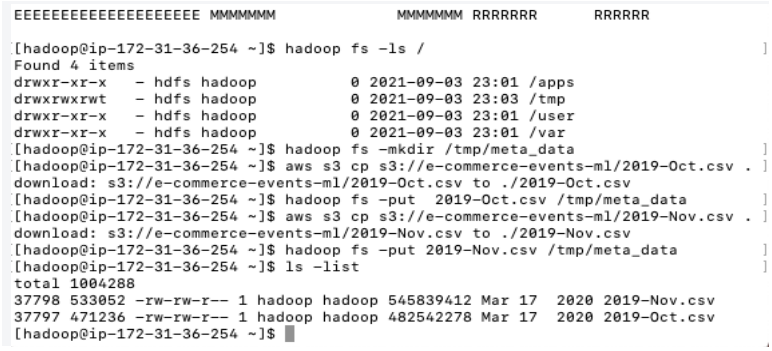
Cosmetics Store case study:

Dataset: We are working with public clickstream dataset of a cosmetics store.

The Implementation Phase 1:

Step 1: After launching an EMR cluster, Move the data from the S3 bucket into the HDFS.

```
hadoop fs -ls /
hadoop fs -mkdir /tmp/meta_data
aws s3 cp s3://e-commerce-events-ml/2019-Oct.csv .
hadoop fs -put  2019-Oct.csv /tmp/meta_data
aws s3 cp s3://e-commerce-events-ml/2019-Nov.csv .
hadoop fs -put 2019-Nov.csv /tmp/meta_data
ls -list
```

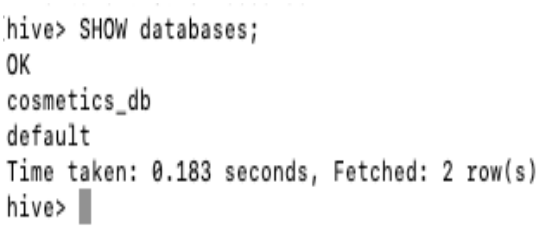


Here we are using CSVSerde with the default properties value for loading the dataset into a Hive table.

Phase 2:

Create a Database:

```
CREATE database if not exists cosmetics_db comment "This is database for
cosmetics store data";
DESCRIBE database extended cosmetics_db;
SHOW databases;
```



```
USE cosmetics_db;

DESCRIBE DATABASE cosmetics_db;
```

```
hive> DESCRIBE DATABASE cosmetics_db;
OK
cosmetics_db    This is database for cosmetics store data      hdfs://ip-172-31-36-254.ec2.internal:8020/user/hive/warehouse/cosme
tics_db.db   hadoop  USER
Time taken: 0.022 seconds, Fetched: 1 row(s)
hive>
```

```
DESCRIBE SCHEMA cosmetics_db;
```

```
hive>
    > DESCRIBE SCHEMA cosmetics_db;
OK
cosmetics_db    This is database for cosmetics store data      hdfs://ip-172-31-36-254.ec2.internal:8020/user/hive/warehouse/cosme
tics_db.db      hadoop  USER
Time taken: 0.023 seconds, Fetched: 1 row(s)
hive>
    >
```

Creating a Table:

```
CREATE EXTERNAL TABLE if not exists test_data
(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' = ',' )
STORED AS TEXTFILE
LOCATION '/tmp/meta_data/'
TBLPROPERTIES ("skip.header.line.count"="1");
```

