HIVE CASE STUDY

(DSC29)

SUBMITTED BY:SWETA SEAL

PROBLEM STATEMENT:

With online sales gaining popularity, tech companies are exploring ways to improve their sales by analyzing customer behavior and gaining insights about product trends. Furthermore, the websites make it easier for customers to find the products they require without much scavenging. Needless to say, the role of big data analysts is among the most sought- after job profiles of this decade. Therefore, as part of this assignment, we will be challenging you, as a big data analyst, to extract data and gather insights from a real-life data set of an e-commerce company.

OBJECTIVE:

The aim is to extract data and gather insights from a real-life data set of an e-commerce company.

DATA:

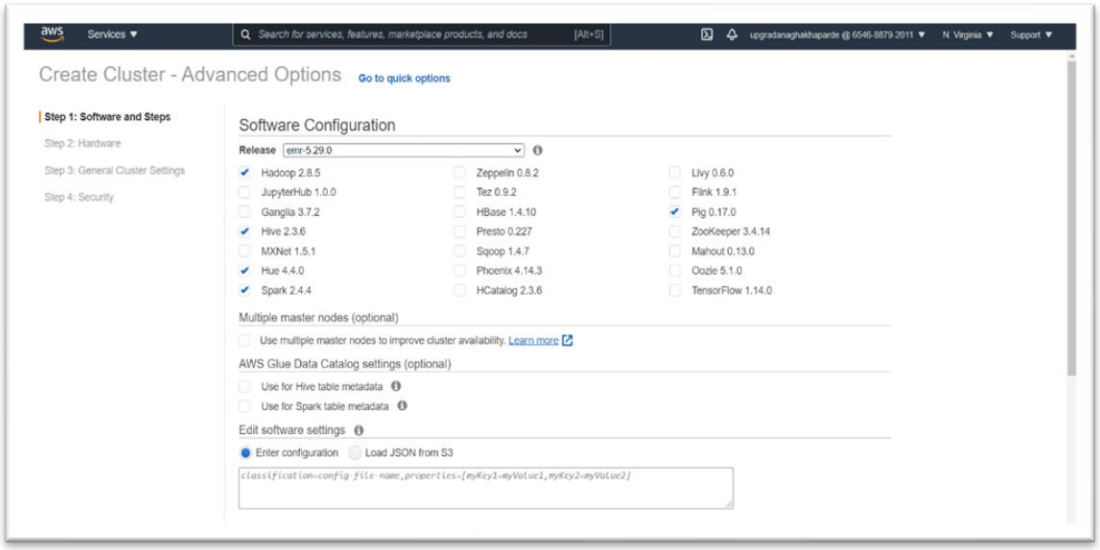
The data used for this assignment is a public clickstream dataset of a cosmetic store. The clickstream data contains all the logs as to how one navigated through the e-commerce website. It also contains other data such as customer time spent on every page, number of clicks made, adding items to the cart, customer id etc.

OVERVIEW OF STEPS:

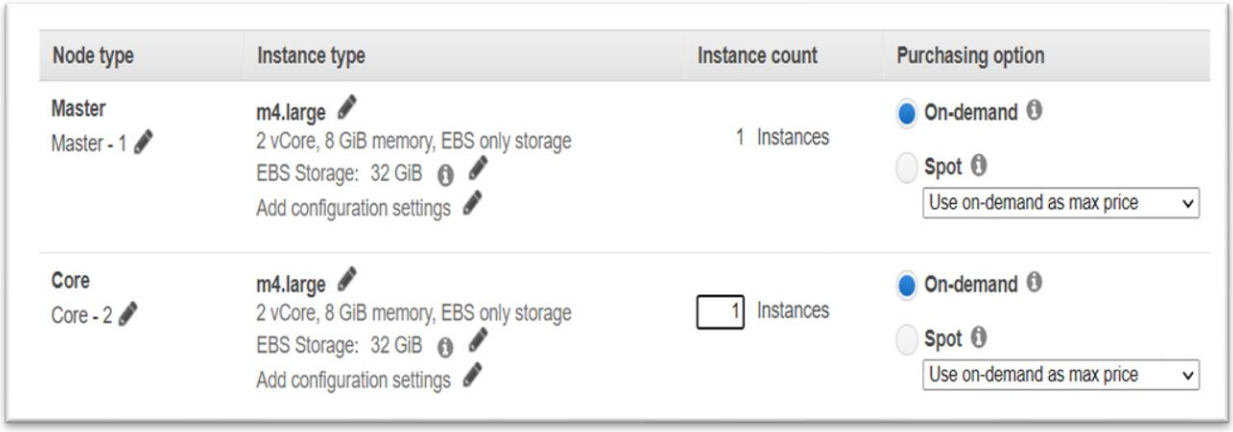
- Copying the data set into HDFS:
  - Launch an EMR cluster that utilizes the hive services, and
  - Move the data from S3 bucket into the HDFS
- Creating the database and launching hive queries on your EMR cluster:
  - Create the structure of your database,
  - Use optimized techniques to run your queries as effectively as possible
  - Show the improvement in performance after optimizing
  - Run hive queries to answer the given questions.
- Cleaning up:
  - Drop your database and
  - Terminate your cluster

❖ CREATING EMR CLUSTER

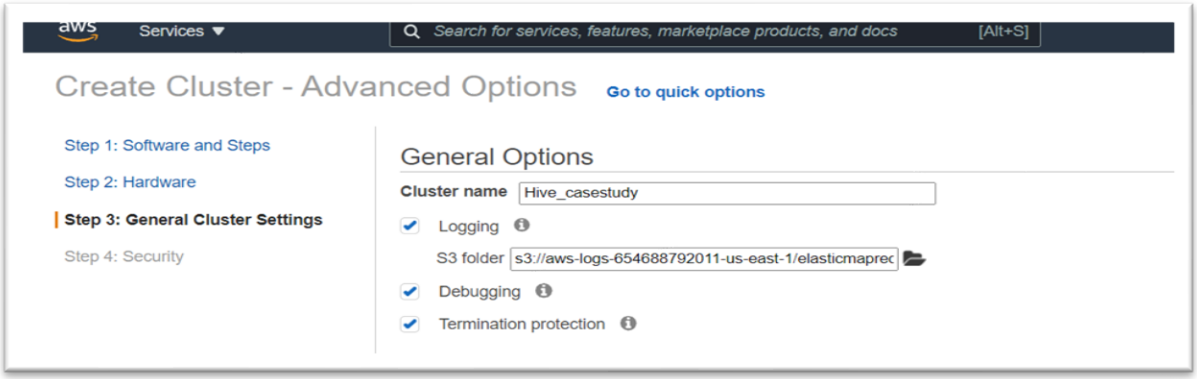
EMR Cluster Landing Page > Create Cluster > Advanced Options >Select the release emr-5.29 and the required services.



