HIVE-Case-Study

PROBLEM STATEMENT:

Using public clickstream data of a cosmetics store dataset, we have to extract some valuable insights which generally data engineers come up within an e-retail company and need to execute queries on some business problems using Hive Query Language (HQL).

Clickstream Data: Data which is collected by tracking our clicks/navigations on the websites and searching for patterns within them. This kind of data is called a clickstream data.

- Ex: E-commerce companies such as Amazon or Flip kart generate clickstream data and give recommendations based on individual's interest.

Datasets:

- https://e-commerce-events-ml.s3.amazonaws.com/2019-Oct.csv
- https://e-commerce-events-ml.s3.amazonaws.com/2019-Nov.csv

EMR Cluster: 5.29.0 release

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

1) Find the total revenue generated due to purchases made in October.
2) Write a query to yield the total sum of purchases per month in a single output.
3) Write a query to find the change in revenue generated due to purchases from October to November.
4) Find distinct categories of products. Categories with null category code can be ignored.
5) Find the total number of products available under each category.
6) Which brand had the maximum sales in October and November combined?
7) Which brands increased their sales from 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.

Implementation phase divided into the following parts:

1) Copying the data set into the HDFS:

i. Launch an EMR cluster that utilizes the Hive services.

Note: We've used a **2-Node EMR Cluster** with both the master and core nodes as **M4.large**.

ii. Move the data from the S3 bucket into the HDFS.

2019-Oct.csv

```
[hadoop@ip-172-31-59-180 ~]$ hadoop distcp s3n://e-commerce-events-ml/2019-Oct.csv /case-study/2019-Oct.csv
```

2019-Nov.csv

```
[hadoop@ip-172-31-59-180 ~]$ hadoop distcp s3n://e-commerce-events-ml/2019-Nov.csv /case-study/2019-Nov.csv
```

iii. Check data location inside Hadoop

```
[hadoop@ip-172-31-59-180 ~]$ hadoop fs -ls /case-study
Found 2 items
-rw-r--r--   1 hadoop hadoop  545839412 2021-09-05 06:56 /case-study/2019-Nov.csv
-rw-r--r--   1 hadoop hadoop  482542278 2021-09-05 06:55 /case-study/2019-Oct.csv
```

2) Creating the database and various tables:

i. Create the structure of database

```
hive> create database if not exists cosmetics;
OK
Time taken: 0.039 seconds
hive> describe database cosmetics;
OK
cosmetics               hdfs://ip-172-31-59-180.ec2.internal:8020/user/hive/warehouse/cosmetics.db      hadoop  USER
Time taken: 0.025 seconds, Fetched: 1 row(s)
```

ii. Create an external Table **ecommerce** using CSVserde properties.

```
hive> create external table if not exists ecommerce(
    > 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'= ',', 'escapeChar'= '\\')
    > stored as textfile
    > TBLPROPERTIES ("skip.header.line.count"="1");
OK
Time taken: 0.45 seconds
hive> describe extended ecommerce;
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

Detailed Table Information      Table(tableName:ecommerce, dbName:cosmetics, owner:hadoop, createTime:1630825435, lastAccessTime:0, retention:0, sd:StorageDescriptor(c
ls:[FieldSchema(name:event_time, type:timestamp, comment:null), FieldSchema(name:event_type, type:string, comment:null), FieldSchema(name:product_id, type:string, comm
nt:null), FieldSchema(name:category_id, type:string, comment:null), FieldSchema(name:category_code, type:string, comment:null), FieldSchema(name:brand, type:string, co
ment:null), FieldSchema(name:price, type:float, comment:null), FieldSchema(name:user_id, type:bigint, comment:null), FieldSchema(name:user_session, type:string, commen
:null)], location:hdfs://ip-172-31-59-180.ec2.internal:8020/user/hive/warehouse/cosmetics.db/ecommerce, inputFormat:org.apache.hadoop.mapred.TextInputFormat, outputFor
at:org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat, compressed:false, numBuckets:-1, serdeInfo:SerDeInfo(name:null, serializationLib:org.apache.hadoop.hive.
erde2.OpenCSVSerde, parameters:{escapeChar=\, separatorChar=,, serialization.format=1}), bucketCols:[], sortCols:[], parameters:{}, skewedInfo:SkewedInfo(skewedColName
:[], skewedColValues:[], skewedColValueLocationMaps:{}), storedAsSubDirectories:false), partitionKeys:[], parameters:{skip.header.line.count=1, totalSize=0, EXTERNAL=T
UE, numRows=0, rawDataSize=0, COLUMN_STATS_ACCURATE={"BASIC_STATS":"true"}, numFiles=0, transient_lastDdlTime=1630825435}, viewOriginalText:null, viewExpandedText:null
 tableType:EXTERNAL_TABLE, rewriteEnabled:false)
Time taken: 0.102 seconds, Fetched: 11 row(s)
hive> load data inpath 'hdfs:///case-study' into table ecommerce;
Loading data to table cosmetics.ecommerce
OK
Time taken: 0.78 seconds
```