Check the structure of the table:

```
Desc test_data;
DESCRIBE FORMATTED test_data; OR DESCRIBE EXTENDED test_data;
```

```
    > DESCRIBE FORMATTED test_data;
OK
# col_name              data_type               comment

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
Database:               cosmetics_db
Owner:                  hadoop
CreateTime:             Fri Sep 03 23:19:44 UTC 2021
LastAccessTime:         UNKNOWN
Retention:              0
Location:               hdfs://ip-172-31-36-254.ec2.internal:8020/tmp/meta_data
Table Type:             EXTERNAL_TABLE
Table Parameters:
        EXTERNAL                TRUE
        numFiles                2
        skip.header.line.count  1
        totalSize               1028381690
        transient_lastDdlTime   1630711184

# Storage Information
SerDe Library:          org.apache.hadoop.hive.serde2.OpenCSVSerde
InputFormat:            org.apache.hadoop.mapred.TextInputFormat
OutputFormat:           org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat
Compressed:             No
Num Buckets:            -1
Bucket Columns:         []
Sort Columns:           []
Storage Desc Params:
        separatorChar           ,
        serialization.format    1
Time taken: 0.157 seconds, Fetched: 38 row(s)
hive>
```

Here if we observe all the data types are in ‘string’by default . This is one of the limitations in serde. We need to cast the data types into desired ones.

```
CREATE EXTERNAL TABLE if not exists store_data (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);
```

```
insert into store_data select
cast(from_unixtime(unix_timestamp(event_time,'yyyy-MM-dd HH:mm:ss Z'),'yyyy-
MM-dd HH:mm:ss') as timestamp) as event_time,event_type,
product_id,category_id,category_code,brand,cast(price as float) as
price,cast(user_id as bigint) as user_id,user_session from test_data;
```

```
Describe store_data;
```

```
hive> describe store_data;
OK
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
Time taken: 0.067 seconds, Fetched: 9 row(s)
hive>
```

```
SHOW tables in cosmetics_db;
```

```
hive> SHOW tables in cosmetics_db;
OK
store_data
test_data
Time taken: 0.059 seconds, Fetched: 2 row(s)
hive>
```

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

```
select event_type,count(event_type) as count from store_data group by
event_type;
```

```
hive> select event_type,count(event_type) as count from store_data group by event_type;
Query ID = hadoop_20210904045219_f88da5dd-f191-4d3c-83e7-c8fe6c25f0f0
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1630710139025_0018)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED     12         12        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      4          4        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 26.20 s
--------------------------------------------------------------------------------
OK
event_type      count
view    3938296
purchase        568041
cart    2544192
remove_from_cart        1687591
Time taken: 26.765 seconds, Fetched: 4 row(s)
hive>
```

‘View’ count is more compared to ‘purchase’ count.

Queries:

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

```
select sum(price) as oct_revenue from store_table where
month(event_time)='10' and event_type='purchase';
```

```
hive> select sum(price) as oct_revenue from store_data  where month(event_time)='10' and event_type='purchase';
Query ID = hadoop_20210904003506_c8259347-d9e5-483d-af67-0b6521f4bb3a
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_1630710139025_0005)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED     12         12        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 31.89 s
--------------------------------------------------------------------------------
OK
oct_revenue
1211538.4295325726
Time taken: 41.9 seconds, Fetched: 1 row(s)
hive>
```

One of the optimization technique to increase the performance is partitioning we can apply that here and compare the execution time.