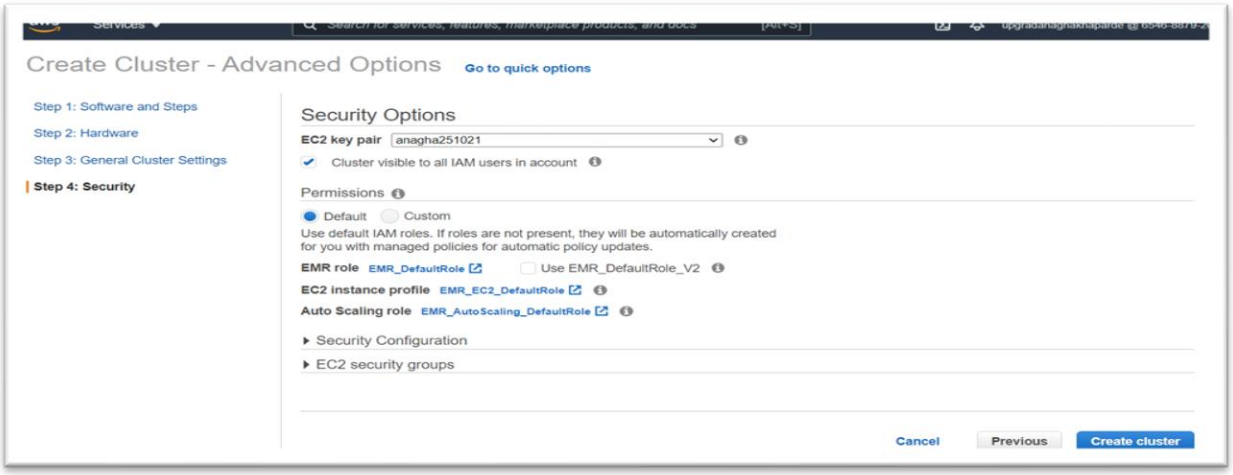
Hardware Configuration Page > To Define the Cluster & Nodes: Instance type for both master and core nodes are M4. large



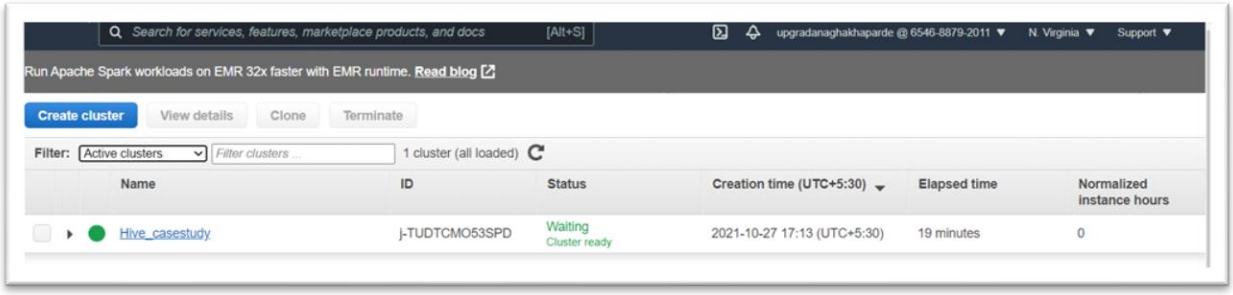
Naming the Cluster uniquely.



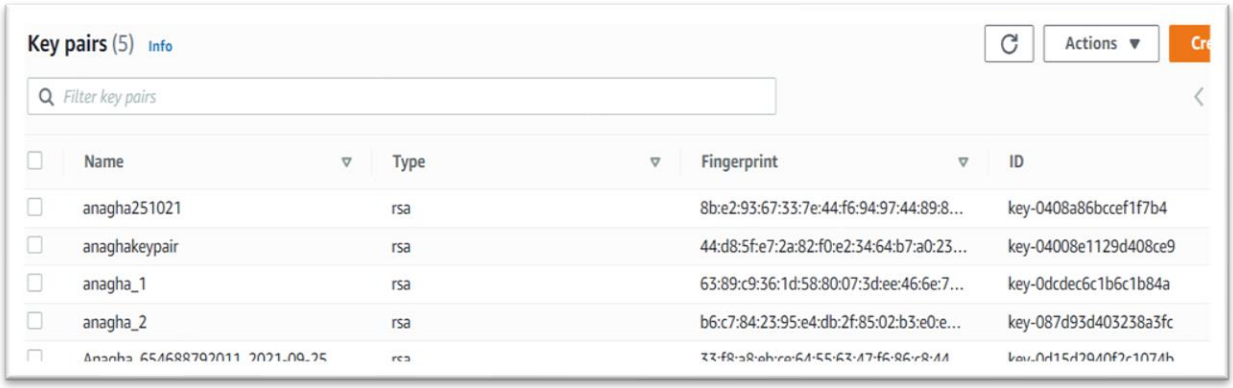
Selecting the key-pair (created before creating the cluster)



Cluster “Hive_casestudy” is successfully created and launched.

