

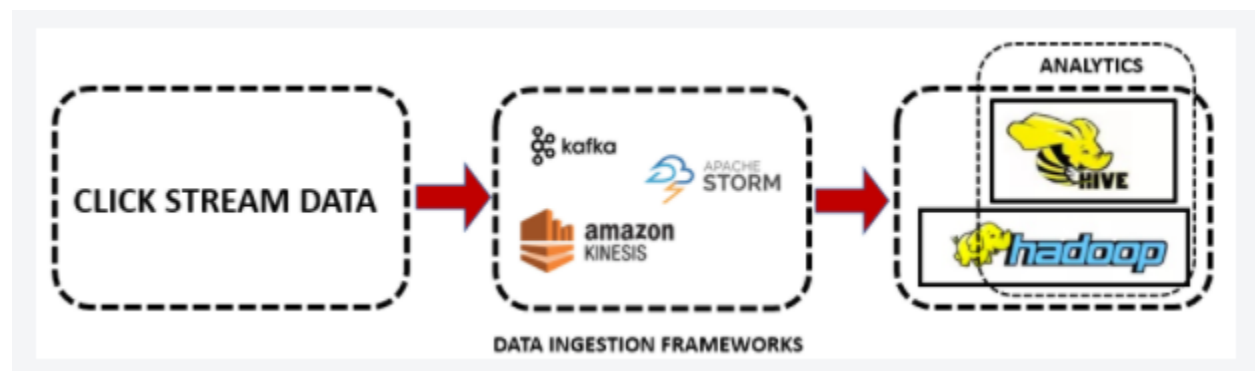
Public Clickstream Data Analysis

Manaswi Kamila

Problem Statement

With online sales gaining popularity, tech companies are exploring ways to improve their sales by analyzing customer behavior and gaining insights about product trends. Furthermore, the websites make it easier for customers to find the products they require without much scavenging.

The objective of this case study is to extract data and gather insights from a real-life public clickstream dataset of a cosmetics store for the months of October and November 2019 which generally data engineers come up within an e-retail company by executing queries using Hive Query Language (HQL).



Clickstream data

Data which is collected by tracking our clicks on websites and searching for patterns within them. E-commerce companies make use of the data to give product recommendations.

Datasets provided

- 2019-Oct.csv - <https://e-commerce-events-ml.s3.amazonaws.com/2019-Oct.csv>
- 2019-Nov.csv - <https://e-commerce-events-ml.s3.amazonaws.com/2019-Nov.csv>
- Attribute description – Excel file which contains attribute details

Implementation

Task 1: Importing the data from S3 to HDFS

Launched an EMR 5.29.0 cluster that utilizes the Hive services.