Static partitioning:

```
CREATE EXTERNAL TABLE if not exists purchase_data (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)
row format delimited fields terminated by "," lines terminated by "\n" stored
as textfile ;
insert into table purchase_data partition(event_type = 'purchase') select
event_time , product_id , category_id , category_code , brand , price ,
user_id , user_session from store_data where event_type = 'purchase';
```

```
hive> show partitions purchase_data;
OK
partition
event_type=purchase
```

```
hive> show tables;
OK
purchase_data
store_data
test_data
Time taken: 0.038 seconds, Fetched: 3 row(s)
hive>
```

```
select * from purchase_data limit 5;
```

```
hive> select * from purchase_data limit 5;
OK
purchase_data.event_time        purchase_data.product_id        purchase_data.category_id       purchase_data.category_code     pur
chase_data.brand        purchase_data.price     purchase_data.user_id   purchase_data.user_session      purchase_data.event_type
2019-10-06 16:17:54     5847742 1924049110428549877     NULL    oniq    7.14    491806426       45f1051b-7f66-4e83-aa77-6b243b0055e
8       purchase
2019-10-06 16:17:54     5847720 1924049110428549877     NULL    oniq    7.14    491806426       45f1051b-7f66-4e83-aa77-6b243b0055e
8       purchase
2019-10-06 16:17:54     5840434 1487580006484804506     NULL    oniq    18.25   491806426       45f1051b-7f66-4e83-aa77-6b243b0055e
8       purchase
2019-10-06 16:19:14     5838726 1487580007365608384     NULL    bluesky 3.97    463428510       fc761144-8d07-4f83-91e2-02566046ac2
7       purchase
2019-10-06 16:19:14     5834653 1487580008145748965     NULL    pole    1.65    463428510       fc761144-8d07-4f83-91e2-02566046ac2
7       purchase
Time taken: 0.193 seconds, Fetched: 5 row(s)
hive>
```

```
select sum(price) as oct_revenue from purchase_data where month(event_time)
='10';
```

```
hive> select sum(price) as oct_revenue from purchase_data  where month(event_time)='10';
Query ID = hadoop_20210904035138_af8afa9e-c053-460b-b60f-3e2b61cad2fe
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1630710139025_0014)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED     12         12        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 46.18 s
--------------------------------------------------------------------------------
OK
1211538.4295325726
Time taken: 47.015 seconds, Fetched: 1 row(s)
hive>
```

Dynamic partition:

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

```
CREATE EXTERNAL TABLE if not exists mnth_dyn_data (event_type
string,product_id string,category_id string,category_code string,brand
string,price float,user_id bigint,user_session string)
PARTITIONed BY (event_time string)
row format delimited fields terminated by "," lines terminated by "\n" stored
as textfile;
```

```
insert into mnth_dyn_data partition(event_time) select event_type,
product_id, category_id , category_code , brand , price , user_id ,
user_session,SUBSTR(event_time,6,2) from store_data;
```

```
show tables;
```

```
hive>
    > show tables;
OK
mnth_dyn_data
purchase_data
store_data
test_data
Time taken: 0.025 seconds, Fetched: 4 row(s)
hive>
```

```
Show partitions mnth_dyn_data;
```

```
hive> show partitions mnth_dyn_data;
OK
partition
event_time=10
event_time=11
```

```
select sum(price) as oct_revnue from mnth_dyn_data where event_time=10
and event_type='purchase';
```

```
hive> select sum(price) as oct_revnue from mnth_dyn_data where event_time=10 and event_type='purchase';
Query ID = hadoop_20210904013307_d2411a60-ddda-4523-9ded-317e2e00e24b
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1630710139025_0007)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      8          8        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 24.50 s
--------------------------------------------------------------------------------
OK
oct_revnue
1211538.4295325726
Time taken: 25.081 seconds, Fetched: 1 row(s)
hive>
```

After Dynamic partition the execution time reduced almost 50%.

