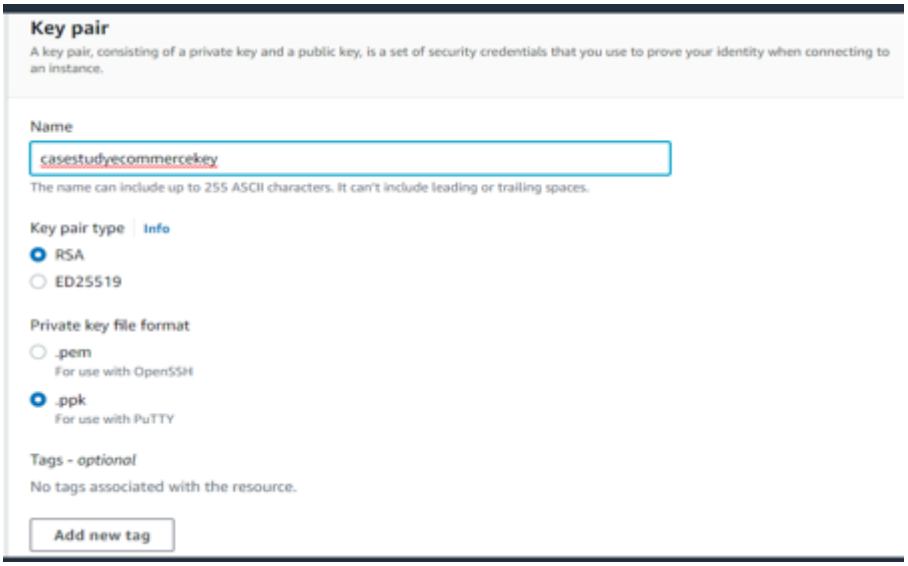
HIVE CASE STUDY

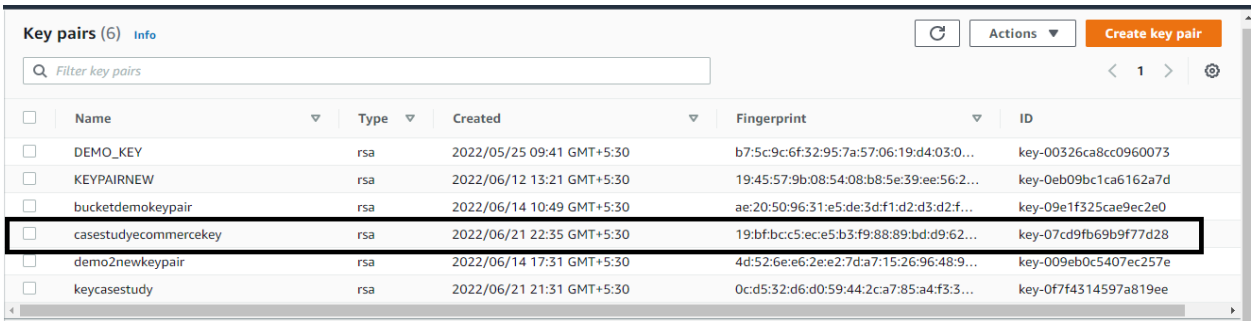
(DS_C37)

Submitted by:- Pallavi Thakur & Anamika Nayak

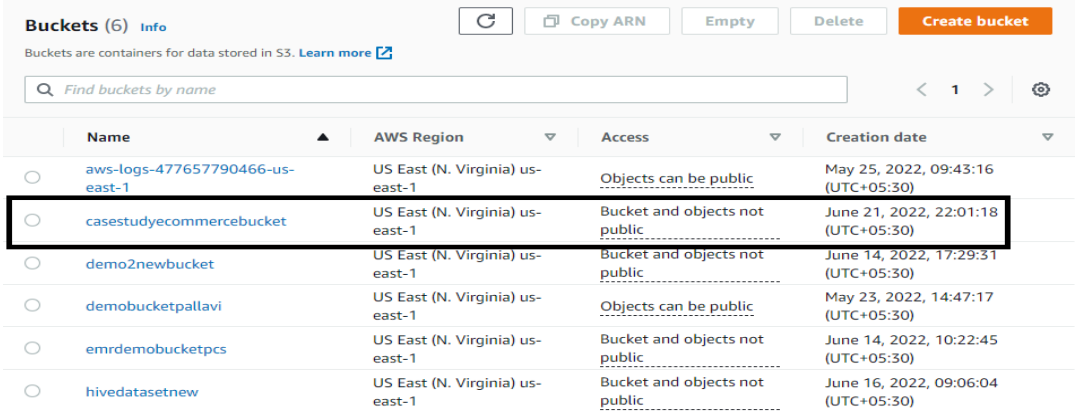
PRE-STEPS BEFORE COPING DATA INTO HDFS:

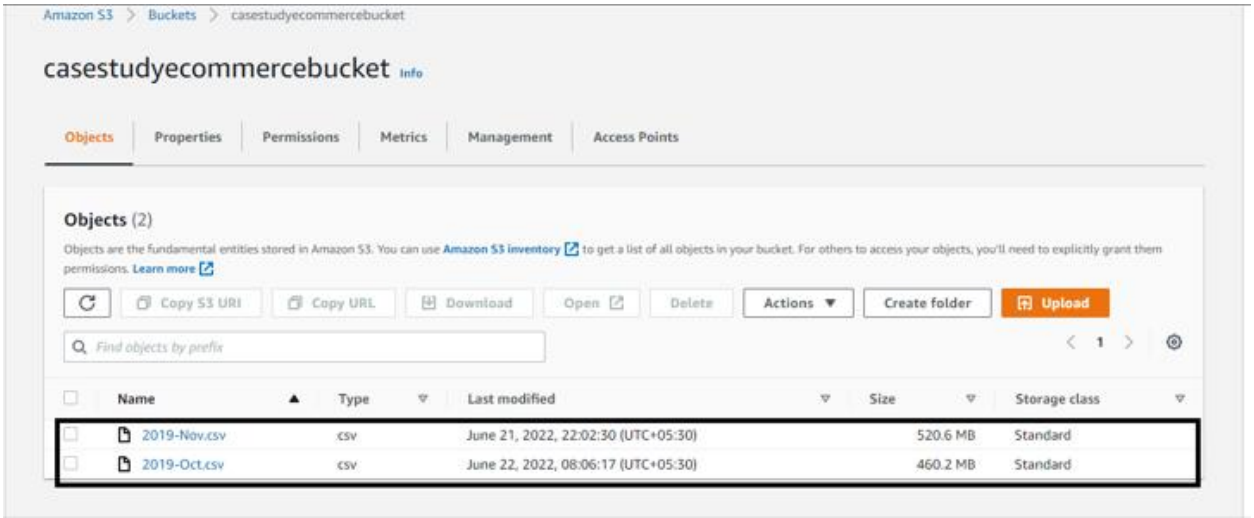
1. KEY PAIR CREATION: "casestudyecommercekey" is the Key-Pair created for this case study.