Created a HDFS directory “cosmetics_store” and imported data from public S3 bucket to HDFS directory using distcp command.

```
hadoop fs -mkdir /cosmetics_store
```

```
hadoop distcp s3n://e-commerce-events-ml/* /cosmetics_store
```

```
hadoop fs -ls /cosmetics_store
```

```
login as: hadoop
Authenticating with public key "imported-openssh-key"

  _ | _ | _ )
 _ | ( _ | _ /   Amazon Linux AMI

https://aws.amazon.com/amazon-linux-ami/2018.03-release-notes/
68 package(s) needed for security, out of 107 available
Run "sudo yum update" to apply all updates.

EEEEEEEEEEEEEEEEEEEE MMMMMMM      MMMMMMM RRRRRRRRRRRRRRR
E::::::::::::::::::::E M::::::::M      M::::::::M R:::::::::R
EE::::::::EEEEEEEE::E M::::::::M      M::::::::M R::::RRRRR:::R
E::::E      EEEEE M::::::::M      M::::::::M RR::::R      R::::R
E::::E      M::::::::M M:::M M:::M M:::M M:::M R:::R      R:::R
E::::EEEEEEEEEE M:::M M:::M M:::M M:::M M:::M R::RRRRR:::R
E:::::::::::::E M:::M M:::M M:::M M:::M M:::M R:::::::::RR
E::::EEEEEEEEEE M:::M M:::M M:::M M:::M R::RRRRR:::R
E::::E      M:::M M:::M M:::M M:::M R:::R      R:::R
E::::E      EEEEE M:::M M:::M M:::M M:::M R:::R      R:::R
EE::::::::EEEEEE::E M:::M M:::M M:::M M:::M R:::R      R:::R
E:::::::::::::E M:::M M:::M M:::M RR::::R      R:::R
EEEEEEEEEEEEEEEEEEEE MMMMMMM      MMMMMMM RRRRRR      RRRRRR

[hadoop@ip-172-31-18-44 ~]$ hadoop fs -mkdir /cosmetics_store
[hadoop@ip-172-31-18-44 ~]$
```

```
distcp counters
Bandwidth in Bbytes=31719532
Bytes Copied=1028381690
Bytes Expected=1028381690
Files Copied=2
[hadoop@ip-10-21-80-62 ~]$ hadoop fs -ls /cosmetics_store
found 2 items
-rw-r--r-- 2 hadoop hadoop 545839412 2021-09-06 16:00 /cosmetics_store/2019-Nov.csv
-rw-r--r-- 2 hadoop hadoop 482542278 2021-09-06 16:00 /cosmetics_store/2019-Oct.csv
[hadoop@ip-10-21-80-62 ~]$
```

Task 2: Creating database and Hive tables

Logged in to Hive CLI and created database “cosmetics_clickstream_data”

```
hadoop@ip-172-31-18-44 ~]$ hive
Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j2.properties Async: false
hive> create schema cosmetics_clickstream_data;
OK
Time taken: 0.901 seconds
hive> show databases;
OK
cosmetics_clickstream_data
default
Time taken: 0.163 seconds, Fetched: 2 row(s)
hive>
```

```
hive>
> use cosmetics_clickstream_data;
OK
Time taken: 0.021 seconds
hive> show tables;
OK
Time taken: 0.03 seconds
hive> describe database cosmetics_clickstream_data;
OK
cosmetics_clickstream_data      hdfs://ip-172-31-18-44.ec2.internal:8020/user/hive/warehouse/cosmetics_clickstream_data.db
Time taken: 0.025 seconds, Fetched: 1 row(s)
hive>
```

Created a table “OctNov2019_data” using CSV Serde and loaded data from the CSV files on HDFS into the tables.

```
hive> create external table OctNov2019_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" = ",", "quoteChar" = "\"")
> TBLPROPERTIES ("skip.header.line.count"="1");
OK
Time taken: 0.921 seconds
hive>
> LOAD DATA INPATH '/cosmetics_store' INTO table OctNov2019_data;
Loading data to table cosmetics_clickstream_data.octnov2019_data
OK
Time taken: 7.506 seconds
```

```

> describe OctNov2019_data;
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: 0.453 seconds, Fetched: 9 row(s)

```

Created tables “cosmetic_OctNov2019_data”, “part_cosmetic_OctNov2019_data”, “bucket_part_cosmetic_OctNov2019_data”

- Derived a new column, "month", by extracting month from event_time column.
- Handled timestamp data by using SUBSTR function (to_date function caused loss of data – time of transaction was lost)

```

hive> create table cosmetic_OctNov2019_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,
  > month int)
  > ROW FORMAT DELIMITED
  > FIELDS TERMINATED BY ',' ;
OK
Time taken: 1.25 seconds

```

To create dynamic partitioned and bucketed tables, set the below properties in hive CLI.

```

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

