HIVE CASE STUDY

By Rakhee Kumari & Steven

Problem Statement

With online sales gaining popularity, tech companies are exploring ways to improve their sales by analysing customer behaviour 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.

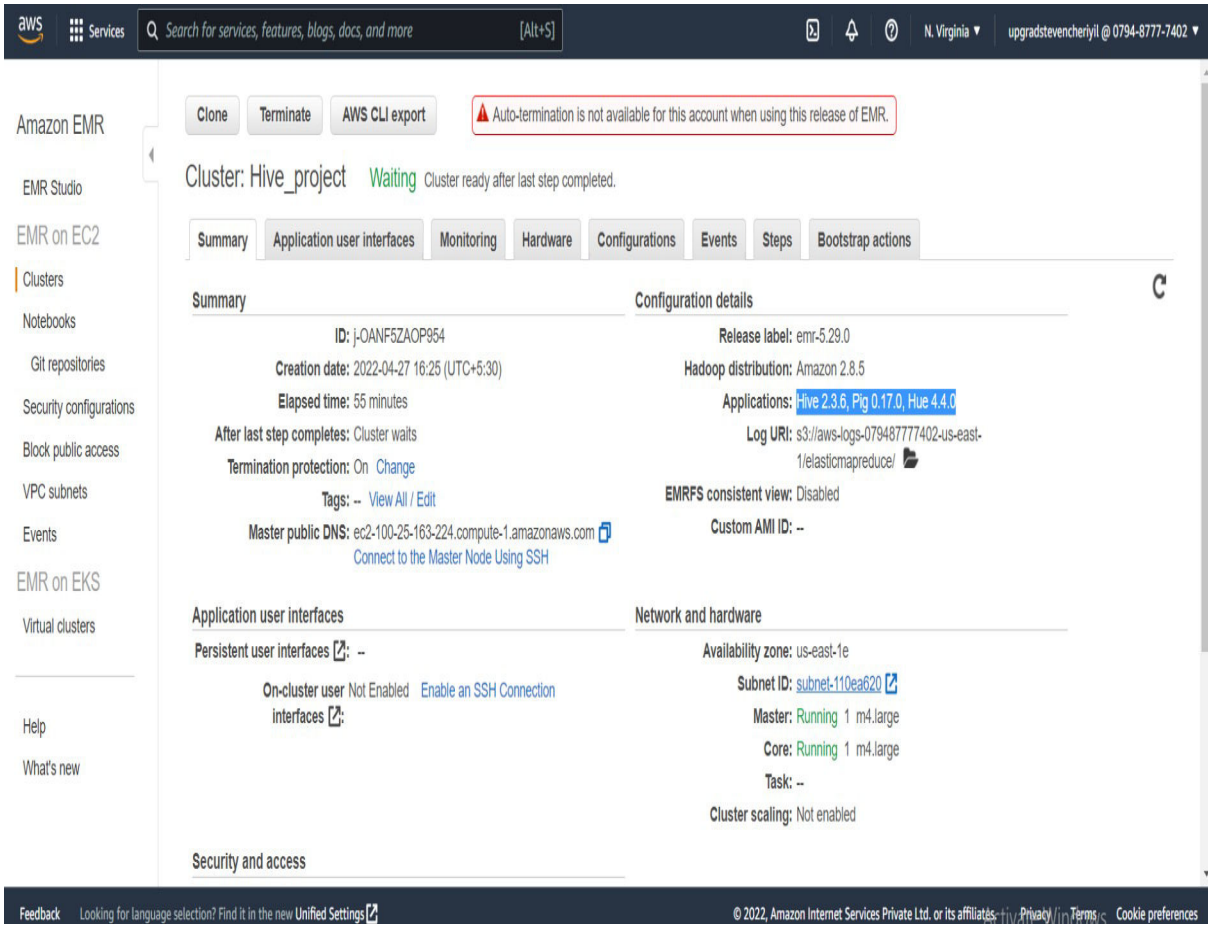
Data for the case study is in the link given below.
https://e-commerce-events-ml.
https://e-commerce-events-ml.

We used a **2-node** EMR cluster with both the master and core nodes as **M4.large**.

IMPORTING THE DATA INTO HDFS

We login to Nuvepro dashboard, go to the console and then to EMR home page → Click on Create Cluster → select release EMR 5.29.0 and select required service for the case study.