Bucketing:

```
set hive.enforce.bucketing = true;
set hive.exec.max.dynamic.partitions.pernode=1000;
```

```
CREATE EXTERNAL TABLE if not exists bucket_data (event_type string,product_id
string,category_id string,category_code string,brand string,price
float,user_id bigint,user_session string)
PARTITIONed BY(event_time string)
Clustered by(event_type) into 3 buckets
row format delimited fields terminated by "," lines terminated by "\n" stored
as textfile ;

insert into bucket_data partition(event_time) select event_type, product_id,
category_id , category_code , brand , price , user_id ,
user_session,SUBSTR(event_time,6,2) from store_data;

show tables;
```

```
hive> show tables;
OK
bucket_data
mnth_dyn_data
purchase_data
store_data
test_data
Time taken: 0.022 seconds, Fetched: 5 row(s)
hive>
```

```
[hadoop@ip-172-31-36-254 ~]$ hadoop fs -ls /user/hive/warehouse/cosmetics_db.db/bucket_data/event_time=10/
Found 3 items
-rwxrwxrwt   1 hadoop hadoop  111941154 2021-09-04 02:43 /user/hive/warehouse/cosmetics_db.db/bucket_data/event_time=10/000000_0
-rwxrwxrwt   1 hadoop hadoop  193438498 2021-09-04 02:43 /user/hive/warehouse/cosmetics_db.db/bucket_data/event_time=10/000001_0
-rwxrwxrwt   1 hadoop hadoop   78349533 2021-09-04 02:42 /user/hive/warehouse/cosmetics_db.db/bucket_data/event_time=10/000002_0
[hadoop@ip-172-31-36-254 ~]$ hive;
```

```
select sum(price) from bucket_data where event_type='purchase' and
event_time=10;
```

```
hive> select sum(price) from bucket_data where event_type='purchase' and event_time=10;
Query ID = hadoop_20210904031628_ee970727-eabc-4779-8da4-eddf9fa83c21
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_1630710139025_0012)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 ..... container     SUCCEEDED      8          8        0        0       0       0
Reducer 2 ..... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 24.35 s
--------------------------------------------------------------------------------
OK
1211538.4295325726
Time taken: 35.478 seconds, Fetched: 1 row(s)
hive>
```

We can create direct static partition table from 2019_Oct.csv for this Query.

```
create external table if not exists oct_data_1 ( 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) row format delimited fields terminated by "," lines terminated by "\n" stored as textfile ;
```

```
insert into table oct_data_1 partition(event_type = 'purchase') select event_time , product_id , category_id , category_code , brand , price , user_id , user_session from store_table where event_type = 'purchase';
```

```
OK
bucket_data
mnth_dyn_data
oct_data_1
purchase_data
store_data
test_data
Time taken: 0.021 seconds, Fetched: 6 row(s)
hive>
```

```
hive> select * from oct_data_1 limit 5;
OK
oct_data_1.event_time   oct_data_1.product_id   oct_data_1.category_id  oct_data_1.category_code        oct_data_1.brand        oct
_data_1.price   oct_data_1.user_id      oct_data_1.user_session oct_data_1.event_type
2019-10-06 16:17:54     5847742 1924049110428549877     NULL    oniq    7.14    491806426       45f1051b-7f66-4e83-aa77-6b243b0055e
8       purchase
2019-10-06 16:17:54     5847720 1924049110428549877     NULL    oniq    7.14    491806426       45f1051b-7f66-4e83-aa77-6b243b0055e
8       purchase
2019-10-06 16:17:54     5840434 1487580006484804506     NULL    oniq    18.25   491806426       45f1051b-7f66-4e83-aa77-6b243b0055e
8       purchase
2019-10-06 16:19:14     5838726 1487580007365608384     NULL    bluesky 3.97    463428510       fc761144-8d07-4f83-91e2-02566046ac2
7       purchase
2019-10-06 16:19:14     5834653 1487580008145748965     NULL    pole    1.65    463428510       fc761144-8d07-4f83-91e2-02566046ac2
7       purchase
Time taken: 0.144 seconds, Fetched: 5 row(s)
```

```
select sum(price) as oct_revenue from oct_data_1 where month(event_time)= 10;
```

```
hive> select sum(price) as oct_revenue from oct_data_1 where month(event_time)= 10;
Query ID = hadoop_20210903044105_c5afeffb-8625-4032-808f-f8ff260e789d
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1630627700358_0016)

--------------------------------------------------------------------------------
        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: 12.88 s
--------------------------------------------------------------------------------
OK
1211538.4295325726
Time taken: 13.803 seconds, Fetched: 1 row(s)
hive>
```

Here the execution time changed from 43 sec to 13 sec.