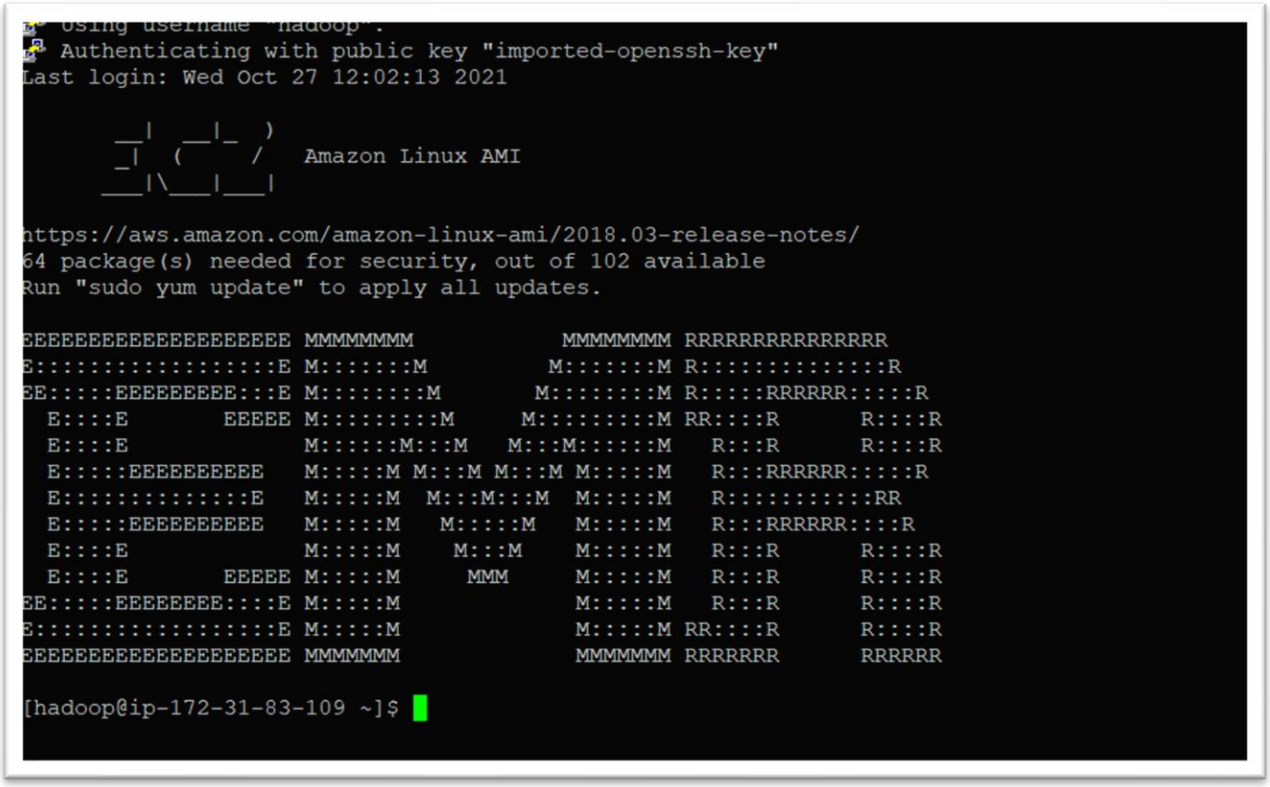


“anagha251021” is the key-pair created for this case study.



- HADOOP & HIVE QUERIES:

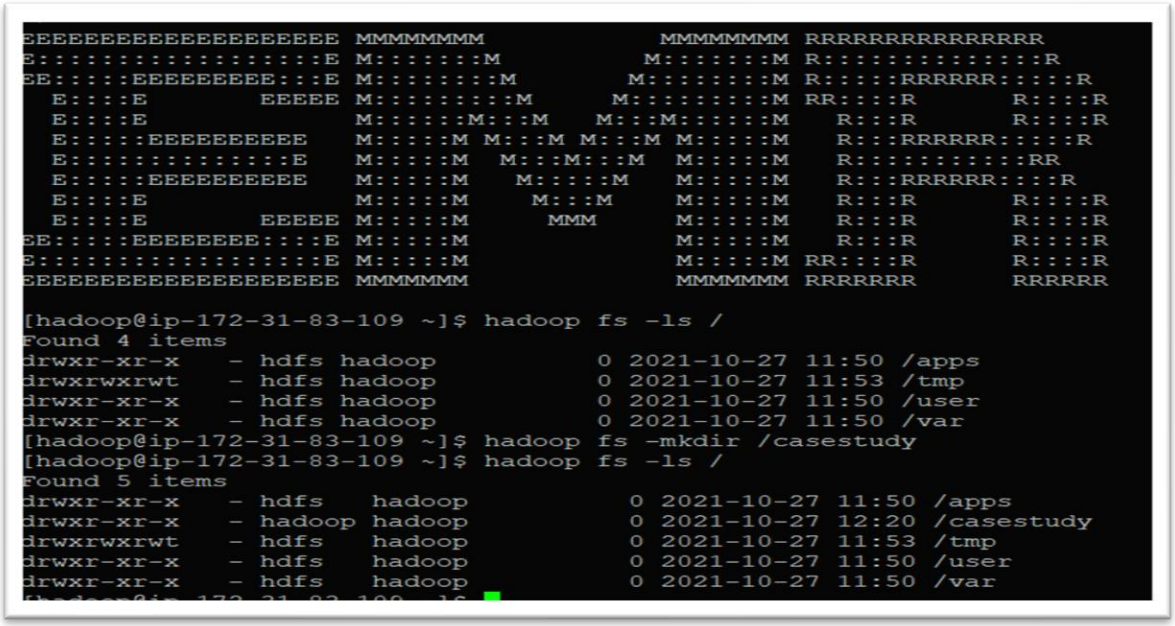
Terminal > connecting to EMR Cluster using SSH



