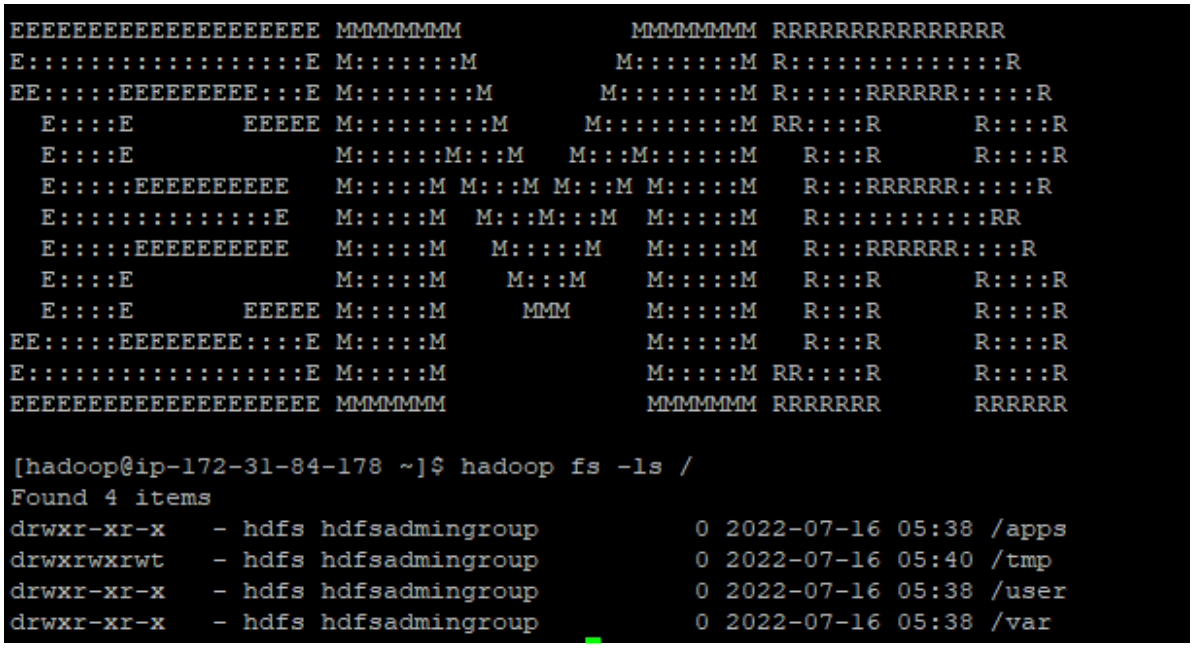
HIVE CASE STUDY

Creating Directory and loading data in HDFS in EMR cluster

- Checking file system and directory :-

  [hadoop@ip-172-31-84-178 ~]$ hadoop fs -ls /

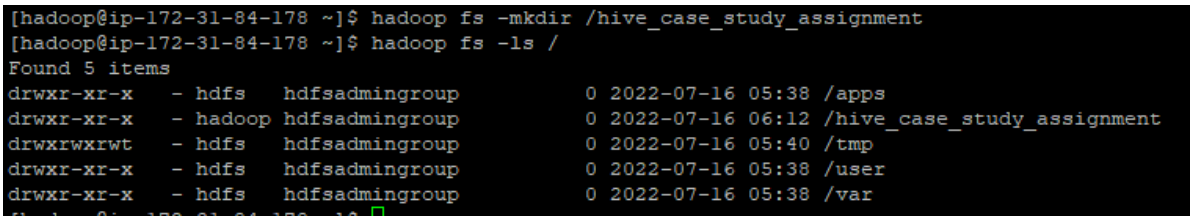


- Creating and checking directory hive_case_study_assignment

  [hadoop@ip-172-31-84-178 ~]$ hadoop fs -mkdir /hive_case_study_assignment

  [hadoop@ip-172-31-84-178 ~]$ hadoop fs -ls /

  We can see that a directory hive_case_study_assignment has been created.