iii. Load data into table

```
load data inpath 'hdfs:///case-study' into table ecommerce;
```

iv. Command to check data in the warehouse.

```
hive> [hadoop@ip-172-31-59-180 ~]$ hadoop fs -ls /user/hive/warehouse/cosmetics.db/ecommerce
Found 2 items
-rwxrwxrwt   1 hadoop hadoop  545839412 2021-09-05 06:56 /user/hive/warehouse/cosmetics.db/ecommerce/2019-Nov.csv
-rwxrwxrwt   1 hadoop hadoop  482542278 2021-09-05 06:55 /user/hive/warehouse/cosmetics.db/ecommerce/2019-Oct.csv
```

v. Create table **cosmetic_data** and insert records from **ecommerce** table.

```
hive> create external table if not exists cosmetic_data(
    > event_time string,
    > event_type string,
    > product_id string,
    > category_id string,
    > category_code string,
    > brand string,
    > price float,
    > user_id bigint,
    > user_session string )
    > row format delimited fields terminated by '|' lines terminated by '\n' stored as textfile;
OK
Time taken: 0.383 seconds
hive> insert into table cosmetic_data
    > select event_time, event_type, product_id, category_id, category_code, brand, price, user_id, user_session
    > from ecommerce;
Query ID = hadoop_20210905070953_1c3fecd9-9390-4a09-b2c4-375f5b231219
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1630821305393_0011)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      2          2        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 01/01  [==========================>>] 100%  ELAPSED TIME: 84.64 s
--------------------------------------------------------------------------------
Loading data to table cosmetics.cosmetic_data
OK
Time taken: 87.591 seconds
```

vi. Create dynamic partition table **dyn_cosmetic_data** and insert records from **cosmetic_data**

- **Command to set dynamic partition**

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

- Command to create dynamic partition table and insert records from cosmetic_data

```
hive> set hive.exec.dynamic.partition=true;
hive> set hive.exec.dynamic.partition.mode=nonstrict;
hive> create external table if not exists dyn_cosmetic_data(
    > event_time string,
    > product_id string,
    > category_id string,
    > category_code string,
    > brand string,
    > price float,
    > user_id bigint,
    > user_session string )
    > partitioned by (event_type string)
    > row format delimited fields terminated by '|' lines terminated by '\n' stored as textfile;
OK
Time taken: 0.072 seconds
hive> insert into table dyn_cosmetic_data
    > partition(event_type)
    > select event_time, product_id, category_id, category_code, brand, price, user_id, user_session, event_type
    > from cosmetic_data;
Query ID = hadoop_20210905072044_ac8cc18c-966b-421d-bf25-3c1de476ee04
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_1630821305393_0012)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      8          8        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      4          4        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 90.60 s
--------------------------------------------------------------------------------
Loading data to table cosmetics.dyn_cosmetic_data partition (event_type=null)

Loaded : 4/4 partitions.
         Time taken to load dynamic partitions: 0.26 seconds
         Time taken for adding to write entity : 0.001 seconds
OK
Time taken: 100.471 seconds
```