1 Launching an EMR cluster that utilizes Hive services



2 Creating a folder on the Hadoop file system

hadoop@ip-172-31-61-99:~

```
Using username "hadoop".
Authenticating with public key "imported-openssh-key"
Last login: Wed Apr 27 11:07:43 2022

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

https://aws.amazon.com/amazon-linux-ami/2018.03-release-notes/
72 package(s) needed for security, out of 102 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-172-31-61-99 ~]$ ls
[hadoop@ip-172-31-61-99 ~]$ hadoop fs -mkdir /test-folder
[hadoop@ip-172-31-61-99 ~]$ hadoop fs -ls /
Found 5 items
drwxr-xr-x   - hdfs   hadoop          0 2022-04-27 11:02 /apps
drwxr-xr-x   - hadoop hadoop          0 2022-04-27 11:11 /test-folder
drwxrwxrwt   - hdfs   hadoop          0 2022-04-27 11:05 /tmp
drwxr-xr-x   - hdfs   hadoop          0 2022-04-27 11:02 /user
drwxr-xr-x   - hdfs   hadoop          0 2022-04-27 11:02 /var
```

3 Moving the data from S3 bucket into the HDFS

- **Importing 2019-Nov.csv file from S3 to HDFS**

hadoop@ip-172-31-61-99:~

```
-bash: distcp: command not found
[hadoop@ip-172-31-61-99 ~]$ hadoop distcp s3n://upgradprojectbucket/2019-Nov.csv
 /tmp/test-folder/2019-Nov.csv
22/04/27 11:19:37 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=[s3n://upgradprojectbucket/2019-Nov.csv], targetPath=/tmp/te
st-folder/2019-Nov.csv, targetPathExists=false, filtersFile='null'}
22/04/27 11:19:38 INFO client.RMProxy: Connecting to ResourceManager at ip-172-3
1-61-99.ec2.internal/172.31.61.99:8032
22/04/27 11:19:44 INFO tools.SimpleCopyListing: Paths (files+dirs) cnt = 1; dirC
nt = 0
22/04/27 11:19:44 INFO tools.SimpleCopyListing: Build file listing completed.
22/04/27 11:19:44 INFO Configuration.deprecation: io.sort.mb is deprecated. Inst
ead, use mapreduce.task.io.sort.mb
22/04/27 11:19:44 INFO Configuration.deprecation: io.sort.factor is deprecated.
Instead, use mapreduce.task.io.sort.factor
22/04/27 11:19:44 INFO tools.DistCp: Number of paths in the copy list: 1
22/04/27 11:19:44 INFO tools.DistCp: Number of paths in the copy list: 1
22/04/27 11:19:44 INFO client.RMProxy: Connecting to ResourceManager at ip-172-3
1-61-99.ec2.internal/172.31.61.99:8032
22/04/27 11:19:45 INFO mapreduce.JobSubmitter: number of splits:1
22/04/27 11:19:45 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_16
51057437121_0001
22/04/27 11:19:46 INFO impl.YarnClientImpl: Submitted application application_16
51057437121_0001
22/04/27 11:19:47 INFO mapreduce.Job: The url to track the job: http://ip-172-31
-61-99.ec2.internal:20888/proxy/application_1651057437121_0001/
22/04/27 11:19:47 INFO tools.DistCp: DistCp job-id: job_1651057437121_0001
22/04/27 11:19:47 INFO mapreduce.Job: Running job: job_1651057437121_0001
22/04/27 11:19:58 INFO mapreduce.Job: Job job_1651057437121_0001 running in uber
 mode : false
22/04/27 11:19:58 INFO mapreduce.Job:  map 0% reduce 0%
22/04/27 11:20:16 INFO mapreduce.Job:  map 100% reduce 0%
22/04/27 11:20:19 INFO mapreduce.Job: Job job_1651057437121_0001 completed succe
ssfully
22/04/27 11:20:19 INFO mapreduce.Job: Counters: 38
        File System Counters
                FILE: Number of bytes read=0
                FILE: Number of bytes written=172471
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=359
                HDFS: Number of bytes written=545839412
                HDFS: Number of read operations=12
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=4
```

