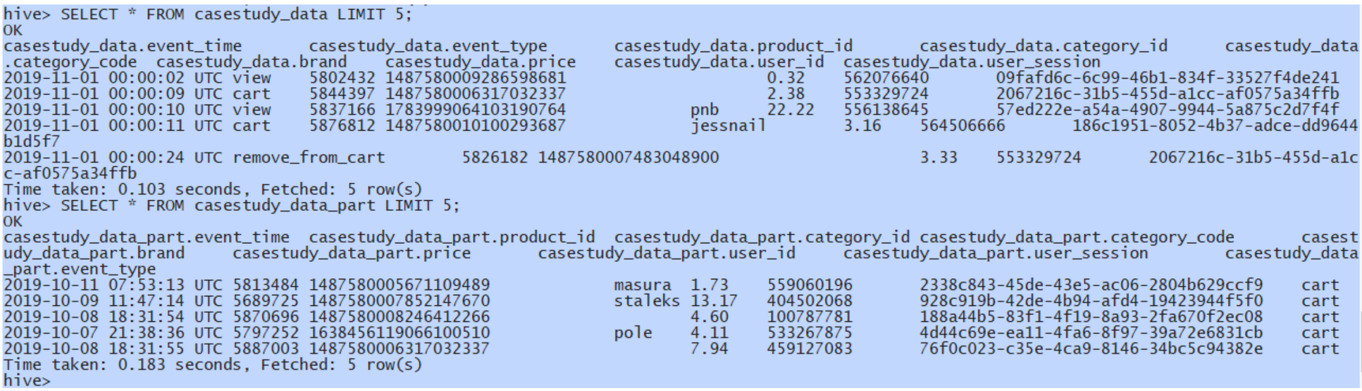
```
hive> CREATE TABLE IF NOT EXISTS casestudy_data_part2 (event_time timestamp, product_id string , category_id string , category_code string ,b
and string , price float, user_id bigint , user_session string )
    > PARTITIONED BY (event_type string)
    > CLUSTERED BY (category_id) INTO 10 BUCKETS
    > ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.OpenCSVSerde'
    > STORED AS TEXTFILE;
OK
Time taken: 0.058 seconds
hive> INSERT INTO TABLE casestudy_data_part2
    > PARTITION(event_type)
    > SELECT event_time, product_id, category_id, category_code, brand, price, user_id, user_session, event_type
    > FROM casestudy_data;
Query ID = hadoop_20210815115649_41f67515-70e0-4207-ac1f-de899b4ee64e
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1629012534257_0010)

----------------------------------------------------------------------------------------------
        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: 154.35 s
----------------------------------------------------------------------------------------------
Loading data to table casestudy.casestudy_data_part2 partition (event_type=null)

Loaded : 4/4 partitions.
         Time taken to load dynamic partitions: 0.283 seconds
         Time taken for adding to write entity : 0.001 seconds
OK
Time taken: 155.705 seconds
```

- Checking if the partitions and buckets are created correctly -



```
[hadoop@ip-172-31-45-173 ~]$ hadoop fs -ls /user/hive/warehouse/casestudy.db
Found 1 items
drwxrwxrwt   - hadoop hadoop          0 2021-08-15 11:39 /user/hive/warehouse/casestudy.db/casestudy_data_part
[hadoop@ip-172-31-45-173 ~]$ hadoop fs -ls /user/hive/warehouse/casestudy.db/casestudy_data_part/
Found 4 items
drwxrwxrwt   - hadoop hadoop          0 2021-08-15 11:39 /user/hive/warehouse/casestudy.db/casestudy_data_part/event_type=cart
drwxrwxrwt   - hadoop hadoop          0 2021-08-15 11:39 /user/hive/warehouse/casestudy.db/casestudy_data_part/event_type=purchase
drwxrwxrwt   - hadoop hadoop          0 2021-08-15 11:39 /user/hive/warehouse/casestudy.db/casestudy_data_part/event_type=remove_from_cart
drwxrwxrwt   - hadoop hadoop          0 2021-08-15 11:39 /user/hive/warehouse/casestudy.db/casestudy_data_part/event_type=view
[hadoop@ip-172-31-45-173 ~]$ hadoop fs -ls /user/hive/warehouse/casestudy.db/casestudy_data_part/event_type=cart/
Found 7 items
-rwxrwxrwt   1 hadoop hadoop  316847184 2021-08-15 11:39 /user/hive/warehouse/casestudy.db/casestudy_data_part/event_type=cart/000000_0
-rwxrwxrwt   1 hadoop hadoop      65648 2021-08-15 11:38 /user/hive/warehouse/casestudy.db/casestudy_data_part/event_type=cart/000002_0
-rwxrwxrwt   1 hadoop hadoop    1256602 2021-08-15 11:38 /user/hive/warehouse/casestudy.db/casestudy_data_part/event_type=cart/000004_0
-rwxrwxrwt   1 hadoop hadoop    1699319 2021-08-15 11:38 /user/hive/warehouse/casestudy.db/casestudy_data_part/event_type=cart/000007_0
-rwxrwxrwt   1 hadoop hadoop       6178 2021-08-15 11:38 /user/hive/warehouse/casestudy.db/casestudy_data_part/event_type=cart/000008_0
-rwxrwxrwt   1 hadoop hadoop      53766 2021-08-15 11:39 /user/hive/warehouse/casestudy.db/casestudy_data_part/event_type=cart/000010_0
-rwxrwxrwt   1 hadoop hadoop     319731 2021-08-15 11:38 /user/hive/warehouse/casestudy.db/casestudy_data_part/event_type=cart/000011_0
[hadoop@ip-172-31-45-173 ~]$ hadoop fs -ls /user/hive/warehouse/casestudy.db/casestudy_data_part2/
Found 4 items
drwxrwxrwt   - hadoop hadoop          0 2021-08-15 11:59 /user/hive/warehouse/casestudy.db/casestudy_data_part2/event_type=cart
drwxrwxrwt   - hadoop hadoop          0 2021-08-15 11:59 /user/hive/warehouse/casestudy.db/casestudy_data_part2/event_type=purchase
drwxrwxrwt   - hadoop hadoop          0 2021-08-15 11:59 /user/hive/warehouse/casestudy.db/casestudy_data_part2/event_type=remove_from_cart
drwxrwxrwt   - hadoop hadoop          0 2021-08-15 11:59 /user/hive/warehouse/casestudy.db/casestudy_data_part2/event_type=view
[hadoop@ip-172-31-45-173 ~]$ hadoop fs -ls /user/hive/warehouse/casestudy.db/casestudy_data_part2/event_type=cart/
Found 10 items
-rwxrwxrwt   1 hadoop hadoop   24725467 2021-08-15 11:58 /user/hive/warehouse/casestudy.db/casestudy_data_part2/event_type=cart/000000_0
-rwxrwxrwt   1 hadoop hadoop   23660628 2021-08-15 11:58 /user/hive/warehouse/casestudy.db/casestudy_data_part2/event_type=cart/000001_0
-rwxrwxrwt   1 hadoop hadoop   44564732 2021-08-15 11:59 /user/hive/warehouse/casestudy.db/casestudy_data_part2/event_type=cart/000002_0
-rwxrwxrwt   1 hadoop hadoop   20811245 2021-08-15 11:59 /user/hive/warehouse/casestudy.db/casestudy_data_part2/event_type=cart/000003_0
-rwxrwxrwt   1 hadoop hadoop   50583729 2021-08-15 11:58 /user/hive/warehouse/casestudy.db/casestudy_data_part2/event_type=cart/000004_0
-rwxrwxrwt   1 hadoop hadoop   39322444 2021-08-15 11:58 /user/hive/warehouse/casestudy.db/casestudy_data_part2/event_type=cart/000005_0
-rwxrwxrwt   1 hadoop hadoop   22380135 2021-08-15 11:58 /user/hive/warehouse/casestudy.db/casestudy_data_part2/event_type=cart/000006_0
-rwxrwxrwt   1 hadoop hadoop   48580056 2021-08-15 11:59 /user/hive/warehouse/casestudy.db/casestudy_data_part2/event_type=cart/000007_0
-rwxrwxrwt   1 hadoop hadoop   28400755 2021-08-15 11:59 /user/hive/warehouse/casestudy.db/casestudy_data_part2/event_type=cart/000008_0
-rwxrwxrwt   1 hadoop hadoop   17219237 2021-08-15 11:58 /user/hive/warehouse/casestudy.db/casestudy_data_part2/event_type=cart/000009_0
[hadoop@ip-172-31-45-173 ~]$
```