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

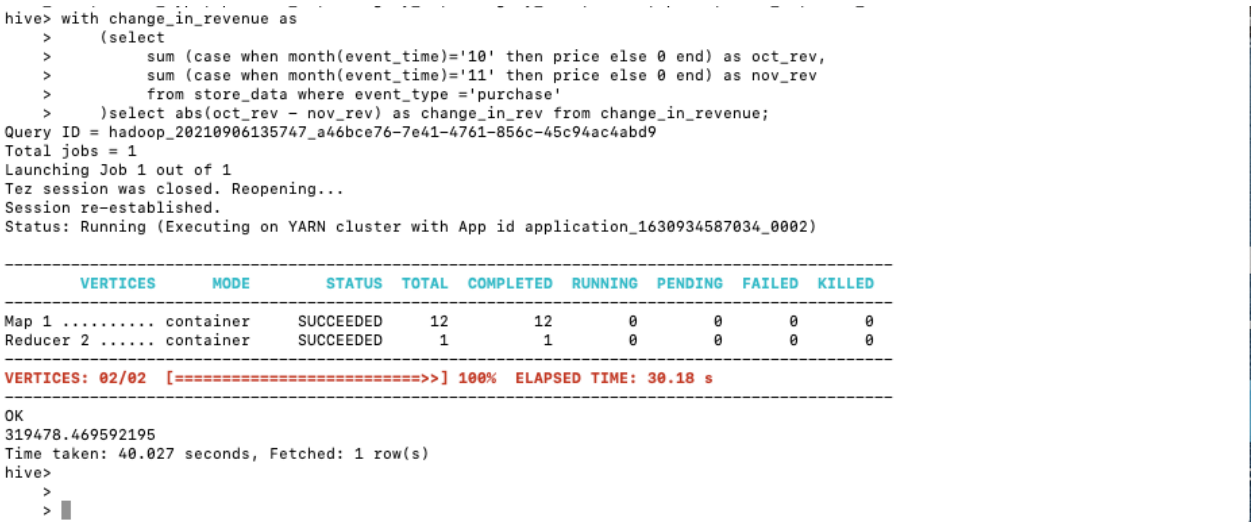
```
select month(event_time) as month, sum(price)as revenue from store_data where
event_type='purchase' group by month(event_time);
```

```
e);
Query ID = hadoop_20210904032637_7e986f6c-4789-424c-ad0d-3033679b634d
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1630710139025_0012)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED     12         12        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      2          2        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 26.44 s
--------------------------------------------------------------------------------
OK
month   revenue
10      1211538.4295325726
11      1531016.8991247676
Time taken: 27.067 seconds, Fetched: 2 row(s)
hive>
```

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

```
with change_in_revenue as
    (select
         sum (case when month(event_time)='10' then price else 0 end) as
oct_rev,
         sum (case when month(event_time)='11' then price else 0 end) as
nov_rev
         from store_data where event_type ='purchase'
    )select abs(oct_rev - nov_rev) as change_in_rev from change_in_revenue;
```



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

```
select distinct category_code as product_category from store_data
where category_code is not null;
```

we have some empty values in our dataset. Based on our requirements while doing queries it is better to change them as NULL values.

For example in brand column we have empty values. Like below.

```
select distinct brand from store_data;
```

```
strong
thuya
uno
uskusi
yoko
zab
zinger

airnails
andrea
balbcare
beauugreen
benovy
bergamo
bosnic
cnd
```

Change empty values to NULL values:

```
Alter table store_data
set tblproperties('serialization.null.format'='');
```

```
hive>
    > Alter table store_data
    > set tblproperties('serialization.null.format'='');
OK
Time taken: 0.092 seconds
hive> select distinct category_code as product_category from store_data where category_code is not null;
Query ID = hadoop_20210904033225_f0f1364e-caca-485f-a719-5aa1bedcf1ef
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1630710139025_0012)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED     12         12        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      4          4        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 26.80 s
--------------------------------------------------------------------------------
OK
product_category
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: 27.4 seconds, Fetched: 11 row(s)
hive>
```