- Loading Data : Since the size of the data is large we'll load the data into HDFS from S3 an into the local storage.

  [hadoop@ip-172-31-84-178 ~]$ hadoop distcp 's3://data-hive/2019-Oct.csv'

  /hive_case_study_assignment/2019-Oct.csv

  [hadoop@ip-172-31-84-178 ~]$ hadoop distcp 's3://data-hive/2019-Nov.csv'

  /hive_case_study_assignment/2019-Nov.csv

```
[hadoop@ip-172-31-84-178 ~]$ hadoop distcp 's3://data-hive/2019-Oct.csv' /hive_case_study_assignment/2019-Oct.csv
22/07/16 06:22:54 INFO tools.OptionsParser: parseChunkSize: blocksperchunk false
22/07/16 06:22:56 INFO tools.DistCp: Input Options: DistCpOptions{atomicCommit=false, syncFolder=false, deleteMissing=false, ignoreFailures=false, overwrite=false, append=false, useDiff=false, useRdiff=false, fromSnapshot=null, toSnapsho
t=null, skipCRC=false, blocking=true, numListstatusThreads=0, maxMaps=20, mapBandwidth=100, sslConfigurationFile='null', copyStrategy='uniformsize', preserveStatus=[], preserveRawXattrs=false, atomicWorkPath=null, logPath=null, sourceFil
eListing=null, sourcePaths=[s3://data-hive/2019-Oct.csv], targetPath=/hive_case_study_assignment/2019-Oct.csv, targetPathExists=false, filtersFile='null', blocksPerChunk=0, copyBufferSize=8192, verboseLog=false}
22/07/16 06:22:56 INFO client.RMProxy: Connecting to ResourceManager at ip-172-31-84-178.ec2.internal/172.31.84.178:8032
22/07/16 06:22:56 INFO client.AHSProxy: Connecting to Application History server at ip-172-31-84-178.ec2.internal/172.31.84.178:10200
22/07/16 06:23:00 INFO tools.SimpleCopyListing: Paths (files+dirs) cnt = 1; dirCnt = 0
22/07/16 06:23:00 INFO tools.SimpleCopyListing: Build file listing completed.
22/07/16 06:23:00 INFO Configuration.deprecation: io.sort.mb is deprecated. Instead, use mapreduce.task.io.sort.mb
22/07/16 06:23:00 INFO Configuration.deprecation: io.sort.factor is deprecated. Instead, use mapreduce.task.io.sort.factor
22/07/16 06:23:00 INFO tools.DistCp: Number of paths in the copy list: 1
22/07/16 06:23:00 INFO tools.DistCp: Number of paths in the copy list: 1
22/07/16 06:23:00 INFO client.RMProxy: Connecting to ResourceManager at ip-172-31-84-178.ec2.internal/172.31.84.178:8032
22/07/16 06:23:00 INFO client.AHSProxy: Connecting to Application History server at ip-172-31-84-178.ec2.internal/172.31.84.178:10200
22/07/16 06:23:01 INFO mapreduce.JobSubmitter: number of splits:1
22/07/16 06:23:01 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1657949956621_0001
22/07/16 06:23:01 INFO conf.Configuration: resource-types.xml not found
22/07/16 06:23:01 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
22/07/16 06:23:01 INFO resource.ResourceUtils: Adding resource type - name = memory-mb, units = Mi, type = COUNTABLE
22/07/16 06:23:01 INFO resource.ResourceUtils: Adding resource type - name = vcores, units = , type = COUNTABLE
22/07/16 06:23:02 INFO impl.YarnClientImpl: Submitted application application_1657949956621_0001
22/07/16 06:23:02 INFO mapreduce.Job: The url to track the job: http://ip-172-31-84-178.ec2.internal:20888/proxy/application_1657949956621_0001/
22/07/16 06:23:02 INFO tools.DistCp: DistCp job-id: job_1657949956621_0001
22/07/16 06:23:02 INFO mapreduce.Job: Running job: job_1657949956621_0001
22/07/16 06:23:11 INFO mapreduce.Job: Job job_1657949956621_0001 running in uber mode : false
22/07/16 06:23:11 INFO mapreduce.Job:  map 0% reduce 0%
22/07/16 06:23:29 INFO mapreduce.Job:  map 100% reduce 0%
22/07/16 06:23:34 INFO mapreduce.Job: Job job_1657949956621_0001 completed successfully
22/07/16 06:23:35 INFO mapreduce.Job: Counters: 38
        File System Counters
                FILE: Number of bytes read=0
                FILE: Number of bytes written=224583
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=363
                HDFS: Number of bytes written=482542278
                HDFS: Number of read operations=11
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=5
                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)=649824
                Total time spent by all reduces in occupied slots (ms)=0
                Total time spent by all map tasks (ms)=20307
                Total vcore-milliseconds taken by all map tasks=20307
                Total megabyte-milliseconds taken by all map tasks=20794368
        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)=463
                CPU time spent (ms)=20350
```

