

HIVE CASE STUDY

BY PRANAV JOSHI and DARSHAN SAWANT

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.

EMR CLI is launched

```
hadoop@ip-172-31-85-6~$ Authenticating with public key "Hive-Key-pair"
Last login: Wed Mar 2 08:22:34 2022
[...]
https://aws.amazon.com/amazon-linux-ami/2018.03-release-notes/
71 package(s) needed for security, out of 106 available
Run "sudo yum update" to apply all updates.

[...]
[hadoop@ip-172-31-85-6 ~]$ hadoop fs -ls /
[hadoop@ip-172-31-85-6 ~]$ hadoop fs -ls /
Found 4 items
drwxr-x-x - hdfs hadoop 0 2022-03-02 08:10 /apps
drwxrwxrwt - hdfs hadoop 0 2022-03-02 08:13 /tmp
drwxr-x-x - hdfs hadoop 0 2022-03-02 08:10 /user
drwxr-x-x - hdfs hadoop 0 2022-03-02 08:10 /var
[hadoop@ip-172-31-85-6 ~]$ hadoop fs -ls /user/
Found 6 items
drwxrwxrwx - hadoop hadoop 0 2022-03-02 08:10 /user/hadoop
drwxr-x-x - mapred mapred 0 2022-03-02 08:10 /user/history
drwxrwxrwx - hdfs hadoop 0 2022-03-02 08:10 /user/hive
drwxrwxrwx - hue hue 0 2022-03-02 08:10 /user/hue
drwxrwxrwx - oozie oozie 0 2022-03-02 08:11 /user/oozie
drwxrwxrwx - root hadoop 0 2022-03-02 08:10 /user/root
Found 4 items
drwxr-x-x - hdfs hadoop 0 2022-03-02 08:10 /user/hadoop
drwxrwxrwt - hdfs hadoop 0 2022-03-02 08:13 /tmp
drwxr-x-x - hdfs hadoop 0 2022-03-02 08:10 /user
drwxr-x-x - hdfs hadoop 0 2022-03-02 08:10 /var
[hadoop@ip-172-31-85-6 ~]$ hadoop fs -ls /user/hive/
Found 1 items
drwxrwxrwt - hdfs hadoop 0 2022-03-02 08:10 /user/hive/warehouse
[hadoop@ip-172-31-85-6 ~]$
```

CREATING A NEW DIRECTORY FOR HIVE CASE STUDY:

Commands:

```
hadoop fs -mkdir /user/hive/hivecasestudy
```

```
hadoop fs -ls /user/hive/
```

```

hadoop@ip-172-31-85-6:~ 
  [-createSnapshot <snapshotDir> [<snapshotName>]]
  [-deleteSnapshot <snapshotDir> [<snapshotName>]
  [-df [-h] [<path> ...]]
  [-du [-s] [-x] <path> ...]
  [-expunge]
  [-find <path> ... <expression> ...]
  [-get [-f] [-p] [-ignoreCrc] [-crc] <src> ... <localdst>]
  [-getfacl [-R] <path>]
  [-getfattr [-R] {-n name | -d} {-e en} <path>]
  [-getmerge [-n1] [-skip-empty-file] <src> <localdst>]
  [-help [cmd ...]]
  [-ls [-c] [-d] [-h] [-q] [-R] [-t] [-S] [-r] [-u] [<path> ...])
  [-mkdtemp [-p] <path> ...]
  [-moveFromLocal <localsrc> ... <dst>]
  [-moveToLocal <src> <localdst>]
  [-mv <src> ... <dst>]
  [-put [-f] [-p] [-l] [-d] <localsrc> ... <dst>]
  [-renameSnapshot <snapshotDir> <oldName> <newName>]
  [-rm [-f] [-r|-R] [-skipTrash] [-safely] <src> ...]
  [-rmdir [-ignore-fail-on-non-empty] <dir> ...]
  [-setfacl [-R] [(|-b|-k) (-m|-x <acl_spec>) <path>]]|([-set <acl_spec> <path>])
  [-setfattr {-n name [-v value] | -x name} <path>]
  [-setrep [-R] [-w] <rep> <path> ...]
  [-stat [format] <path> ...]
  [-tail [-f] <file>]
  [-test -[dofsz] <path>]
  [-text [-ignoreCrc] <src> ...]
  [-touchz <path> ...]
  [-truncate [-w] <length> <path> ...]
  [-usage [cmd ...]]

Generic options supported are
--conf <configuration file>      specify an application configuration file
-D <property>=<value>           use value for given property
--fs <file:///hdfs://namenode:port> specify default filesystem URL to use, overrides 'fs.defaultFS' property from configurations.
--jt <local|resourcemanager>:port> specify a ResourceManager
--files <comma separated list of files>  specify comma separated files to be copied to the map reduce cluster
--libjars <comma separated list of jars>   specify comma separated jar files to include in the classpath.
--archives <comma separated list of archives>  specify comma separated archives to be unarchived on the compute machines.

The general command line syntax is
command [genericOptions] [commandOptions]

[hadoop@ip-172-31-85-6 ~]$ hadoop fs -mkdir /user/hive/hivecasestudy
[hadoop@ip-172-31-85-6 ~]$ hadoop fs -ls /user/hive
Found 2 items
drwxr-xr-x  - hadoop hadoop          0 2022-03-02 08:41 /user/hive/hivecasestudy
drwxrwxwt  - hdfs  hadoop          0 2022-03-02 08:10 /user/hive/warehouse
[hadoop@ip-172-31-85-6 ~]$ 
```

New directory is successfully created

LOADING THE DATA FROM S3 BUCKET to HDFS:

Copy the data from S3 to HDFS

For 2019 October:

```
hadoop distcp s3://hivecasestudy-pranav/2019-Oct.csv /user/hive/hivecasestudy/2019-Oct.csv
```

For 2019 November:

```
hadoop distcp s3://hivecasestudy-pranav/2019-Nov.csv /user/hive/hivecasestudy/2019-Nov.csv
```

```

hadoop@ip-172-31-85-6:~ 
[.hadoop@ip-172-31-85-6 ~] hadoop distcp s3://hivecasestudy-pranav/2019-Oct.csv /user/hive/hivecasestudy/2019-Oct.csv
22/03/02 08:48:17 INFO tools.DistCp: Input Options: DistCpOptions(atomicCommit=false, syncFolder=false, deleteMissing=false, ignoreFailures=false, overwrite=false, skipCRC=false, blocking=true, numListStatusThreads=0, maxAmps=20, mapBandwidth=100, sslConfigurationFile='null', copyStrategy='uniformsize', preserveStatus={}, preserveRawAttrs=false, atomicWorkPath=null, logPath=null, sourceFileListing=null, sourcePaths=[s3://hivecasestudy-pranav/2019-Oct.csv], targetPath=/user/hive/hivecasestudy/2019-Oct.csv, targetPathExists=false, filtersFile='null')
22/03/02 08:48:17 INFO client.RMProxy: Connecting to ResourceManager at ip-172-31-85-6.ec2.internal/172.31.85.6:8032
22/03/02 08:48:23 INFO tools.SimpleCopyListing: Paths (files+dirs) cmr = 1; dircnt = 0
22/03/02 08:48:23 INFO Configuration.deprecation: io.sort.mb is deprecated. Instead, use mapreduce.task.io.sort.mb
22/03/02 08:48:23 INFO Configuration.deprecation: io.sort.factor is deprecated. Instead, use mapreduce.task.io.sort.factor
22/03/02 08:48:23 INFO tools.DistCp: Number of paths in the copy list: 1
22/03/02 08:48:23 INFO tools.DistCp: Number of paths in the copy list: 1
22/03/02 08:48:23 INFO client.RMProxy: Connecting to ResourceManager at ip-172-31-85-6.ec2.internal/172.31.85.6:8032
22/03/02 08:48:24 INFO mapred.JobSubmitter: number of splits:1
22/03/02 08:48:24 INFO mapred.JobSubmissionHandler: Submitted application application_1646208725012_0003
22/03/02 08:48:24 INFO impl.YarnClientImpl: Submitted application application_1646208725012_0003
22/03/02 08:48:24 INFO mapred.Job: The url to track the job: http://ip-172-31-85-6.ec2.internal:20888/proxy/application_1646208725012_0003/
22/03/02 08:48:24 INFO mapred.Job: Running job: job_1646208725012_0003
22/03/02 08:48:33 INFO mapred.Job: Job job_1646208725012_0003 running in uber mode : false
22/03/02 08:48:52 INFO mapred.Job: map 0% reduce 0%
22/03/02 08:48:57 INFO mapred.Job: Job job_1646208725012_0003 completed successfully
22/03/02 08:48:57 INFO mapred.Job: Counters: 38
File System Counters
FILE: Number of bytes read=0
FILE: Number of bytes written=172470
FILE: Number of read operations=0
FILE: Number of large read operations=0
FILE: Number of write operations=0
HDFS: Number of bytes read=360
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
S3: Number of bytes read=482542278
S3: Number of bytes written=0
S3: Number of read operations=0
S3: Number of large read operations=0
S3: Number of write operations=0
Job Counters
Launched map tasks=1
Other local map tasks=1
Total time spent by all maps in occupied slots (ms)=639392
Total time spent by all reduces in occupied slots (ms)=0
Total time spent by all map tasks (ms)=19981
Total vcore-milliseconds taken by all map tasks=19981
Total megabyte-milliseconds taken by all map tasks=20460544
Map-Reduce Framework