- Command to check partition information i.e. check dynamic partitions & their location in HDFS.

```
hive> describe extended dyn_cosmetic_data;
OK
event_time              string
product_id              string
category_id             string
category_code           string
brand                   string
price                   float
user_id                 bigint
user_session            string
event_type              string

# Partition Information
# col_name              data_type               comment

event_type              string

Detailed Table Information      Table(tableName:dyn_cosmetic_data, dbName:cosmetics, owner:hadoop, createTime:1630826433, lastAccessTime:0, retention:0, sd:StorageDescr
iptor(cols:[FieldSchema(name:event_time, type:string, comment:null), FieldSchema(name:product_id, type:string, comment:null), FieldSchema(name:category_id, type:string,
 comment:null), FieldSchema(name:category_code, type:string, comment:null), FieldSchema(name:brand, type:string, comment:null), FieldSchema(name:price, type:float, comm
ent:null), FieldSchema(name:user_id, type:bigint, comment:null), FieldSchema(name:user_session, type:string, comment:null), FieldSchema(name:event_type, type:string, co
mment:null)], location:hdfs://ip-172-31-59-180.ec2.internal:8020/user/hive/warehouse/cosmetics.db/dyn_cosmetic_data, inputFormat:org.apache.hadoop.mapred.TextInputForma
t, outputFormat:org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat, compressed:false, numBuckets:-1, serdeInfo:SerDeInfo(name:null, serializationLib:org.apache.
hadoop.hive.serde2.lazy.LazySimpleSerDe, parameters:{serialization.format=|, line.delim=
, field.delim=|}), bucketCols:[], sortCols:[], parameters:{}, skewedInfo:SkewedInfo(skewedColNames:[], skewedColValues:[], skewedColValueLocationMaps:{}), storedAsSubDi
rectories:false), partitionKeys:[FieldSchema(name:event_type, type:string, comment:null)], parameters:{totalSize=961284701, EXTERNAL=TRUE, numRows=8738120, rawDataSize=
952546581, COLUMN_STATS_ACCURATE={"BASIC_STATS":"true"}, numFiles=4, numPartitions=4, transient_lastDdlTime=1630826433}, viewOriginalText:null, viewExpandedText:null, t
ableType:EXTERNAL_TABLE, rewriteEnabled:false)
Time taken: 0.165 seconds, Fetched: 17 row(s)
hive> show partitions dyn_cosmetic_data;
OK
event_type=cart
event_type=purchase
event_type=remove_from_cart
event_type=view
Time taken: 0.05 seconds, Fetched: 4 row(s)
hive> [hadoop@ip-172-31-59-180 ~]$ hadoop fs -ls /user/hive/warehouse/cosmetics.db/dyn_cosmetic_data
Found 4 items
drwxrwxrwt   - hadoop hadoop          0 2021-09-05 07:22 /user/hive/warehouse/cosmetics.db/dyn_cosmetic_data/event_type=cart
drwxrwxrwt   - hadoop hadoop          0 2021-09-05 07:22 /user/hive/warehouse/cosmetics.db/dyn_cosmetic_data/event_type=purchase
drwxrwxrwt   - hadoop hadoop          0 2021-09-05 07:22 /user/hive/warehouse/cosmetics.db/dyn_cosmetic_data/event_type=remove_from_cart
drwxrwxrwt   - hadoop hadoop          0 2021-09-05 07:22 /user/hive/warehouse/cosmetics.db/dyn_cosmetic_data/event_type=view
```

vii. Create bucketed-partitioned table **buck_cosmetic_data** based on event type and insert records from cosmetic_data table.

```
hive> create table if not exists buck_cosmetic_data(
    > event_time string,
    > 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 (event_time) into 7 buckets
    > row format delimited fields terminated by '|' lines terminated by '\n' stored as textfile;
OK
Time taken: 0.057 seconds
hive> insert into table buck_cosmetic_data
    > partition(event_type)
    > select event_time, product_id, category_id, category_code, brand, price, user_id, user_session, event_type
    > from cosmetic_data;
Query ID = hadoop_20210905080214_99f76f30-418e-4a98-b2f9-d9950867415f
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1630821305393_0015)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      8          8        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      4          4        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 127.38 s
----------------------------------------------------------------------------------------------
Loading data to table cosmetics.buck_cosmetic_data partition (event_type=null)

Loaded : 4/4 partitions.
         Time taken to load dynamic partitions: 0.227 seconds
         Time taken for adding to write entity : 0.001 seconds
OK
Time taken: 128.623 seconds
```

- Partition description, check 7 buckets and their location inside HDFS

```
hive> describe extended buck_cosmetic_data;
OK
event_time              string
product_id              string
category_id             string
category_code           string
brand                   string
price                   float
user_id                 bigint
user_session            string
event_type              string

# Partition Information
# col_name              data_type               comment

event_type              string

Detailed Table Information      Table(tableName:buck_cosmetic_data, dbName:cosmetics, owner:hadoop, createTime:1630828928, lastAccessTime:0, retention:0, sd:StorageDesc
riptor(cols:[FieldSchema(name:event_time, type:string, comment:null), FieldSchema(name:product_id, type:string, comment:null), FieldSchema(name:category_id, type:string
, comment:null), FieldSchema(name:category_code, type:string, comment:null), FieldSchema(name:brand, type:string, comment:null), FieldSchema(name:price, type:float, com
ment:null), FieldSchema(name:user_id, type:bigint, comment:null), FieldSchema(name:user_session, type:string, comment:null), FieldSchema(name:event_type, type:string, c
omment:null)], location:hdfs://ip-172-31-59-180.ec2.internal:8020/user/hive/warehouse/cosmetics.db/buck_cosmetic_data, inputFormat:org.apache.hadoop.mapred.TextInputFor
mat, outputFormat:org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat, compressed:false, numBuckets:7, serdeInfo:SerDeInfo(name:null, serializationLib:org.apache
.hadoop.hive.serde2.lazy.LazySimpleSerDe, parameters:{serialization.format=|, line.delim=
, field.delim=|}), bucketCols:[event_time], sortCols:[], parameters:{}, skewedInfo:SkewedInfo(skewedColNames:[], skewedColValues:[], skewedColValueLocationMaps:{}), sto
redAsSubDirectories:false), partitionKeys:[FieldSchema(name:event_type, type:string, comment:null)], parameters:{totalSize=961284701, numRows=8738120, rawDataSize=95254
6581, COLUMN_STATS_ACCURATE={"BASIC_STATS":"true"}, numFiles=28, numPartitions=4, transient_lastDdlTime=1630828928}, viewOriginalText:null, viewExpandedText:null, table
Type:MANAGED_TABLE, rewriteEnabled:false)
Time taken: 0.094 seconds, Fetched: 17 row(s)
hive> [hadoop@ip-172-31-59-180 ~]$ hadoop fs -ls /user/hive/warehouse/cosmetics.db/buck_cosmetic_data/event_type=remove_from_cart
Found 7 items
-rwxrwxrwt   1 hadoop hadoop   26510418 2021-09-05 08:03 /user/hive/warehouse/cosmetics.db/buck_cosmetic_data/event_type=remove_from_cart/000000_0
-rwxrwxrwt   1 hadoop hadoop   26504529 2021-09-05 08:04 /user/hive/warehouse/cosmetics.db/buck_cosmetic_data/event_type=remove_from_cart/000001_0
-rwxrwxrwt   1 hadoop hadoop   26419250 2021-09-05 08:03 /user/hive/warehouse/cosmetics.db/buck_cosmetic_data/event_type=remove_from_cart/000002_0
-rwxrwxrwt   1 hadoop hadoop   26380602 2021-09-05 08:03 /user/hive/warehouse/cosmetics.db/buck_cosmetic_data/event_type=remove_from_cart/000003_0
-rwxrwxrwt   1 hadoop hadoop   26566389 2021-09-05 08:03 /user/hive/warehouse/cosmetics.db/buck_cosmetic_data/event_type=remove_from_cart/000004_0
-rwxrwxrwt   1 hadoop hadoop   26382146 2021-09-05 08:04 /user/hive/warehouse/cosmetics.db/buck_cosmetic_data/event_type=remove_from_cart/000005_0
-rwxrwxrwt   1 hadoop hadoop   26432611 2021-09-05 08:03 /user/hive/warehouse/cosmetics.db/buck_cosmetic_data/event_type=remove_from_cart/000006_0
```