- Checking loaded files :-

  [hadoop@ip-172-31-84-178 ~]$ hadoop fs -ls /hive_case_study_assignment

  We can see that files are loaded successfully

```
[hadoop@ip-172-31-84-178 ~]$ hadoop fs -ls /hive_case_study_assignment
Found 2 items
-rw-r--r--   1 hadoop hdfsadmingroup  545839412 2022-07-16 06:26 /hive_case_study_assignment/2019-Nov.csv
-rw-r--r--   1 hadoop hdfsadmingroup  482542278 2022-07-16 06:23 /hive_case_study_assignment/2019-Oct.csv
```

- Launching Hive now :-

  [hadoop@ip-172-31-84-178 ~]$ hive

```
[hadoop@ip-172-31-84-178 ~]$ hive

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

- Creating database case_Study :-

  hive> create database if not exists case_study ;

  We can see that case_Study database has been created

```
hive> create database if not exists case_study ;
OK
Time taken: 2.373 seconds
hive> show databases ;
OK
case_study
default
Time taken: 0.271 seconds, Fetched: 2 row(s)
```

- Creating an External table i.e., 'ecom_sales' which will hold the data for both the data files stored in temporary directory of HDFS.

```
hive> use case_study ;

hive> create External table if not exists ecom_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 ;
```

```
hive> create External table if not exists ecom_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.apac
he.hadoop.hive.serde2.OpenCSVSerde' WITH SERDEPROPERTIES ("separatorChar"=",","quoteChar"="\"","escapeChar"="\\") stored as textfile ;
OK
Time taken: 0.093 seconds
```

- Validating if table created or not :-

```
hive> show databases ;

hive> use case_study ;

hive> desc ecom_sales ;
```

We can see that the ecom_sales table created successfully

```
hive> show databases ;
OK
case_study
default
Time taken: 0.939 seconds, Fetched: 2 row(s)
hive> use case_study ;
OK
Time taken: 0.273 seconds
hive> show tables ;
OK
ecom_sales
Time taken: 0.122 seconds, Fetched: 1 row(s)
hive> desc ecom_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: 0.19 seconds, Fetched: 9 row(s)
```

- Loading the data into table ecom_sales :-

```
hive> load data inpath '/hive_case_study_assignment/2019-Oct.csv' into table ecom_sales ;

hive> load data inpath '/hive_case_study_assignment/2019-Nov.csv' into table ecom_sales ;
```

We can see that data loaded successfully

```
hive> load data inpath '/hive_case_study_assignment/2019-Oct.csv' into table ecom_sales ;
Loading data to table case_study.ecom_sales
OK
Time taken: 3.266 seconds
hive> load data inpath '/hive_case_study_assignment/2019-Nov.csv' into table ecom_sales ;
Loading data to table case_study.ecom_sales
OK
Time taken: 0.652 seconds
```

- Validating if data loaded in the table or not for five rows :-

  hive> select * from ecom_sales limit 5 ;

  We can see that the data loaded successfully in ecom_sales table