```

Verifying whether the data is successfully copied into HDFS from S3 buckets

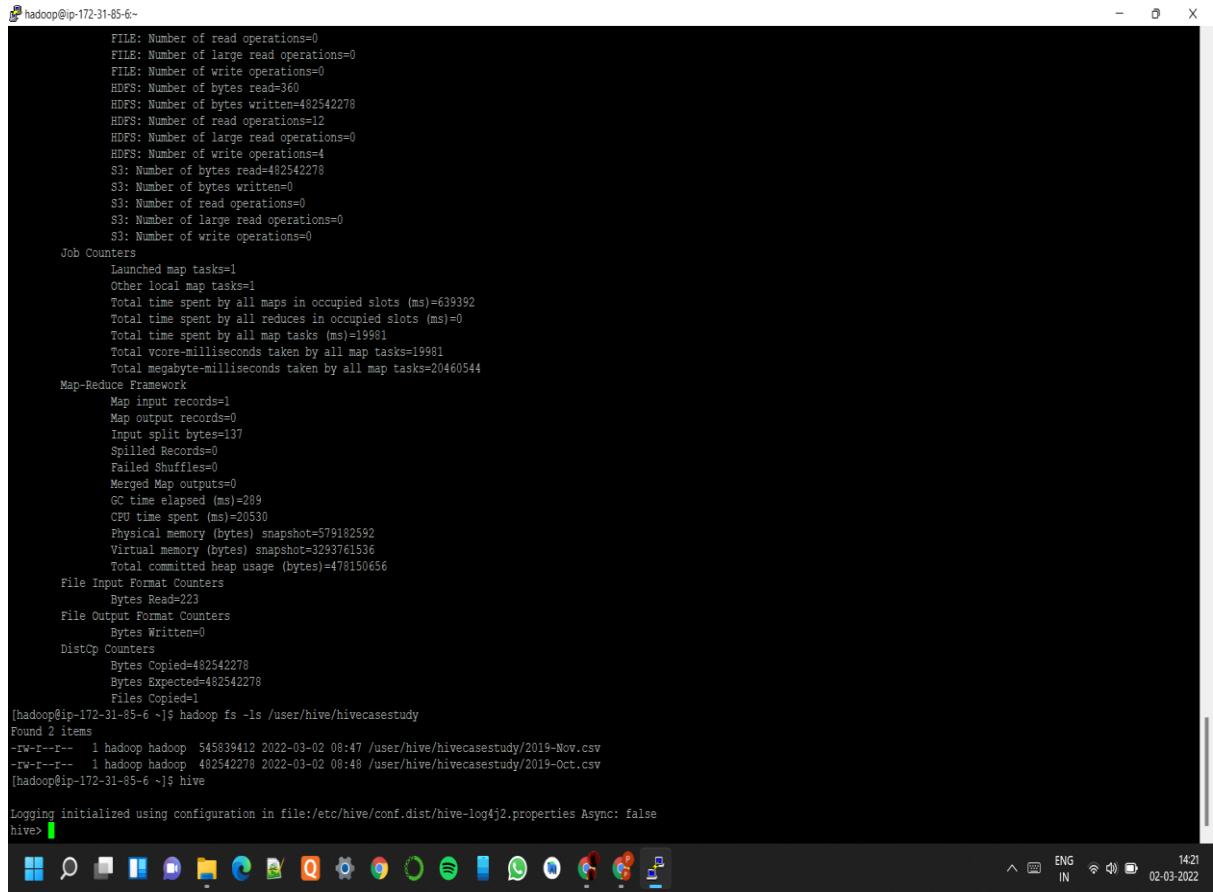
Command: hadoop fs -ls /user/hive/hivecasestudy

```

hadoop@ip-172-31-85-6:~ 
[.hadoop@ip-172-31-85-6 ~] hadoop fs -ls /user/hive/hivecasestudy
File System Counters
FILE: Number of bytes read=0
FILE: Number of bytes written=172470
FILE: Number of read operations=0
FILE: Number of large read operations=0
FILE: Number of write operations=0
HDFS: Number of bytes read=360
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
S3: Number of bytes read=482542278
S3: Number of bytes written=0
S3: Number of read operations=0
S3: Number of large read operations=0
S3: Number of write operations=0
Job Counters
Launched map tasks=1
Other local map tasks=1
Total time spent by all maps in occupied slots (ms)=639392
Total time spent by all reduces in occupied slots (ms)=0
Total time spent by all map tasks (ms)=19981
Total vcore-milliseconds taken by all map tasks=19981
Total megabyte-milliseconds taken by all map tasks=20460544
Map-Reduce Framework
Map input records=1
Map output records=0
Input split bytes=137
Spilled Records=0
Failed Shuffles=0
Merged Map outputs=0
GC time elapsed (ms)=289
CPU time spent (ms)=20530
Physical memory (bytes) snapshot=579182592
Virtual memory (bytes) snapshot=3293761536
Total committed heap usage (bytes)=478150656
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
[hadoop@ip-172-31-85-6 ~]$ hadoop fs -ls /user/hive/hivecasestudy
Found 2 items
-rw-r--r-- 1 hadoop hadoop 545839412 2022-03-02 08:47 /user/hive/hivecasestudy/2019-Nov.csv
-rw-r--r-- 1 hadoop hadoop 462542278 2022-03-02 08:48 /user/hive/hivecasestudy/2019-Oct.csv
[hadoop@ip-172-31-85-6 ~]$ 

```

Moving to hive:



```
[hadoop@ip-172-31-85-6 ~]$ hadoop fs -ls /user/hive/hivecasestudy
Found 2 items
-rw-r--r-- 1 hadoop hadoop 545039412 2022-03-02 08:47 /user/hive/hivecasestudy/2019-Nov.csv
-rw-r--r-- 1 hadoop hadoop 482542278 2022-03-02 08:48 /user/hive/hivecasestudy/2019-Oct.csv
[hadoop@ip-172-31-85-6 ~]$ hive
Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j2.properties Async: false
hive>
```

CREATING AN EXTERNAL TABLE IN HIVE:

```
CREATE EXTERNAL TABLE IF NOT EXISTS Retailstore (event_time timestamp, event_type string,
product_id string, category_id string, category_code string, brand string, price float, user_id bigint,
user_session string) ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.OpenCSVSerde' STORED
AS TEXTFILE LOCATION '/user/hive/hivecasestudy' tblproperties("skip.header.line.count"="1");
```

```
hadoop@ip-172-31-85-6:~$ hadoop fs -ls /user/hive/hivecasestudy
Found 2 items
-rw-r--r-- 1 hadoop hadoop 545839412 2022-03-02 08:47 /user/hive/hivecasestudy/2019-Nov.csv
-rw-r--r-- 1 hadoop hadoop 462542278 2022-03-02 08:48 /user/hive/hivecasestudy/2019-Oct.csv
[hadoop@ip-172-31-85-6 ~]$ hive
Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j2.properties Async: false
hive> CREATE EXTERNAL TABLE IF NOT EXISTS Retailstore(event_time timestamp,event_type string,product_id string,category_id string,category_code string,brand string,price float,user_id bigint,t,user_session string)ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.OpenCSVSerde'STORED AS TEXTFILE LOCATION '/user/hivecasestudy'tblproperties("skip.header.line.count"="1");
OK
Time taken: 1.547 seconds
hive>
```

APPLYING OPTIMIZATION TECHNIQUES - PARTITIONING AND BUCKETING:

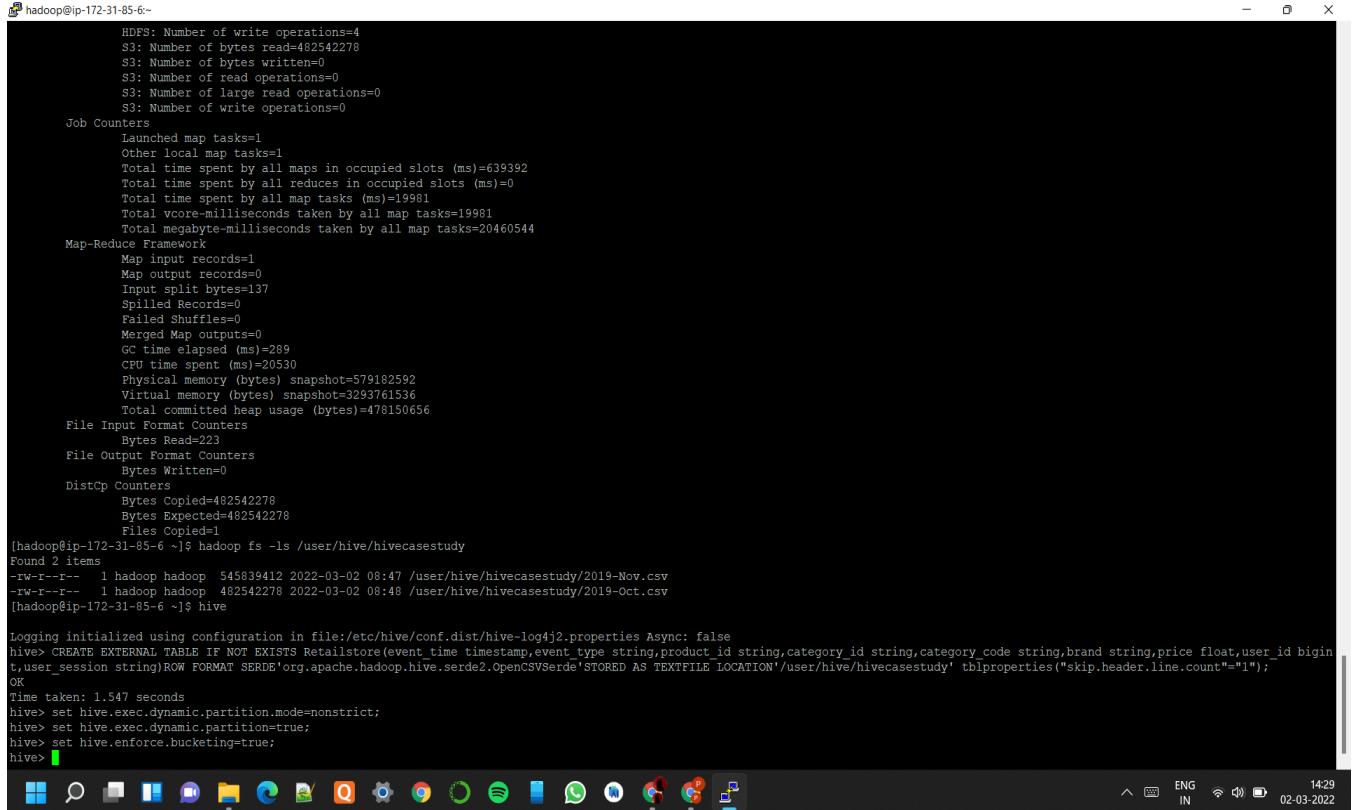
Below commands are to enable the dynamic partitioning and bucketing