2. <u>S3 BUCKET CREATION:</u> “casestudyecommercebucket” is the bucket created for this case study.





3. <u>EMR CLUSTER CREATION:</u> EMR Cluster Landing Page > Create Cluster > Advanced Options > Selecting the release emr-5.29.0 and the required services

Cluster: ecommercecasestudycluster Starting

Summary | Application user interfaces | Monitoring | Hardware | Config

Summary

ID:	j-QRQ4UDY17ZTN
Creation date:	2022-06-24 08:47 (UTC+5:30)
Elapsed time:	0 seconds
After last step completes:	Cluster waits
Termination protection:	Off Change
Tags:	-- View All / Edit
Master public DNS:	--

Configuration details

Release label:	emr-5.29.0
Hadoop distribution:	Amazon 2.8.5
Applications:	Ganglia 3.7.2, Hive 2.3.6, Hue 4.4.0, Mahout 0.13.0, Pig 0.17.0, Tez 0.9.2
Log URI:	s3://casestudyecommercebucket/

- Network and Hardware Page > to define the cluster & nodes: Instance type for both master &core nodes are M4.large

Application user interfaces

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

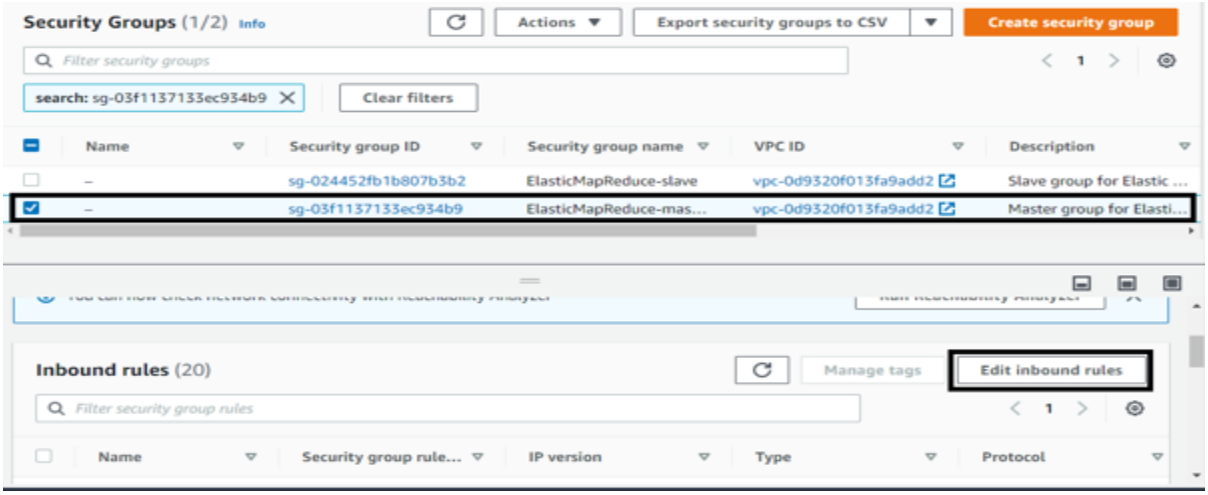
Network and hardware

Availability zone:	us-east-1e
Subnet ID:	subnet-0ae138c31f3c10d44
Master:	Running 1 m4.large
Core:	Running 1 m4.large
Task:	--
Cluster scaling:	Not enabled

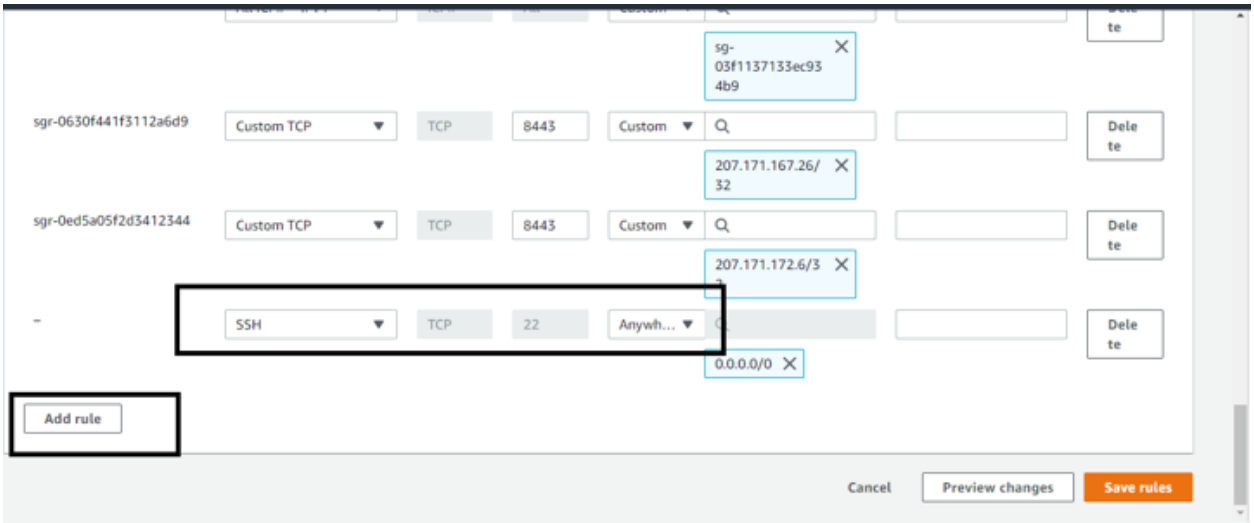
Security and access

Key name:	casestudyecommercekey
EC2 instance profile:	EMR_EC2_DefaultRole
EMR role:	EMR_DefaultRole
Visible to all users:	All Change
Security groups for Master:	sg-03f1137133ec934b9 (ElasticMapReduce-master)