set hive.cli.print.header=true;
*SELECT * FROM casestudy_data LIMIT 5;*
*SELECT * FROM casestudy_data_part LIMIT 5;*



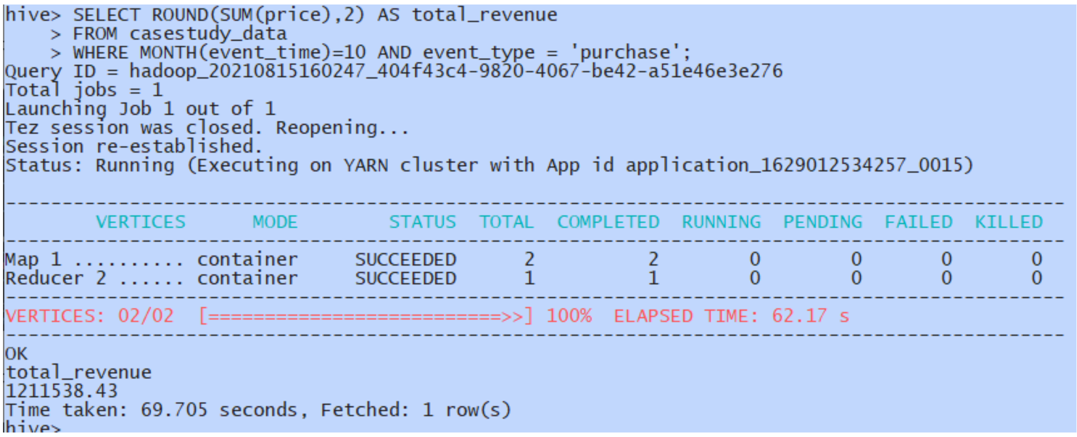
```
hive> SELECT * FROM casestudy_data LIMIT 5;
OK
casestudy_data.event_time       casestudy_data.event_type       casestudy_data.product_id        casestudy_data.category_id      casestudy_data
.category_code  casestudy_data.brand    casestudy_data.price    casestudy_data.user_id  casestudy_data.user_session
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-dd9644
b1d5f7
2019-11-01 00:00:24 UTC remove_from_cart        5826182 1487580007483048900                      3.33    553329724       2067216c-31b5-455d-a1c
c-af0575a34ffb
Time taken: 0.103 seconds, Fetched: 5 row(s)
hive> SELECT * FROM casestudy_data_part LIMIT 5;
OK
casestudy_data_part.event_time  casestudy_data_part.product_id  casestudy_data_part.category_id casestudy_data_part.category_code        casest
udy_data_part.brand     casestudy_data_part.price       casestudy_data_part.user_id     casestudy_data_part.user_session         casestudy_data
_part.event_type
2019-10-11 07:53:13 UTC 5813484 1487580005671109489             masura  1.73    559060196       2338c843-45de-43e5-ac06-2804b629ccf9     cart
2019-10-09 11:47:14 UTC 5689725 1487580007852147670             staleks 13.17   404502068       928c919b-42de-4b94-afd4-19423944f5f0     cart
2019-10-08 18:31:54 UTC 5870696 1487580008246412266                     4.60    100787781       188a44b5-83f1-4f19-8a93-2fa670f2ec08     cart
2019-10-07 21:38:36 UTC 5797252 1638456119066100510             pole    4.11    533267875       4d44c69e-ea11-4fa6-8f97-39a72e6831cb     cart
2019-10-08 18:31:55 UTC 5887003 1487580006317032337                     7.94    459127083       76f0c023-c35e-4ca9-8146-34bc5c94382e     cart
Time taken: 0.183 seconds, Fetched: 5 row(s)
hive>
```

Hive Queries

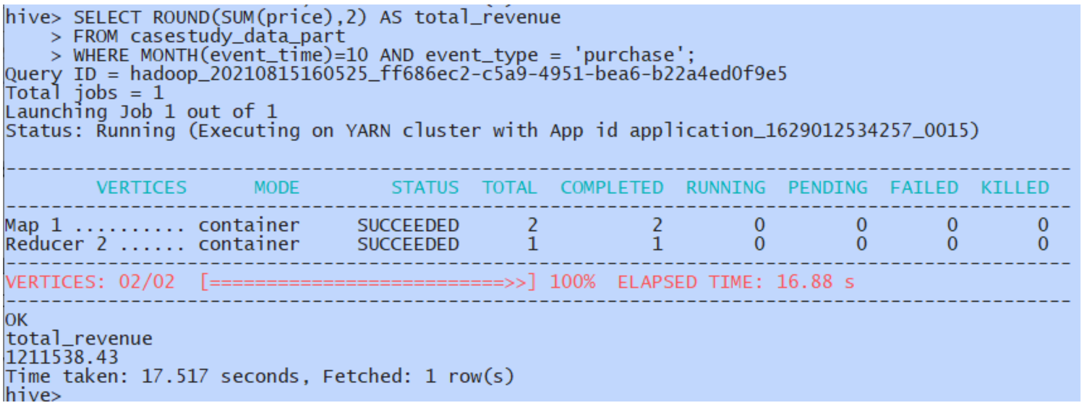
- **Question 1: Find the total revenue generated due to purchases made in October. -**
  SELECT ROUND(SUM(price),2) AS total_revenue
  FROM casestudy_data_part2
  WHERE MONTH(event_time)=10 AND event_type = 'purchase';