***Note*: *Screenshots of queries of each question along with their outputs are attached below*:**

3) Launching Hive queries on our EMR cluster to answer the questions given below:

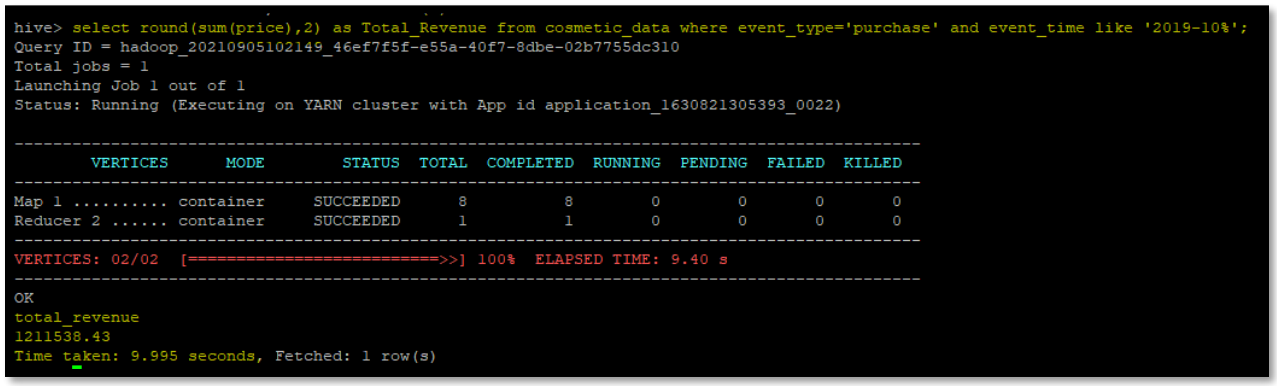
- Used optimized techniques such as partitioned table (based on event_type) or bucketed-partitioned table (based on event_time) to run queries as efficiently as possible

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

Output: Total_Revenue: 1211538.43

- **Without-Partitioned:**

```
select round(sum(price),2) as Total_Revenue
from cosmetic_data
where event_type='purchase' and event_time like '2019-10%' ;
```



***Before Optimization*:- Execution Time: 9.995 seconds**

- **With-Partitioned:**

```
select round(sum(price),2) as Total_Revenue
from dyn_cosmetic_data
where event_type='purchase' and event_time like '2019-10%';
```

```
hive> select round(sum(price),2) as Total_Revenue from dyn_cosmetic_data where event_type='purchase' and event_time like '2019-10%';
Query ID = hadoop_20210905102451_fb9aeaa3-3a0d-4b92-be69-6d250ebe72f8
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1630821305393_0022)

--------------------------------------------------------------------------------
        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: 13.13 s
--------------------------------------------------------------------------------
OK
total_revenue
1211538.43
Time taken: 14.106 seconds, Fetched: 1 row(s)
```