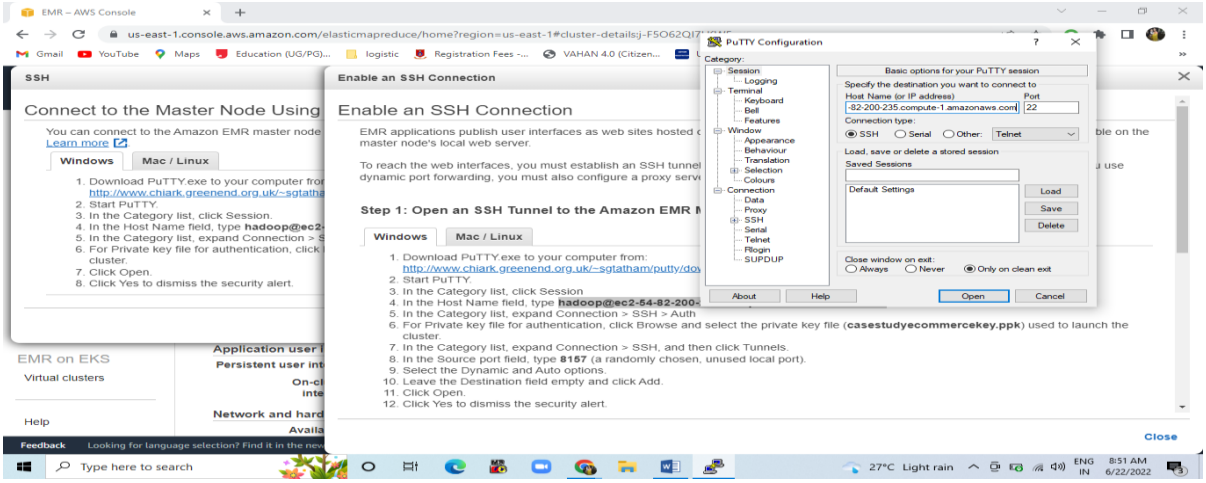
4. <u>ADD SECURITY INBOUND RULES FOR MASTER NODE</u>



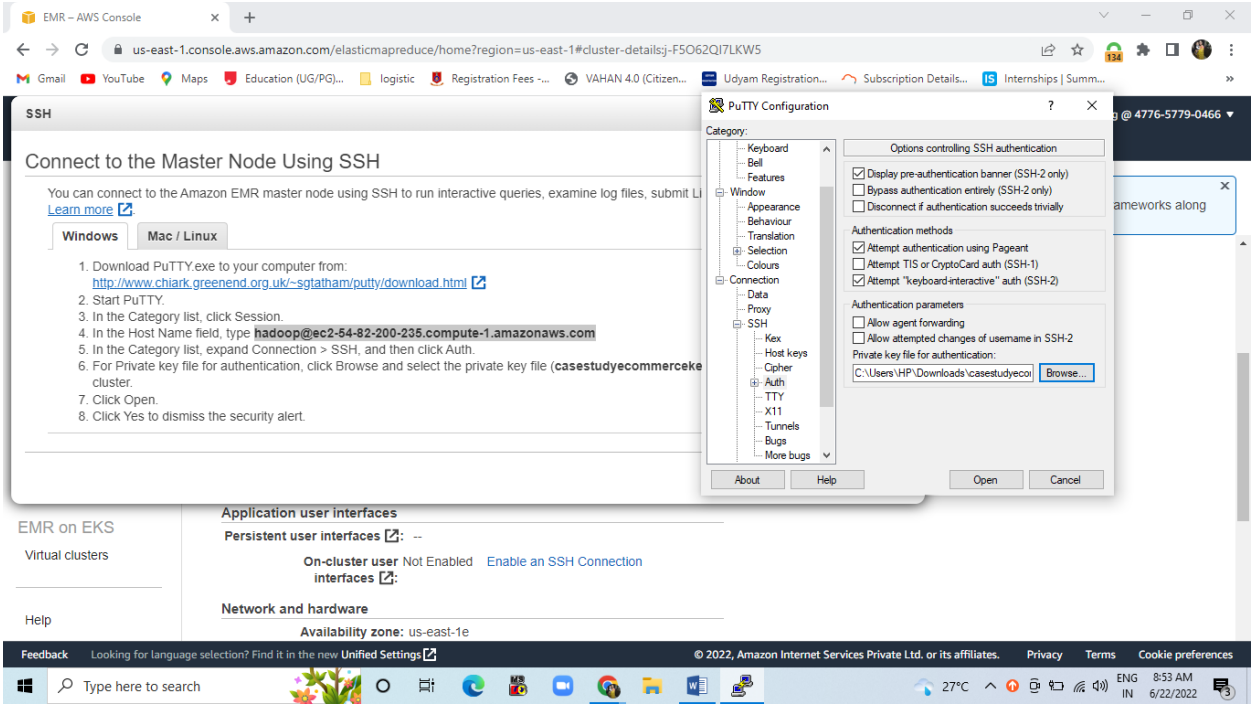
- Rules are added at the end by using the tab ADD RULE > SAVE RULES



5. ENABLE SSH CONNECTION FOR MASTER NODE USING PUTTY



Add key pair “casestudyecommercekey” as a private key for authentication



Once the above setup is completed, we connected to the Master Node to perform Hive Queries:

1. Terminal > Connecting to EMR Cluster using SSH.

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

https://aws.amazon.com/amazon-linux-ami/2018.03-release-notes/
68 package(s) needed for security, out of 97 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-172-31-60-228 ~]$
```

2. Creating a directory “hivecasestudyanamikaandpallavi”

hadoop fs -mkdir /hivecasestudyanamikaandpallavi
hadoop fs -ls /

```
[hadoop@ip-172-31-51-250 ~]$ hadoop fs -ls /
Found 4 items
drwxr-xr-x   - hdfs hadoop          0 2022-06-22 02:57 /apps
drwxrwxrwt   - hdfs hadoop          0 2022-06-22 02:59 /tmp
drwxr-xr-x   - hdfs hadoop          0 2022-06-22 02:57 /user
drwxr-xr-x   - hdfs hadoop          0 2022-06-22 02:57 /var
[hadoop@ip-172-31-51-250 ~]$ hadoop fs -mkdir /hivecasestudyanamikaandpallavi
[hadoop@ip-172-31-51-250 ~]$ hadoop fs -ls /
Found 5 items
drwxr-xr-x   - hdfs   hadoop          0 2022-06-22 02:57 /apps
drwxr-xr-x   - hadoop hadoop          0 2022-06-22 03:35 /hivecasestudyanamikaandpallavi
drwxrwxrwt   - hdfs   hadoop          0 2022-06-22 02:59 /tmp
drwxr-xr-x   - hdfs   hadoop          0 2022-06-22 02:57 /user
drwxr-xr-x   - hdfs   hadoop          0 2022-06-22 02:57 /var
[hadoop@ip-172-31-51-250 ~]$
```

3. Loading the October dataset into HDFS from S3:

hadoop distcp 's3://casestudyecommercebucket/2019-Oct.csv' /hivecasestudyanamikaandpallavi/oct_2019.csv