Comparing the performance of the base table with the optimized tables –

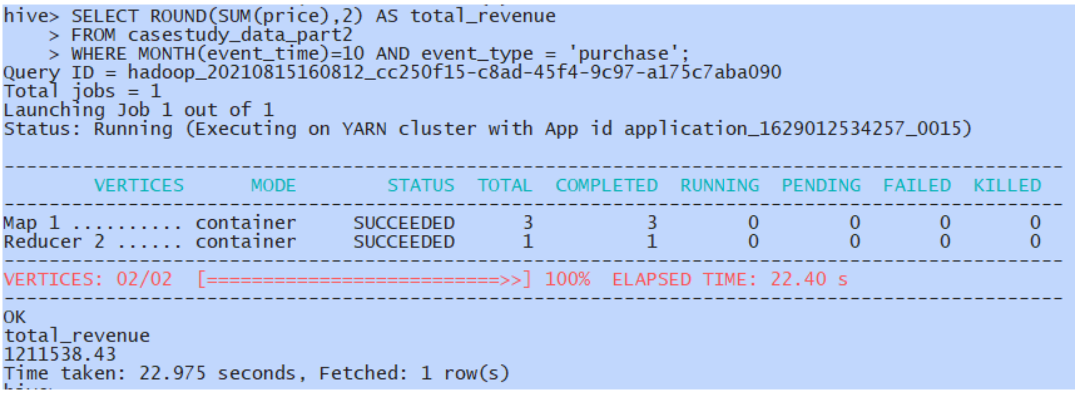
The below screenshots are of the same query from the base table and the bucketed table. The bucketed table takes less time to query the result than the base table. This is the use of partitioning and bucketing the data.



```
hive> SELECT ROUND(SUM(price),2) AS total_revenue
    > FROM casestudy_data
    > WHERE MONTH(event_time)=10 AND event_type = 'purchase';
Query ID = hadoop_20210815160247_404f43c4-9820-4067-be42-a51e46e3e276
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_1629012534257_0015)

----------------------------------------------------------------------------------------------
        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: 62.17 s
----------------------------------------------------------------------------------------------
OK
total_revenue
1211538.43
Time taken: 69.705 seconds, Fetched: 1 row(s)
hive>
```



```
hive> SELECT ROUND(SUM(price),2) AS total_revenue
    > FROM casestudy_data_part
    > WHERE MONTH(event_time)=10 AND event_type = 'purchase';
Query ID = hadoop_20210815160525_ff686ec2-c5a9-4951-bea6-b22a4ed0f9e5
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1629012534257_0015)

----------------------------------------------------------------------------------------------
        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: 16.88 s
----------------------------------------------------------------------------------------------
OK
total_revenue
1211538.43
Time taken: 17.517 seconds, Fetched: 1 row(s)
hive>
```



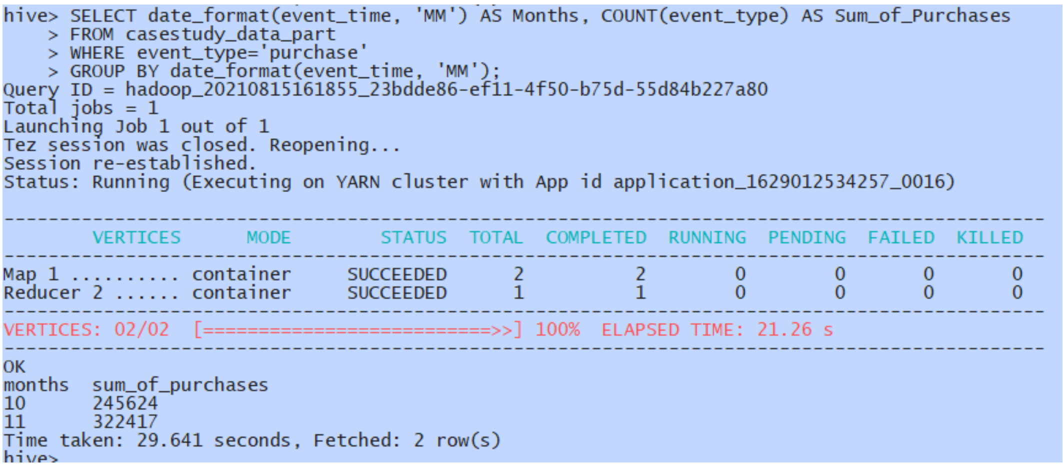
```
hive> SELECT ROUND(SUM(price),2) AS total_revenue
    > FROM casestudy_data_part2
    > WHERE MONTH(event_time)=10 AND event_type = 'purchase';
Query ID = hadoop_20210815160812_cc250f15-c8ad-45f4-9c97-a175c7aba090
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1629012534257_0015)

----------------------------------------------------------------------------------------------
        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: 22.40 s
----------------------------------------------------------------------------------------------
OK
total_revenue
1211538.43
Time taken: 22.975 seconds, Fetched: 1 row(s)
```

Findings:

The Total revenue generated based on Purchase in the month of October 2019 was 1,211,538.43/-. **Casestudy_data_part** took 17.5 seconds and **casestudy_data_part2** took 22.98 seconds whereas the base table took 69.7 seconds. As **casestudy_data_part** has a better performance, we will continue using this table for all the questions.

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

SELECT date_format(event_time, 'MM') AS Months, COUNT(event_type) AS Sum_of_Purchases
FROM casestudy_data_part
WHERE event_type='purchase'
GROUP BY date_format(event_time, 'MM');



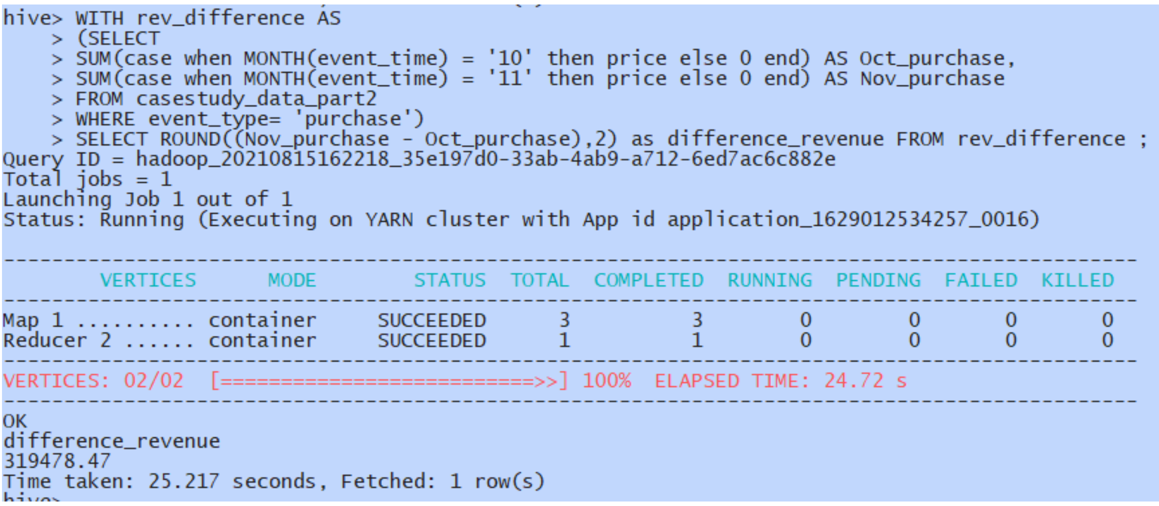
```
hive> SELECT date_format(event_time, 'MM') AS Months, COUNT(event_type) AS Sum_of_Purchases
    > FROM casestudy_data_part
    > WHERE event_type='purchase'
    > GROUP BY date_format(event_time, 'MM');
Query ID = hadoop_20210815161855_23bdde86-ef11-4f50-b75d-55d84b227a80
Total jobs = 1
Launching Job 1 out of 1
Tez session was closed. Reopening...
Session re-established.
Status: Running (Executing on YARN cluster with App id application_1629012534257_0016)

----------------------------------------------------------------------------------------------
        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: 21.26 s
----------------------------------------------------------------------------------------------
OK
months  sum_of_purchases
10      245624
11      322417
Time taken: 29.641 seconds, Fetched: 2 row(s)
hive>
```

Findings:

- There was more purchase made in the month of November (11), 322,417 than in the month of October (10), 245,624.
- The month of November is more profitable than the month of October.