hadoop@ip-172-31-61-99:~

```
22/04/27 11:19:58 INFO mapreduce.Job:  map 0% reduce 0%
22/04/27 11:20:16 INFO mapreduce.Job:  map 100% reduce 0%
22/04/27 11:20:19 INFO mapreduce.Job: Job job_1651057437121_0001 completed succe
ssfully
22/04/27 11:20:19 INFO mapreduce.Job: Counters: 38
        File System Counters
                FILE: Number of bytes read=0
                FILE: Number of bytes written=172471
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=359
                HDFS: Number of bytes written=545839412
                HDFS: Number of read operations=12
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=4
                S3N: Number of bytes read=545839412
                S3N: Number of bytes written=0
                S3N: Number of read operations=0
                S3N: Number of large read operations=0
                S3N: Number of write operations=0
        Job Counters
                Launched map tasks=1
                Other local map tasks=1
                Total time spent by all maps in occupied slots (ms)=590112
                Total time spent by all reduces in occupied slots (ms)=0
                Total time spent by all map tasks (ms)=18441
                Total vcore-milliseconds taken by all map tasks=18441
                Total megabyte-milliseconds taken by all map tasks=18883584
        Map-Reduce Framework
                Map input records=1
                Map output records=0
                Input split bytes=136
                Spilled Records=0
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=370
                CPU time spent (ms)=20990
                Physical memory (bytes) snapshot=563040256
                Virtual memory (bytes) snapshot=3299827712
                Total committed heap usage (bytes)=460324864
        File Input Format Counters
                Bytes Read=223
        File Output Format Counters
                Bytes Written=0
        DistCp Counters
                Bytes Copied=545839412
                Bytes Expected=545839412
                Files Copied=1
```

- **Importing 2019-Oct.csv file from S3 to HDFS**

hadoop@ip-172-31-61-99:~

```
ov.csv
[hadoop@ip-172-31-61-99 ~]$ hadoop distcp s3n://upgradprojectbucket/2019-Oct.csv
 /tmp/test-folder/2019-Oct.csv
22/04/27 11:28:28 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=[s3n://upgradprojectbucket/2019-Oct.csv], targetPath=/tmp/te
st-folder/2019-Oct.csv, targetPathExists=false, filtersFile='null'}
22/04/27 11:28:29 INFO client.RMProxy: Connecting to ResourceManager at ip-172-3
1-61-99.ec2.internal/172.31.61.99:8032
22/04/27 11:28:34 INFO tools.SimpleCopyListing: Paths (files+dirs) cnt = 1; dirC
nt = 0
22/04/27 11:28:34 INFO tools.SimpleCopyListing: Build file listing completed.
22/04/27 11:28:34 INFO Configuration.deprecation: io.sort.mb is deprecated. Inst
ead, use mapreduce.task.io.sort.mb
22/04/27 11:28:34 INFO Configuration.deprecation: io.sort.factor is deprecated.
Instead, use mapreduce.task.io.sort.factor
22/04/27 11:28:34 INFO tools.DistCp: Number of paths in the copy list: 1
22/04/27 11:28:34 INFO tools.DistCp: Number of paths in the copy list: 1
22/04/27 11:28:34 INFO client.RMProxy: Connecting to ResourceManager at ip-172-3
1-61-99.ec2.internal/172.31.61.99:8032
22/04/27 11:28:34 INFO mapreduce.JobSubmitter: number of splits:1
22/04/27 11:28:35 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_16
51057437121_0003
22/04/27 11:28:35 INFO impl.YarnClientImpl: Submitted application application_16
51057437121_0003
22/04/27 11:28:35 INFO mapreduce.Job: The url to track the job: http://ip-172-31
-61-99.ec2.internal:20888/proxy/application_1651057437121_0003/
22/04/27 11:28:35 INFO tools.DistCp: DistCp job-id: job_1651057437121_0003
22/04/27 11:28:35 INFO mapreduce.Job: Running job: job_1651057437121_0003
22/04/27 11:28:45 INFO mapreduce.Job: Job job_1651057437121_0003 running in uber
 mode : false
22/04/27 11:28:45 INFO mapreduce.Job:  map 0% reduce 0%
22/04/27 11:29:02 INFO mapreduce.Job:  map 100% reduce 0%
22/04/27 11:29:05 INFO mapreduce.Job: Job job_1651057437121_0003 completed succe
ssfully
22/04/27 11:29:06 INFO mapreduce.Job: Counters: 38
        File System Counters
                FILE: Number of bytes read=0
                FILE: Number of bytes written=172471
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=359
                HDFS: Number of bytes written=482542278
                HDFS: Number of read operations=12
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=4
```