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

```

hadoop@ip-172-31-85-6:~$ hadoop fs -ls /user/hive/hivecasestudy
Found 2 items
-rw-r--r-- 1 hadoop hadoop 545839412 2022-03-02 08:47 /user/hive/hivecasestudy/2019-Nov.csv
-rw-r--r-- 1 hadoop hadoop 482542278 2022-03-02 08:49 /user/hive/hivecasestudy/2019-Oct.csv
[hadoop@ip-172-31-85-6 ~]$ hive
Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j2.properties Async: false
hive> CREATE EXTERNAL TABLE IF NOT EXISTS Retailstore(event_time timestamp, event_type string, product_id string, category_id string, category_code string, brand string, price float, user_id bigint, user_session string) ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.OpenCSVSerde' STORED AS TEXTFILE LOCATION '/user/hive/hivecasestudy' tblproperties("skip.header.line.count"="1");
OK
Time taken: 1.547 seconds
hive> set hive.exec.dynamic.partition.mode=nonstrict;
hive> set hive.exec.dynamic.partition=true;
hive> set hive.enforce.bucketing=true;
hive> [REDACTED]

```



Creating an optimized table by applying partitioning on “event_type” and bucketing on “price”

```

CREATE TABLE IF NOT EXISTS Dynamic_Retailstore(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 (price) INTO 10 BUCKETS ROW FORMAT SERDE
'org.apache.hadoop.hive.serde2.OpenCSVSerde' STORED AS TEXTFILE LOCATION
'/user/hive/hivecasestudy' tblproperties('skip.header.line.count' = '1');

```

```

hadoop@ip-172-31-85-6:~$ S3: Number of write operations=0
Job Counters
    Launched map tasks=1
    Other local map tasks=1
    Total time spent by all maps in occupied slots (ms)=639392
    Total time spent by all reduces in occupied slots (ms)=0
    Total time spent by all map tasks (ms)=19981
    Total vcore-milliseconds taken by all map tasks=19981
    Total megabyte-milliseconds taken by all map tasks=20460544
Map-Reduce Framework
    Map input records=1
    Map output records=0
    Input split bytes=137
    Spilled Records=0
    Failed Shuffles=0
    Merged Map outputs=0
    GC time elapsed (ms)=289
    CPU time spent (ms)=20530
    Physical memory (bytes) snapshot=579182592
    Virtual memory (bytes) snapshot=3293761536
    Total committed heap usage (bytes)=478150656
File Input Format Counters
    Bytes Read=223
File Output Format Counters
    Bytes Written=0
DistCp Counters
    Bytes Copied=462542278
    Bytes Expected=462542278
    Files Copied=1
[hadoop@ip-172-31-85-6 ~]$ hadoop fs -ls /user/hive/hivecasestudy
Found 2 items
-rw-r--r-- 1 hadoop hadoop 545839412 2022-03-02 08:47 /user/hive/hivecasestudy/2019-Nov.csv
-rw-r--r-- 1 hadoop hadoop 482542278 2022-03-02 08:48 /user/hive/hivecasestudy/2019-Oct.csv
[hadoop@ip-172-31-85-6 ~]$ hive
Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j2.properties Async: false
hive> CREATE EXTERNAL TABLE IF NOT EXISTS Retailstore(event_time timestamp,event_type string,product_id string,category_id string,category_code string,brand string,price float,user_id bigint,user_session string)ROW FORMAT SERDE'org.apache.hadoop.hive.serde2.OpenCSVSerde'STORED AS TEXTFILE LOCATION'/user/hive/hivecasestudy'tblproperties("skip.header.line.count"="1");
OK
Time taken: 1.547 seconds
hive> set hive.exec.dynamic.partition.mode=nonstrict;
hive> set hive.exec.dynamic.partition=true;
hive> set hive.enforce.bucketing=true;
hive> CREATE TABLE IF NOT EXISTS Dynamic_Retailstore(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(price)INTO 10 BUCKETS ROW FORMAT SERDE'org.apache.hadoop.hive.serde2.OpenCSVSerde'STORED AS TEXTFILE LOCATION'/user/hive/hivecasestudy'tblproperties("skip.header.line.count"="1");
OK
Time taken: 0.104 seconds
hive> 

```

INSERTING THE DATA INTO NEWLY CREATED OPTIMIZED TABLE (Dynamic_Retailstore) FROM EXISTING TABLE(Retailstore):

INSERT INTO TABLE Dynamic_Retailstore PARTITION (event_type) SELECT event_time, product_id, category_id, category_code, brand, price, user_id, user_session, event_type FROM Retailstore;

```

hadoop@ip-172-31-85-6:~$ Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j2.properties Async: false
hive> CREATE EXTERNAL TABLE IF NOT EXISTS Retailstore(event_time timestamp,event_type string,product_id string,category_id string,category_code string,brand string,price float,user_id bigint,user_session string)ROW FORMAT SERDE'org.apache.hadoop.hive.serde2.OpenCSVSerde'STORED AS TEXTFILE LOCATION'/user/hive/hivecasestudy'tblproperties("skip.header.line.count"="1");
OK
Time taken: 1.547 seconds
hive> set hive.exec.dynamic.partition.mode=nonstrict;
hive> set hive.exec.dynamic.partition=true;
hive> set hive.enforce.bucketing=true;
hive> CREATE TABLE IF NOT EXISTS Dynamic_Retailstore(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(price)INTO 10 BUCKETS ROW FORMAT SERDE'org.apache.hadoop.hive.serde2.OpenCSVSerde'STORED AS TEXTFILE LOCATION'/user/hive/hivecasestudy'tblproperties("skip.header.line.count"="1");
OK
Time taken: 0.104 seconds
hive> show table
>
[1]+  Stopped                  hive
[hadoop@ip-172-31-85-6 ~]$ hive
Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j2.properties Async: false
hive> show tables;
OK
dynamic_retailstore
retailstore
Time taken: 0.663 seconds, Fetched: 2 row(s)
hive> INSERT INTO TABLE Dynamic_Retailstore PARTITION(event_type)SELECT event_time,product_id,category_id,category_code,brand,price,user_id,user_session,event_type FROM Retailstore;
FAILED: SemanticException [Error 10096]: Dynamic partition strict mode requires at least one static partition column. To turn this off set hive.exec.dynamic.partition.mode=nonstrict
hive> set hive.exec.dynamic.partition.mode=nonstrict;
hive> INSERT INTO TABLE Dynamic_Retailstore PARTITION(event_type)SELECT event_time,product_id,category_id,category_code,brand,price,user_id,user_session,event_type FROM Retailstore;
Query ID = hadoop_20220302090910_f74e2100-406d-4aa7-8272-e720f49e3807
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1646208725012_0005)

-----  

  VERTICES      MODE      STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED  

Map 1 ..... container  SUCCEEDED   2      2      0      0      0      0      0  

Reduce 2 ..... container  SUCCEEDED   5      0      0      0      0      0      0  

-----  

VERTICES: 0/2  [=====>>>] 100% ELAPSED TIME: 171.03 s  

Loading data to table default.dynamic_retailstore partition (event_type=null)  

Loaded : 4/4 partitions.  

  Time taken to load dynamic partitions: 1.106 seconds  

  Time taken for adding to write entity : 0.003 seconds
OK
Time taken: 177.869 seconds
hive> 

```

Output: Based on the above results, it partitioned into 4

ANSWERING GIVEN QUESTIONS:

find the total revenue generated due to purchases made in October

Base table:

```
SELECT SUM(price) AS total_revenue_oct FROM Retailstore WHERE  
MONTH(event_time) = '10' AND event_type = 'purchase';
```

Time taken is 138.77 seconds

Optimized table:

```
SELECT SUM(price) AS total_revenue_oct FROM Dynamic_Retailstore WHERE  
MONTH(event_time) = 10 AND event_type = 'purchase';
```

```

hadoop@ip-172-31-85-6:~ 
B::::E      M::::M   M:::M   M::::M   R:::R      R:::R
B::::E      EEEE M::::M   M:::M   M::::M   R:::R      R:::R
EE:::::EEEEEE E:::B M::::M   M:::M   M::::M   R:::R      R:::R
E:::::::E:::::B M::::M   M:::M   M::::M RR:::R      R:::R
EEEEE:::::EEEEE M:::::::M   M::::::M RRRRRR      RRRRRR

(hadoop@ip-172-31-85-6 ~)$ hive
Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j2.properties Async: false
hive> show tables;
OK
dynamic_retailstore
retailstore
Time taken: 0.75 seconds, Fetched: 2 row(s)
hive> SELECT SUM(price)AS total_revenue_oct FROM Retailstore WHERE MONTH(event_time)='10' AND event_type='purchase';
Query ID = hadoop_2022030209221_lc5fc4f1-c619-4b85-afea-e9849669d8c
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1646208725012_0006)

-----  

VERTICES    MODE      STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED  

-----  

Map 1 ..... container    SUCCEEDED    5      5      0      0      0      0  

Reducer 2 .... container    SUCCEEDED    1      1      0      0      0      0  

-----  

VERTICES: 02/02 [=====>>] 100% ELAPSED TIME: 134.26 s  

-----  

OK
1211538.4299997438
Time taken: 138.777 seconds, Fetched: 1 row(s)
hive> SELECT SUM(price)AS total_revenue_oct FROM Dynamic_Retailstore WHERE MONTH(event_time)='10' AND event_type='purchase';
Query ID = hadoop_20220302092536_b763cc2e-23d4-4816-a4e5-99fed9bd23ce
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1646208725012_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: 26.65 s  

-----  

OK
1211532.4500002791
Time taken: 28.051 seconds, Fetched: 1 row(s)
hive> 