Creating a directory "casestudy"

hadoop fs -mkdir /casestudy

hadoop fs -ls /



Loading the datasets into HDFS from S3:

hadoop distcp s3://anagha1/2019-Oct.csv /casestudy/2019_Oct.csv

```
[hadoop@ip-172-31-83-109 ~]$ hadoop distcp s3://anagha1/2019-Oct.csv /casestudy/2019_Oct.csv
21/10/27 12:31:28 INFO tools.DistCp: Input Options: DistCpOptions{atomicCommit=false, syncFolder=fa
lse, deleteMissing=false, ignoreFailures=false, overwrite=false, skipCRC=false, blocking=true, numL
iststatusThreads=0, maxMaps=20, mapBandwidth=100, sslConfigurationFile='null', copyStrategy='unifor
msize', preserveStatus=[], preserveRawXattrs=false, atomicWorkPath=null, logPath=null, sourceFileLi
sting=null, sourcePaths=[s3://anagha1/2019-Oct.csv], targetPath=/casestudy/2019_Oct.csv, targetPath
Exists=false, filtersFile='null'}
21/10/27 12:31:28 INFO client.RMProxy: Connecting to ResourceManager at ip-172-31-83-109.ec2.intern
al/172.31.83.109:8032
21/10/27 12:31:33 INFO tools.SimpleCopyListing: Paths (files+dirs) cnt = 1; dirCnt = 0
21/10/27 12:31:33 INFO tools.SimpleCopyListing: Build file listing completed.
21/10/27 12:31:33 INFO Configuration.deprecation: io.sort.mb is deprecated. Instead, use mapreduce.
task.io.sort.mb
```

```
        File Input Format Counters
                Bytes Read=210
        File Output Format Counters
                Bytes Written=0
        DistCp Counters
                Bytes Copied=482542278
                Bytes Expected=482542278
                Files Copied=1
[hadoop@ip-172-31-83-109 ~]$
```

hadoop distcp s3://anagha1/2019-Nov.csv /casestudy/2019_Nov.csv

```
[hadoop@ip-172-31-83-109 ~]$ hadoop distcp s3://anagha1/2019-Nov.csv /casestudy/2019_Nov.csv
21/10/27 12:41:00 INFO tools.DistCp: Input Options: DistCpOptions{atomicCommit=false, syncFolder=fals
e, deleteMissing=false, ignoreFailures=false, overwrite=false, skipCRC=false, blocking=true, numLists
tatusThreads=0, maxMaps=20, mapBandwidth=100, sslConfigurationFile='null', copyStrategy='uniformsize'
, preserveStatus=[], preserveRawXattrs=false, atomicWorkPath=null, logPath=null, sourceFileListing=nu
ll, sourcePaths=[s3://anagha1/2019-Nov.csv], targetPath=/casestudy/2019_Nov.csv, targetPathExists=fal
se, filtersFile='null'}
21/10/27 12:41:00 INFO client.RMProxy: Connecting to ResourceManager at ip-172-31-83-109.ec2.internal
/172.31.83.109:8032
21/10/27 12:41:04 INFO tools.SimpleCopyListing: Paths (files+dirs) cnt = 1; dirCnt = 0
21/10/27 12:41:04 INFO tools.SimpleCopyListing: Build file listing completed.
21/10/27 12:41:04 INFO Configuration.deprecation: io.sort.mb is deprecated. Instead, use mapreduce.ta
sk.io.sort.mb
```

```
        File Input Format Counters
                Bytes Read=210
        File Output Format Counters
                Bytes Written=0
        DistCp Counters
                Bytes Copied=545839412
                Bytes Expected=545839412
                Files Copied=1
[hadoop@ip-172-31-83-109 ~]$
```

Viewing the data

hadoop fs -cat /casestudy/2019_Oct.csv | head

```
[hadoop@ip-172-31-83-109 ~]$ hadoop fs -cat /casestudy/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-4778-8d
2c-92e149dab885
2019-10-01 00:00:03 UTC,cart,5773353,1487580005134238553,,runail,2.62,463240011,26dd6e6e-4dac-4778-8d
2c-92e149dab885
2019-10-01 00:00:07 UTC,cart,5881589,2151191071051219817,,lovely,13.48,429681830,49e8d843-adf3-428b-a
2c3-fe8bc6a307c9
2019-10-01 00:00:07 UTC,cart,5723490,1487580005134238553,,runail,2.62,463240011,26dd6e6e-4dac-4778-8d
2c-92e149dab885
2019-10-01 00:00:15 UTC,cart,5881449,1487580013522845895,,lovely,0.56,429681830,49e8d843-adf3-428b-a2
c3-fe8bc6a307c9
2019-10-01 00:00:16 UTC,cart,5857269,1487580005134238553,,runail,2.62,430174032,73dea1e7-664e-43f4-8b
30-d32b9d5af04f
2019-10-01 00:00:19 UTC,cart,5739055,1487580008246412266,,kapous,4.75,377667011,81326ac6-daa4-4f0a-b4
88-fd0956a78733
2019-10-01 00:00:24 UTC,cart,5825598,1487580009445982239,,,0.56,467916806,2f5b5546-b8cb-9ee7-7ecd-842
76f8ef486
2019-10-01 00:00:25 UTC,cart,5698989,1487580006317032337,,,1.27,385985999,d30965e8-1101-44ab-b45d-cc1
bb9fae694
```

hadoop fs -cat /casestudy/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-33527f4
de241
2019-11-01 00:00:09 UTC,cart,5844397,1487580006317032337,,,2.38,553329724,2067216c-31b5-455d-a1cc-af0575a
34ffb
2019-11-01 00:00:10 UTC,view,5837166,1783999064103190764,,pnb,22.22,556138645,57ed222e-a54a-4907-9944-5a8
75c2d7f4f
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-bd36-0fc039e
77580
2019-11-01 00:00:34 UTC,remove_from_cart,5870838,1487580007675986893,,milv,0.79,429913900,2f0bff3c-252f-4
fe6-afcd-5d8a6a92839a
```

DATASETS ARE SUCCESSFULLY LOADED.

LAUNCH HIVE

```
[hadoop@ip-172-31-83-109 ~]$ hive

Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j2.properties Async: false
hive> show databases;
OK
default
Time taken: 0.605 seconds, Fetched: 1 row(s)
hive>
```

Creating new database “hive_assignment”

Hive> create database if not exists hive_assignmnet;

Hive> show databases;

Hive > describe database hive_assignmnet;