hadoop@ip-172-31-61-99:~

```
                FILE: Number of bytes read=0
                FILE: Number of bytes written=172471
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=359
                HDFS: Number of bytes written=482542278
                HDFS: Number of read operations=12
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=4
                S3N: Number of bytes read=482542278
                S3N: Number of bytes written=0
                S3N: Number of read operations=0
                S3N: Number of large read operations=0
                S3N: Number of write operations=0
        Job Counters
                Launched map tasks=1
                Other local map tasks=1
                Total time spent by all maps in occupied slots (ms)=551072
                Total time spent by all reduces in occupied slots (ms)=0
                Total time spent by all map tasks (ms)=17221
                Total vcore-milliseconds taken by all map tasks=17221
                Total megabyte-milliseconds taken by all map tasks=17634304
        Map-Reduce Framework
                Map input records=1
                Map output records=0
                Input split bytes=136
                Spilled Records=0
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=410
                CPU time spent (ms)=18980
                Physical memory (bytes) snapshot=543019008
                Virtual memory (bytes) snapshot=3295203328
                Total committed heap usage (bytes)=445644800
        File Input Format Counters
                Bytes Read=223
        File Output Format Counters
                Bytes Written=0
        DistCp Counters
                Bytes Copied=482542278
                Bytes Expected=482542278
                Files Copied=1
```

4 Checking if the files are correctly imported to HDFS

hadoop fs -ls /hiveassignment

```
[hadoop@ip-172-31-49-96 ~]$ hadoop fs -ls /hiveassignment
Found 2 items
-rw-r--r--   1 hadoop hadoop  545839412 2022-01-02 06:21 /hiveassignment/2019-Nov.csv
-rw-r--r--   1 hadoop hadoop  482542278 2022-01-02 06:17 /hiveassignment/2019-Oct.csv
```

We can confirm the databases were loaded successfully.

```
[hadoop@ip-172-31-49-96 ~]$ hive

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

Creating database “upgrad_assignment”:

Create database if not exists upgrad_assignment;

use upgrad_assignment;

```
hive> Create database if not exists upgrad_assignment;
OK
Time taken: 0.065 seconds
```

```
hive> use upgrad_assignment;
OK
Time taken: 0.052 seconds
```

Creating an External Table, Sales:

create External table if not exists sales(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"=","."quoteChar"="\""."escapeChar"="\\") stored as textfile

Location '/hiveassignment' TBLPROPERTIES("skip.header.line.count"="1");

```
hive> create External table if not exists sales(event_time timestamp,event_type
    > string,product_id string,category_id string,category_code string,brand string,price f
    > user_id bigint,user_session string) ROW FORMAT SERDE
    > 'org.apache.hadoop.hive.serde2.OpenCSVSerde'
    > WITH SERDEPROPERTIES ("separatorChar"=","."quoteChar"="\""."escapeChar"="\\")
    > stored as textfile
    > Location '/hiveassignment' TBLPROPERTIES("skip.header.line.count"="1");
OK
Time taken: 0.386 seconds
```

desc sales;

```
hive> desc sales;
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: 0.418 seconds, Fetched: 9 row(s)
```

Loading the Data ointo the table:

hive> load data inpath '/hiveassignment/2019-Oct.csv' into table sales;

hive> load data inpath '/hiveassignment/2019-Nov.csv' into table sales;

```
hive>  load data inpath '/hiveassignment/2019-Oct.csv' into table sales;
Loading data to table upgrad_assignment.sales
OK
Time taken: 3.264 seconds
```

```
hive> load data inpath '/hiveassignment/2019-Nov.csv' into table sales;
Loading data to table upgrad_assignment.sales
OK
Time taken: 1.149 seconds
```

We are required to provide answers to the questions given below:

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

hive> set hive.cli.print.header=true;

hive> select sum(price) from sales where Month(event_time)=10 and event_type='purchase';

```
hive> set hive.cli.print.header=true;
hive> select sum(price) from sales where Month(event_time)=10 and event_type='purchase';
Query ID = hadoop_20220102064210_59e412ed-d3e3-4769-b292-a6c50c8e59cb
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_1641103761279_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: 40.41 s
--------------------------------------------------------------------------------
OK
_c0
1211530.4299997438
Time taken: 56.901 seconds, Fetched: 1 row(s)
hive>
```

Here the query takes 56.90 seconds which can be optimized by creating dynamic partition and then compare the execution time.

Dynamic Partitioning and Bucketing:

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;
```

Creating a table by name sales_dp to store the dataset which we partitioned by using 'event_type' and clustered by 'user_id'.