```
hive> select * from ecom_sales limit 5 ;
OK
event_time      event_type      product_id      category_id     category_code   brand   price   user_id user_session
2019-11-01 00:00:02 UTC view    5802432 1487580009286598681                     0.32    562076640       09fafd6c-6c99-46b1-834f-33527f4de241
2019-11-01 00:00:09 UTC cart    5844397 1487580006317032337                     2.38    553329724       2067216c-31b5-455d-a1cc-af0575a34ffb
2019-11-01 00:00:10 UTC view    5837166 1783999064103190764             pnb     22.22   556138645       57ed222e-a54a-4907-9944-5a875c2d7f4f
2019-11-01 00:00:11 UTC cart    5876812 1487580010100293687             jessnail        3.16    564506666       186c1951-8052-4b37-adce-dd9644b1d5f7
Time taken: 3.341 seconds, Fetched: 5 row(s)
```

==========

HIVE QUERY AND ANALYSIS

Assignment Questions and Solution

Q1 > Find the total revenue generated due to purchases made in October.

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

```
hive> select sum(price) from ecom_sales where Month(event_time)=10 and event_type='purchase';
Query ID = hadoop_20220716100811_3ae55e41-7a29-492a-aabc-fccb0bc117b2
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_1657962535224_0009)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      6          6        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 127.38 s
--------------------------------------------------------------------------------
OK
2423076.859999421
Time taken: 137.375 seconds, Fetched: 1 row(s)
```

→ As we can see here that the query takes 137.375 seconds which can be optimized by creating dynamic partition and bucketing and then we can compare the execution time.

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;
```

→ Let's Create a new table by name ecom_sales_dp to store the dataset which we will partitioned by using 'event_type' and clustered by 'user_id'.

hive> create External table if not exists ecom_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 ;
hive> desc ecom_sales_dp ;

```
hive> create External table if not exists ecom_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) clust
red by(user_id) into 5 buckets ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.OpenCSVSerde' stored as textfile ;
OK
Time taken: 0.141 seconds
hive> desc ecom_sales_dp ;
OK
event_time              string                  from deserializer
product_id              string                  from deserializer
category_id             string                  from deserializer
category_code           string                  from deserializer
brand                   string                  from deserializer
price                   string                  from deserializer
user_id                 string                  from deserializer
user_session            string                  from deserializer
event_type              string

# Partition Information
# col_name              data_type               comment

event_type              string
Time taken: 0.463 seconds, Fetched: 14 row(s)
```

→ Let's Load the data into the new table ecom_sales_dp from the old ecom_sales table:-
hive> insert into ecom_sales_dp partition(event_type) select event_time, product_id, category_id,category_code, brand, price, user_id, user_session, event_type from ecom_sales;

```
hive> insert into ecom_sales_dp partition(event_type) select event_time, product_id, category_id,category_code, brand, price, user_id, user_session, event_type from ecom_sales;
Query ID = hadoop_20220716104343_e50b8b35-bd8b-4f7c-81dd-8d01587cd641
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_1657962535224_0010)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      6          6        0        0       0       0
Reducer 2 ...... container     SUCCEEDED     18         18        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 325.70 s
--------------------------------------------------------------------------------
Loading data to table case_study.ecom_sales_dp partition (event_type=null)

Loaded : 5/5 partitions.
        Time taken to load dynamic partitions: 0.537 seconds
        Time taken for adding to write entity : 0.003 seconds
OK
Time taken: 335.068 seconds
```

→ Now let's execute the same query again for comparison :-
hive> select sum(price) from ecom_sales_dp where Month(event_time)=10 and event_type='purchase';

```
hive> select sum(price) from ecom_sales_dp where Month(event_time)=10 and event_type='purchase';
Query ID = hadoop_20220716105242_40ce71a4-e2ed-4138-8bae-a2d9e4504244
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1657962535224_0010)

--------------------------------------------------------------------------------
        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: 29.75 s
--------------------------------------------------------------------------------
OK
2423076.8599997694
Time taken: 30.951 seconds, Fetched: 1 row(s)
```

→ We can now see how the time taken has reduced drastically due to partitioning and bucketing. Earlier it took almost 137.375 seconds for the query to run however now it took only 30.951 seconds.