```
hive> create database if not exists hive_assignmnet;
OK
Time taken: 0.306 seconds
hive> show databases;
OK
default
hive_assignmnet
Time taken: 0.016 seconds, Fetched: 2 row(s)
hive> describe database hive_assignmnet;
OK
hive_assignmnet         hdfs://ip-172-31-83-109.ec2.internal:8020/user/hive/warehouse/hive_assignmnet.db  h
adoop   USER
Time taken: 0.046 seconds, Fetched: 1 row(s)
hive>
```

Creating new table “retail”

CREATE EXTERNAL TABLE IF NOT EXISTS retail (event_time timestamp, event_type string, product_id string, category_id string, category_code string, brand string, price decimal(10,3), user_id bigint, user_session string) ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.OpenCSVSerde' WITH SERDEPROPERTIES ("separatorChar" = "," , "quoteChar" = "\"", "escapeChar" = "\\") stored as textfile LOCATION '/casestudy' TBLPROPERTIES ("skip.header.line.count"="1") ;

```
hive>  CREATE EXTERNAL TABLE IF NOT EXISTS retail (event_time timestamp, event_t
ype string,
    > product_id string, category_id string, category_code string, brand string,
 price decimal(10,3), user_id bigint,
    > user_session string) ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.OpenC
SVSerde' WITH
    > SERDEPROPERTIES ("separatorChar" = "," , "quoteChar" = "\"", "escapeChar"
= "\\") stored as textfile
    > LOCATION '/casestudy' TBLPROPERTIES ("skip.header.line.count"="1") ;
OK
Time taken: 0.419 seconds
hive>
```

Hive> describe retail;

```
hive> describe retail;
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: 0.202 seconds, Fetched: 9 row(s)
hive>
```

LOADING DATA INTO TABLE “retail”

Hive> LOAD DATA INPATH '/casestudy/2019_Oct.csv' INTO TABLE retail ;

```
hive> LOAD DATA INPATH '/casestudy/2019_Oct.csv' INTO TABLE retail ;
Loading data to table default.retail
OK
Time taken: 1.137 seconds
```

hive> LOAD DATA INPATH '/casestudy/2019_Nov.csv' INTO TABLE retail;

```
hive>  LOAD DATA INPATH '/casestudy/2019_Nov.csv' INTO TABLE retail;
Loading data to table default.retail
OK
Time taken: 0.649 seconds
```

PERFORMING DATA CHECK:

Hive> SELECT* FROM retail WHERE MONTH(event_time)=11 limit 5;

Hive> SELECT* FROM retail WHERE MONTH(event_time)=10 limit 5;

```
hive> SELECT* FROM retail WHERE MONTH(event_time)=11 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     22.2
2       556138645       57ed222e-a54a-4907-9944-5a875c2d7f4f
2019-11-01 00:00:11 UTC cart    5876812 1487580010100293687             jessnail    3
.16     564506666       186c1951-8052-4b37-adce-dd9644b1d5f7
2019-11-01 00:00:24 UTC remove_from_cart        5826182 1487580007483048900         3
.33     553329724       2067216c-31b5-455d-a1cc-af0575a34ffb
Time taken: 2.42 seconds, Fetched: 5 row(s)
hive> SELECT* FROM retail WHERE MONTH(event_time)=10 limit 5;
OK
2019-10-01 00:00:00 UTC cart    5773203 1487580005134238553             runail  2.62    4
63240011        26dd6e6e-4dac-4778-8d2c-92e149dab885
2019-10-01 00:00:03 UTC cart    5773353 1487580005134238553             runail  2.62    4
63240011        26dd6e6e-4dac-4778-8d2c-92e149dab885
2019-10-01 00:00:07 UTC cart    5881589 2151191071051219817             lovely  13.48   4
29681830        49e8d843-adf3-428b-a2c3-fe8bc6a307c9
2019-10-01 00:00:07 UTC cart    5723490 1487580005134238553             runail  2.62    4
63240011        26dd6e6e-4dac-4778-8d2c-92e149dab885
2019-10-01 00:00:15 UTC cart    5881449 1487580013522845895             lovely  0.56    4
29681830        49e8d843-adf3-428b-a2c3-fe8bc6a307c9
Time taken: 0.4 seconds, Fetched: 5 row(s)
```

QUESTION 1:

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

> SELECT SUM(price) FROM retail WHERE MONTH(event_time)=10 AND event_type='purchase' ;

```
hive> SELECT SUM(price) FROM retail WHERE MONTH(event_time)=10 AND event_type='purchase' ;
Query ID = hadoop_20211027155528_32d2afcc-228a-44f5-8f1a-75d9fa96b3da
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_1635335523172_0005)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      2          2        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 63.44 s
--------------------------------------------------------------------------------
OK
1211538.4299997438
Time taken: 73.857 seconds, Fetched: 1 row(s)
```

Time taken to execute the above querry is 73.85 sec.

This is very high. Hence, to reduce this execution time, we will dynamically partition the table “retail” and add bucket to create an optimized table.

DYNAMIC PARTITIONING:

Hive> set hive.exec.dynamic.partition=true;

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

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

PARTITION TABLE 1: retail_part_1

Partition on: event_type (there are 4 types and all questions are related to ‘purchase’)

> CREATE EXTERNAL TABLE IF NOT EXISTS retail_part_1 (event_time timestamp, product_id string, category_id string, category_code string, brand string, price decimal(10,3), 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 ;

>describe retail_part_1 ;

```
hive> CREATE EXTERNAL TABLE IF NOT EXISTS retail_part_1 (event_time timestamp, product_id string, category_id string
, category_code string, brand string, price decimal(10,3), 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' ST
ORED AS textfile ;
OK
Time taken: 0.087 seconds
hive> describe retail_part_1 ;
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
Time taken: 0.117 seconds, Fetched: 14 row(s)
```

>INSERT INTO TABLE retail_part_1 PARTITION (event_type) SELECT event_time,product_id,category_id, category_code, brand, price, user_id, user_session, event_type FROM retail;

```
hive> INSERT INTO TABLE retail_part_1 PARTITION (event_type) SELECT event_time,product_id,category_id, category_code
, brand, price, user_id, user_session, event_type FROM retail;
Query ID = hadoop_20211027171727_b55afe67-b3f8-4ba2-a481-6e2f3f5d9edb
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_1635335523172_0006)