```

```

hive>
> set hive.exec.dynamic.partition=true;
hive> set hive.exec.dynamic.partition.mode=nonstrict;
hive> set hive.enforce.bucketing=true;
hive>
> create table part_cosmetic_OctNov2019_data(
> event_time timestamp,
> product_id string,
> category_id string,
> category_code string,
> brand string,
> price float,
> user_id bigint,
> user_session string,
> month int)
> PARTITIONED BY (event_type string)
> ROW FORMAT DELIMITED
> FIELDS TERMINATED BY ',' ;
OK
Time taken: 4.593 seconds
hive>
> create external table bucket_part_cosmetic_OctNov2019_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 (month int,event_type string )
> CLUSTERED BY (brand) INTO 20 BUCKETS
> ROW FORMAT DELIMITED
> FIELDS TERMINATED BY ',' ;
OK
Time taken: 1.008 seconds
hive>

```

Inserted records from OctNov2019_data and cosmetic_OctNov2019_data tables.


```

hive>
> INSERT INTO cosmetic_OctNov2019_data SELECT substr(event_time, 1, 20), event_type,
> product_id, category_id, category_code, brand, price, user_id, user_session, month(event_time) FROM OctNov2019_data ;
Query ID = hadoop_20210906102311_4cf6e30b-608f-455c-bf33-fd1f954b67b2
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1630501538063_0183)

-----
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: 68.79 s
-----
Loading data to table cosmetics_clickstream_data.cosmetic_octnov2019_data
K
Time taken: 84.851 seconds
hive>
> INSERT INTO part_cosmetic_OctNov2019_data PARTITION(event_type) SELECT substr(event_time, 1, 20),
> product_id, category_id, category_code, brand, price, user_id, user_session, month, event_type FROM cosmetic_OctNov2019_data ;
Query ID = hadoop_20210906102436_a436aff4-51e6-46c0-9cf7-228883d7a279
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1630501538063_0183)

-----
VERTICES      MODE        STATUS      TOTAL   COMPLETED   RUNNING   PENDING   FAILED   KILLED
-----
Map 1 ..... container    SUCCEEDED     61        61         0         0         0         0
Reducer 2 ..... container    SUCCEEDED      4         4         0         0         0         0
-----
VERTICES: 02/02 [=====] 100% ELAPSED TIME: 49.57 s
-----
Loading data to table cosmetics_clickstream_data.part_cosmetic_octnov2019_data partition (event_type=null)
K
Loaded : 4/4 partitions.
Time taken to load dynamic partitions: 5.751 seconds
Time taken for adding to write entity : 0.0 seconds
K
Time taken: 72.914 seconds
hive>

```

```

hive>
> INSERT INTO bucket_part_cosmetic_OctNov2019_data PARTITION(event_type, month) SELECT substr(event_time, 1, 20),
> product_id, category_id, category_code, brand, price, user_id, user_session, month, event_type FROM cosmetic_OctNov2019_data ;
Query ID = hadoop_20210906102610_f01fcd4a-e2d1-474e-9bde-9c7953b148d1
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1630501538063_0183)

-----
VERTICES      MODE        STATUS      TOTAL   COMPLETED   RUNNING   PENDING   FAILED   KILLED
-----
Map 1 ..... container    SUCCEEDED     61        61         0         0         0         0
Reducer 2 ..... container    SUCCEEDED      4         4         0         0         0         0
-----
VERTICES: 02/02 [=====] 100% ELAPSED TIME: 77.08 s
-----
Loading data to table cosmetics_clickstream_data.bucket_part_cosmetic_octnov2019_data partition (month=null, event_type=null)
K
Loaded : 8/8 partitions.
Time taken to load dynamic partitions: 6.475 seconds
Time taken for adding to write entity : 0.0 seconds
K
Time taken: 148.882 seconds
hive>