```

Time taken with optimized table is 28.051 seconds

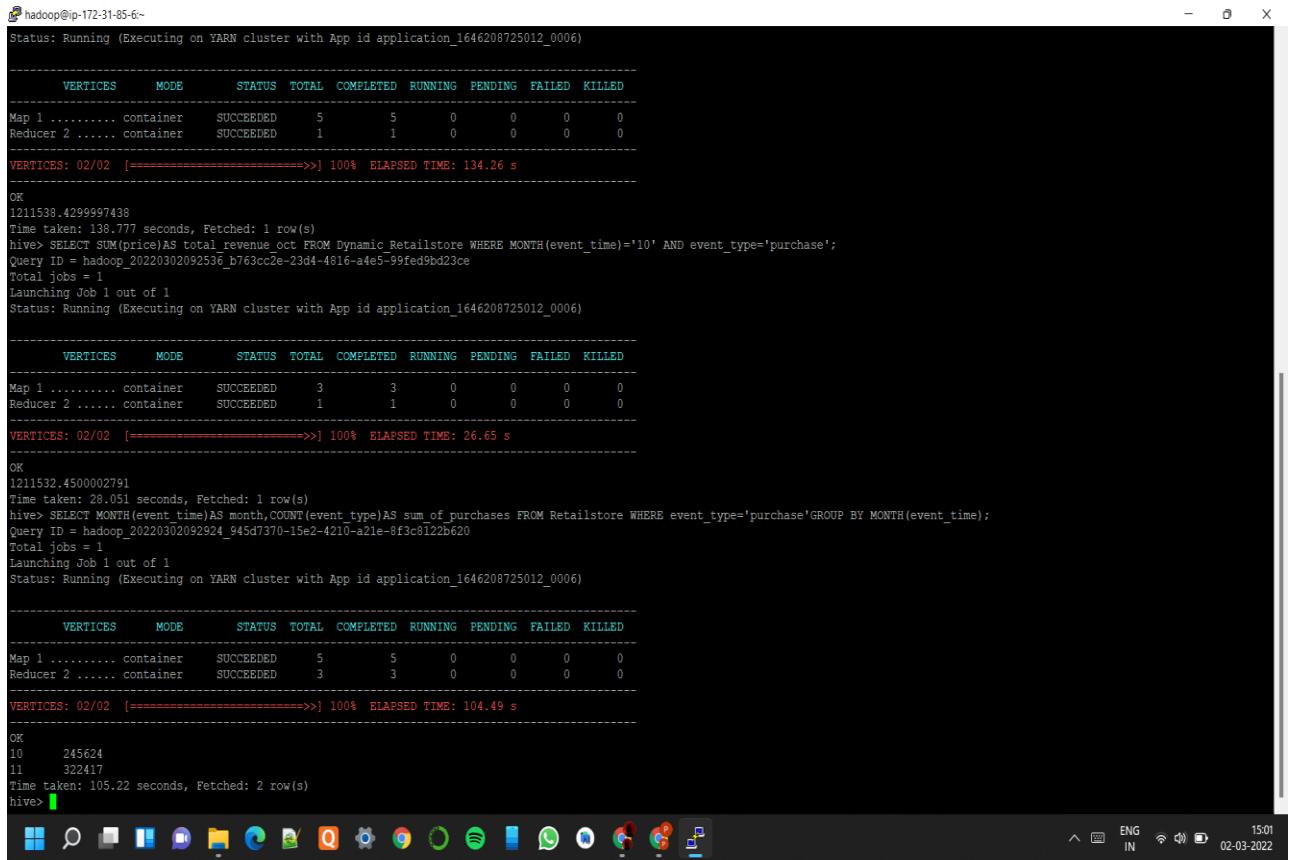
Insights:

1. The total revenue generated based on Purchase made in the month of October is 1,211,532.45 /-
2. Hence, optimized table gives better performance in execution time.

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

Base Query:

```
SELECT MONTH(event_time) AS month, COUNT(event_type) AS sum_of_purchases
FROM Retailstore WHERE event_type = 'purchase' GROUP BY MONTH(event_time);
```



The screenshot shows a terminal window with the following content:

```
hadoop@ip-172-31-85-6:~$ Status: Running (Executing on YARN cluster with App id application_1646208725012_0006)

-----  
 VERTICES  MODE      STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED  
Map 1 ..... container  SUCCEEDED   5      5      0      0      0      0  
Reducer 2 ..... container  SUCCEEDED   1      1      0      0      0      0  
-----  
VERTICES: 02/02 [=====>>] 100% ELAPSED TIME: 134.26 s  
  
OK  
1211538.4299997438  
Time taken: 138.777 seconds, Fetched: 1 row(s)  
hive> SELECT SUM(price)AS total_revenue Oct FROM Dynamic_Retailstore WHERE MONTH(event_time)='10' AND event_type='purchase';  
Query ID = hadoop_20220302092536_b763cc2e-23d4-4816-a4e5-99fed9bd23ce  
Total jobs = 1  
Launching Job 1 out of 1  
Status: Running (Executing on YARN cluster with App id application_1646208725012_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: 26.65 s  
  
OK  
1211532.4500002791  
Time taken: 28.051 seconds, Fetched: 1 row(s)  
hive> SELECT MONTH(event_time)AS month,COUNT(event_type)AS sum_of_purchases FROM Retailstore WHERE event_type='purchase'GROUP BY MONTH(event_time);  
Query ID = hadoop_20220302092924_945d7370-15e2-4210-a21e-8f3c8122b620  
Total jobs = 1  
Launching Job 1 out of 1  
Status: Running (Executing on YARN cluster with App id application_1646208725012_0006)  
  
-----  
 VERTICES  MODE      STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED  
Map 1 ..... container  SUCCEEDED   5      5      0      0      0      0  
Reducer 2 ..... container  SUCCEEDED   3      3      0      0      0      0  
-----  
VERTICES: 02/02 [=====>>] 100% ELAPSED TIME: 104.49 s  
  
OK  
10      245624  
11      322417  
Time taken: 105.22 seconds, Fetched: 2 row(s)  
hive>
```

Time taken is 105.22 seconds

Optimized table:

```
SELECT MONTH(event_time) AS month, COUNT(event_type) AS sum_of_purchases
FROM Dynamic_Retailstore WHERE event_type = 'purchase' GROUP BY
MONTH(event_time);
```

```

hadoop@ip-172-31-85-6:~
```

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: 26.65 s

OK
1211532.4500002791
Time taken: 28.051 seconds, Fetched: 1 row(s)
hive> SELECT MONTH(event_time)AS month,COUNT(event_type)AS sum_of_purchases FROM Retailstore WHERE event_type='purchase'GROUP BY MONTH(event_time);
Query ID = hadoop_20220302092924_945d7370-15e2-4210-a21e-8f1c8122b620
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1646208725012_0006)

VERTICES	MODE	STATUS	TOTAL	COMPLETED	RUNNING	PENDING	FAILED	KILLED
Map 1	container	SUCCEEDED	5	5	0	0	0	0
Reducer 2	container	SUCCEEDED	3	3	0	0	0	0

VERTICES: 02/02 [=====>>>] 100% ELAPSED TIME: 104.49 s

OK
10 245624
11 322417
Time taken: 105.22 seconds, Fetched: 2 row(s)
hive> SELECT MONTH(event_time)AS month,COUNT(event_type)AS sum_of_purchases FROM Dynamic_Retailstore WHERE event_type='purchase'GROUP BY MONTH(event_time);
Query ID = hadoop_20220302093146_4ca32d7d-c93e-4193-8b06-644db5b76c30
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1646208725012_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: 25.16 s

OK
10 245619
11 322412
Time taken: 26.037 seconds, Fetched: 2 row(s)
hive>



ENG IN 02-03-2022 15:02

Time taken is 26.037 seconds

Insights:

- Sum of purchases made in the month of October is 245619 and in the month of November 322412, which means number of purchases are increased in November month
- With proper partitioning and bucketing on table we can reduce execution time.