Desc sales_dp;

```
hive> create External table if not exists sales_dp(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;
OK
Time taken: 0.089 seconds
hive> desc sales_dp;
OK
col_name        data_type       comment
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
```

Loading the data into the new table:

insert into sales_dp partition(event_type) select event_time, product_id, category_id,category_code, brand, price, user_id, user_session, event_type from sales;

```
hive> insert into sales_dp partition(event_type) select event_time, product_id, category_id, category_code, brand, price, user_id, user_session, event_type from sales;
Query ID = hadoop_20220102065934_89df5fa6-8044-41b8-a08f-b495018e87e8
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_1641103761278_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: 113.54 s
--------------------------------------------------------------------------------
Loading data to table upgrad_assignment.sales_dp partition (event_type=null)

Loaded : 4/4 partitions.
         Time taken to load dynamic partitions: 0.617 seconds
         Time taken for adding to write entity : 0.007 seconds
OK
event_time      product_id      category_id     category_code   brand   price   user_id user_session    event_type
Time taken: 126.946 seconds
hive>
```

<u>Now executing the same Q1:</u>

select sum(price) from sales_dp where Month(event_time)=10 and event_type='purchase';

```
hive> select sum(price) from sales_dp where Month(event_time)=10 and
    > event_type='purchase';
Query ID = hadoop_20220102070525_e507749a-b0fa-4f5b-afc8-ce22feeffbd2
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1641103761278_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: 17.53 s
--------------------------------------------------------------------------------
OK
_c0
1211538.4299999105
Time taken: 18.169 seconds, Fetched: 1 row(s)
```

We can notice how the time taken reduced drastically due to partitioning and bucketing. Now it took only 18.17 sec.

The total sales in the month of October is 1211538.42.

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

select Month(event_time) as Month, sum(price) as sum, count(event_type) as cnt from sales where event_type='purchase' group by Month(event_time);

```
hive> select Month(event_time) as Month, sum(price) as sum, COUNT(event_type) as cnt
    > from sales where event_type='purchase' group by Month(event_time);
Query ID = hadoop_20220102070705_5d14246c-98bc-4352-bc60-454a103166c0
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1641103761278_0005)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      2          2        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      3          3        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 37.67 s
--------------------------------------------------------------------------------
OK
month   sum     cnt
10      1211538.4299997438      245624
11      1531016.900000122       322417
Time taken: 38.284 seconds, Fetched: 2 row(s)
hive>
```

In the month of October, the total purchase is 245624 and sales is 1211538.42

In the month of November, the total purchase is 322417 and sales is 1531016.90

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

with CTE1 as (select sum(case when Month(event_time)=10 then price else 0 end) as Oct, sum(case when Month(event_time)=11 then price else 0 end) as Nov from sales_dp where event_type='purchase' and Month(event_time) in (10,11)) select Oct,Nov,(Nov-Oct) as diff from CTE1;

```
hive> with CTE1 as (select sum(case when Month(event_time)=10 then price else 0 end ) as Oct,sum(case when Month(event_time)=11 then price else 0 end) as Nov from sales_dp where event_type=
'purchase' and Month(event_time) in (10,11) ) select Oct,Nov,(Nov-Oct) as diff from CTE1;
Query ID = hadoop_20220102070857_d1393f68-239b-4fbd-889c-3351b447ea16
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1641103761278_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: 19.94 s
--------------------------------------------------------------------------------
OK
oct     nov     diff
1211538.4299999105      1531016.8999999287      319478.47000001813
Time taken: 20.694 seconds, Fetched: 1 row(s)
hive>
```

We can see the difference in the revenue is 319478.47

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

select distinct split(category_code,'\\.')[0] as Categories from sales_dp where split(category_code,'\\.')[0]<>' ';

```
hive> select distinct split(category_code,'\\.')[0] as Categories from sales_dp where
    > split(category_code,'\\.')[0]<>'';
Query ID = hadoop_20220102071052_af73873d-0f35-414f-8bf1-a299f35b0985
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1641103761278_0005)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED     14         14        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      5          5        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 57.04 s
--------------------------------------------------------------------------------
OK
categories
furniture
appliances
accessories
apparel
sport
stationery
Time taken: 57.803 seconds, Fetched: 6 row(s)
```

We can see the distinct categories are Furniture, Appliances, Accessories, Apparel, Sport, Stationery

5 **<u>Find the total number of products available under each category.</u>**

select split(category_code,'\\.')[0] as category, count(product_id) as Prodcount from sales group by split(category_code,'\\.')[0] order by Prodcount desc;

```
hive> select split(category_code,'\\.')[0] as category , count(product_id) as Prodcount from  sales group by split(category_code,'\\.')[0] order by Prodcount desc;
Query ID = hadoop_20220102071551_600ba1f7-0eb8-45e9-9a7d-b5bb28f6c043
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1641103761278_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
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 52.56 s
--------------------------------------------------------------------------------
OK
category        prodcount
        8594895
appliances      61736
stationery      26722
furniture       23604
apparel 18232
accessories     12929
sport   2
Time taken: 53.145 seconds, Fetched: 7 row(s)
hive>
```