```
drwxr-xr-x   - hadoop hadoop          0 2022-06-23 06:36 /hivecasestudyanamikaan
dpallavi
drwxrwxrwt   - hdfs   hadoop          0 2022-06-23 06:22 /tmp
drwxr-xr-x   - hdfs   hadoop          0 2022-06-23 06:19 /user
drwxr-xr-x   - hdfs   hadoop          0 2022-06-23 06:19 /var
[hadoop@ip-172-31-60-242 ~]$ hadoop distcp 's3://casestudyecommercebucket/2019-O
ct.csv' /hivecasestudyanamikaandpallavi/oct_2019.csv
22/06/23 06:36:46 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://casestudyecommercebucket/2019-Oct.csv], targetPath=/hi
vecasestudyanamikaandpallavi/oct_2019.csv, targetPathExists=false, filtersFile='
null'}
```

4. Loading the November datasets into HDFS from S3:

hadoop distcp 's3://casestudyecommercebucket/2019-Nov.csv' /hivecasestudyanamikaandpallavi/nov_2019.csv

```
                Bytes Read=227
        File Output Format Counters
                Bytes Written=0
        DistCp Counters
                Bytes Copied=482542278
                Bytes Expected=482542278
                Files Copied=1
[hadoop@ip-172-31-60-242 ~]$ hadoop distcp 's3://casestudyecommercebucket/2019-Nov.csv' /hivecasestu
dyanamikaandpallavi/nov_2019.csv
22/06/23 06:40:24 INFO tools.DistCp: Input Options: DistCpOptions{atomicCommit=false, syncFolder=fal
se, deleteMissing=false, ignoreFailures=false, overwrite=false, skipCRC=false, blocking=true, numLis
tstatusThreads=0, maxMaps=20, mapBandwidth=100, sslConfigurationFile='null', copyStrategy='uniformsi
ze', preserveStatus=[], preserveRawXattrs=false, atomicWorkPath=null, logPath=null, sourceFileListin
g=null, sourcePaths=[s3://casestudyecommercebucket/2019-Nov.csv], targetPath=/hivecasestudyanamikaan
dpallavi/nov_2019.csv, targetPathExists=false, filtersFile='null'}
```

Files are copied into directory

```
        File Input Format Counters
                Bytes Read=227
        File Output Format Counters
                Bytes Written=0
        DistCp Counters
                Bytes Copied=545839412
                Bytes Expected=545839412
                Files Copied=1
[hadoop@ip-172-31-60-242 ~]$
```

5. Check the files in directory

   hadoop fs -ls /hivecasestudyanamikaandpallavi

```
                Bytes Expected=545839412
                Files Copied=1
[hadoop@ip-172-31-60-242 ~]$ hadoop fs -ls /hivecasestudyanamikaandpallavi
Found 2 items
-rw-r--r--   1 hadoop hadoop  545839412 2022-06-23 06:40 /hivecasestudyanamikaandpallavi/nov_2019.csv
-rw-r--r--   1 hadoop hadoop  482542278 2022-06-23 06:37 /hivecasestudyanamikaandpallavi/oct_2019.csv
[hadoop@ip-172-31-60-242 ~]$
```

6. Viewing the data for both the dataset:

hadoop fs -cat /hivecasestudyanamikaandpallavi/oct_2019.csv | head

```
[hadoop@ip-172-31-60-228 ~]$ hadoop fs -cat /hivecasestudyanamikaandpallavi/oct_2019.csv | head
event_time,event_type,product_id,category_id,category_code,brand,price,user_id,user_session
2019-10-01 00:00:00 UTC,cart,5773203,1487580005134238553,,runail,2.62,463240011,26dd6e6e-4dac-4778-8d2c-92e149dab885
2019-10-01 00:00:03 UTC,cart,5773353,1487580005134238553,,runail,2.62,463240011,26dd6e6e-4dac-4778-8d2c-92e149dab885
2019-10-01 00:00:07 UTC,cart,5881589,2151191071051219817,,lovely,13.48,429681830,49e8d843-adf3-428b-a2c3-fe8bc6a307c9
2019-10-01 00:00:07 UTC,cart,5723490,1487580005134238553,,runail,2.62,463240011,26dd6e6e-4dac-4778-8d2c-92e149dab885
2019-10-01 00:00:15 UTC,cart,5881449,1487580013522845895,,lovely,0.56,429681830,49e8d843-adf3-428b-a2c3-fe8bc6a307c9
2019-10-01 00:00:16 UTC,cart,5857269,1487580005134238553,,runail,2.62,430174032,73deale7-664e-43f4-8b30-d32b9d5af04f
2019-10-01 00:00:19 UTC,cart,5739055,1487580008246412266,,kapous,4.75,377667011,81326ac6-daa4-4f0a-b488-fd0956a78733
2019-10-01 00:00:24 UTC,cart,5825598,1487580009445982239,,,0.56,467916806,2f5b5546-b8cb-9ee7-7ecd-84276f8ef486
2019-10-01 00:00:25 UTC,cart,5698989,1487580006317032337,,,1.27,385985999,d30965e8-1101-44ab-b45d-cc1bb9fae694
```

hadoop fs -cat /hivecasestudyanamikaandpallavi/nov_2019.csv | head

```
[hadoop@ip-172-31-60-228 ~]$ hadoop fs -cat /hivecasestudyanamikaandpallavi/nov_2019.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
```

Datasets are successfully loaded.

Once the Data Set is successfully loaded, we will connect to hive trigger our Hive Query Language:

1. Launch Hive and check existing databases. Creating new database : “hivecasestudyforap”

hive> create database if not exists hivecasestudyforap;

```
hive> show databases;
OK
default
Time taken: 0.034 seconds, Fetched: 1 row(s)
hive> create database if not exists hivecasestudyforap;
OK
Time taken: 0.491 seconds
hive>
```

2. Creating new table: “retail_ap”:

hive>create external table if not exists retail_ap(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"=",","guoteChar" = "\"","escapeChar"="\\") stored as textfile location '/hivecasestudyanamikaandpallavi' TBLPROPERTIES("skip.header.line.count"="1");

```
Time taken: 0.631 seconds
hive> create external table if not exists retail_ap(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 SERDEPRO
PERTIES("separatorChar"=",","guoteChar" = "\"","escapeChar"="\\") stored as textfile location '/hivecasestudyanamikaandpallavi' TBLPROPERTIES(
"skip.header.line.count"="1");
OK
Time taken: 0.755 seconds
```

hive>describe retail_ap;

```
Time taken: 0.755 seconds
hive> describe retail_ap;
OK
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: 0.151 seconds, Fetched: 9 row(s)
```

3. Loading data into table : “retail_ap”

hive>load data inpath '/hivecasestudyanamikaandpallavi/oct_2019.csv' into table retail_ap;
hive>load data inpath '/hivecasestudyanamikaandpallavi/nov_2019.csv' into table retail_ap;