Using Optimized table from below questions onwards:

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

```
SELECT (SUM(CASE WHEN MONTH(event_time)=11 THEN price ELSE 0 END) - SUM(CASE WHEN MONTH(event_time)=10 THEN price ELSE 0 END)) AS change_in_rev FROM Dynamic_Retailstore WHERE event_type = 'purchase' AND MONTH(event_time) in ('10','11');
```

```

hadoop@ip-172-31-85-6:~
```

	Map 1	container	SUCCEEDED	5	5	0	0	0	0
Reducer 2	container	SUCCEEDED	3	3	0	0	0	0	0

VERTICES: 02/02 [=====>>] 100% ELAPSED TIME: 104.49 s

OK
10 245624
11 322417
Time taken: 105.22 seconds, Fetched: 2 row(s)
hive> SELECT MONTH(event_time)AS month,COUNT(event_type)AS sum_of_purchases FROM Dynamic_Retailstore WHERE event_type='purchase'GROUP BY MONTH(event_time);
Query ID = hadoop_20220302093146_4ca32d7d-c93e-4193-8b06-f44db5b76c30
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1646208725012_0006)

	VERTICES	MODE	STATUS	TOTAL	COMPLETED	RUNNING	PENDING	FAILED	KILLED
Map 1	container	SUCCEEDED	3	3	0	0	0	0	0
Reducer 2	container	SUCCEEDED	1	1	0	0	0	0	0

VERTICES: 02/02 [=====>>] 100% ELAPSED TIME: 25.16 s

OK
10 245619
11 322412
Time taken: 26.037 seconds, Fetched: 2 row(s)
hive> SELECT (SUM(CASE WHEN MONTH(event_time)=11 THEN price ELSE 0 END)-SUM(CASE WHEN MONTH(event_time)=10 THEN price ELSE 0 END))AS change_in_rev FROM Dynamic_Retailstore WHERE event_type='purchase' AND MONTH(event_time)in('10','11');
Query ID = hadoop_20220302093918_cae74cb6-ecd8-41c3-bd36-2fa0296fa822
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_1646208725012_0007)

	VERTICES	MODE	STATUS	TOTAL	COMPLETED	RUNNING	PENDING	FAILED	KILLED
Map 1	container	SUCCEEDED	3	3	0	0	0	0	0
Reducer 2	container	SUCCEEDED	1	1	0	0	0	0	0

VERTICES: 02/02 [=====>>] 100% ELAPSED TIME: 29.31 s

OK
319437.7899997565
Time taken: 39.513 seconds, Fetched: 1 row(s)
hive>



15:10
ENG IN 02-03-2022

Insights:

1. Time taken to execute the query is 39.513 seconds
2. Revenue increased in November by 319437.789 from October

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

```
SELECT DISTINCT SPLIT(category_code,'\\.')[0] AS Category FROM dynpart_buck_retailstore
WHERE category_code != '';
```

```

hadoop@ip-172-31-85-6:~ 
at org.apache.hadoop.hive ql.parse.HiveParser.regularBody(HiveParser.java:36633)
at org.apache.hadoop.hive ql.parse.HiveParser.queryStatementExpressionBody(HiveParser.java:35822)
at org.apache.hadoop.hive ql.parse.HiveParser.queryStatementExpression(HiveParser.java:35710)
at org.apache.hadoop.hive ql.parse.HiveParser.execCStatement(HiveParser.java:2284)
at org.apache.hadoop.hive ql.parse.HiveParser.statement(HiveParser.java:1333)
at org.apache.hadoop.hive ql.parse.ParseDriver.parse(ParseDriver.java:208)
at org.apache.hadoop.hive ql.parse.ParseUtils.parse(ParseUtils.java:77)
at org.apache.hadoop.hive ql.parse.ParseUtils.parse(ParseUtils.java:70)
at org.apache.hadoop.hive ql.Driver.compile(Driver.java:468)
at org.apache.hadoop.hive ql.Driver.compileInternal(Driver.java:1317)
at org.apache.hadoop.hive ql.Driver.runInternal(Driver.java:1457)
at org.apache.hadoop.hive ql.Driver.run(Driver.java:1237)
at org.apache.hadoop.hive ql.Driver.run(Driver.java:1227)
at org.apache.hadoop.hive cli.CliDriver.processLocalCmd(CliDriver.java:233)
at org.apache.hadoop.hive cli.CliDriver.processCmd(CliDriver.java:184)
at org.apache.hadoop.hive cli.CliDriver.processLine(CliDriver.java:403)
at org.apache.hadoop.hive cli.CliDriver.executeDriver(CliDriver.java:821)
at org.apache.hadoop.hive cli.CliDriver.run(CliDriver.java:759)
at org.apache.hadoop.hive cli.CliDriver.main(CliDriver.java:86)
at sun.reflect.NativeMethodAccessorImpl.invoke0(Native Method)
at sun.reflect.NativeMethodAccessorImpl.invoke(NativeMethodAccessorImpl.java:62)
at sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccessorImpl.java:43)
at java.lang.reflect.Method.invoke(Method.java:498)
at org.apache.hadoop.util.RunJar.run(RunJar.java:239)
at org.apache.hadoop.util.RunJar.main(RunJar.java:153)
FAILED: ParseException line 1:104 cannot recognize input near ',' '<EOF>' '<EOF>' in expression specification
hive> SELECT DISTINCT SPLIT(category_code,'\\.')[0] AS Category FROM Dynamic_Retailstore WHERE category_code!="";
Query ID = hadoop_20220302094306_61838fe6-dc59-06a5-bbf47la09c0e
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1646208725012_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: 67.38 s  

-----  

OK  

furniture  

appliances  

accessories  

apparel  

sport  

stationery  

Time taken: 68.184 seconds, Fetched: 6 row(s)
hive>

```

Insights:

1. Time taken to execute the query is 68.184 seconds
2. Total we got 6 distinct categories are – furniture, appliances, accessories, apparel, sport, stationary.

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

```

SELECT SPLIT(category_code,'\\.')[0] AS Category, COUNT(product_id) AS num_of_prod
FROM Dynamic_Retailstore WHERE category_code != " " GROUP BY
SPLIT(category_code,'\\.')[0] ORDER BY num_of_prod DESC;

```

```

hadoop@ip-172-31-85-6:~ 
at org.apache.hadoop.hive.ql.parse.HiveParser.regularBody(HiveParser.java:36633)
at org.apache.hadoop.hive.ql.parse.HiveParser.queryStatementExpressionBody(HiveParser.java:35822)
at org.apache.hadoop.hive.ql.parse.HiveParser.queryStatementExpression(HiveParser.java:35710)
at org.apache.hadoop.hive.ql.parse.HiveParser.executeQueryStatement(HiveParser.java:2284)
at org.apache.hadoop.hive.ql.parse.HiveParser.executeStatement(HiveParser.java:1333)
at org.apache.hadoop.hive.ql.parse.ParseDriver.parse(ParseDriver.java:168)
at org.apache.hadoop.hive.ql.parse.ParseDriver.parse(ParseDriver.java:77)
at org.apache.hadoop.hive.ql.parse.ParseUtils.parse(ParseUtils.java:70)
at org.apache.hadoop.hive.ql.Driver.compile(Driver.java:468)
at org.apache.hadoop.hive.ql.Driver.compileInternal(Driver.java:1317)
at org.apache.hadoop.hive.ql.Driver.runInternal(Driver.java:1457)
at org.apache.hadoop.hive.ql.Driver.run(Driver.java:1227)
at org.apache.hadoop.hive.cli.CliDriver.processLocalCmd(CliDriver.java:233)
at org.apache.hadoop.hive.cli.CliDriver.processCmd(CliDriver.java:194)
at org.apache.hadoop.hive.cli.CliDriver.processLine(CliDriver.java:403)
at org.apache.hadoop.hive.cli.CliDriver.executeDriver(CliDriver.java:821)
at org.apache.hadoop.hive.cli.CliDriver.main(CliDriver.java:759)
at org.apache.hadoop.hive.cli.CliDriver.main(CliDriver.java:868)
at sun.reflect.NativeMethodAccessorImpl.invoke(NativeMethodAccessorImpl.java:62)
at sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccessorImpl.java:43)
at java.lang.reflect.Method.invoke(Method.java:498)
at org.apache.hadoop.util.RunJar.run(RunJar.java:239)
at org.apache.hadoop.util.RunJar.main(RunJar.java:153)

FAILED: ParseException line 1:104 cannot recognize input near ';' '<EOF>' '<EOF>' in expression specification
hive> SELECT DISTINCT $PLIT(category_code,'\\.')[0]AS category FROM Dynamic_Retailstore WHERE category_code!="";
Query ID : 20220302094306_61d30fe6-dc59-4935-0e83-bbf471a89c0e
Total jobs : 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1646208725012_0007)

-----  

VERTICES  MODE      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED  

Map 1 ..... container  SUCCESSED   6      6       0       0       0       0  

Reducer 2 ..... container  SUCCESSED   5      5       0       0       0       0  

-----  

VERTICES: 02/02 [=====>>>] 100% ELAPSED TIME: 67.38 s  

-----  

OK  

furniture  

appliances  

accessories  

apparel  

sport  

stationery  

Time taken: 68.184 seconds, Fetched: 6 row(s)  

hives: 1

```

Insights:

1. Time taken to execute the query is 68.111 seconds
2. Appliances are having highest number of products available with 61736 compared to other categories.
3. Stationary and Furniture categories are almost equally registered with available ranges from 23000 to 27000.
4. Sports category is least available with 2 products

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

```

WITH tot_sales AS( SELECT brand, (SUM(CASE WHEN MONTH(event_time)=10 THEN price ELSE 0 END) + SUM(CASE WHEN MONTH(event_time)=11 THEN PRICE ELSE 0 END)) AS total_sales FROM Dynamic_Retailstore WHERE event_type = 'purchase' AND MONTH(event_time) in ('10','11') AND brand != " GROUP BY brand) SELECT brand, total_sales FROM tot_sales ORDER BY total_sales DESC LIMIT 1;