- **With Bucketed-Partitioned**:

```
select round(sum(price),2) as Total_Revenue
from buck_cosmetic_data
where event_type='purchase' and event_time like '2019-10%';
```

```
hive> select round(sum(price),2) as Total_Revenue from buck_cosmetic_data where event_type='purchase' and event_time like '2019-10%';
Query ID = hadoop_20210905102737_bc1c48f0-7473-40e8-bacf-a2a426b37a66
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1630821305393_0022)

--------------------------------------------------------------------------------
        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: 2.41 s
--------------------------------------------------------------------------------
OK
total_revenue
1211538.43
Time taken: 2.949 seconds, Fetched: 1 row(s)
```

***After Optimization*: - Execution Time: 2.949 seconds**

Optimization Note:

Here, execution time of the bucketed-partitioned (buck_cosmetic_data) table is reduced to 2.949 seconds whereas non-partitioned (cosmetic_data) table executed the same query in 9.995 seconds and partitioned table (dyn_cosmetic_data) executed in 14.106 seconds. Hence, performance improved after using optimization on this query.

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

Output:

total_sum	month
1211538.43	10
1531016.9	11

- **Without-Partitioned**:

```
select round(sum(price),2) as total_sum, month(event_time) as month
from cosmetic_data
where event_type='purchase'
group by month(event_time);
```

***Before Optimization*:- Execution Time: 11.681 seconds**

- **With-Partitioned**:

```
select round(sum(price),2) as total_sum, month(event_time) as month
from dyn_cosmetic_data
where event_type='purchase'
group by month(event_time);
```

```
hive> select round(sum(price),2) as total_sum, month(event_time) as month from dyn_cosmetic_data where event_type='purchase' group by month(event_time);
Query ID = hadoop_20210905103634_f1f7139d-45dc-4e04-8386-7424768f82cd
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1630821305393_0022)

----------------------------------------------------------------------------------------------
        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: 2.29 s
----------------------------------------------------------------------------------------------
OK
total_sum       month
1211538.43      10
1531016.9       11
Time taken: 2.784 seconds, Fetched: 2 row(s)
```

After Optimization:- Execution Time: 2.784 seconds

- **With-Bucketed-Partitioned:**

```
select round(sum(price),2) as total_sum, month(event_time) as month
from buck_cosmetic_data
where event_type='purchase'
group by month(event_time);
```

```
hive> select round(sum(price),2) as total_sum, month(event_time) as month from buck_cosmetic_data where event_type='purchase' group by month(event_time);
Query ID = hadoop_20210905104531_7e57264f-7c3f-4b4b-baef-ccde4ade46f6
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1630821305393_0022)

----------------------------------------------------------------------------------------------
        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: 1.97 s
----------------------------------------------------------------------------------------------
OK
total_sum       month
1211538.43      10
1531016.9       11
Time taken: 2.46 seconds, Fetched: 2 row(s)
```

After Optimization: - Execution Time: 2.46 seconds

Optimization Note:

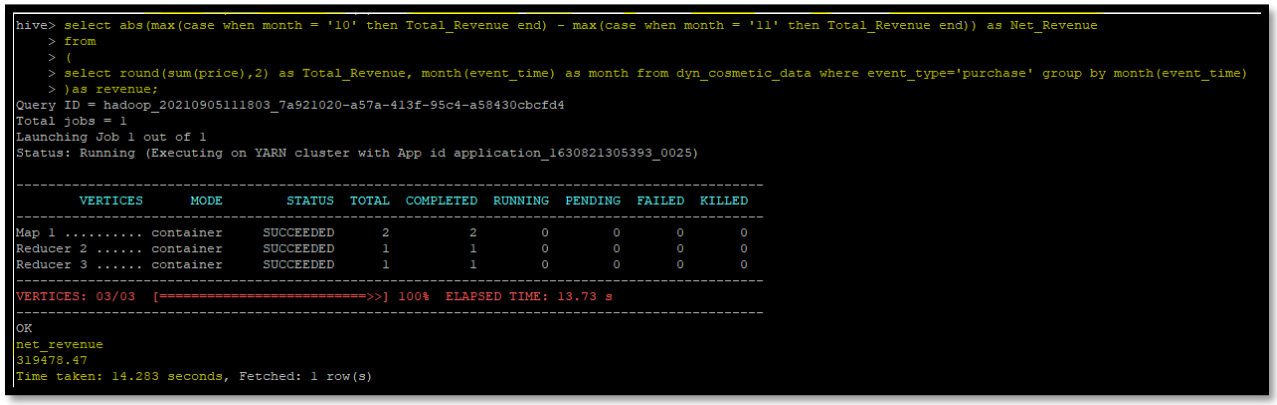
Here, execution time of the bucketed-partitioned (buck_cosmetic_data) table is reduced to 2.46 seconds whereas non-partitioned (cosmetic_data) table executed the same query in 11.681 seconds and partitioned table (dyn_cosmetic_data) executed in 2.784 seconds. Hence, performance improved after using optimization on this query.