The total number of products under each category is as follows: Appliances 61736, Stationery 26722, Furniture 23604, Apparel 18232, Accessories 12929, Sport 2

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

select brand, sum(price) as totalsales from sales_dp where brand <>' ' and event_type ='purchase' group by brand order by totalsales desc limit 1;

```
hive> select brand, sum(price) as totalsales from sales_dp where brand <>'' and
    > event_type='purchase' group by brand order by totalsales desc limit 1;
Query ID = hadoop_20220102071804_975b0c72-6115-47de-86e3-12979ff0bd27
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1641103761278_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
Reducer 3 ......  container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 16.75 s
--------------------------------------------------------------------------------
OK
brand   totalsales
runail  148297.93999999954
Time taken: 17.426 seconds, Fetched: 1 row(s)
```

We can see that Runail is the brand with the maximum sales for oct and nov. Total sales is 148297.94

7 Which brands increased their sales from October to November?

with CTE2 as (select brand, sum(case when Month(event_time)=10 then price else 0 end) as Oct, sum(case when Month(event_time)=11 then price else 0 end) as Nov from sales_dp where event_type='purchase' and group by brand) select brand, Oct,Nov,(Nov-Oct) as diff from CTE2 where (Nov-Oct)>0 ORDER BY diff;

```
hive> with CTE2 as(select brand, sum(case when month(event_time)=10 then price else 0
    > end) as Oct,sum(case when month(event_time)=11 then price else 0 end) as Nov from
    > sales_dp where event_type='purchase' group by brand) select brand , Oct,Nov,(Nov-Oct) as
    > diff from CTE2 where (Nov-Oct)>0 ORDER BY diff;
Query ID = hadoop_20220102071955_2f02bd7f-96ec-468f-a5fb-29217a6cf54f
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1641103761278_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
Reducer 3 ......  container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 10.82 s
--------------------------------------------------------------------------------
OK
brand   oct     nov     diff
ovale   2.54    3.1     0.56
cosima  20.229999999999997      20.929999999999996      0.699999999999993
grace   100.91999999999997      102.61  1.690000000000261
helloganic      0.0     3.1     3.1
skinity 8.88    12.440000000000001      3.560000000000005
bodyton 1276.240000000006       1280.640000000003       4.299999999999727
moyou   5.71    10.280000000000001      4.570000000000001
neoleor 43.41   51.7    8.290000000000006
soleo   204.1999999999997       212.52999999999986      8.330000000000155
jaguar  1102.110000000001       1110.65 8.539999999999964
tertio  236.16000000000003      245.79999999999998      9.639999999999958
fly     17.14   27.17   10.030000000000001
rasyan  18.799999999999997      28.939999999999998      10.14
deoproce        316.84  329.16999999999996      12.329999999999984
barbie  0.0     12.39   12.39
supertan        50.36999999999999       66.51   16.140000000000015
treaclemoon     163.37  181.49000000000004      18.120000000000033
kamill  63.00999999999999       81.49000000000001       18.480000000000018
```