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

WITH rev_difference AS
    (SELECT
    SUM(case when MONTH(event_time) = '10' then price else 0 end) AS Oct_purchase,
    SUM(case when MONTH(event_time) = '11' then price else 0 end) AS Nov_purchase
    FROM casestudy_data_part2
    WHERE event_type= 'purchase')
SELECT ROUND((Nov_purchase - Oct_purchase),2) as difference_revenue FROM rev_difference ;



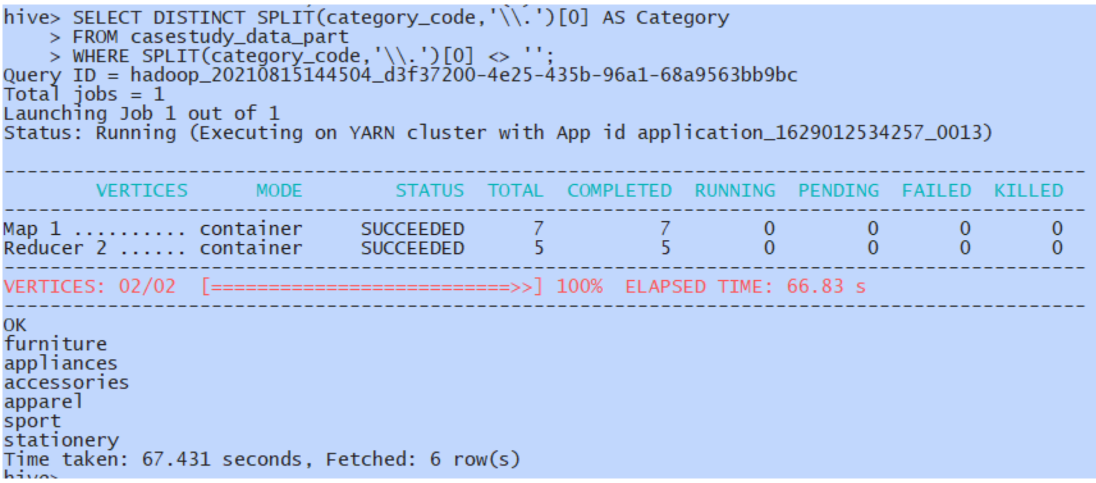
```
hive> WITH rev_difference AS
    > (SELECT
    > SUM(case when MONTH(event_time) = '10' then price else 0 end) AS Oct_purchase,
    > SUM(case when MONTH(event_time) = '11' then price else 0 end) AS Nov_purchase
    > FROM casestudy_data_part2
    > WHERE event_type= 'purchase')
    > SELECT ROUND((Nov_purchase - Oct_purchase),2) as difference_revenue FROM rev_difference ;
Query ID = hadoop_20210815162218_35e197d0-33ab-4ab9-a712-6ed7ac6c882e
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1629012534257_0016)

----------------------------------------------------------------------------------------------
        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: 24.72 s
----------------------------------------------------------------------------------------------
OK
difference_revenue
319478.47
Time taken: 25.217 seconds, Fetched: 1 row(s)
```

Findings:

- The difference in revenue between October and November month is 319478.47.
- The revenue generated in November of 2019 was more than the revenue generated in the month of October.

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

SELECT DISTINCT SPLIT(category_code,'\\.')[0] AS Category
FROM casestudy_data_part
WHERE SPLIT(category_code,'\\.')[0] <> '';



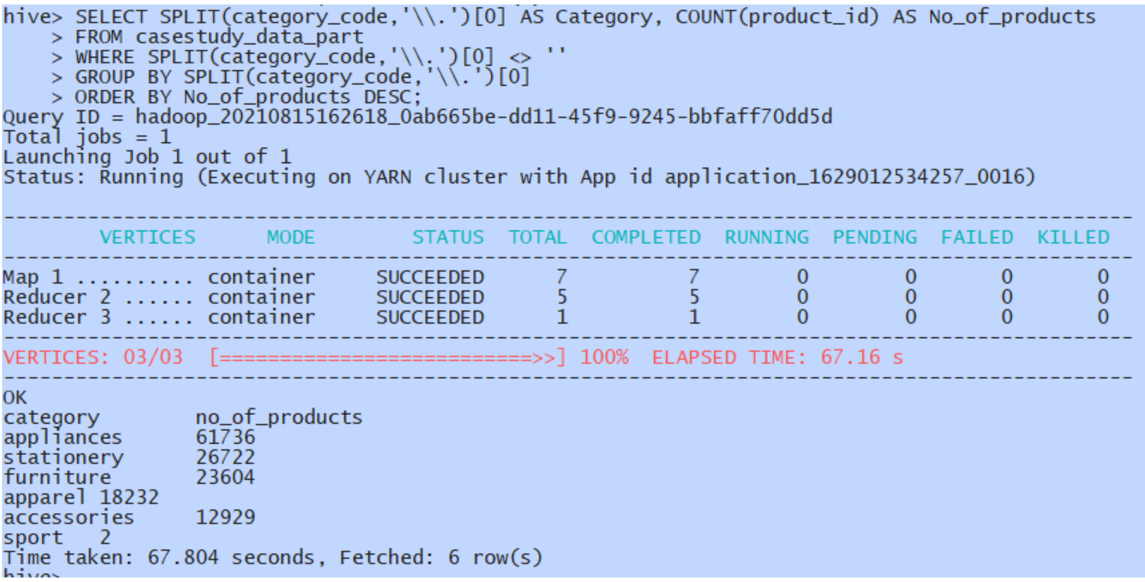
```
hive> SELECT DISTINCT SPLIT(category_code,'\\.')[0] AS Category
    > FROM casestudy_data_part
    > WHERE SPLIT(category_code,'\\.')[0] <> '';
Query ID = hadoop_20210815144504_d3f37200-4e25-435b-96a1-68a9563bb9bc
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1629012534257_0013)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      7          7        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      5          5        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 66.83 s
----------------------------------------------------------------------------------------------
OK
furniture
appliances
accessories
apparel
sport
stationery
Time taken: 67.431 seconds, Fetched: 6 row(s)
hive>
```

Findings:

There are 6 different categories under which company sells their different products. i.e accessories, apparel, appliances, furniture, sport and stationery.

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

SELECT SPLIT(category_code,'\\.')[0] AS Category, COUNT(product_id) AS No_of_products
FROM casestudy_data_part
WHERE SPLIT(category_code,'\\.')[0] <> ''
GROUP BY SPLIT(category_code,'\\.')[0]
ORDER BY No_of_products DESC;