```

```

hadoop@ip-172-31-85-6:~ 
at org.apache.hadoop.hive ql.parse.HiveParser.queryStatementExpression(HiveParser.java:35710)
at org.apache.hadoop.hive ql.parse.HiveParser.exeStatement(HiveParser.java:2284)
at org.apache.hadoop.hive ql.parse.HiveParser.statement(HiveParser.java:1333)
at org.apache.hadoop.hive ql.parse.ParseDriver.parse(ParseDriver.java:208)
at org.apache.hadoop.hive ql.parse.ParseUtils.parse(ParseUtils.java:77)
at org.apache.hadoop.hive ql.parse.ParseUtils.parse(ParseUtils.java:70)
at org.apache.hadoop.hive ql.parse.ParseUtils.parse(ParseUtils.java:70)
at org.apache.hadoop.hive ql.Driver.compile(Driver.java:468)
at org.apache.hadoop.hive ql.Driver.compileInternal(Driver.java:1317)
at org.apache.hadoop.hive ql.Driver.runInternal(Driver.java:1457)
at org.apache.hadoop.hive ql.Driver.run(Driver.java:1237)
at org.apache.hadoop.hive ql.Driver.run(Driver.java:1227)
at org.apache.hadoop.hive cli.CliDriver.processLocalCmd(CliDriver.java:233)
at org.apache.hadoop.hive cli.CliDriver.processed(CliDriver.java:184)
at org.apache.hadoop.hive cli.CliDriver.processLine(CliDriver.java:403)
at org.apache.hadoop.hive cli.CliDriver.executeDriver(CliDriver.java:821)
at org.apache.hadoop.hive cli.CliDriver.run(CliDriver.java:759)
at org.apache.hadoop.hive cli.CliDriver.main(CliDriver.java:686)
at sun.reflect.NativeMethodAccessorImpl.invoke0(Native Method)
at sun.reflect.NativeMethodAccessorImpl.invoke(NativeMethodAccessorImpl.java:62)
at sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccessorImpl.java:43)
at java.lang.reflect.Method.invoke(Method.java:498)
at org.apache.hadoop.util.RunJar.run(RunJar.java:239)
at org.apache.hadoop.util.RunJar.main(RunJar.java:153)
FAILED: ParseException line 2:45 cannot recognize input near 'OREDR' 'BY' 'total_sales' in table source
hive> SELECT brand, total_sales FROM tot_sale ORDER BY total_sales DESC LIMIT 1;
FAILED: SemanticException [Error 10001]: Line 1:30 Table not found 'tot_sale'
hive> WITH tot_sale AS (SELECT brand, (SUM(CASE WHEN MONTH(event_time)=10 THEN price ELSE 0 END)+SUM(CASE WHEN MONTH(event_time)=11 THEN PRICE ELSE 0 END))AS total_sales FROM Dynamic_Retailist
ore WHERE event_type='purchase' AND MONTH(event_time)in('10','11')AND brand='%' GROUP BY brand)
> SELECT brand, total_sales FROM tot_sale ORDER BY total_sales DESC LIMIT 1;
Query ID = hadoop_20220302095820_25e55b62-c91e-46ac-9abd-618343552f57
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_1646208725012_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: 30.97 s  

-----  

OK  

runmail 148292.46000001638  

Time taken: 41.757 seconds, Fetched: 1 row(s)  

hive> [  ]
```

Insights:

1. Runail is the brand that has the highest sales in total of both the months October and November

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

```

WITH brand_sales AS( SELECT brand, SUM(CASE WHEN MONTH(event_time)=10 THEN price ELSE 0 END) AS oct_sales, SUM(CASE WHEN MONTH(event_time)=11 THEN price ELSE 0 END) AS nov_sales
FROM Dynamic_Retailstore WHERE event_type = 'purchase' AND MONTH(event_time) in ('10','11')
AND brand != " GROUP BY brand) SELECT brand, oct_sales, nov_sales, nov_sales-oct_sales AS
sale_diff FROM brand_sales WHERE nov_sales-oct_sales > 0 ORDER BY sale_diff DESC;

```

```
[hadoop@ip-172-31-85-6 ~]$ hive

Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j2.properties Async: false
hive> WITH brand_sales AS(SELECT brand,SUM(CASE WHEN MONTH(event_time)=10 THEN price ELSE 0 END)AS oct_sales,SUM(CASE WHEN MONTH(event_time)=11 THEN price ELSE 0 END)AS nov_sales FROM Dynamic Retailstore WHERE event_type = 'purchase' AND MONTH(event_time) in ('10','11') AND brand!="" GROUP BY brand)
   > SELECT brand oct_sales,nov_sales,nov_sales-oct_sales AS sale_diff FROM brand_sales WHERE nov_sales-oct_sales>0 ORDER BY sale_diff DESC;
FAILED: ParseException line 2:111 missing EOF at 'OREDR' near '0'
hive> WITH brand_sales AS(SELECT brand,SUM(CASE WHEN MONTH(event_time)=10 THEN price ELSE 0 END)AS oct_sales,SUM(CASE WHEN MONTH(event_time)=11 THEN price ELSE 0 END)AS nov_sales FROM Dynamic Retailstore WHERE event_type = 'purchase' AND MONTH(event_time) in ('10','11') AND brand!="" GROUP BY brand)
   > SELECT brand oct_sales,nov_sales,nov_sales-oct_sales AS sale_diff FROM brand_sales WHERE nov_sales-oct_sales>0 ORDER BY sale_diff DESC;
Query ID : hadoop_202203020101008_2d207adc-7acb-4828-8de6-e6c80701a263
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_1646208725012_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: 31.93 s  
-----  
OK  
grattol 71472.7100000044 36027.1700000042  
uno 51039.7500000047 15737.720000000285  
lianail 16394.24000000544 10501.40000000052  
ingarden 33566.21000000625 10404.82000000425  
strong 38671.26999999975 9474.6399999997  
jessnail 3345.22999999992 7057.389999999839  
cosmoprofi 14536.98999999909 6214.18000000008  
polarus 11371.93000000004 5358.210000000055  
runail 76754.69000000473 5216.919999993101  
freedecor 7671.80000000062 4250.020000000052  
staleks 11875.61000000015 3355.880000000001  
hpw.style 14837.440000000377 3265.289999999917  
lovely 11939.05999999967 3234.679999999982  
marathon 10273.09999999999 2992.350000000002  
haruyama 12352.910000000145 2962.2200000000576  
yoko 11707.879999999976 2950.369999999992  
italwax 24799.369999999864 2859.130000000183  
benovy 3259.970000000002 2850.3500000000002  
kaypro 3268.6999999999985 2387.3599999999988  
estel 24142.67000000027 2385.9200000000002  
concept 13380.399999999978 2348.26  
kapous 14093.079999999864 2165.9200000001256
```

hadoop@ip-172-31-85-6:~

italwan	24799.369999999986	2859.1300000000183
benovoy	3259.970000000002	2850.350000000002
kaypro	3268.699999999985	2387.359999999998
estel	24142.670000000027	2385.520000000002
concept	13380.399999999978	2348.26
kapous	14093.079999999864	2165.52000000001256
f.o.x	8577.279999999979	1953.049999999972
masura	33058.4699999992	1792.3900000006033
milv	5642.0100000001285	1737.0700000001025
beautix	12222.95	1729.00000000031
artere	4327.24999999996	1596.6099999999942
dcmix	12009.16999999851	1537.1199999999317
shlik	4839.72000000001	1498.519999999977
smart	5902.13999999998	1444.88000000001
roublöff	4913.770000000003	1422.4100000000026
levrania	3664.10000000004	1420.5400000000018
onig	9841.6499999988	1416.2400000000325
irisk	46946.0399999916	1354.079999994633
severina	6120.47999999956	1344.6000000000013
jicojo	2015.10000000006	1309.5800000000006
zeitun	2009.631300.96999999998	
beauty-free	1782.859999999983	1228.689999999986
swarovski	3043.159999999794	1155.229999999896
de.lux	2775.50999999973	1115.809999999945
metzger	6457.16000000005	1083.709999999955
markell	2834.42999999994	1065.679999999942
sanoto	1209.679999999998	1052.54
nagaraku	5327.6800000000285	957.939999999986
ecolab	1214.300000000009	951.4500000000008
art-visage	2997.800000000056	905.0900000000079
levissime	3085.309999999854	857.809999999979
missha	2150.279999999997	856.449999999998
solomeya	2685.759999999996	786.099999999949
rosi	3841.56000000001	764.519999999991
refectocil	3475.580000000036	759.400000000005
kaaral	5086.07673.64000000003	
kosmekka	1813.37631.930000000003	
kinetics	6945.25999999998	611.009999999992
browxenna	14916.7300000007	585.3600000000133
aerials	5691.52000000021	572.620000000063
uskusi	5690.31000000031	548.0400000000509
coifin	1428.489999999998	525.489999999996
s.care	913.07500.390000000004	
limoni	1796.60000000004	487.699999999998
matrix	3726.740000000016	483.490000000007
gehwol	1557.66468.609999999999	
greemy	489.49460.280000000003	
bioqua	1398.120000000001	455.23
farmavita	1291.97454.6	