juno	0.0	21.08	21.08
veraclara	50.11	71.21000000000001	21.10000000000001
glysolid	69.72999999999999	91.59	21.860000000000014
godefroy	401.22	425.12	23.899999999999977
binacil	0.0	24.259999999999998	24.259999999999998
blixr	38.95	63.400000000000006	24.450000000000003
profepil	93.36	118.02000000000002	24.660000000000025
estelere	444.81000000000006	471.8700000000002	27.060000000000116
orly	902.38	931.09	28.710000000000036
biore	60.650000000000006	90.31	29.659999999999997
beautyblender	78.74000000000001	109.41	30.669999999999987
vilenta	197.59999999999997	231.20999999999993	33.60999999999996
mavala	409.0400000000001	446.32000000000005	37.27999999999997
likato	296.06	340.97	44.910000000000025
ladykin	125.65	170.57	44.91999999999999
foamie	35.04	80.49	45.449999999999996
elskin	251.08999999999997	307.65	56.56
balbcare	155.33000000000004	212.38000000000005	57.05000000000001
koelcia	55.5	112.75000000000001	57.250000000000014
profhenna	679.2299999999998	736.8499999999997	57.61999999999989
kares	0.0	59.449999999999996	59.449999999999996
marutaka-foot	49.22	109.33	60.11
dewal	0.0	61.28999999999999	61.28999999999999
inm	288.02	351.21	63.19
laboratorium	246.49999999999997	312.52	66.02000000000001
cutrin	299.37	361.62	68.25
egomania	77.47	146.04	68.57
konad	739.8299999999999	810.6700000000002	70.84000000000026
hirvel	163.04000000000002	234.33	71.28999999999999
koelf	422.7299999999999	507.28999999999996	84.56000000000006
plazan	101.37	194.01	92.63999999999999
aura	83.95	177.51	93.55999999999999
kerasys	430.9100000000001	525.1999999999999	94.28999999999905
enjoy	41.349999999999994	136.57000000000002	95.22000000000003
depilflax	2707.069999999974	2803.779999999998	96.71000000000049
eos	54.33999999999999	152.60999999999999	98.27
carmex	145.07999999999998	243.36	98.28000000000003
batiste	772.4000000000001	874.17	101.76999999999987
asmo	645.5800000000002	762.3100000000001	116.729999999999
dizao	819.13	945.5100000000003	126.38000000000034
igrobeauty	513.6600000000002	645.07	131.40999999999985
finish	98.38	230.38	132.0
nefertiti	233.52000000000004	366.64	133.11999999999995
elizavecca	70.53	204.3	133.77
miskin	158.04000000000002	293.06999999999994	135.02999999999992
latinoil	249.52	384.59000000000003	135.07000000000002
farmona	1692.4600000000005	1843.43	150.96999999999957
cristalinas	427.63	584.95	157.32000000000005
chi	358.93999999999994	538.6100000000001	179.6700000000002

chi	358.93999999999994	538.6100000000001	179.6700000000002
matreshka	0.0	182.67000000000002	182.67000000000002
freshbubble	318.7	502.34000000000003	183.64000000000004
mane	66.78999999999999	260.26	193.47
keen	236.34999999999997	435.62000000000006	199.2700000000001
ecocraft	41.160000000000004	241.95	200.79
fedua	52.38	263.81	211.43
provoc	827.9899999999994	1063.8199999999993	235.8299999999998
skinlite	651.9400000000003	890.4500000000004	238.5100000000001
entity	479.7100000000009	719.2599999999995	239.54999999999965
trind	298.07000000000005	542.96	244.89
protokeratin	201.25	456.79	255.54000000000002
beauugreen	511.51000000000005	768.3499999999999	256.83999999999986
bluesky	10307.239999999882	10565.529999999948	258.29000000006636
candy	534.9599999999999	799.3799999999999	264.41999999999998
insight	1443.7000000000003	1721.9600000000005	278.2600000000002
kocostar	310.84999999999997	594.93	284.08
happyfons	801.9200000000003	1091.5900000000006	289.6700000000003
kims	330.03999999999996	632.0400000000001	302.0000000000001
shary	871.9599999999996	1176.4899999999989	304.5299999999993
nitrile	847.2799999999999	1162.6799999999998	315.4
lowence	242.84000000000003	567.75	324.9099999999997
jas	3318.960000000002	3657.430000000017	338.4699999999998
ellips	245.85000000000002	606.0399999999998	360.1899999999998
ledor	2093.6100000000024	2471.5300000000025	387.9200000000001
naomi	0.0	389.0	389.0
kiss	421.55	817.3299999999997	395.7799999999997
yu-r	271.40999999999997	673.71	402.30000000000007
sophin	1067.8600000000004	1515.5200000000002	447.65999999999985
farmavita	837.3700000000001	1291.9699999999998	454.5999999999997
bioaqua	942.8900000000001	1398.1200000000003	455.23000000000025
greymy	29.21	489.49	460.28000000000003
gehwol	1089.0700000000002	1557.6800000000003	468.6100000000001
matrix	3243.250000000001	3726.7400000000007	483.4899999999998
limoni	1308.9	1796.6000000000006	487.7000000000005
s.care	412.68	913.07	500.39000000000004
coifin	903.0	1428.4899999999998	525.4899999999998
uskusi	5142.270000000011	5690.309999999995	548.0399999999881
airnails	5118.899999999999	5691.520000000004	572.6200000000053
browxenna	14331.369999999992	14916.730000000005	585.3600000000133
kinetics	6334.250000000018	6945.260000000015	611.0099999999966
kosmekka	1181.44	1813.37	631.9299999999998
kaaral	4412.430000000003	5086.070000000004	673.6400000000012
refectocil	2716.180000000002	3475.580000000005	759.4000000000028
rosi	3077.040000000001	3841.560000000002	764.5200000000013
solomeya	1899.7	2685.799999999984	786.0999999999983
missha	1293.83	2150.2799999999997	856.4499999999998
levissime	2227.500000000004	3085.310000000104	857.8100000000063
art-visage	2092.710000000002	2997.800000000007	905.0900000000051