```
Time taken: 0.151 seconds, Fetched: 9 row(s)
hive> load data inpath '/hivecasestudyanamikaandpallavi/oct_2019.csv' into table retail_ap;
Loading data to table default.retail_ap
OK
Time taken: 4.846 seconds
hive> load data inpath '/hivecasestudyanamikaandpallavi/nov_2019.csv' into table retail_ap;
Loading data to table default.retail_ap
OK
```

Performing data check:

```
hive> select * from retail_ap 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-dd9644b1d
5f7
2019-11-01 00:00:24 UTC remove_from_cart        5826182 1487580007483048900                     3.33    553329724       2067216c-31b5-455d-a1cc-a
f0575a34ffb
```

As we can see that the datasets are properly loaded into the retail_ap table we can trigger our queries and check the output

QUESTION 1:

Find the total revenue generated due to purchases made in October

Solution Query:

select sum(price) from retail_ap where month(event_time) = 10 and event_type = "purchase" ;

```
:2
hive> select sum(price) from retail_ap where month(event_time)=10 and event_type="purchase";
Query ID = hadoop_20220625120003_13c2c033-321b-47bd-9129-c0219c12101c
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1655983803851_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: 70.22 s
--------------------------------------------------------------------------------
OK
1211538.4299997438
Time taken: 71.484 seconds, Fetched: 1 row(s)
```

Time Taken to execute the above query is 71.484 seconds.

This time is very high. In order to reduce the execution time of query, we create dynamic partition of the table “retail_ap” and add buckets for query optimization.

DYNAMIC PARTITIONING:

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

```
Time taken: 71.484 seconds, Fetched: 1 row(s)
hive> set hive.exec.dynamic.partition=true;
hive> set hive.exec.dynamic.partition.mode=nonstrict;
hive>
```

PARTITION TABLE 1: retail_ap_eventtype

hive>create external table if not exists retail_ap_eventtype(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(user_id) INTO 5 buckets row format SERDE 'org.apache.hadoop.hive.serde2.OpenCSVSerde' stored as textfile;

We have taken event_type as the partition key as it has 4 distinct values and if we check through the question sets that is given it will be used in most of the where condition. Hence it will be efficient to take event_type as the partitioning key for better query optimization.

hive>DESCRIBE retail_ap_eventtype;

```
Time taken: 0.101 seconds
hive> create external table if not exists retail_ap_eventtype(event_time timestamp,product_id string,category_id string,category_code string,brand st
ring,price float,user_id bigint,user_session string)PARTITIONED BY(event_type string)CLUSTERED BY(user_id) INTO 5 buckets row format SERDE 'org.apach
e.hadoop.hive.serde2.OpenCSVSerde' stored as textfile;
OK
Time
hive> DESCRIBE retail_ap_eventtype;
OK
event_time              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
event_type              string

# Partition Information
# col_name              data_type               comment

event_type              string
```

INSERT INTO TABLE retail_ap_eventtype PARTITION(event_type) SELECT event_time,product_id,category_id,category_code,brand,price,user_id,user_session ,event_type from retail_ap;

```
hive> set hive.exec.dynamic.partition.mode=nonstrict;
hive> INSERT INTO TABLE retail_ap_eventtype PARTITION(event_type) SELECT event_time,product_id,category_id,category_code,brand,price,user_id,user_session
 ,event_type from retail_ap;
Query ID = hadoop_20220623121957_21e56cfa-7827-4453-a64b-47699a2fda71
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1655983803851_0005)

--------------------------------------------------------------------------------
        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: 183.40 s
--------------------------------------------------------------------------------
Loading data to table default.retail_ap_eventtype partition (event_type=null)

Loaded : 4/4 partitions.
        Time taken to load dynamic partitions: 0.842 seconds
        Time taken for adding to write entity : 0.008 seconds
OK
Time taken: 190.738 seconds
hive>
```

Executing the same query with the new table “retail_ap_eventtype” partition table.

Query: select sum(price) from retail_ap_eventtype where month(event_time) = 10 and event_type = "purchase" ;

Output: 1211538.43

```
hive> select sum(price) from retail_ap_eventtype where month(event_time)=10 and event_type="purchase";
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1655983803851_0006)

--------------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      3          3        0        0       0       0
Reduce 2 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 22.30 s
--------------------------------------------------------------------------------------------------
OK
1211538.429999988
Time taken: 28.669 seconds, Fetched: 1 row(s)
```

Time Taken to execute the above query is 28.669 sec.

QUESTION 2:

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

Query: SELECT MONTH(event_time),SUM(price)AS amountpurchase, COUNT(event_type)as eventcount from retail_ap_eventtype where event_type="purchase" GROUP BY MONTH(event_time);

Output:

10	1211538.4300000465	245624
11	1531016.8999999745	322417

```
hive> SELECT MONTH(event_time),SUM(price)AS amountpurchase,COUNT(event_type)as eventcount from retail_ap_eventtype where event_type="purchase"GROUP BY M
NTH(event_time);
Query ID = hadoop_20220623123720_7e2e90b7-445c-41bb-952f-097aebc96b74
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_1655983803851_0008)

--------------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      3          3        0        0       0       0
Reduce 2 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 29.17 s
--------------------------------------------------------------------------------------------------
OK
10      1211538.4299998982      245624
11      1531016.8999999384      322417
Time taken: 45.343 seconds, Fetched: 2 row(s)
```

QUESTION 3:

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

WITH diffrev AS (SELECT SUM (CASE WHEN date_format (event_time,'MM') =10 THEN price ELSE 0 END) AS October, SUM(CASE WHEN date_format (event_time,'MM')=11 THEN price ELSE 0 END) AS November FROM retail_ap_eventtype WHERE date_format(event_time,'MM') IN (10,11) AND event_type='purchase') SELECT October, November, (November - October) as Differenceinrevenue FROM diffrev ;

Output: 1211538.429999989 1531016.8999999384 319478.47000000405

```
hive> WITH diffrev AS (SELECT SUM (CASE WHEN date_format (event_time,'MM') =10 THEN price ELSE 0 END) AS October, SUM(CASE WHEN date_format (event_time,'
MM')=11 THEN price ELSE 0 END) AS November FROM retail_ap_eventtype WHERE date_format(event_time,'MM') IN (10,11) AND event_type='purchase') SELECT Octob
er, November, (November - October) as Differenceinrevenue FROM diffrev ;
Query ID = hadoop_20220623124342_6a31290c-8638-495c-a0c9-5b2c5061b8cb
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1655983803851_0008)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      3          3        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 47.03 s
--------------------------------------------------------------------------------
OK
1211538.429999898       1531016.8999999384      319478.4700000405
Time taken: 47.73 seconds, Fetched: 1 row(s)
```

QUESTION 4:

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

Query: SELECT DISTINCT split (category_code,'\\.')[0] AS category FROM retail_ap_eventtype WHERE split (category_code,'\\.')[0]!='';

Output :

Furniture
appliances
accessories
apparel
sport
stationery