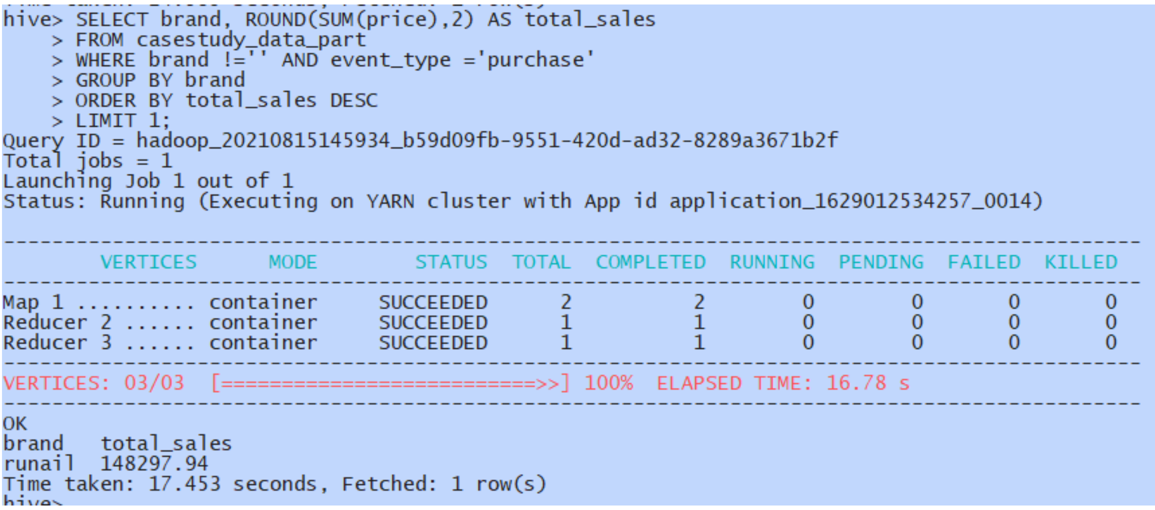
```
hive> SELECT SPLIT(category_code,'\\.')[0] AS Category, COUNT(product_id) AS No_of_products
    > FROM casestudy_data_part
    > WHERE SPLIT(category_code,'\\.')[0] <> ''
    > GROUP BY SPLIT(category_code,'\\.')[0]
    > ORDER BY No_of_products DESC;
Query ID = hadoop_20210815162618_0ab665be-dd11-45f9-9245-bbfaff70dd5d
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1629012534257_0016)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      7          7        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: 67.16 s
----------------------------------------------------------------------------------------------
OK
category        no_of_products
appliances      61736
stationery      26722
furniture       23604
apparel 18232
accessories     12929
sport   2
Time taken: 67.804 seconds, Fetched: 6 row(s)
hive>
```

Findings:

- Appliances with 61,736 products is the leading category, followed by stationery and furniture as second and third respectively.
- Sports category has only 2 products registered.

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

SELECT brand, ROUND(SUM(price),2) AS total_sales
FROM casestudy_data_part
WHERE brand !='' AND event_type ='purchase'
GROUP BY brand
ORDER BY total_sales DESC
LIMIT 1;

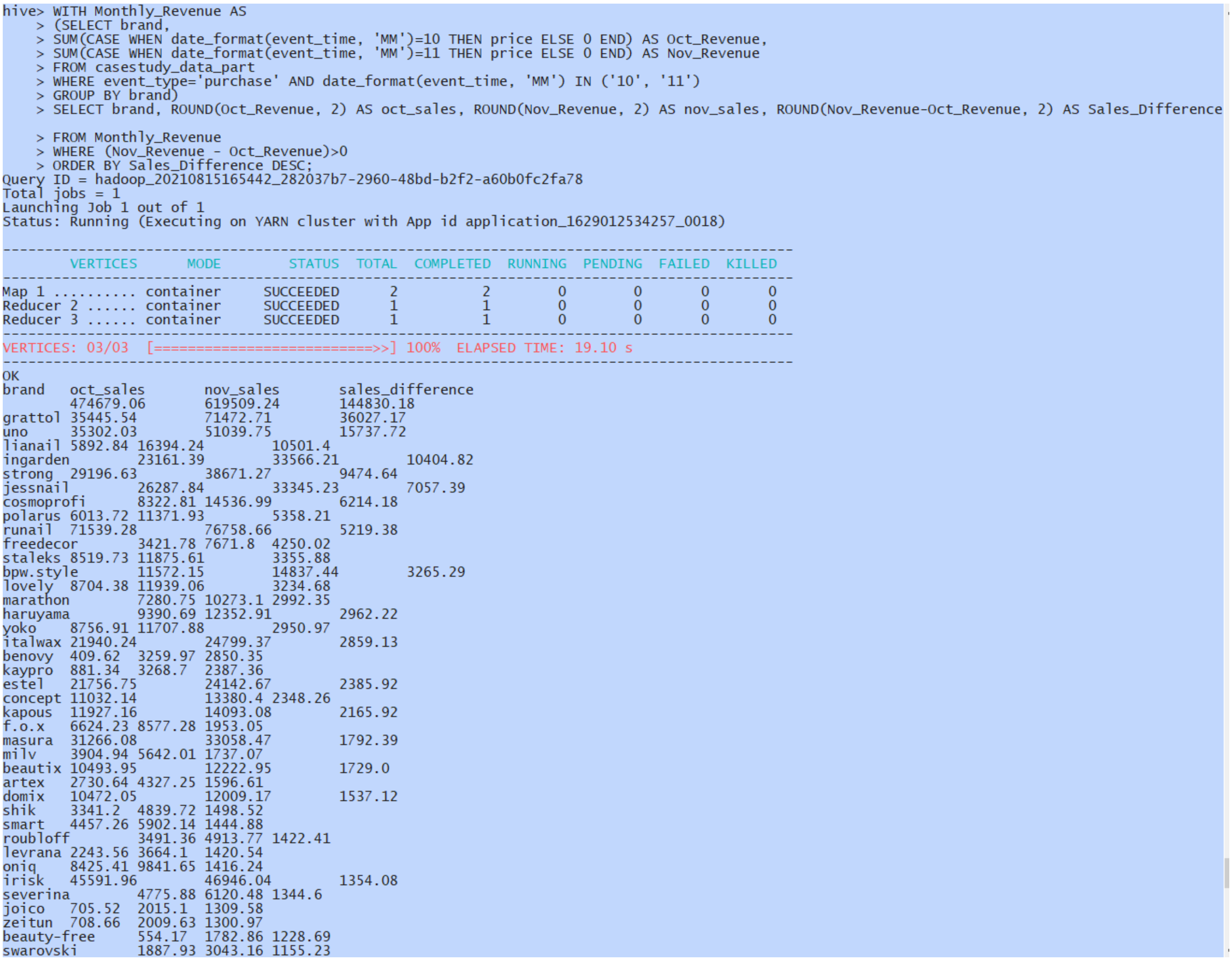


```
hive> SELECT brand, ROUND(SUM(price),2) AS total_sales
    > FROM casestudy_data_part
    > WHERE brand !='' AND event_type ='purchase'
    > GROUP BY brand
    > ORDER BY total_sales DESC
    > LIMIT 1;
Query ID = hadoop_20210815145934_b59d09fb-9551-420d-ad32-8289a3671b2f
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1629012534257_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
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 16.78 s
----------------------------------------------------------------------------------------------
OK
brand   total_sales
runail  148297.94
Time taken: 17.453 seconds, Fetched: 1 row(s)
hive>
```