ecolab	262.84999999999997	1214.3000000000002	951.4500000000003	
nagaraku	4369.740000000011	5327.680000000004	957.9399999999932	
yanoto	157.14	1209.6799999999998	1052.54	
markell	1768.7499999999982	2834.4299999999994	1065.6800000000012	
metzger	5373.450000000001	6457.159999999996	1083.7099999999955	
de.lux	1659.6999999999966	2775.509999999988	1115.8099999999913	
swarovski	1887.9299999999955	3043.159999999917	1155.2299999999962	
beauty-free	554.1700000000001	1782.8600000000022	1228.690000000002	
zeitun	708.6600000000002	2009.6299999999997	1300.9699999999993	
joico	705.52	2015.1000000000001	1309.5800000000002	
severina	4775.879999999865	6120.47999999998	1344.599999999993	
irisk	4559.959999999848	4694[illegible]	1354.0800000000454	
oniq	8425.41	9841.649999999996	1416.2399999999961	
levrana	2243.5599999999995	3664.100000000002	1420.5400000000027	
roubloff	3491.3600000000006	4913.770000000005	1422.4100000000044	
smart	4457.259999999995	5902.139999999995	1444.88	
shik	3341.2000000000016	4839.720000000003	1498.5200000000013	
domix	10472.05	12009.169999999982	1537.1199999999826	
artex	2730.6399999999994	4327.250000000001	1596.6100000000015	
beautix	10493.949999999986	12222.94999999999	1729.0000000000036	
milv	3904.9399999999707	5642.009999999994	1737.0700000000152	
masura	31266.079999999012	33058.46999999964	1792.3900000006252	
f.o.x	6624.23	8577.279999999992	1953.049999999992	
kopous	11927.159999999974	14093.000000000005	2165.9200000000348	
concept	11032.139999999989	13380.399999999932	2348.259999999944	
estel	21756.74999999993	24142.67000000002	2385.920000000089	
kaypro	881.34	3268.6999999999994	2387.359999999999	
benovy	409.61999999999995	3259.9700000000003	2850.3500000000004	
italwax	21940.239999999918	24799.369999999904	2859.1299999999865	
yoko	8756.91	11707.879999999986	2950.9699999999866	
haruyama	9390.689999999888	12352.910000000014	2962.2200000001267	
marathon	7280.750000000002	10273.1	2992.3499999999985	
lovely	8704.379999999994	11939.059999999974	3234.6799999999803	
bpw.style	11572.150000000067	14837.440000000197	3265.29000000013	
staleks	8519.730000000018	11875.610000000015	3355.8799999999974	
freedecor	3621.179999999866	7671.79999999995	4230.019999999964	
runail	71539.27999999898	76758.65999999987	5219.380000000892	
polarus	6013.72	11371.930000000004	5358.210000000004	
cosmoprofi	8322.909999999992	14536.990000000038	6214.180000000046	
jessnail	26287.840000000127	33345.23000000015	7057.390000000021	
strong	29196.629999999997	38671.27000000002	9474.640000000021	
ingarden	23161.38999999977	33566.21000000009	10404.820000000316	
lianail	5892.839999999984	16394.239999999972	10501.399999999989	
uno	35302.029999999984	51039.75000000009	15737.720000000103	
grattol	35445.53999999996	71472.71000000335	36027.170000003396	
	474679.0600000162	619509.2400000163	144830.1800000001	

Time taken: 19.658 seconds, Fetched: 161 row(s)

From the output we can see that 161 brand were able to increase their sales from the month of October to November.

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, sum(price) as Totalpurchases from sales_dp where event_type ='purchase' group by user_id order by Totalpurchases DESC limit 10;

```
hive> select user_id, sum(price) as Totalpurchases from sales_dp where event_type='purchase'
    > group by user_id order by Totalpurchases DESC limit 10;
Query ID = hadoop_20220102072427_ccd6ccd4-2a39-42d8-9acb-e42b30f9bee0
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1641103761278_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
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 17.59 s
--------------------------------------------------------------------------------
OK
user_id totalpurchases
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: 18.289 seconds, Fetched: 10 row(s)
hive>
```

We can see the top 10 users in the output.

Finishing Up

Once we are done, we can drop the databases, quit the hive and then terminate the EMR cluster.