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

```
Select category_code as category,count(product_id) as products from
store_data where category_code is not null group by category_code;
```

```
hive> Select category_code as category,count(product_id) as products from store_data where category_code is not null group by category_code;
FAILED: SemanticException [Error 10001]: Line 1:68 Table not found 'store_data'
hive> Select category_code as category,count(product_id) as products from store_table where category_code is not null group by category_code;
Query ID = hadoop_20210903014344_ff966578-64ef-4409-a160-043fb222b3ff
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1630627700358_0006)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 ..... container     SUCCEEDED     12        12        0        0       0       0
Reducer 2 ..... container     SUCCEEDED      4          4        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 24.96 s
--------------------------------------------------------------------------------
OK
accessories.bag 11681
appliances.environment.vacuum   59761
appliances.personal.hair_cutter 1643
sport.diving    2
apparel.glove   18232
furniture.bathroom.bath 9857
furniture.living_room.cabinet   13439
stationery.cartrige     26722
accessories.cosmetic_bag        1248
appliances.environment.air_conditioner  332
furniture.living_room.chair     308
Time taken: 25.468 seconds, Fetched: 11 row(s)
hive>
```

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

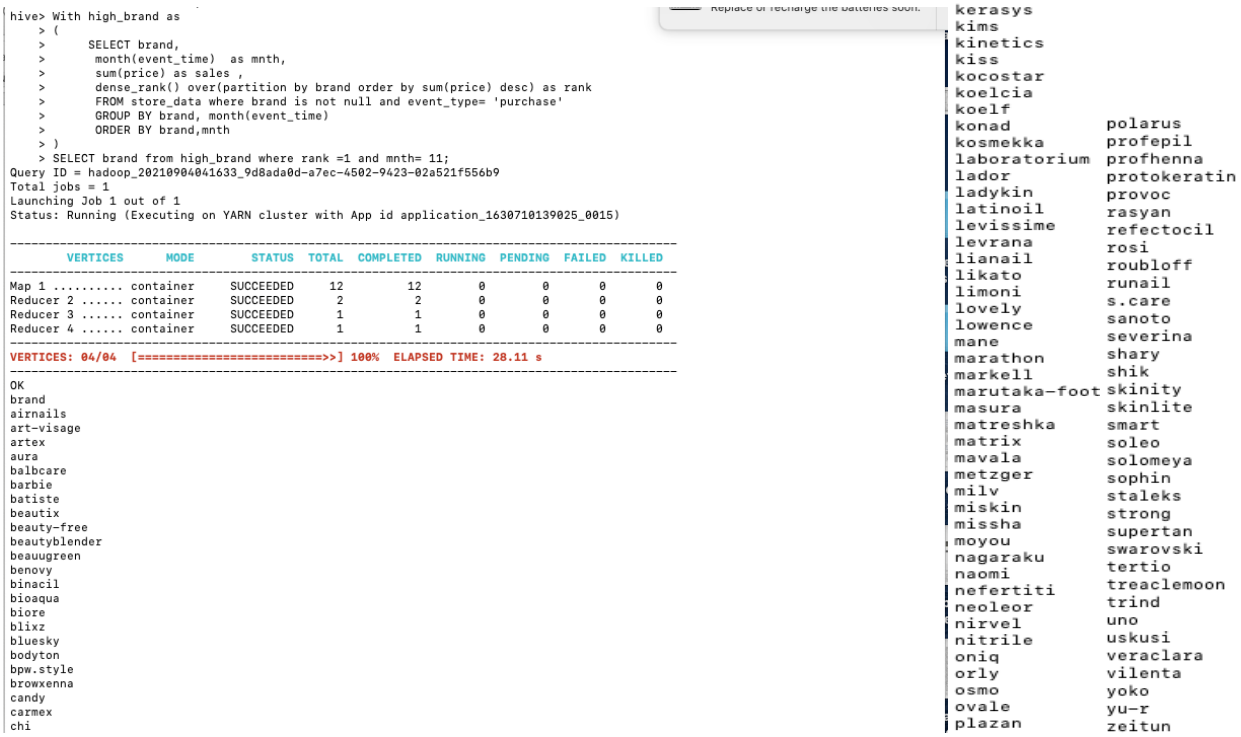
```
Select brand, round(sum(price),2) as max_sales from store_data where brand is
not null and event_type = 'purchase' group by brand order by max_sales desc
limit 1;
```

```
hive> Select brand,round(sum(price),2) as max_sales from store_data where brand is not null and event_type = 'purchase' group by br
and order by max_sales desc limit 1;
Query ID = hadoop_20210904041330_5b0cc43e-4fe5-42e8-86c9-67ae956a2ab7
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1630710139025_0015)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED     12         12        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: 27.19 s
--------------------------------------------------------------------------------
OK
brand   max_sales
runail  148297.94
Time taken: 28.022 seconds, Fetched: 1 row(s)
hive>
```