--------------------------------------------------------------------------------
        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: 161.94 s
--------------------------------------------------------------------------------
Loading data to table default.retail_part_1 partition (event_type=null)

Loaded : 4/4 partitions.
        Time taken to load dynamic partitions: 0.412 seconds
        Time taken for adding to write entity : 0.002 seconds
OK
Time taken: 171.816 seconds
hive>
```

Executing the same query with the new table “retail_part_1” to check the time.

> SELECT SUM(price) FROM retail_part_1 WHERE MONTH(event_time)=10 AND event_type='purchase' ;

```
hive> SELECT SUM(price) FROM retail_part_1 WHERE MONTH(event_time)=10 AND
    > event_type='purchase' ;
Query ID = hadoop_20211027172928_d7195305-c4c8-4b13-802e-a4f489afb7f7
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1635335523172_0007)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      3          3        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 23.37 s
--------------------------------------------------------------------------------
OK
1211538.429999898
Time taken: 24.502 seconds, Fetched: 1 row(s)
hive>
```

Time taken to execute the above query is 24.502 sec.

PARTITION TABLE 2: retail_part_2

Partition on : month

Hive> > CREATE EXTERNAL TABLE IF NOT EXISTS retail_part_2 (event_time string, event_type string, product_id string, category_id string, category_code string, brand string, price decimal(10,3), user_id bigint, user_session string) PARTITIONED BY(month int) CLUSTERED BY (brand) INTO 5 buckets ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.OpenCSVSerde' STORED AS textfile ;

>describe retail_part_2 ;

```
hive> CREATE EXTERNAL TABLE IF NOT EXISTS retail_part_2 (event_time timestamp, event_type string, product_id string,
 category_id string, category_code string, brand string, price decimal(10,3), user_id bigint, user_session string) P
ARTITIONED BY(month int) CLUSTERED BY (brand) INTO 5 buckets ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.OpenCSV
Serde' STORED AS textfile ;
OK
Time taken: 0.159 seconds
hive> describe retail_part_2 ;
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
month                   int

# Partition Information
# col_name              data_type               comment

month                   int
Time taken: 0.287 seconds, Fetched: 15 row(s)
```

>INSERT INTO TABLE retail_part_2 PARTITION (month) SELECT event_time, event_type, product_id, category_id, category_code, brand, price, user_id, user_session, MONTH(CAST(REPLACE(event_time,'UTC',’’) AS timestamp)) FROM retail ;

```
hive> INSERT INTO TABLE retail_part_2 PARTITION (month) SELECT event_time, event_type, product
_id, category_id, category_code, brand, price, user_id, user_session, MONTH(CAST(REPLACE(event
_time,'UTC','') AS timestamp)) FROM retail ;
Query ID = hadoop_20211028105359_7567cb1f-c087-405a-b4e5-09cb57a81a66
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_1635416412840_0004)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      2          2        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      5          5        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 178.50 s
--------------------------------------------------------------------------------
Loading data to table default.retail_part_2 partition (month=null)

Loaded : 2/2 partitions.
         Time taken to load dynamic partitions: 0.203 seconds
         Time taken for adding to write entity : 0.0 seconds
OK
Time taken: 187.45 seconds
```

Executing the same query with the new table ‘retail_part_2’ to check the time.

>Select sum(price) from retail_part_2 where month(event_type) = 10 and event_type =’purchase’ ;

```
hive> SELECT SUM(price) FROM retail_part_2 WHERE MONTH(event_time)=10 AND
    > event_type='purchase' ;
Query ID = hadoop_20211028105851_4d3683ce-1769-4a28-8706-7a969fb98f26
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1635416412840_0004)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      8          8        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 76.62 s
--------------------------------------------------------------------------------
OK
1211538.4299999713
Time taken: 77.334 seconds, Fetched: 1 row(s)
```

Time taken to execute the above query is 77.85 sec.

We get an optimized table by partitioning on ‘event_type’ and clustering by ‘user_id’.

Hence, for all the following analysis, we will be using the optimized table’retail_part_1’.

QUESTION 1:

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

> SELECT SUM(price) FROM retail_part_1 WHERE MONTH(event_time)=10 AND event_type='purchase' ;

```
hive>  SELECT SUM(price) FROM retail_part_1 WHERE MONTH(event_time)=10 AND
    > event_type='purchase' ;
Query ID = hadoop_20211028113022_ea6e7a22-0fce-4dc2-a604-6b9c75155bf7
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1635416412840_0005)

--------------------------------------------------------------------------------
        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: 22.95 s
--------------------------------------------------------------------------------
OK
1211538.4299998982
Time taken: 23.581 seconds, Fetched: 1 row(s)
```

The total revenue generated in the month of October is 1211538.429.

QUESTION 2:

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

>SELECT MONTH(event_time), SUM(price) as sum_purchase, COUNT(event_type) as cnt FROM retail_part_1 WHERE event_type='purchase' GROUP BY MONTH(event_time) ;

```
hive>  SELECT MONTH(event_time), SUM(price) as sum_purchase, COUNT(event_type) as cnt FROM
    > retail_part_1 WHERE event_type='purchase' GROUP BY MONTH(event_time) ;
Query ID = hadoop_20211028113556_e90eb61a-909d-4728-8a76-aa6bbf665091
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1635416412840_0005)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      3          3        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 23.19 s
--------------------------------------------------------------------------------
OK
10      1211538.4299998982      245624
11      1531016.8999999384      322417
Time taken: 23.827 seconds, Fetched: 2 row(s)
```

In the month of October 245624 purchases generated revenue of 1211538.4299. Similarly in the month of November 322417 purchases generated revenue of 1531016.899.

QUESTION 3:

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

>WITH diff 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_part_1 WHERE date_format(event_time,'MM') IN (10,11) AND event_type='purchase') SELECT October, November, (November - October) as Difference FROM diff ;