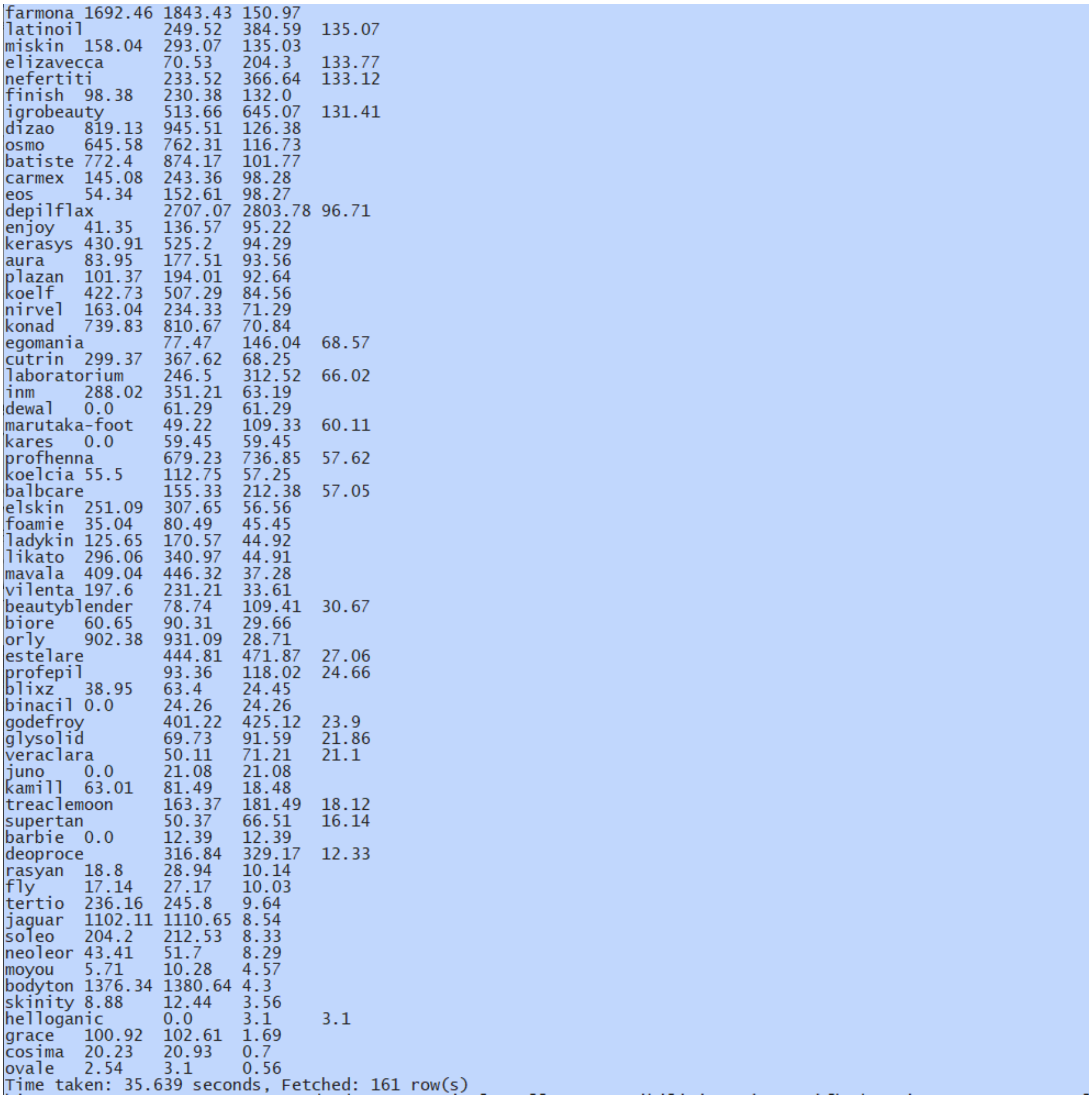
Findings:
Runail is the brand that has the highest / maximum sales in the month of October and November of 2019 combined.

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

WITH Monthly_Revenue AS
(SELECT brand,
SUM(CASE WHEN date_format(event_time, 'MM')=10 THEN price ELSE 0 END) AS Oct_Revenue,
SUM(CASE WHEN date_format(event_time, 'MM')=11 THEN price ELSE 0 END) AS Nov_Revenue
FROM casestudy_data_part
WHERE event_type='purchase' AND date_format(event_time, 'MM') IN ('10', '11')
GROUP BY brand)
SELECT brand, ROUND(Oct_Revenue, 2) AS oct_sales, ROUND(Nov_Revenue, 2) AS nov_sales, ROUND(Nov_Revenue-Oct_Revenue, 2) AS Sales_Difference
FROM Monthly_Revenue
WHERE (Nov_Revenue - Oct_Revenue)>0
ORDER BY Sales_Difference DESC;



```
hive> WITH Monthly_Revenue AS
    > (SELECT brand,
    > SUM(CASE WHEN date_format(event_time, 'MM')=10 THEN price ELSE 0 END) AS Oct_Revenue,
    > SUM(CASE WHEN date_format(event_time, 'MM')=11 THEN price ELSE 0 END) AS Nov_Revenue
    > FROM casestudy_data_part
    > WHERE event_type='purchase' AND date_format(event_time, 'MM') IN ('10', '11')
    > GROUP BY brand)
    > SELECT brand, ROUND(Oct_Revenue, 2) AS oct_sales, ROUND(Nov_Revenue, 2) AS nov_sales, ROUND(Nov_Revenue-Oct_Revenue, 2) AS Sales_Difference

    > FROM Monthly_Revenue
    > WHERE (Nov_Revenue - Oct_Revenue)>0
    > ORDER BY Sales_Difference DESC;
Query ID = hadoop_20210815165442_282037b7-2960-48bd-b2f2-a60b0fc2fa78
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1629012534257_0018)

----------------------------------------------------------------------------------------------
        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: 19.10 s
----------------------------------------------------------------------------------------------
OK
brand   oct_sales       nov_sales       sales_difference
        474679.06       619509.24       144830.18
grattol 35445.54        71472.71        36027.17
uno     35302.03        51039.75        15737.72
lianail 5892.84 16394.24        10501.4
ingarden        23161.39        33566.21        10404.82
strong  29196.63        38671.27        9474.64
jessnail        26287.84        33345.23        7057.39
cosmoprofi      8322.81 14536.99        6214.18
polarus 6013.72 11371.93        5358.21
runail  71539.28        76758.66        5219.38
freedecor       3421.78 7671.8  4250.02
staleks 8519.73 11875.61        3355.88
bpw.style       11572.15        14837.44        3265.29
lovely  8704.38 11939.06        3234.68
marathon        7280.75 10273.1 2992.35
haruyama        9390.69 12352.91        2962.22
yoko    8756.91 11707.88        2950.97
italwax 21940.24        24799.37        2859.13
benovy  409.62  3259.97 2850.35
kaypro  881.34  3268.7  2387.36
estel   21756.75        24142.67        2385.92
concept 11032.14        13380.4 2348.26
kapous  11927.16        14093.08        2165.92
f.o.x   6624.23 8577.28 1953.05
masura  31266.08        33058.47        1792.39
milv    3904.94 5642.01 1737.07
beautix 10493.95        12222.95        1729.0
artex   2730.64 4327.25 1596.61
domix   10472.05        12009.17        1537.12
shik    3341.2  4839.72 1498.52
smart   4457.26 5902.14 1444.88
roubloff        3491.36 4913.77 1422.41
levrana 2243.56 3664.1  1420.54
oniq    8425.41 9841.65 1416.24
irisk   45591.96        46946.04        1354.08
severina        4775.88 6120.48 1344.6
joico   705.52  2015.1  1309.58
zeitun  708.66  2009.63 1300.97
beauty-free     554.17  1782.86 1228.69
swarovski       1887.93 3043.16 1155.23
```