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

```
With high_brand as
(
     SELECT brand,
      month(event_time)  as mnth,
      sum(price) as sales ,
      dense_rank() over(partition by brand order by sum(price) desc) as rank
      FROM store_data where brand is not null and event_type= 'purchase'
      GROUP BY brand, month(event_time)
      ORDER BY brand,mnth
)
SELECT brand from high_brand where rank =1 and mnth= 11;
```



```
hive> With high_brand as
    > (
    >      SELECT brand,
    >       month(event_time)  as mnth,
    >       sum(price) as sales ,
    >       dense_rank() over(partition by brand order by sum(price) desc) as rank
    >       FROM store_data where brand is not null and event_type= 'purchase'
    >       GROUP BY brand, month(event_time)
    >       ORDER BY brand,mnth
    > )
    > SELECT brand from high_brand where rank =1 and mnth= 11;
Query ID = hadoop_20210904041633_9d8ada0d-a7ec-4502-9423-02a521f556b9
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1630710139025_0015)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED     12         12        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      2          2        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
Reducer 4 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 04/04  [==========================>>] 100%  ELAPSED TIME: 28.11 s
--------------------------------------------------------------------------------
OK
brand
airnails
art-visage
artex
aura
balbcare
barbie
batiste
beautix
beauty-free
beautyblender
beauugreen
benovy
binacil
bioaqua
biore
blixz
bluesky
bodyton
bpw.style
browxenna
candy
carmex
chi
juno
kaaral
kamill
kapous
kares
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
matreshka
matrix
mavala
metzger
milv
miskin
missha
moyou
nagaraku
naomi
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: 28.728 seconds, Fetched: 160 row(s)
```

```
With high_brand as
(
     SELECT brand,
      month(event_time)  as mnth,
      sum(price) as sales ,
      dense_rank() over(partition by brand order by sum(price) desc) as rank
      FROM oct_data_1 where brand is not null and event_type= 'purchase'
      GROUP BY brand, month(event_time)
      ORDER BY brand,mnth
)
SELECT brand from high_brand where rank =1 and mnth= 11;
```

```
hive> With high_brand as
    > (
    >      SELECT brand,
    >       month(event_time)  as mnth,
    >       sum(price) as sales ,
    >       dense_rank() over(partition by brand order by sum(price) desc) as rank
    >       FROM oct_data_1 where brand is not null and event_type= 'purchase'
    >       GROUP BY brand, month(event_time)
    >       ORDER BY brand,mnth
    > )
    > SELECT brand from high_brand where rank =1 and mnth= 11;
Query ID = hadoop_20210903044906_722d8e13-0219-42e0-b7e2-a83ab027cab3
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_1630627700358_0017)

--------------------------------------------------------------------------------
        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
Reducer 4 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 04/04  [==========================>>] 100%  ELAPSED TIME: 14.61 s
--------------------------------------------------------------------------------
OK
airnails
art-visage
artex
aura
balbcare
barbie
batiste
beautix
beauty-free
beautyblender
beauugreen
benovy
binacil
bioaqua
biore
blixz
bluesky
bodyton
bpw.style
browxenna
candy
carmex
chi
coifin
concept
cosima
cosmoprofi
cristalinas
cutrin
de.lux
deoproce
depilflax
dewal
dizao
domix
ecocraft
ecolab
egomania
elizavecca
ellips
elskin
enjoy
entity
eos
estel
estelare
f.o.x
lianail
likato
limoni
lovely
lowence
mane
marathon
markell
marutaka-foot
masura
matreshka
matrix
mavala
metzger
milv
miskin
missha
moyou
nagaraku
naomi
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: 23.158 seconds, Fetched: 160 row(s)
hive>
```

```
estelare
f.o.x
farmavita
farmona
fedua
finish
fly
foamie
freedecor
freshbubble
gehwol
glysolid
godefroy
grace
grattol
greymy
happyfons
haruyama
helloganic
igrobeauty
ingarden
inm
insight
irisk
italwax
jaguar
jas
jessnail
joico
juno
kaaral
kamill
kapous
kares
kaypro
keen
kerasys
kims
kinetics
kiss
kocostar
koelcia
koelf
konad
kosmekka
laboratorium
lador
ladykin
latinoil
levissime
```