```
hive> WITH diff 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_part_1 WHERE date_format(event_time,'MM') IN (10,11) AND
    > event_type='purchase') SELECT October, November, (November - October) as Difference FROM
 diff ;
Query ID = hadoop_20211028114342_53e37404-299c-48e7-b0c0-5775e42e7084
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_1635416412840_0006)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      3          3        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 38.47 s
--------------------------------------------------------------------------------
OK
1211538.429999898       1531016.8999999384      319478.4700000405
Time taken: 47.861 seconds, Fetched: 1 row(s)
```

The change in revenue generated from October to November is 319478.47.

QUESTION 4:

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

>SELECT DISTINCT split(category_code,'\\.')[0] AS category FROM retail_part_1 WHERE split(category_code,'\\.')[0]<>'' ;

```
hive> SELECT DISTINCT split(category_code,'\\.')[0] AS category FROM retail_part_1 WHERE split
(category_code,'\\.')[0]<>'' ;
Query ID = hadoop_20211028115551_cdbd10cb-d523-4267-aa58-32995ac0c395
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1635416412840_0007)

--------------------------------------------------------------------------------
        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: 67.16 s
--------------------------------------------------------------------------------
OK
furniture
appliances
accessories
apparel
sport
stationery
Time taken: 67.791 seconds, Fetched: 6 row(s)
```

There are 6 types of products. They are: Furniture, Appliances, Accessories, Apparel, Sport, Stationery.

QUESTION 5:

Find the total number of products available under each category.

hive>SELECT split(category_code,'\\.')[0] AS category, COUNT(product_id) AS prd FROM retail_part_1 GROUP BY split(category_code,'\\.')[0] ORDER BY prd DESC ;

```
hive> SELECT split(category_code,'\\.')[0] AS category, COUNT(product_id) AS prd FROM retail_p
art_1 GROUP BY split(category_code,'\\.')[0] ORDER BY prd DESC ;
Query ID = hadoop_20211028120053_b11d5416-f32c-44b3-a0f6-7c9e6be0bf22
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1635416412840_0007)

--------------------------------------------------------------------------------
        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: 70.13 s
--------------------------------------------------------------------------------
OK
        8594895
appliances      61736
stationery      26722
furniture       23604
apparel 18232
accessories     12929
sport   2
Time taken: 70.79 seconds, Fetched: 7 row(s)
```

'Sport' category has the least number of products, whereas 'appliances' has 61736 products.

Question 6:

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

>SELECT brand, SUM(price) AS Sales FROM retail_part_1 WHERE brand <>'' AND event_type='purchase' GROUP BY brand ORDER BY Sales DESC LIMIT 1 ;

```
hive> SELECT brand, SUM(price) AS Sales FROM retail_part_1 WHERE brand <>'' AND event_type='pu
rchase'
    > GROUP BY brand ORDER BY Sales DESC LIMIT 1 ;
Query ID = hadoop_20211028120554_b837b3f8-e130-40c2-9f04-4f8148a8fb15
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1635416412840_0007)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      3          3        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 22.97 s
--------------------------------------------------------------------------------
OK
runail  148297.93999999898
Time taken: 23.903 seconds, Fetched: 1 row(s)
```

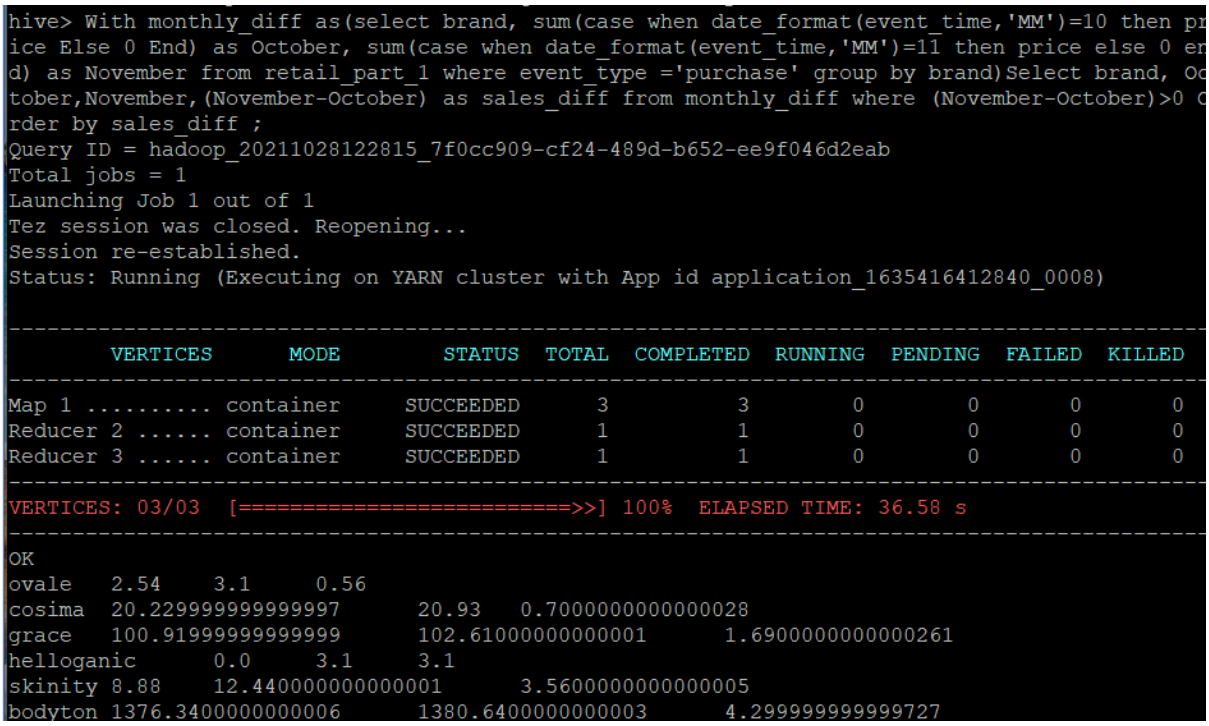
Brand 'runail' has the maximum sales for both months combined.

QUESTION 7:

Which brands increased their sales from October to November?

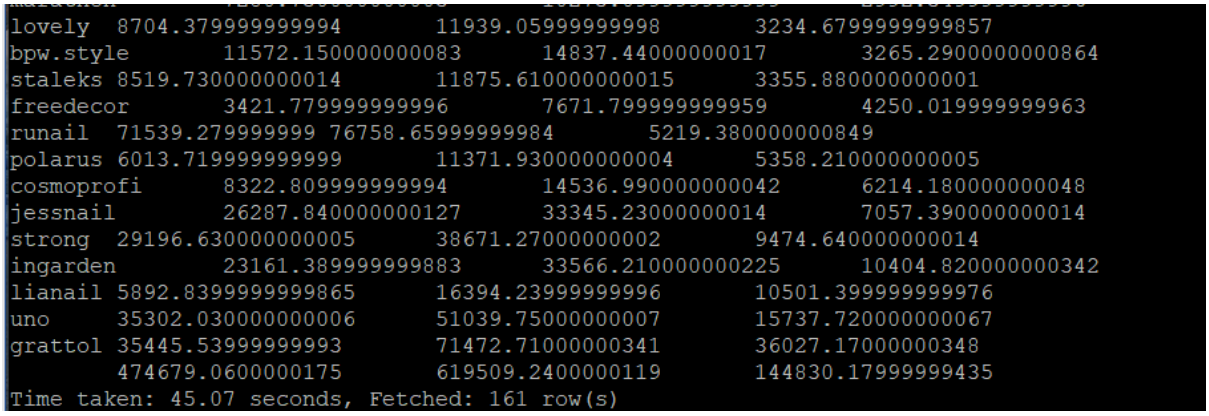
hive>WITH monthly_diff AS (SELECT brand, SUM(CASE WHEN date_format(event_time,'MM')=10 THEN price ELSE 0 END) AS October, SUM(CASE WHEN date_format(event_time,'MM')=11 THEN price ELSE 0 END) AS November FROM retail_part_1 WHERE event_type='purchase' GROUP BY brand) SELECT brand, October,

November, (November-October) as Sales_diff FROM monthly_diff WHERE (November-October) >0 ORDER BY Sales_diff ;