```
hive> SELECT DISTINCT split (category_code,'\\.')[0] AS category FROM retail_ap_eventtype WHERE split (category_code,'\\.')[0]!='';
Query ID = hadoop_20220623150426_ecad3b71-4171-4113-8ad2-45119d0d3c71
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1655983803851_0009)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      6          6        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      5          5        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 79.88 s
--------------------------------------------------------------------------------
OK
furniture
appliances
accessories
apparel
sport
stationery
```

QUESTION 5:

Find the total number of products available under each category.

Query: SELECT split (category_code,'\\.')[0] AS category, COUNT (product_id) AS prd FROM retail_ap_eventtype GROUP BY split (category_code,'\\.')[0] ORDER BY prd DESC;

Output:

appliances 61736

stationery 26722

furniture 23604

apparel 18232

accessories 12929

sport 2

```
hive> SELECT split (category_code,'\\.')[0] AS category, COUNT (product_id) AS prd FROM retail_ap_eventtype GROUP BY split (category_code,'\\.')[0] ORDER
 BY prd DESC;
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1655983803851_0009)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      6          6        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: 81.11 s
--------------------------------------------------------------------------------
OK
        8594895
appliances      61736
stationery      26722
furniture       23604
apparel 18232
accessories     12929
sport   2
```

QUESTION 6:

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

Query: SELECT brand, SUM (price) AS Sales FROM retail_ap_eventtype WHERE brand <>'' AND event_type='purchase' GROUP BY brand ORDER BY Sales DESC LIMIT 1 ;

Output : runail 148297.93999999898

```
hive> SELECT brand, SUM (price) AS Sales FROM retail_ap_eventtype WHERE brand <>'' AND event_type='purchase' GROUP BY brand ORDER BY Sales DESC LIMIT 1 ;

Query ID = hadoop_20220623131201_7eddd4cf-7f86-4828-97e5-f0cb0ac9f74d
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1655983803851_0009)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      3          3        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 25.37 s
--------------------------------------------------------------------------------
OK
runail  148297.93999999898
```

QUESTION 7:

Which brands increased their sales from October to November?

Query: WITH MOM_Sales as (select brand, round (sum (case when date_format (event_time, 'MM') = 10 then price else 0 end),2) as Sales_Oct_19, round (sum (case when date_format (event_time, 'MM') =11 then price else 0 end),2) as Sales_Nov_19 from retail_ap_eventtype where event_type = 'purchase' and date_format (event_time, 'MM') in ('10', '11') group by brand) select brand, Sales_Oct_19, Sales_Nov_19, (Sales_Nov_19 - Sales_Oct_19) as MOM_Sales_Difference from MOM_Sales where (Sales_Nov_19-Sales_Oct_19) > 0 order by MOM_Sales_Difference desc;

Output :

	474679.06	619509.24	144830.18
grattol	35445.54	71472.71	36027.170000000006
uno	35302.03	51039.75	15737.720000000001
lianail	5892.84	16394.24	10501.400000000001
ingarden	23161.39	33566.21	10404.82
strong	29196.63	38671.27	9474.639999999996
jessnail	26287.84	33345.23	7057.390000000003
cosmoprofi	8322.81	14536.99	6214.18
polarus	6013.72	11371.93	5358.21
runail	71539.28	76758.66	5219.380000000005
freedecor	3421.78	7671.8	4250.02
staleks	8519.73	11875.61	3355.880000000001
bpw.style	11572.15	14837.44	3265.290000000001
lovely	8704.38	11939.06	3234.6800000000003
marathon	7280.75	10273.1	2992.3500000000004
haruyama	9390.69	12352.91	2962.2199999999993
yoko	8756.91	11707.88	2950.9699999999993
italwax	21940.24	24799.37	2859.1299999999974
benovy	409.62	3259.97	2850.35
kaypro	881.34	3268.7	2387.3599999999997
estel	21756.75	24142.67	2385.9199999999983
concept	11032.14	13380.4	2348.26
kapous	11927.16	14093.08	2165.92
f.o.x	6624.23	8577.28	1953.050000000001
masura	31266.08	33058.47	1792.3899999999994
milv	3904.94	5642.01	1737.0700000000002
beautix	10493.95	12222.95	1729.0
artex	2730.64	4327.25	1596.6100000000001
domix	10472.05	12009.17	1537.1200000000008
shik	3341.2	4839.72	1498.5200000000004
smart	4457.26	5902.14	1444.88
roubloff	3491.36	4913.77	1422.4100000000003
levrana	2243.56	3664.1	1420.54

oniq 8425.41 9841.65 1416.2399999999998
irisk 45591.96 46946.04 1354.0800000000017