Answer: The total sales in the month of October is *2423076.8599997694*

Q2 > Write a query to yield the total sum of purchases per month in a single output.

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

→ hive> select Month(event_time) as Month, sum(price) as total_sum_of_purchase from ecom_sales_dp where event_type='purchase' group by Month(event_time);

```
hive> set hive.cli.print.header=true;
hive> select Month(event_time) as Month, sum(price) as total_sum_of_purchase from ecom_sales_dp where event_type='purchase' group by Month(event_time);
Query ID = hadoop_20220716110634_1ec0a443-958e-43c9-941c-735c65cb1bea
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1657962535224_0011)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      5          5        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      2          2        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 30.05 s
--------------------------------------------------------------------------------
OK
month   total_sum_of_purchase
10      2423076.8599997694
11      3062033.7999997106
Time taken: 30.927 seconds, Fetched: 2 row(s)
```

Answer: In the month of October the total sum of purchases was *2423076.8599997694*
In the month of November the total sum of purchases was *3062033.7999997106*

Q3 > Write a query to find the change in revenue generated due to purchases from October to November.

→ hive> with revenue 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 ecom_sales_dp where event_type='purchase' and Month(event_time) in (10,11)) select Oct,Nov,(Nov-Oct) as difference from revenue;

```
hive> with revenue 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 ecom_sales_dp where event_type='purchase' and Month(event_time) in (10,
11) ) select Oct,Nov,(Nov-Oct) as difference from revenue;
Query ID = hadoop_20220716112248_fc59252b-d51e-4c24-b7a2-f1fc7b8350ed
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1657962535224_0012)

--------------------------------------------------------------------------------
        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: 34.06 s
--------------------------------------------------------------------------------
OK
2423076.8599997694      3062033.799999711       638956.9399999417
Time taken: 34.79 seconds, Fetched: 1 row(s)
```

Answer: The difference in the revenue is *638956.9399999417*

Q4 > Find distinct categories of products. Categories with null category code can be ignored.

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

```
hive> select distinct split(category_code,'\\.')[0] as Categories from ecom_sales_dp where split(category_code,'\\.')[0]<>'';
Query ID = hadoop_20220716112910_8634dd00-88a6-4335-8468-6f03dbb606cf
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_1657962535224_0013)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      6          6        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      2          2        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 121.20 s
--------------------------------------------------------------------------------
OK
apparel
category_code
furniture
stationery
accessories
appliances
sport
Time taken: 130.947 seconds, Fetched: 7 row(s)
```

Answer: We can see that the distinct categories are *apparel, category_code, furniture, stationery, accessories, appliances and sport*.

Q5 > Find the total number of products available under each category.

→ hive> select split(category_code,'\\.')[0] as category , count(product_id) as count_prod from ecom_sales group by split(category_code,'\\.')[0] order by count_prod desc;

```
hive> select split(category_code,'\\.')[0] as category , count(product_id) as count_prod from ecom_sales group by split(category_code,'\\.')[0] order by count_prod desc;
Query ID = hadoop_20220716114406_608be72e-fea3-4778-9e09-a2c3639d4e9b
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1657962535224_0014)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      6          6        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      2          2        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 115.02 s
--------------------------------------------------------------------------------
OK
        17189790
appliances      123472
stationery      53444
furniture       47208
apparel 36464
accessories     25858
category_code   4
sport   4
Time taken: 116.028 seconds, Fetched: 8 row(s)
```

Answer: Total number of products under each category is as follows: *Appliances – 123472, Stationery – 53444, Furniture – 47208, Apparel – 36464, Accessories – 25858 and sport – 4.*

Q6> Which brand had the maximum sales in October and November combined?