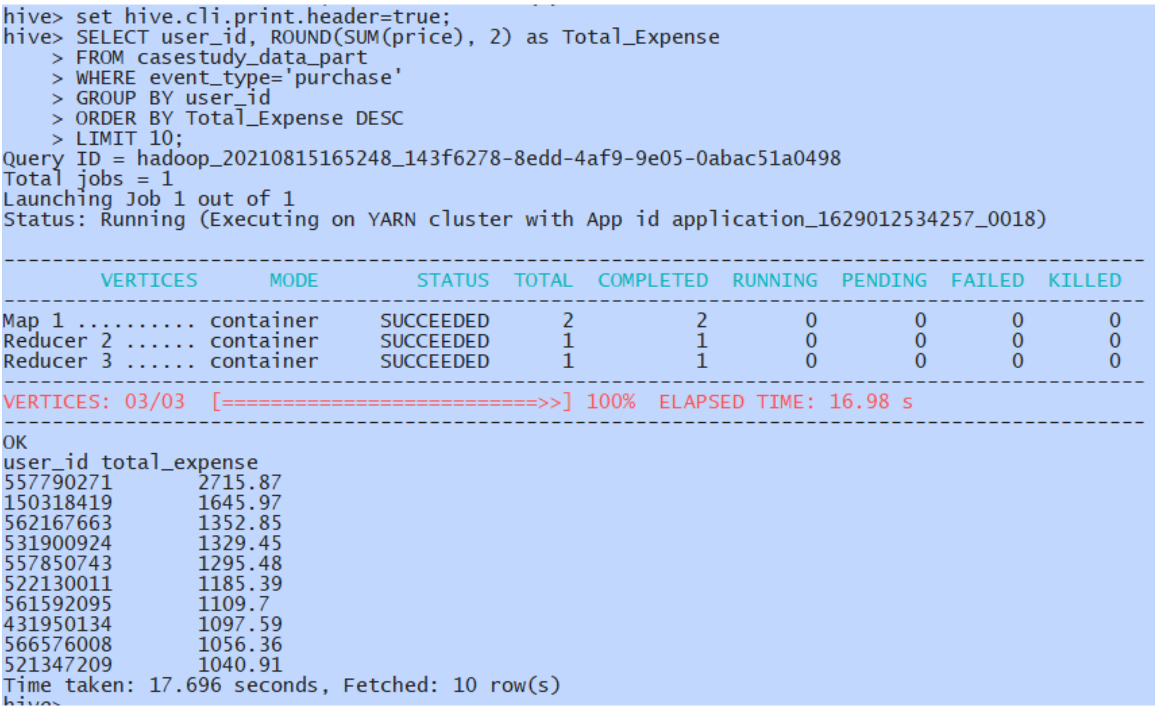
```
farmona 1692.46 1843.43 150.97
latinoil        249.52  384.59  135.07
miskin  158.04  293.07  135.03
elizavecca      70.53   204.3   133.77
nefertiti       233.52  366.64  133.12
finish  98.38   230.38  132.0
igrobeauty      513.66  645.07  131.41
dizao   819.13  945.51  126.38
osmo    645.58  762.31  116.73
batiste 772.4   874.17  101.77
carmex  145.08  243.36  98.28
eos     54.34   152.61  98.27
depilflax       2707.07 2803.78 96.71
enjoy   41.35   136.57  95.22
kerasys 430.91  525.2   94.29
aura    83.95   177.51  93.56
plazan  101.37  194.01  92.64
koelf   422.73  507.29  84.56
nirvel  163.04  234.33  71.29
konad   739.83  810.67  70.84
egomania        77.47   146.04  68.57
cutrin  299.37  367.62  68.25
laboratorium    246.5   312.52  66.02
inm     288.02  351.21  63.19
dewal   0.0     61.29   61.29
marutaka-foot   49.22   109.33  60.11
kares   0.0     59.45   59.45
profhenna       679.23  736.85  57.62
koelcia 55.5    112.75  57.25
balbcare        155.33  212.38  57.05
elskin  251.09  307.65  56.56
foamie  35.04   80.49   45.45
ladykin 125.65  170.57  44.92
likato  296.06  340.97  44.91
mavala  409.04  446.32  37.28
vilenta 197.6   231.21  33.61
beautyblender   78.74   109.41  30.67
biore   60.65   90.31   29.66
orly    902.38  931.09  28.71
estelare        444.81  471.87  27.06
profepil        93.36   118.02  24.66
blixz   38.95   63.4    24.45
binacil 0.0     24.26   24.26
godefroy        401.22  425.12  23.9
glysolid        69.73   91.59   21.86
veraclara       50.11   71.21   21.1
juno    0.0     21.08   21.08
kamill  63.01   81.49   18.48
treaclemoon     163.37  181.49  18.12
supertan        50.37   66.51   16.14
barbie  0.0     12.39   12.39
deoproce        316.84  329.17  12.33
rasyan  18.8    28.94   10.14
fly     17.14   27.17   10.03
tertio  236.16  245.8   9.64
jaguar  1102.11 1110.65 8.54
soleo   204.2   212.53  8.33
neoleor 43.41   51.7    8.29
moyou   5.71    10.28   4.57
bodyton 1376.34 1380.64 4.3
skinity 8.88    12.44   3.56
helloganic      0.0     3.1     3.1
grace   100.92  102.61  1.69
cosima  20.23   20.93   0.7
ovale   2.54    3.1     0.56
Time taken: 35.639 seconds, Fetched: 161 row(s)
```

Findings:

- Total of 161 brands had an increase in the selling from October to November.
- ‘Grattol’ brand has the highest total increment and ‘Ovale’ seems to have least increment from October to November.

- **Question 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.

SELECT user_id, ROUND(SUM(price), 2) as Total_Expense
FROM casestudy_data_part
WHERE event_type='purchase'
GROUP BY user_id
ORDER BY Total_Expense DESC
LIMIT 10;

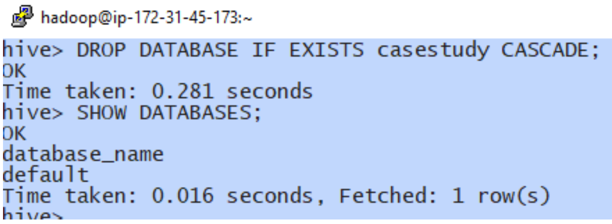


```
hive> set hive.cli.print.header=true;
hive> SELECT user_id, ROUND(SUM(price), 2) as Total_Expense
    > FROM casestudy_data_part
    > WHERE event_type='purchase'
    > GROUP BY user_id
    > ORDER BY Total_Expense DESC
    > LIMIT 10;
Query ID = hadoop_20210815165248_143f6278-8edd-4af9-9e05-0abac51a0498
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1629012534257_0018)

----------------------------------------------------------------------------------------------
        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: 16.98 s
----------------------------------------------------------------------------------------------
OK
user_id total_expense
557790271       2715.87
150318419       1645.97
562167663       1352.85
531900924       1329.45
557850743       1295.48
522130011       1185.39
561592095       1109.7
431950134       1097.59
566576008       1056.36
521347209       1040.91
Time taken: 17.696 seconds, Fetched: 10 row(s)
hive>
```

Findings:
The above is the list of the top 10 users or buyers who have spent the most and could be rewarded with a Golden Customer plan to attract more people in the coming future.