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

```
select user_id,sum(price) as purchase, dense_rank() over( order by sum(price) desc) as rank from store_data where event_type= 'purchase' group by user_id limit 10;
```

```
hive>
    > select user_id,sum(price) as purchase, dense_rank() over( order by sum(price) desc) as rank from store_data where event_type=
 'purchase' group by user_id limit 10;
Query ID = hadoop_20210904041940_b7a61a31-eadf-4ebf-9858-b439e0fec5ef
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1630710139025_0015)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED     12         12        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: 28.21 s
--------------------------------------------------------------------------------
OK
user_id purchase        rank
557790271       2715.8699957430363      1
150318419       1645.970008611679       2
562167663       1352.8499938696623      3
531900924       1329.4499949514866      4
557850743       1295.4800310581923      5
522130011       1185.3899966478348      6
561592095       1109.700007289648       7
431950134       1097.5900000333786      8
566576008       1056.3600097894669      9
521347209       1040.9099964797497      10
Time taken: 28.897 seconds, Fetched: 10 row(s)
hive>
```

```
select user_id,sum(price) as purchase, dense_rank() over( order by sum(price) desc) as rank from oct_data_1 group by user_id limit 10;
```

```
hive> select user_id,sum(price) as purchase, dense_rank() over( order by sum(price) desc) as rank from oct_data_1 group by user_id limit 10;
Query ID = hadoop_20210903051309_1e60c5dd-d8ed-4cb6-a6c8-476a0c8ebd7c
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1630627700358_0018)

--------------------------------------------------------------------------------
        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: 14.93 s
--------------------------------------------------------------------------------
OK
557790271       2715.8699957430363      1
150318419       1645.970008611679       2
562167663       1352.8499938696623      3
531900924       1329.4499949514866      4
557850743       1295.4800310581923      5
522130011       1185.3899966478348      6
561592095       1109.700007289648       7
431950134       1097.5900000333786      8
566576008       1056.3600097894669      9
521347209       1040.9099964797497      10
Time taken: 15.457 seconds, Fetched: 10 row(s)
hive>
```

Here also we used partition table oct_table_1 which we created before the execution time reduced almost 50%.

Observations:

1. The performance wise Partition is effective for low volume data. In our data set also we observed that performance rate increased when we use partitioning.
2. For larger data creating a bucketing gives you 2-3x better query performance than a non-bucket table.

Insights:

3. Depend on the data views and cart event type are more compared to purchase ones.
4. The total revenue is high in November month than October month.
5. Highest number of products available under appliances. environment.vaccume category.
6. Runail brand has highest sales compared with other brands.
7. Over all 43% brands increased their sales from October to November.
8. The user_id: 557790271 spent most in two months.