severina 4775.88 6120.48 1344.5999999999995
joico 705.52 2015.1 1309.58
zeitun 708.66 2009.63 1300.9700000000003
beauty-free 554.17 1782.86 1228.69
swarovski 1887.93 3043.16 1155.2299999999998
de.lux 1659.7 2775.51 1115.8100000000002
metzger 5373.45 6457.16 1083.71
markell 1768.75 2834.43 1065.6799999999998
sanoto 157.14 1209.68 1052.54
nagaraku 4369.74 5327.68 957.9400000000005
ecolab 262.85 1214.3 951.4499999999999
art-visage 2092.71 2997.8 905.0900000000001
levissime 2227.5 3085.31 857.81
missha 1293.83 2150.28 856.4500000000003
solomeya 1899.7 2685.8 786.1000000000001
rosi 3077.04 3841.56 764.52
refectocil 2716.18 3475.58 759.4000000000001
kaaral 4412.43 5086.07 673.6399999999994
kosmekka 1181.44 1813.37 631.9299999999998
kinetics 6334.25 6945.26 611.0100000000002
browxenna 14331.37 14916.73 585.3599999999988
airnails 5118.9 5691.52 572.6200000000008
uskusi 5142.27 5690.31 548.04
coifin 903.0 1428.49 525.49
s.care 412.68 913.07 500.39000000000004
limoni 1308.9 1796.6 487.6999999999998
matrix 3243.25 3726.74 483.4899999999998
gehwol 1089.07 1557.68 468.6100000000001
greymy 29.21 489.49 460.28000000000003
bioaqua 942.89 1398.12 455.2299999999999
farmavita 837.37 1291.97 454.6
sophin 1067.86 1515.52 447.6600000000001
yu-r 271.41 673.71 402.3
kiss 421.55 817.33 395.78000000000003
naomi 0.0 389.0 389.0
lador 2083.61 2471.53 387.9200000000001
ellips 245.85 606.04 360.18999999999994
jas 3318.96 3657.43 338.4699999999998
lowence 242.84 567.75 324.90999999999997
nitrile 847.28 1162.68 315.4000000000001
shary 871.96 1176.49 304.53
kims 330.04 632.04 301.99999999999994

happyfons	801.92	1091.59	289.66999999999996
kocostar	310.85	594.93	284.0799999999999
insight	1443.7	1721.96	278.26
candy	534.96	799.38	264.41999999999996
bluesky	10307.24	10565.53	258.2900000000009
beauugreen	511.51	768.35	256.84000000000003
protokeratin	201.25	456.79	255.54000000000002
trind	298.07	542.96	244.89000000000004
entity	479.71	719.26	239.55
skinlite	651.94	890.45	238.51
provoc	827.99	1063.82	235.82999999999993
fedua	52.38	263.81	211.43
ecocraft	41.16	241.95	200.79
keen	236.35	435.62	199.27
mane	66.79	260.26	193.46999999999997
freshbubble	318.7	502.34	183.64
matreshka	0.0	182.67	182.67
chi	358.94	538.61	179.67000000000002
cristalinas	427.63	584.95	157.32000000000005
farmona	1692.46	1843.43	150.97000000000003
latinoil	249.52	384.59	135.06999999999996
miskin	158.04	293.07	135.03
elizavecca	70.53	204.3	133.77
nefertiti	233.52	366.64	133.11999999999998
finish	98.38	230.38	132.0
igrobeauty	513.66	645.07	131.41000000000008
dizao	819.13	945.51	126.38
osmo	645.58	762.31	116.7299999999999
batiste	772.4	874.17	101.76999999999998
carmex	145.08	243.36	98.28
eos	54.34	152.61	98.27000000000001
depilflax	2707.07	2803.78	96.71000000000004
enjoy	41.35	136.57	95.22
kerasys	430.91	525.2	94.29000000000002
aura	83.95	177.51	93.55999999999999
plazan	101.37	194.01	92.63999999999999
koelf	422.73	507.29	84.56
nirvel	163.04	234.33	71.29000000000002
konad	739.83	810.67	70.83999999999992
egomania	77.47	146.04	68.57
cutrin	299.37	367.62	68.25
laboratorium	246.5	312.52	66.01999999999998
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 57.620000000000005
koelcia 55.5 112.75 57.25
balbcare 155.33 212.38 57.04999999999998
elskin 251.09 307.65 56.559999999999974
foamie 35.04 80.49 45.449999999999996
ladykin 125.65 170.57 44.91999999999999
likato 296.06 340.97 44.910000000000025
mavala 409.04 446.32 37.27999999999997
vilenta 197.6 231.21 33.610000000000014
beautyblender 78.74 109.41 30.67
biore 60.65 90.31 29.660000000000004
orly 902.38 931.09 28.710000000000036
estelare 444.81 471.87 27.060000000000002
profepil 93.36 118.02 24.659999999999997
blixz 38.95 63.4 24.449999999999996
binacil 0.0 24.26 24.26
godefroy 401.22 425.12 23.899999999999977
glysolid 69.73 91.59 21.86
veraclara 50.11 71.21 21.099999999999994
juno 0.0 21.08 21.08
kamill 63.01 81.49 18.479999999999997
treaclemoon 163.37 181.49 18.120000000000005
supertan 50.37 66.51 16.140000000000008
barbie 0.0 12.39 12.39
deoproce 316.84 329.17 12.330000000000041
rasyan 18.8 28.94 10.14
fly 17.14 27.17 10.030000000000001
tertio 236.16 245.8 9.640000000000015
jaguar 1102.11 1110.65 8.540000000000191
soleo 204.2 212.53 8.330000000000013
neoleor 43.41 51.7 8.290000000000006
moyou 5.71 10.28 4.569999999999999
bodyton 1376.34 1380.64 4.300000000000182
skinity 8.88 12.44 3.5599999999999987
helloganic 0.0 3.1 3.1
grace 100.92 102.61 1.6899999999999977
cosima 20.23 20.93 0.6999999999999993
ovale 2.54 3.1 0.56