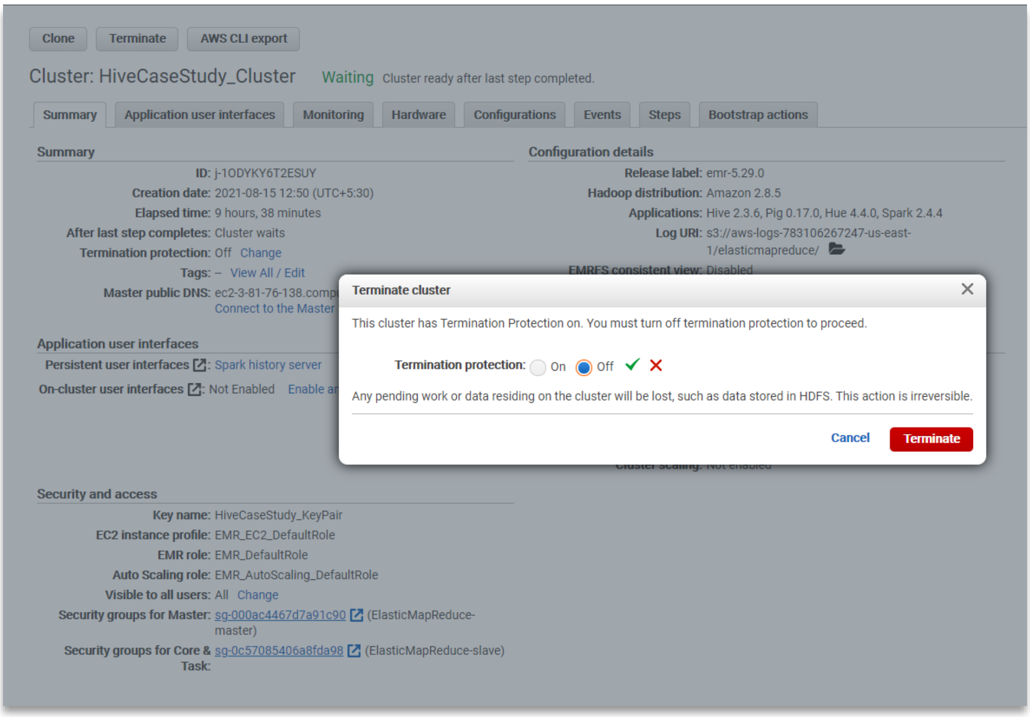
Cleaning up

- Drop the database



```
hive> DROP DATABASE IF EXISTS casestudy CASCADE;
OK
Time taken: 0.281 seconds
hive> SHOW DATABASES;
OK
database_name
default
Time taken: 0.016 seconds, Fetched: 1 row(s)
hive>
```

Once the operations are done, terminate the cluster by changing the Termination protection from ON to OFF and then click on the terminate button.



Cluster: HiveCaseStudy_Cluster Waiting Cluster ready after last step completed.

Summary

ID: j-1ODYKY6T2ESUY
Creation date: 2021-08-15 12:50 (UTC+5:30)
Elapsed time: 9 hours, 38 minutes
After last step completes: Cluster waits
Termination protection: Off Change
Tags: -- View All / Edit
Master public DNS: ec2-3-81-76-138.comp
Connect to the Master

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-783106267247-us-east-1/elasticmapreduce/
EMRFS consistent view: Disabled

Application user interfaces

Persistent user interfaces: Spark history server
On-cluster user interfaces: Not Enabled Enable a

Terminate cluster

This cluster has Termination Protection on. You must turn off termination protection to proceed.

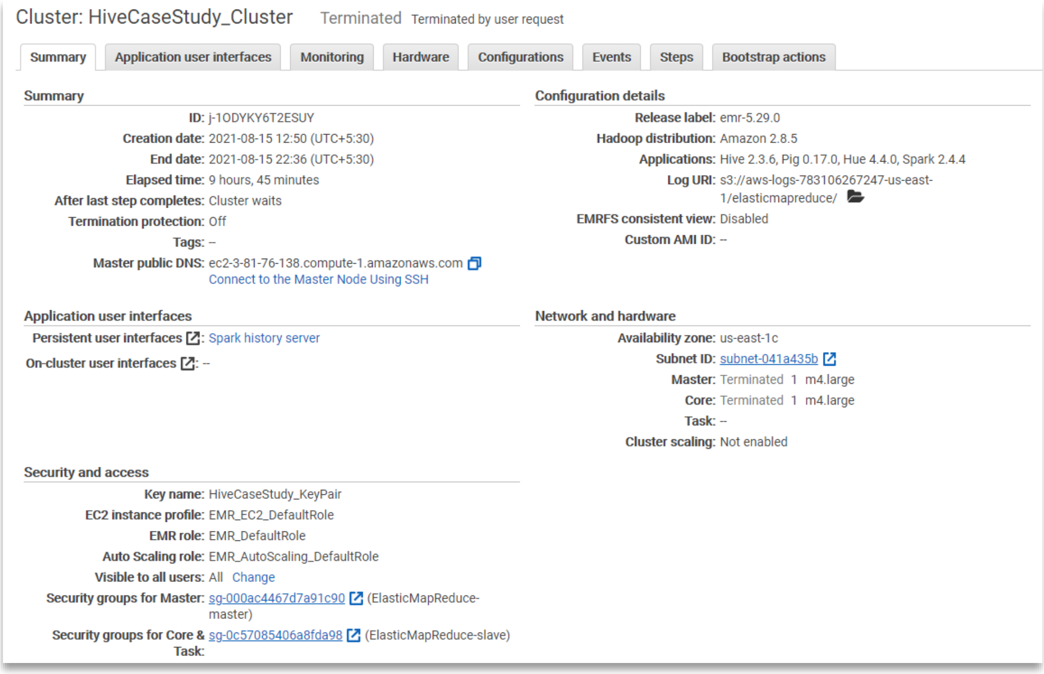
Termination protection: ○ On ● Off ✓ ✗

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

Security and access

Key name: HiveCaseStudy_KeyPair
EC2 instance profile: EMR_EC2_DefaultRole
EMR role: EMR_DefaultRole
Auto Scaling role: EMR_AutoScaling_DefaultRole
Visible to all users: All Change
Security groups for Master: sg-000ac4467d7a91c90 (ElasticMapReduce-master)
Security groups for Core & Task: sg-0c57085406a8fda98 (ElasticMapReduce-slave)

- Click on Terminate.



Cluster: HiveCaseStudy_Cluster Terminated Terminated by user request

Summary

ID: j-1ODYKY6T2ESUY
Creation date: 2021-08-15 12:50 (UTC+5:30)
End date: 2021-08-15 22:36 (UTC+5:30)
Elapsed time: 9 hours, 45 minutes
After last step completes: Cluster waits
Termination protection: Off
Tags: --
Master public DNS: ec2-3-81-76-138.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-783106267247-us-east-1/elasticmapreduce/
EMRFS consistent view: Disabled
Custom AMI ID: --

Application user interfaces

Persistent user interfaces: Spark history server
On-cluster user interfaces: --

Network and hardware

Availability zone: us-east-1c
Subnet ID: subnet-041a435b
Master: Terminated 1 m4.large
Core: Terminated 1 m4.large
Task: --
Cluster scaling: Not enabled

Security and access

Key name: HiveCaseStudy_KeyPair
EC2 instance profile: EMR_EC2_DefaultRole
EMR role: EMR_DefaultRole
Auto Scaling role: EMR_AutoScaling_DefaultRole
Visible to all users: All Change
Security groups for Master: sg-000ac4467d7a91c90 (ElasticMapReduce-master)
Security groups for Core & Task: sg-0c57085406a8fda98 (ElasticMapReduce-slave)