hadoop@ip-172-31-85-6:~

```
s.cart 913.07 500.3900000000000004
limon 1796.6000000000000004 487.6999999999998
matrix 3726.74000000000016 493.4900000000007
gehwol 1357.68 280.9999999999999
magni 169.49 460.2800000000003
blisqua 1398.120000000001 455.23
farmavita 1291.97 454.6
sophin 1515.52000000000045
yu-r 673.7099999999999 402.2999999999999
kiss 817.330000000004 395.78000000000037
naomi 389.0 389.0
lador 2471.530000000016 387.9199999999996
ellips 606.04 360.19
jas 3657.430000000026 338.4700000000007
lowence 567.7499999999999 324.9099999999999
marielle 162.6799999999999 330.1000000000001
shay 3176.489999999996 304.5300000000002
kins 632.0400000000001 302.0000000000006
happyfons 1091.590000000008 289.6700000000004
kostocart 594.9299999999998 284.0799999999999
insight 1721.960000000001 278.26000000000045
candy 799.379999999994 264.4199999999995
blueksy 10565.529999999784 258.28999999628
beaugreen 768.349999999999 256.8399999999975
protokeratin 456.79 255.5400000000002
trind 542.96 244.0
entry 19.259999999999991 239.5499999999975
entryt 19.259999999999991 238.510000000001
provoc 1063.820000000024 235.8300000000021
fedua 263.81 211.43
ecocrافت 241.9499999999996 200.7899999999996
keen 435.6199999999995 199.2699999999995
mane 260.26 193.47
freshbubble 502.3399999999975 183.6399999999987
matreshka 182.6700000000004 182.6700000000004
chi 538.6100000000001 179.670000000000002
cristalinas 584.9499999999995 157.3200000000005
familius 1843.429999999998 150.9699999999831
kalinich 384.5800000000013 135.7700000000014
niskin 293.069999999994 135.0299999999992
elizavecca 204.3000000000004 133.7700000000004
neferetiti 366.64 133.1199999999992
finish 230.38 132.0
igrobeauty .645.0700000000006 131.410000000000008
dizao 945.510000000014 126.38000000000102
osmo 762.31 116.7299999999999
batiste 874.1699999999996 101.76999999999975
carmex 243.36 98.27999999999975
eos 152.61 98.2700000000001
```

```
hadoop@p-172-31-85-6~  
nefertiti 366.64 133.11999999999992  
finish 230.38 132.0  
igrobeauty 645.0700000000006 131.41000000000008  
diao 945.5100000000014 126.38000000000102  
osm 7.1231 116.72309999999999  
muscite 874.1699999999998 101.76999999999975  
carmax 243.36 98.27999999999997  
eos 152.61 98.27000000000001  
depiflax 2803.779999999998 96.71000000000231  
enjoy 136.57000000000002 95.22000000000003  
keras5 525.2 94.28999999999999  
aura 177.50999999999996 93.55099999999996  
plazan 184.0100000000005 92.64000000000005  
kern 80.899999999995 84.55999999999593  
nirvel 234.32999999999987 71.28999999999988  
knomad 810.669999999992 70.83999999999446  
egomania 146.04000000000002 68.57000000000002  
cutrien 367.62 68.25  
laboratorium 312.52 66.01999999999998  
imra 351.210000000001 63.19000000000017  
deval 61.28999999999999 61.28999999999999  
mudakefoot 60.11000000000001 60.110000000000014  
kangs 59.45 59.45  
profenna 736.84999999999999 57.61999999999966  
koelcia 112.75 57.25  
balbcare 212.37999999999997 57.05000000000001  
elskin 307.6500000000015 56.56000000000006  
foamix 80.49 45.44999999999996  
ladykin 170.57 44.92  
likato 340.96999999999987 44.9100000000014  
momo 446.2000000000001 37.2800000000003  
silenta 21.210000000004 33.61000000000007  
beautyblender 109.40999999999999 30.66999999999973  
biore 90.31 29.65999999999997  
orly 931.0900000000004 28.71000000000015  
estelare 471.8700000000023 27.05999999999978  
propofil 118.0200000000001 24.65999999999997  
blixt 63.4 24.44999999999996  
binacil 24.25999999999999 24.25999999999998  
perfector 435.12 23.89999999999977  
glysolid 91.56999999999999 21.86000000000014  
veraclara 71.2100000000001 21.1  
juno 21.08 21.08  
kamill 81.4900000000001 18.48000000000004  
treaclemon 181.4900000000004 18.12000000000005  
supertan 66.5100000000002 16.14000000000008  
barbie 12.39 12.39  
deoproce 329.1700000000001 12.33000000000041  
rasayan 28.93999999999998 10.14
```

```

hadoop@ip-172-31-85-6:~$ hadoop fs -text /user/hadoop/brands.csv | head -n 160
+-----+-----+
| brand_name | october_sales | november_sales |
+-----+-----+
| nirvel | 234.32999999999987 | 71.28999999999988 |
| konad | 810.6699999999992 | 70.83999999999946 |
| egomania | 146.04000000000002 | 68.57000000000002 |
| cutrin | 367.62 | 68.25 |
| laboratorium | 312.52 | 66.01999999999988 |
| im | 381.21000000000001 | 63.19000000000017 |
| oval | 61.28999999999999 | 61.28999999999999 |
| marutaka-foot | 109.33000000000001 | 60.11000000000014 |
| kares | 59.45 | 59.45 |
| profhenna | 736.8499999999999 | 57.61999999999966 |
| koelcia | 112.75 | 57.25 |
| balbcare | 212.37999999999997 | 57.05000000000001 |
| elskin | 307.65000000000015 | 56.56000000000006 |
| foamei | 80.49 | 45.44999999999996 |
| ladykin | 191.57 | 57.25 |
| littles | 140.65999999999997 | 44.91000000000014 |
| manvala | 446.320000000001 | 37.28000000000003 |
| vilenta | 231.21000000000004 | 33.61000000000007 |
| beautyblender | 109.40999999999998 | 30.66999999999973 |
| biore | 90.31 | 29.65999999999997 |
| orly | 931.090000000004 | 28.71000000000015 |
| estelare | 471.8700000000023 | 27.059999999999718 |
| profepil | 118.02000000000001 | 24.659999999999997 |
| blix | 63.4 | 24.44999999999996 |
| binail | 24.25999999999998 | 24.25999999999998 |
| godofredo | 425.12 | 23.859999999999977 |
| glysolid | 91.58999999999999 | 21.86000000000014 |
| veraclar | 71.21000000000001 | 21.1 |
| juno | 21.08 | 21.08 |
| kamill | 81.49000000000001 | 18.48000000000004 |
| treaclemoon | 181.49000000000004 | 18.12000000000005 |
| supertan | 66.51000000000002 | 16.14000000000008 |
| barbie | 12.39 | 12.39 |
| desproce | 304.17000000000001 | 12.33000000000041 |
| mayyan | 28.93999999999998 | 10.14 |
| fly | 27.16999999999998 | 10.02999999999998 |
| tertio | 245.8 | 9.64000000000015 |
| jaguar | 1110.650000000003 | 8.53999999999964 |
| soleo | 212.529999999998 | 8.329999999999814 |
| neoleor | 51.7 | 8.29000000000006 |
| moyou | 10.28000000000001 | 4.57000000000001 |
| bodyton | 1380.639999999999 | 4.30000000000637 |
| skinty | 12.44000000000001 | 3.56000000000005 |
| hellologo | 3.1 | 3.1 |
| proxin | 102.60999999999999 | 1.609999999999693 |
| coima | 20.92999999999993 | 0.659999999999922 |
| ovale | 3.1 | 0.56 |
Time taken: 44.335 seconds, Fetched: 160 row(s)
hive:~
```

Insights:

1. Here are some 160 brands with increment in the selling from October to November.
2. ‘Grattol’ brand has the highest total increment i.e., 36,027/- and ‘Ovale’ seems to have the least increment of 0.56/- from October to November.
3. Among all these brands lists, ‘Runail’ which was the best brand in terms of selling in October and November combined is also in the top 10 brands with high increment for October (71539.28) to November (76758.61) i.e., increment of total 5219.38.
4. This implies that ‘Runail’ is the best and popular brand among all other brands within people.

8. Your company wants to reward the top 10 users of its websites 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 total_amt_spend FROM Dynamic_Retailstore WHERE event_type = 'purchase' GROUP BY user_id ORDER BY total_amt_spend DESC LIMIT 10;
```

```

hadoop@ip-172-31-85-6:~ 
juno 21.08 21.08 18.480000000000004
kamill 81.49000000000001 18.120000000000005
treaclemoon 181.49000000000004 16.140000000000008
supertan 66.51000000000002
barbie 12.39 12.39
deoproce 329.17000000000001 12.33000000000041
rasyan 28.93999999999998 10.14
fly 27.16999999999998 10.029999999999998
tertio 245.8 9.64000000000015
jaguar 1110.65000000000003 8.539999999999964
soleo 212.5299999999998 8.329999999999814
neoleor 51.7 8.29000000000006
moyou 102.28000000000001 4.57000000000001
bodyton 1380.639999999999 4.300000000000637
skinity 12.440000000000001 3.560000000000005
hellologic 3.1 3.1
grace 102.60999999999999 1.6899999999999693
cosima 20.92999999999993 0.699999999999922
ovale 3.1 0.56
Time taken: 44.335 seconds, Fetched: 160 row(s)
hive> SELECT user_id,SUM(price)AS total_amt_spend FROM Dynamic_Retailstore WHERE event_type='purchase' GROUP BY user_id
> ORDER BY total_amt_spend DESC LIMIT 10;
Query ID = hadoop_20220302101425_23c27fe2-1d49-4d95-add7-12c846f1dbf3
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1646208725012_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: 26.93 s  