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

Output: Net_Revenue: 319478.47

- **With-Partitioned:**

```
select abs(max(case when month = '10' then Total_Revenue end) - max(case when month = '11' then Total_Revenue end)) as Net_Revenue
from (
select round(sum(price),2) as Total_Revenue, month(event_time) as month
from dyn_cosmetic_data
where event_type='purchase'
group by month(event_time)
) as revenue;
```



```
hive> select abs(max(case when month = '10' then Total_Revenue end) - max(case when month = '11' then Total_Revenue end)) as Net_Revenue
    > from
    > (
    > select round(sum(price),2) as Total_Revenue, month(event_time) as month from dyn_cosmetic_data where event_type='purchase' group by month(event_time)
    > )as revenue;
Query ID = hadoop_20210905111803_7a921020-a57a-413f-95c4-a58430cbcfd4
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1630821305393_0025)

--------------------------------------------------------------------------------
        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
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 13.73 s
--------------------------------------------------------------------------------
OK
net_revenue
319478.47
Time taken: 14.283 seconds, Fetched: 1 row(s)
```

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

Output:

```
category_code
accessories.bag
appliances.environment.vacuum
appliances.personal.hair_cutter
sport.diving
apparel.glove
furniture.bathroom.bath
furniture.living_room.cabinet
stationery.cartrige
accessories.cosmetic_bag
appliances.environment.air_conditioner
furniture.living_room.chair
```

- **Without-Partitioned**:

```
select distinct category_code

from cosmetic_data

where category_code != '';
```

```
hive> select distinct category_code from cosmetic_data where category_code != '';
Query ID = hadoop_20210905114903_340137b0-f74d-45a7-9f83-fdf487ebf12d
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1630821305393_0025)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      8          8        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      4          4        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 7.15 s
----------------------------------------------------------------------------------------------
OK
category_code
accessories.bag
appliances.environment.vacuum
appliances.personal.hair_cutter
sport.diving
apparel.glove
furniture.bathroom.bath
furniture.living_room.cabinet
stationery.cartrige
accessories.cosmetic_bag
appliances.environment.air_conditioner
furniture.living_room.chair
Time taken: 7.561 seconds, Fetched: 11 row(s)
```

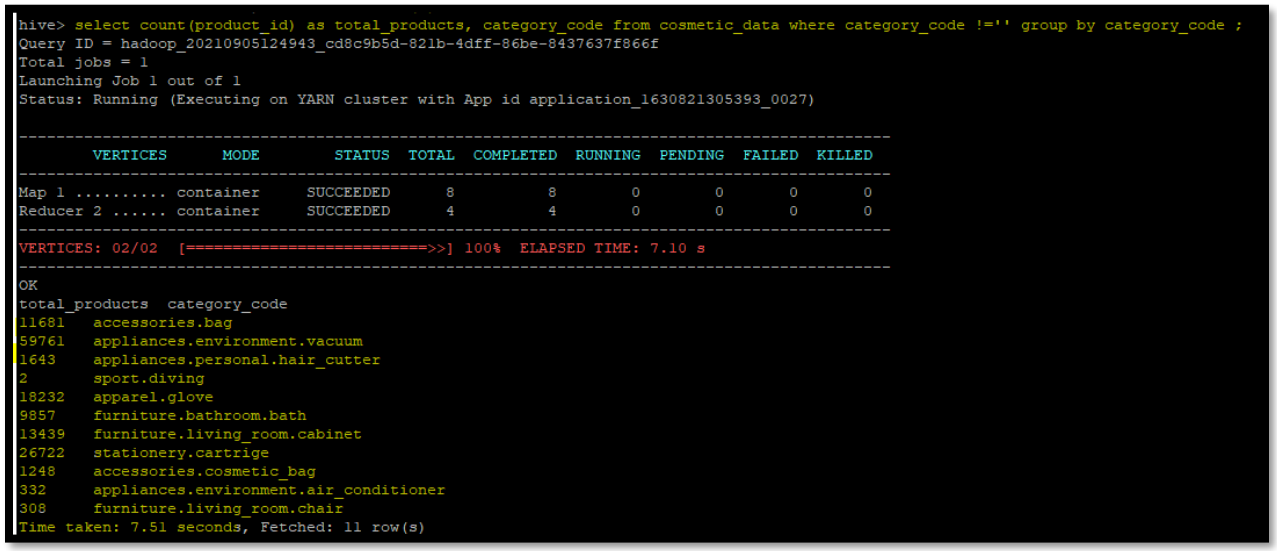
5) **Find the total number of products available under each category**.