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

```
hive> select brand, sum(price) as total_sales from ecom_sales_dp where brand <>'' and event_type='purchase' group by brand order by total_sales desc limit 1;
Query ID = hadoop_20220716123116_06aa7a99-edbd-42d0-9477-47c18f49665d
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1657973679146_0003)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 ..........   container     SUCCEEDED      5          5        0        0       0       0
Reducer 2 ......   container     SUCCEEDED      2          2        0        0       0       0
Reducer 3 ......   container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 30.33 s
--------------------------------------------------------------------------------
OK
runail  148297.9399999995
Time taken: 32.788 seconds, Fetched: 1 row(s)
```

Answer: *Runail* is the brand with total sales of *148297.9399999995* which is the maximum sales for October and November combined.

Q7> Which brands increased their sales from October to November?

→ hive> with Brand 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 ecom_sales_dp where event_type='purchase' group by brand) select brand , Oct,Nov,(Nov-Oct) as difference from Brand where (Nov-Oct)>0 ORDER BY difference;

```
hive> with Brand 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
    > ecom_sales_dp where event_type='purchase' group by brand) select brand , Oct,Nov,(Nov-Oct) as difference from Brand where (Nov-Oct)>0 ORDER BY difference;
Query ID = hadoop_20220716124118_02d87acb-b128-4fca-844d-399f169ede67
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1657973679146_0004)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      5          5        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      2          2        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 29.23 s
----------------------------------------------------------------------------------------------
OK
ovale   2.54    3.1     0.56
cosima  20.229999999999997      20.93   0.7000000000000028
grace   100.91999999999999      102.60999999999999      1.6899999999999977
helloganic      0.0     3.1     3.1
skinity 8.88    12.440000000000001      3.5600000000000005
bodyton 1376.3400000000001      1380.6400000000003      4.300000000000182
moyou   5.71    10.280000000000001      4.570000000000001
neoleor 43.41   51.7    8.290000000000006
soleo   204.20000000000013      212.53000000000011      8.329999999999984
jaguar  1102.1100000000001      1110.6499999999999      8.539999999999736
tertio  236.16000000000003      245.79999999999995      9.63999999999993
fly     17.14   27.17   10.030000000000001
rasyan  18.799999999999997      28.94   10.140000000000004
deoproce        316.84000000000003      329.17  12.329999999999984
barbie  0.0     12.39   12.39
supertan        50.36999999999999       66.50999999999999       16.14
treaclemoon     163.37  181.49000000000004      18.120000000000033
kamill  63.01   81.49000000000001       18.48000000000001
juno    0.0     21.08   21.08
veraclara       50.11   71.21   21.099999999999994
glysolid        69.73   91.58999999999999       21.859999999999985
godefroy        401.22  425.11999999999995      23.899999999999992
binacil 0.0     24.259999999999998      24.259999999999998
blixz   38.95   63.400000000000006      24.450000000000003
profepil        93.36   118.02  24.659999999999997
estelare        444.8100000000001       471.87000000000006      27.059999999999945
orly    902.3799999999998       931.0899999999999       28.71000000000015
biore   60.650000000000006      90.31   29.659999999999997
beautyblender   78.74000000000001       109.41  30.669999999999987
vilenta 197.59999999999997      231.20999999999998      33.610000000000014
mavala  409.04  446.32000000000005      37.28000000000003
likato  296.0599999999995       340.97  44.91000000000008
ladykin 125.65  170.57  44.91999999999999
foamie  35.04   80.49   45.449999999999996
elskin  251.09000000000003      307.65  56.559999999999945
balbcare        155.33  212.38000000000002      57.05000000000001
koelcia 55.49999999999999       112.75  57.25000000000001
profhenna       679.2299999999999       736.8499999999999       57.620000000000005
kares   0.0     59.449999999999996      59.449999999999996
marutaka-foot   49.22   109.33  60.11
dewal   0.0     61.28999999999999       61.28999999999999
inm     288.02000000000004      351.21  63.18999999999994
laboratorium    246.5   312.52  66.01999999999998
cutrin  299.36999999999995      367.61999999999995      68.25
egomania        77.47   146.04  68.57
artex   2730.64 4327.249999999998       1596.6099999999983
beautix 10493.949999999992      12222.950000000003      1729.000000000011
milv    3904.939999999994       5642.009999999988       1737.0699999999943
masura  31266.07999999897       33058.46999999907       1792.3900000001013
f.o.x   6624.229999999997       8577.279999999993       1953.0499999999965
kapous  11927.15999999997       14093.079999999965      2165.9199999999946
concept 11032.140000000018      13380.400000000016      2348.2599999999984
estel   21756.74999999997       24142.66999999998       2385.920000000009
kaypro  881.34  3268.7  2387.3599999999997
benovy  409.61999999999995      3259.970000000002       2850.350000000002
italwax 21940.239999999958      24799.369999999937      2859.129999999979
yoko    8756.909999999993       11707.879999999997      2950.970000000005
haruyama        9390.690000000044       12352.910000000003      2962.2199999999593
marathon        7280.749999999999       10273.099999999999      2992.3499999999995
lovely  8704.38 11939.059999999998      3234.6799999999985
bpw.style       11572.14999999988       14837.439999999844      3265.2899999999645
staleks 8519.730000000005       11875.610000000002      3355.8799999999974
freedecor       3421.779999999992       7671.800000000016       4250.020000000024
runail  71539.27999999975       76758.66000000061       5219.380000000863
polarus 6013.72 11371.929999999998      5358.209999999998
cosmoprofi      8322.809999999996       14536.990000000003      6214.180000000008
jessnail        26287.839999999975      33345.229999999996      7057.390000000021
strong  29196.62999999999       38671.26999999999       9474.64
ingarden        23161.38999999995       33566.20999999988       10404.819999999934
lianail 5892.8399999999965      16394.239999999918      10501.399999999921
uno     35302.0299999998        51039.74999999998       15737.720000000176
grattol 35445.53999999992       71472.70999999953       36027.169999999605
        474679.06000000867      619509.2399999994       144830.17999999074
Time taken: 34.746 seconds, Fetched: 161 row(s)
```