```
hive> With monthly_diff as(select brand, sum(case when date_format(event_time,'MM')=10 then pr
ice Else 0 End) as October, sum(case when date_format(event_time,'MM')=11 then price else 0 en
d) as November from retail_part_1 where event_type ='purchase' group by brand)Select brand, Oc
tober,November,(November-October) as sales_diff from monthly_diff where (November-October)>0 O
rder by sales_diff ;
Query ID = hadoop_20211028122815_7f0cc909-cf24-489d-b652-ee9f046d2eab
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_1635416412840_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
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 36.58 s
--------------------------------------------------------------------------------
OK
ovale   2.54    3.1     0.56
cosima  20.229999999999997      20.93   0.7000000000000028
grace   100.91999999999999      102.61000000000001      1.6900000000000261
helloganic      0.0     3.1     3.1
skinity 8.88    12.440000000000001      3.5600000000000005
bodyton 1376.3400000000006      1380.6400000000003      4.299999999999727
```



```
lovely  8704.379999999994       11939.05999999998       3234.6799999999857
bpw.style       11572.150000000083      14837.44000000017       3265.2900000000864
staleks 8519.730000000014       11875.610000000015      3355.880000000001
freedecor       3421.779999999996       7671.799999999959       4250.019999999963
runail  71539.279999999 76758.65999999984       5219.380000000849
polarus 6013.719999999999       11371.930000000004      5358.210000000005
cosmoprofi      8322.809999999994       14536.990000000042      6214.180000000048
jessnail        26287.840000000127      33345.23000000014       7057.390000000014
strong  29196.630000000005      38671.27000000002       9474.640000000014
ingarden        23161.389999999883      33566.210000000225      10404.820000000342
lianail 5892.8399999999865      16394.23999999996       10501.399999999976
uno     35302.030000000006      51039.75000000007       15737.720000000067
grattol 35445.53999999993       71472.71000000341       36027.17000000348
        474679.0600000175       619509.2400000119       144830.17999999435
Time taken: 45.07 seconds, Fetched: 161 row(s)
```

Total of 161 brands have increased their sales 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.

hive>SELECT user_id, SUM(price) AS expense FROM retail_part_1 WHERE event_type='purchase' GROUP BY user_id ORDER BY expense DESC LIMIT 10 ;

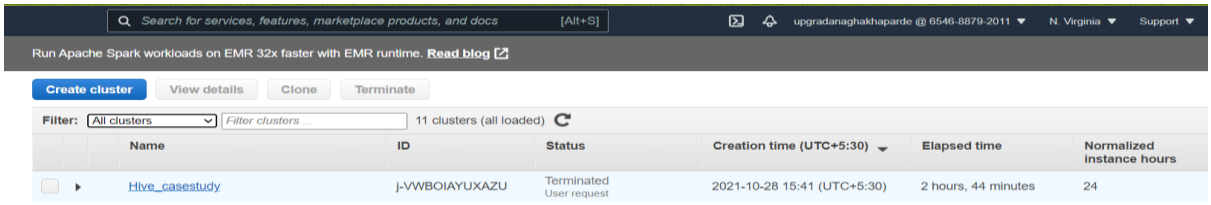
```
hive> SELECT user_id, SUM(price) AS expense FROM retail_part_1 WHERE event_type='purchase' Gro
up by user_id Order by expense desc limit 10 ;
Query ID = hadoop_20211028123701_cc924276-2014-4bfc-a578-ebfe4196f0f5
Total jobs = 1
Launching Job 1 out of 1
Tez session was closed. Reopening...
Session re-established.
Status: Running (Executing on YARN cluster with App id application_1635416412840_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: 24.79 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
Time taken: 32.669 seconds, Fetched: 10 row(s)
```

Above is the list of top 10 users(user_ids along with the amount spent) who spent the most.

- **Cleaning up:**
  Once the analysis is completed, deleting the database and terminating the cluster.

```
hive> drop database hive_assignmnet ;
OK
Time taken: 0.184 seconds
```



THANK YOU.