Output:

total_products	category_code
11681	accessories.bag
59761	appliances.environment.vacuum
1643	appliances.personal.hair_cutter
2	sport.diving
18232	apparel.glove
9857	furniture.bathroom.bath
13439	furniture.living_room.cabinet
26722	stationery.cartrige
1248	accessories.cosmetic_bag
332	appliances.environment.air_conditioner
308	furniture.living_room.chair

- **<u>Without-Partitioned</u>:**

```
select count(product_id) as total_products, category_code
from cosmetic_data
where category_code!= ''
group by category_code ;
```



```
hive> select count(product_id) as total_products, category_code from cosmetic_data where category_code !='' group by category_code ;
Query ID = hadoop_20210905124943_cd8c9b5d-821b-4dff-86be-8437637f866f
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1630821305393_0027)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 ..... container     SUCCEEDED      8          8        0        0       0       0
Reducer 2 ..... container     SUCCEEDED      4          4        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 7.10 s
--------------------------------------------------------------------------------
OK
total_products  category_code
11681   accessories.bag
59761   appliances.environment.vacuum
1643    appliances.personal.hair_cutter
2       sport.diving
18232   apparel.glove
9857    furniture.bathroom.bath
13439   furniture.living_room.cabinet
26722   stationery.cartrige
1248    accessories.cosmetic_bag
332     appliances.environment.air_conditioner
308     furniture.living_room.chair
Time taken: 7.51 seconds, Fetched: 11 row(s)
```

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

Output: Brand: Runail

- **With-Partitioned:**

```
with max_sales as (
select brand, sum(price) as total_sales from dyn_cosmetic_data
where event_type='purchase' and brand!=''
group by brand
order by total_sales desc
)
select brand from max_sales limit 1;
```

```
hive> with max_sales as (
    > select brand, sum(price) as total_sales from dyn_cosmetic_data where event_type='purchase' and brand!=''
    > group by brand order by total_sales desc
    > )
    > select brand from max_sales limit 1;
Query ID = hadoop_20210905180414_022bcbc3-bc0f-4c6d-ba29-737568e694a2
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1630821305393_0030)

--------------------------------------------------------------------------------
        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
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 1.91 s
--------------------------------------------------------------------------------
OK
brand
runail
Time taken: 2.422 seconds, Fetched: 1 row(s)
```

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

Note: 152 brands increased their sales from October to November which are attached below in snapshot.

- **With-Partitioned:**

```
with sales_oct as (
select brand, round(sum(price),2) as oct_sales from dyn_cosmetic_data
where event_type='purchase' and brand!='' and month(event_time)='10'
group by brand
),
sales_nov as(
select brand, round(sum(price),2) as nov_sales from dyn_cosmetic_data
where event_type='purchase' and brand!='' and month(event_time)='11'
group by brand
)
select b.brand  from sales_oct a, sales_nov b
where a.brand = b.brand and round((nov_sales-oct_sales),2)>0;
```

```
hive> with sales_oct as (
    > select brand, round(sum(price),2) as oct_sales from dyn_cosmetic_data
    > where event_type='purchase' and brand!='' and month(event_time)='10' group by brand
    > ),
    > sales_nov as(
    > select brand, round(sum(price),2) as nov_sales from dyn_cosmetic_data
    > where event_type='purchase' and brand!='' and month(event_time)='11' group by brand
    > )
    > select b.brand
    > from sales_oct a, sales_nov b
    > where a.brand = b.brand and round((nov_sales-oct_sales),2)>0;
Query ID = hadoop_20210905182139_c0f71805-3eea-4440-8a8b-639f917e200f
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1630821305393_0030)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      2          2        0        0       0       0
Map 4 .......... container     SUCCEEDED      2          2        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
Reducer 5 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 05/05  [==========================>>] 100%  ELAPSED TIME: 4.72 s
----------------------------------------------------------------------------------------------
OK
b.brand
airnails
art-visage
artex
aura
balbcare
batiste
beautix
beauty-free
beautyblender
beauugreen
benovy
bioaqua
biore
blixz
bluesky
```

```
bodyton
bpw.style
browxenna
candy
carmex
chi
coifin
concept
cosima
cosmoprofi
cristalinas
cutrin
de.lux
deoproce
depilflax
dizao
domix
ecocraft
ecolab
egomania
elizavecca
ellips
elskin
enjoy
entity
eos
estel
estelare
f.o.x
farmavita
farmona
fedua
finish
fly
foamie
freedecor
freshbubble
gehwol
glysolid
godefroy
grace
grattol
greymy
happyfons
```

```
haruyama
igrobeauty
ingarden
inm
insight
irisk
italwax
jaguar
jas
jessnail
joico
kaaral
kamill
kapous
kaypro
keen
kerasys
kims
kinetics
kiss
kocostar
koelcia
koelf
konad
kosmekka
laboratorium
lador
ladykin
latinoil
levissime
levrana
lianail
likato
limoni
lovely
lowence
mane
marathon
markell
marutaka-foot
masura
matrix
mavala
metzger
```

```
milv
miskin
missha
moyou
nagaraku
nefertiti
neoleor
nirvel
nitrile
oniq
orly
osmo
ovale
plazan
polarus
profepil
profhenna
protokeratin
provoc
rasyan
refectocil
rosi
roubloff
runail
s.care
sanoto
severina
shary
shik
skinity
skinlite
smart
soleo
solomeya
sophin
staleks
strong
supertan
swarovski
tertio
treaclemoon
trind
uno
uskusi
```

```
veraclara
vilenta
yoko
yu-r
zeitun
Time taken: 5.587 seconds, Fetched: 152 row(s)
```