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

Q8> Your company wants to reward the top 10 users of its website with a Golden Customer plan. Write a query to generate a list of top 10 users who spend the most.

→ hive> select user_id, sum(price) as total_purchases from ecom_sales_dp where event_type='purchase' group by user_id order by total_purchases desc limit 10;

```
hive> select user_id, sum(price) as total_purchases from ecom_sales_dp where event_type='purchase' group by user_id order by total_purchases desc limit 10;
Query ID = hadoop_20220716125246_87244ebe-61c9-4254-b777-99d8707b8316
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1657973679146_0005)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      5          5        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      2          2        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 26.78 s
--------------------------------------------------------------------------------
OK
557790271       2715.869999999993
150318419       1645.9700000000005
562167663       1352.8500000000001
531900924       1329.4499999999991
557850743       1295.4800000000002
522130011       1185.3899999999994
561592095       1109.700000000001
431950134       1097.59
566576008       1056.360000000001
521347209       1040.9100000000005
Time taken: 32.973 seconds, Fetched: 10 row(s)
```

Answer: Below are the top 10 users and their total respective purchase of the website:-

557790271-2715.869999999993
150318419-1645.9700000000005
562167663-1352.8500000000001
531900924-1329.4499999999991
557850743-1295.4800000000002
522130011-1185.3899999999994
561592095-1109.700000000001
431950134-1097.59
566576008-1056.360000000001
521347209-1040.9100000000005