-----  

OK  

557790271 2715.869999999995  

150318419 1645.969999999998  

562167663 1352.850000000001  

531900924 1329.449999999996  

557850743 1295.480000000007  

522130011 1185.389999999999  

561592095 1109.700000000003  

431950134 1097.589999999997  

566576008 1056.359999999997  

521347209 1040.910000000003  

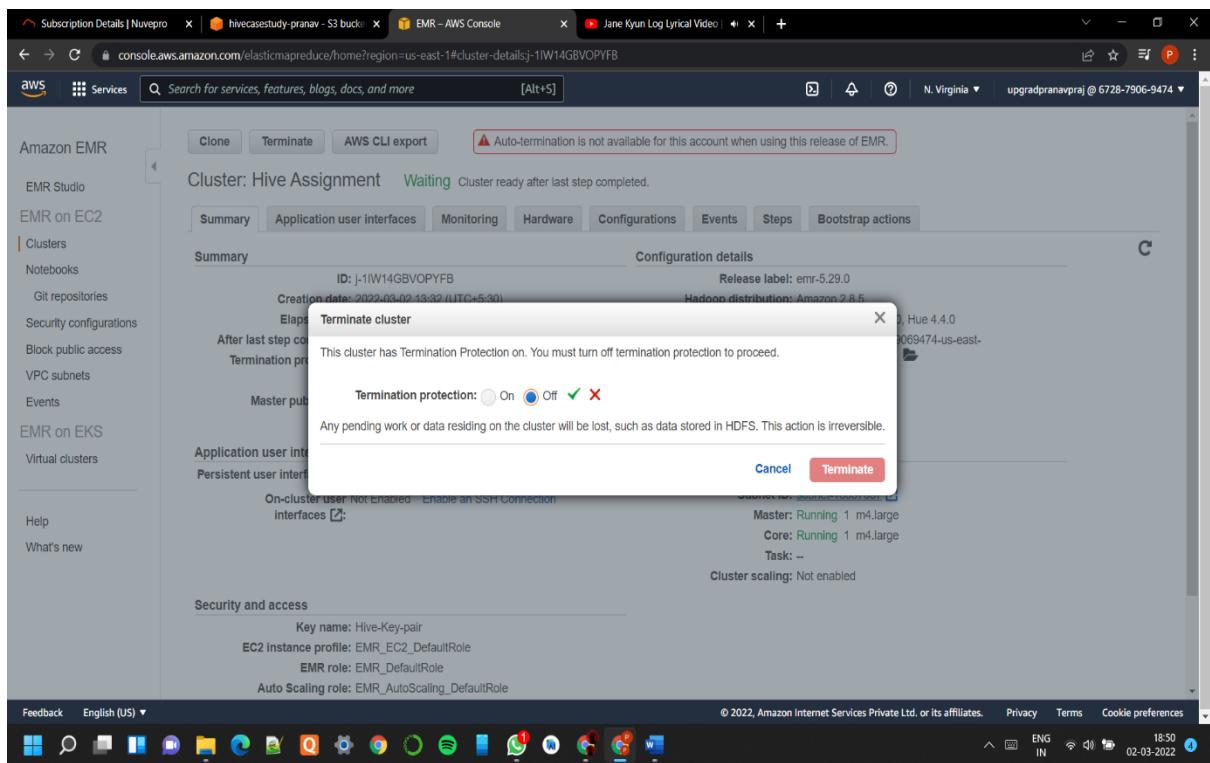
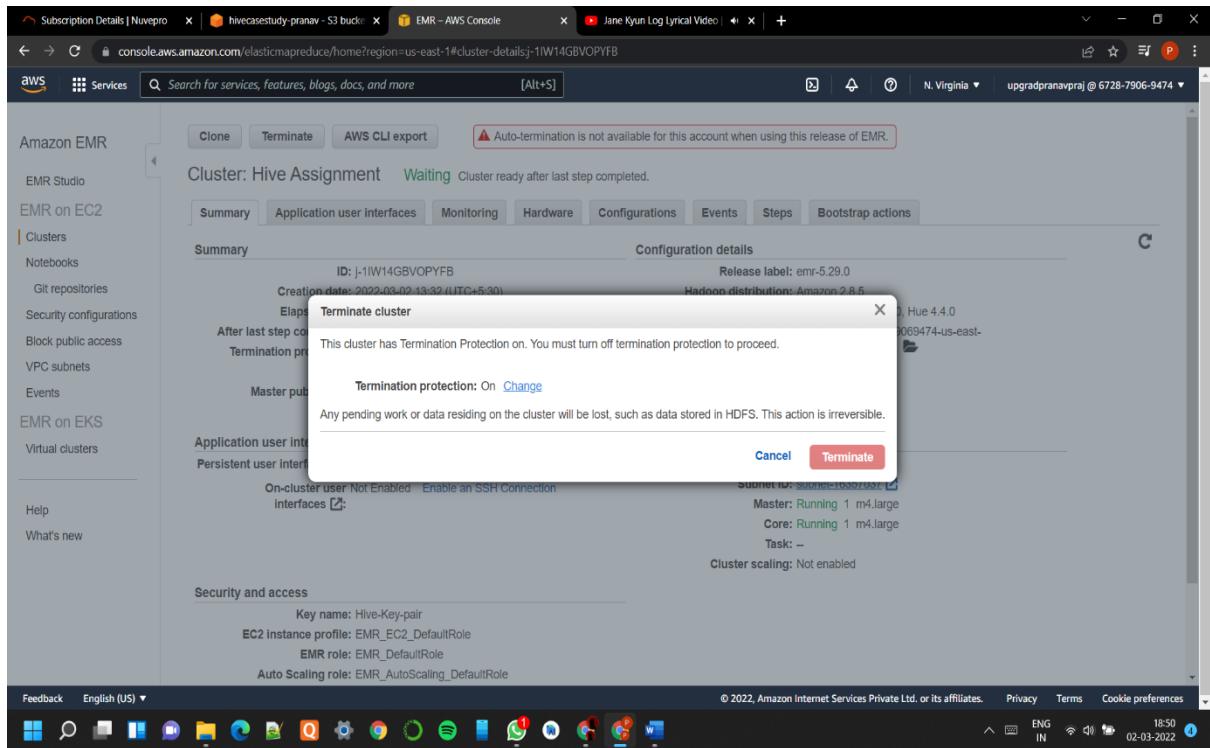
Time taken: 28.658 seconds, Fetched: 10 row(s)
hive> 
```

Insights:

1. Here is the list of the top 10 users or buyers who have spent the most and could be rewarded with a Golden Customer plan to attract more people in the coming future.
2. With the Optimized table the execution time reduced with proper partitioning and bucketing.
3. Time taken to execute this query on optimized table is 28.658 seconds.

TERMINATION PROCESS:

After completing our analysis, we should terminate the EMR cluster



Amazon EMR

EMR Studio

EMR on EC2

Clusters

Notebooks

Git repositories

Security configurations

Block public access

VPC subnets

Events

EMR on EKS

Virtual clusters

Help

What's new

Subscription Details | Nuvepro | hivecasestudy-pranav - S3 bucket | EMR - AWS Console | Jane Kyun Log Lyrical Video | +

console.aws.amazon.com/elasticmapreduce/home?region=us-east-1#cluster-details:j-1IW14GBVOPYFB

Clone Terminate AWS CLI export [Alt+S] Auto-termination is not available for this account when using this release of EMR.

Cluster: Hive Assignment Terminating Terminated by user request

Summary Application user interfaces Monitoring Hardware Configurations Events Steps Bootstrap actions

Summary Configuration details

ID: j-1IW14GBVOPYFB Release label: emr-5.29.0

Creation date: 2022-03-02 13:32 (UTC+5:30) Hadoop distribution: Amazon 2.8.5

Elapsed time: 5 hours, 18 minutes Applications: Hive 2.3.6, Pig 0.17.0, Hue 4.4.0

After last step completes: Cluster waits Log URI: s3://aws-logs-672879069474-us-east-1/elasticmapreduce/

Termination protection: Off EMRFS consistent view: Disabled

Tags: -- Custom AMI ID: --

Master public DNS: ec2-54-89-211-6.compute-1.amazonaws.com Connect to the Master Node Using SSH

Application user interfaces Network and hardware

Persistent user interfaces: -- Availability zone: us-east-1d

On-cluster user interfaces: -- Subnet ID: subnet-16357037

Master: Running 1 m4.large

Core: Running 1 m4.large

Task: --

Cluster scaling: Not enabled

Security and access

Key name: Hive-Key-pair

EC2 instance profile: EMR_EC2_DefaultRole

EMR role: EMR_DefaultRole

Auto Scaling role: EMR_AutoScaling_DefaultRole

Feedback English (US) © 2022, Amazon Internet Services Private Ltd. or its affiliates. Privacy Terms Cookie preferences ENG IN 18:50 02-03-2022

Amazon EMR

EMR Studio

EMR on EC2

Clusters

Notebooks

Git repositories

Security configurations

Block public access

VPC subnets

Events

EMR on EKS

Virtual clusters

Help

What's new

Subscription Details | Nuvepro | EMR - AWS Console | EMR - AWS Console | Wo Ladki Hai Kahan Lyrical | +

console.aws.amazon.com/elasticmapreduce/home?region=us-east-1#cluster-details:j-1IW14GBVOPYFB

Clone Terminate AWS CLI export [Alt+S] Auto-termination is not available for this account when using this release of EMR.

Cluster: Hive Assignment Terminated Terminated by user request

Summary Application user interfaces Monitoring Hardware Configurations Events Steps Bootstrap actions

Summary Configuration details

ID: j-1IW14GBVOPYFB Release label: emr-5.29.0

Creation date: 2022-03-02 13:32 (UTC+5:30) Hadoop distribution: Amazon 2.8.5

End date: 2022-03-02 18:53 (UTC+5:30) Applications: Hive 2.3.6, Pig 0.17.0, Hue 4.4.0

Elapsed time: 5 hours, 20 minutes Log URI: s3://aws-logs-672879069474-us-east-1/elasticmapreduce/

After last step completes: Cluster waits EMRFS consistent view: Disabled

Termination protection: Off Custom AMI ID: --

Tags: --

Master public DNS: ec2-54-89-211-6.compute-1.amazonaws.com Connect to the Master Node Using SSH

Application user interfaces Network and hardware

Persistent user interfaces: -- Availability zone: us-east-1d

On-cluster user interfaces: -- Subnet ID: subnet-16357037

Master: Terminated 1 m4.large

Core: Terminated 1 m4.large

Task: --

Cluster scaling: Not enabled

Security and access

Key name: Hive-Key-pair

EC2 instance profile: EMR_EC2_DefaultRole

EMR role: EMR_DefaultRole

Auto Scaling role: EMR_AutoScaling_DefaultRole

Feedback English (US) © 2022, Amazon Internet Services Private Ltd. or its affiliates. Privacy Terms Cookie preferences ENG IN 18:59 02-03-2022

Cluster terminated!!