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.**

Output: Top10_user_ids

557790271

150318419

562167663

531900924

557850743

522130011

561592095

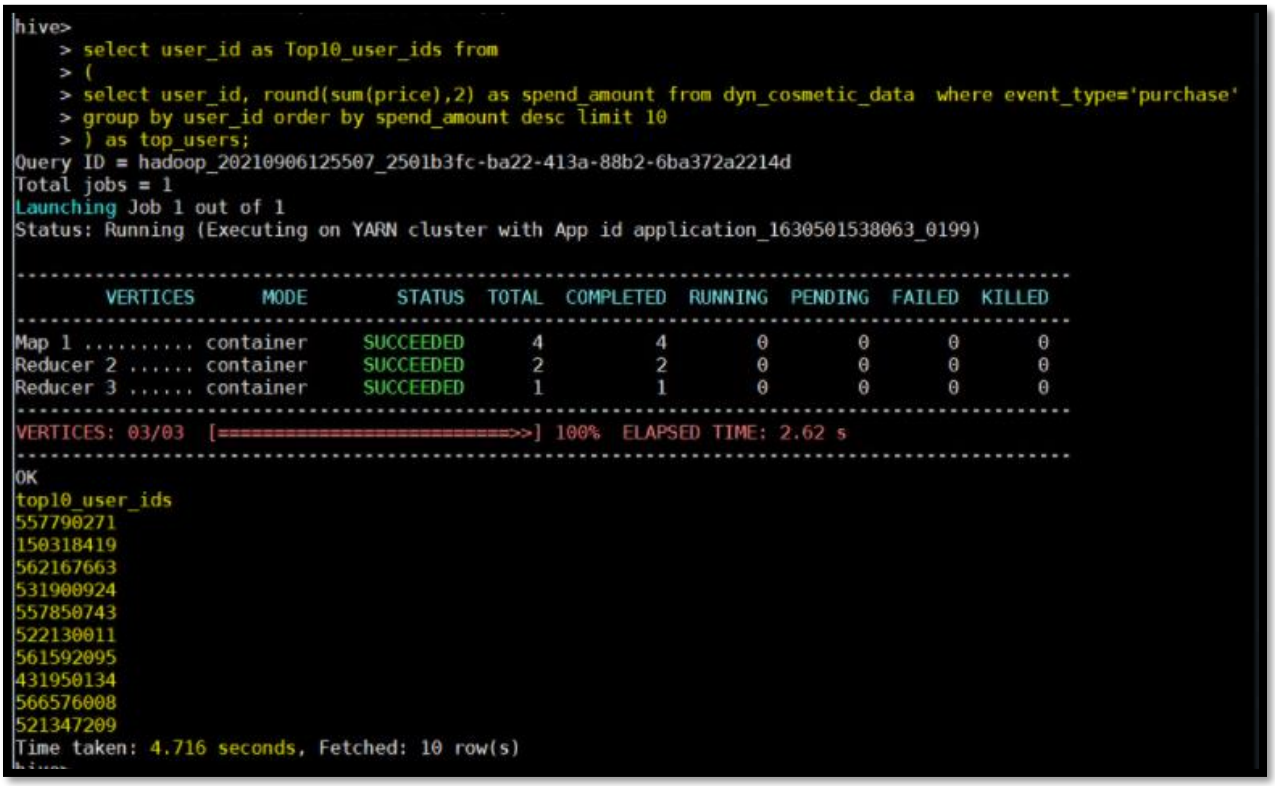
4319550134

566576008

521347209

- **With-Partitioned**:

```
select user_id as Top10_user_ids from
(
select user_id, round(sum(price),2) as spend_amount
from dyn_cosmetic_data
where event_type='purchase'
group by user_id
order by spend_amount desc limit 10
) as top_users;
```


hive>
> select user_id as Top10_user_ids from
> (
> select user_id, round(sum(price),2) as spend_amount from dyn_cosmetic_data where event_type='purchase'
> group by user_id order by spend_amount desc limit 10
>) as top_users;
Query ID = hadoop_20210906125507_2501b3fc-ba22-413a-88b2-6ba372a2214d
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1630501538063_0199)
VERTICES MODE STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED
Map 1 ..... container SUCCEEDED 4 4 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: 2.62 s
OK
top10_user_ids
557790271
150318419
562167663
531900924
557850743
522130011
561592095
431950134
566576008
521347209
Time taken: 4.716 seconds, Fetched: 10 row(s)