```
hive> WITH MOM_Sales as ( select brand, round ( sum (case when date_format (event_time, 'MM') = 10 then price else 0 end),2) as Sales_Oct_19, round (sum (case when date_format (event_ti
'MM') =11 then price else 0 end),2) as Sales_Nov_19 from retail_ap_eventtype where event_type = 'purchase' and date_format (event_time, 'MM') in ('10', '11') group by brand ) select bra
Sales_Oct_19, Sales_Nov_19, (Sales_Nov_19 - Sales_Oct_19) as MOM_Sales_Difference from MOM_Sales where (Sales_Nov_19-Sales_Oct_19) > 0 order by MOM_Sales_Difference desc;
Query ID = hadoop_20220624141953_073ef781-1199-4854-9e32-585d81e30262
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1656074668744_0009)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 ..... container     SUCCEEDED      5          5        0        0       0       0
Reducer 2 ..... container     SUCCEEDED      2          2        0        0       0       0
Reducer 3 ..... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 52.70 s
--------------------------------------------------------------------------------
OK
        474679.06       619509.24       144830.18
grattol 35445.54        71472.71        36027.170000000006
lino    35302.03        51039.75        15737.720000000001
lianail 5892.84 16394.24        10501.400000000001
ingarden        23161.39        33566.21        10404.82
strong  29196.63        38671.27        9474.639999999996
lessnail        26287.84        33345.23        7057.390000000003
cosmoprofi      8322.81 14536.99        6214.18
polarus 6013.72 11371.93        5358.21
runail  71539.28        76758.66        5219.380000000005
freedecor       3421.78 7671.8  4250.02
staleks 8519.73 11875.61        3355.880000000001
bpw.style       11572.15        14837.44        3265.290000000001
lovely  8704.38 11939.06        3234.6800000000003
marathon        7280.75 10273.1 2992.3500000000004
haruyama        9390.69 12352.91        2962.2199999999993
yoko    8756.91 11707.88        2950.969999999993
italwax 21940.24        24799.37        2859.1299999999974
benovy  409.62  3259.97 2850.35
```

benovy	409.62	3259.97	2850.35
kaypro	881.34	3268.7	2387.3599999999997
estel	21756.75	24142.67	2385.9199999999983
concept	11032.14	13380.4	2348.26
kapous	11927.16	14093.08	2165.92
f.o.x	6624.23	8577.28	1953.050000000001
masura	31266.08	33058.47	1792.3899999999994
milv	3904.94	5642.01	1737.0700000000002
beautix	10493.95	12222.95	1729.0
artex	2730.64	4327.25	1596.6100000000001
domix	10472.05	12009.17	1537.1200000000008
shik	3341.2	4839.72	1498.5200000000004
smart	4457.26	5902.14	1444.88
roubloff	3491.36	4913.77	1422.4100000000003
levrana	2243.56	3664.1	1420.54
oniq	8425.41	9841.65	1416.2399999999998
irisk	45591.96	46946.04	1354.0800000000017
severina	4775.88	6120.48	1344.5999999999995
joico	705.52	2015.1	1309.58
zeitun	708.66	2009.63	1300.9700000000003
beauty-free	554.17	1782.86	1228.69
swarovski	1887.93	3043.16	1155.2299999999998
de.lux	1659.7	2775.51	1115.8100000000002
metzger	5373.45	6457.16	1083.71
markell	1768.75	2834.43	1065.6799999999998
sanoto	157.14	1209.68	1052.54
nagaraku	4369.74	5327.68	957.9400000000005
ecolab	262.85	1214.3	951.4499999999999
art-visage	2092.71	2997.8	905.0900000000001
levissime	2227.5	3085.31	857.81
missha	1293.83	2150.28	856.4500000000003
solomeya	1899.7	2685.8	786.1000000000001
rosi	3077.04	3841.56	764.52
refectocil	2716.18	3475.58	759.4000000000001
kaaral	4412.43	5086.07	673.6399999999994
kosmekka	1181.44	1813.37	631.9299999999998
kinetics	6334.25	6945.26	611.0100000000002
browxenna	14331.37	14916.73	585.3599999999988
airnails	5118.9	5691.52	572.6200000000008
uskusi	5142.27	5690.31	548.04
coifin	903.0	1428.49	525.49
s.care	412.68	913.07	500.39000000000004
limoni	1308.9	1796.6	487.6999999999998
matrix	3243.25	3726.74	483.4899999999998
gehwol	1089.07	1557.68	468.6100000000001
greymy	29.21	489.49	460.28000000000003
bioaqua	942.89	1398.12	455.2299999999999
farmavita	837.37	1291.97	454.6
sophin	1067.86	1515.52	447.6600000000001
yu-r	271.41	673.71	402.3
kiss	421.55	817.33	395.78000000000003
naomi	0.0	389.0	389.0
lador	2083.61	2471.53	387.9200000000001
ellips	245.85	606.04	360.18999999999994
jas	3318.96	3657.43	338.4699999999998
lowence	242.84	567.75	324.90999999999997
nitrile	847.28	1162.68	315.4000000000001
shary	871.96	1176.49	304.53
kims	330.04	632.04	301.99999999999994

happyfons	801.92	1091.59	289.66999999999996
kocostar	310.85	594.93	284.0799999999999
insight	1443.7	1721.96	278.26
candy	534.96	799.38	264.41999999999996
bluesky	10307.24	10565.53	258.2900000000009
beauugreen	511.51	768.35	256.84000000000003
protokeratin	201.25	456.79	255.54000000000002
trind	298.07	542.96	244.89000000000004
entity	479.71	719.26	239.55
skinlite	651.94	890.45	238.51
provoc	827.99	1063.82	235.82999999999993
fedua	52.38	263.81	211.43
ecocraft	41.16	241.95	200.79
keen	236.35	435.62	199.27
mane	66.79	260.26	193.46999999999997
freshbubble	318.7	502.34	183.64
matreshka	0.0	182.67	182.67
chi	358.94	538.61	179.67000000000002
cristalinas	427.63	584.95	157.32000000000005
farmona	1692.46	1843.43	150.97000000000003
latinoil	249.52	384.59	135.06999999999996
miskin	158.04	293.07	135.03
elizavecca	70.53	204.3	133.77
nefertiti	233.52	366.64	133.11999999999998
finish	98.38	230.38	132.0
igrobeauty	513.66	645.07	131.41000000000008
dizao	819.13	945.51	126.38
osmo	645.58	762.31	116.7299999999999
batiste	772.4	874.17	101.76999999999998
carmex	145.08	243.36	98.28
eos	54.34	152.61	98.27000000000001
depilflax	2707.07	2803.78	96.71000000000004
enjoy	41.35	136.57	95.22
kerasys	430.91	525.2	94.29000000000002
aura	83.95	177.51	93.55999999999999
plazan	101.37	194.01	92.63999999999999
koelf	422.73	507.29	84.56
nirvel	163.04	234.33	71.29000000000002
konad	739.83	810.67	70.83999999999992
egomania	77.47	146.04	68.57
cutrin	299.37	367.62	68.25
laboratorium	246.5	312.52	66.01999999999998
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	57.620000000000005
koelcia	55.5	112.75	57.25
balbcare	155.33	212.38	57.04999999999998
elskin	251.09	307.65	56.559999999999974
foamie	35.04	80.49	45.449999999999996
ladykin	125.65	170.57	44.91999999999999
likato	296.06	340.97	44.910000000000025
mavala	409.04	446.32	37.27999999999997
vilenta	197.6	231.21	33.610000000000014
beautyblender	78.74	109.41	30.67
biore	60.65	90.31	29.660000000000004

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.

Query: SELECT user_id, SUM(price) AS expense FROM retail_ap_eventtype WHERE event_type='purchase' GROUP BY user_id ORDER BY expense DESC LIMIT 10 ;

Output:

557790271	2715.869999999991
150318419	1645.9700000000005
562167663	1352.8500000000006
531900924	1329.4500000000003
557850743	1295.4799999999996
522130011	1185.3900000000003
561592095	1109.7000000000007
431950134	1097.5899999999997
566576008	1056.3600000000006
521347209	1040.9099999999999

Time taken: 36.364 seconds, Fetched: 10 row(s)

```
hive> SELECT user_id, SUM(price) AS expense FROM retail_ap_eventtype WHERE event_type='purchase' GROUP BY user_id ORDER BY expense DESC LIMIT 10 ;
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_1655983803851_0010)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      3          3        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 29.99 s
--------------------------------------------------------------------------------
OK
557790271       2715.869999999991
150318419       1645.9700000000005
562167663       1352.85
531900924       1329.45
557850743       1295.4800000000005
522130011       1185.3899999999999
561592095       1109.7000000000005
431950134       1097.5899999999997
566576008       1056.3600000000004
521347209       1040.9099999999999
```

Cleaning up :

Once the analysis is completed, deleting the database & terminating the cluster.

```
hive> show databases;
OK
default
hivecasestudyforap
Time taken: 0.025 seconds, Fetched: 2 row(s)
hive> DROP database hivecasestudyforap;
OK
Time taken: 0.356 seconds
hive> SHOW DATABASES;
OK
default
Time taken: 0.011 seconds, Fetched: 1 row(s)
hive>
```