```

Checked the table data and the data format.

```
hive> select * from cosmetic_OctNov2019_data limit 5;
OK
2019-11-01 00:00:02 view 5802432 1487580009286598681 0.32 562076640 09fafd6c-6c99-46b1-834f-33527f4de241 11
2019-11-01 00:00:09 cart 5844397 1487580006317032337 2.38 553329724 20672216c-31b5-455d-alc-af0575a34ffb 11
2019-11-01 00:00:10 view 5837166 1783999064103190764 pnb 22.22 556138645 57ed222e-a54a-4907-9944-5a875c2d7f4f 11
2019-11-01 00:00:11 cart 5876812 1487580010100293687 jessnail 3.16 564506666 186c1951-8052-4b37-adce-dd9644b1d5f7 11
2019-11-01 00:00:24 remove_from_cart 5826182 1487580007483048900 3.33 553329724 20672216c-31b5-455d-alc-af0575a34ffb 11
Time taken: 1.438 seconds, Fetched: 5 row(s)
hive> select * from part_cosmetic_OctNov2019_data limit 5;
OK
2019-11-07 12:01:17 5677166 1487580008246412266 estel 4.29 568555254 e767f1c2-86f4-418c-8c84-99cd18e60e5f 11 view
2019-11-07 12:01:18 5876926 1487580011308253293 domix 2.65 542334607 bd4f3e17-0654-46ee-alc6-4a7ca29e79 11 view
2019-11-07 12:01:18 5726235 1487580005268456287 9.52 565980020 2dba2db7-4096-462d-8377-97e82a38b616 11 view
2019-11-07 12:01:18 5689725 1487580007852147670 staleks 13.17 520935011 ac1d8111-20eb-4099-be51-eff1db63031f 11 view
2019-11-07 12:01:21 5655332 1487580007457083075 5.71 563674973 912c4b3c-c750-43b3-a352-04bb40cc89d0 11 view
Time taken: 1.995 seconds, Fetched: 5 row(s)
hive> select * from bucket_part_cosmetic_OctNov2019_data limit 5;
OK
2019-11-16 11:56:03 5864368 1487580008263189483 elizavecca 11.75 235706079 1956c7fd-92a3-440f-ac5f-bbb161bc414b 11 remove_from_cart
2019-11-16 10:14:09 5846617 1487580012373606546 foamie 12.54 569718531 54ee84b2-87f5-4b9b-b265-efc0a4971ee7 11 remove_from_cart
2019-11-16 15:44:12 5863106 1487580011702517887 elizavecca 15.46 427566381 cfd8049-9cf6-4d8d-9ff5-73184320ada9 11 remove_from_cart
2019-11-16 14:25:31 5586142 1958278551207674674 inn 7.6 557790271 6b0635e8-ebdf-45b3-9494-18fcfc7c210 11 remove_from_cart
2019-11-16 08:21:59 5864335 1783999072332415142 elizavecca 24.92 572121798 6acd6b11-3b28-45eb-b8f0-5f17e8f6d5cc 11 remove_from_cart
Time taken: 1.62 seconds, Fetched: 5 row(s)
```

Checked the count of rows in each table.

```
hive>
> select count(*) from bucket_part_cosmetic_OctNov2019_data;
OK
1
c0
8738120
Time taken: 0.154 seconds, Fetched: 1 row(s)
hive> select count(*) from part_cosmetic_OctNov2019_data;
OK
1
c0
8738120
Time taken: 0.133 seconds, Fetched: 1 row(s)
hive> select count(*) from cosmetic_OctNov2019_data;
OK
1
c0
8738120
Time taken: 0.149 seconds, Fetched: 1 row(s)
hive>
```

Datatype of columns

```
hive>
> describe bucket_part_cosmetic_OctNov2019_data;
OK
event_time timestamp
product_id string
category_id string
category_code string
brand string
price float
server_id bigint
server_session string
month int
event_type string

+ Partition Information
+ col_name data_type comment
+ month int
+ event_type string
Time taken: 1.703 seconds, Fetched: 15 row(s)
```

Checked the partitions in the tables created and the HDFS location of the database and the tables.

Partitions and sub-partitions

“part_cosmetic_OctNov2019_data” – There are 4 partitions with respect to event_type

“bucket_part_cosmetic_OctNov2019_data” – There are 8 partitions with respect to event_type and month columns.

```
hive> show partitions part_cosmetic_OctNov2019_data;
OK
event_type=view
event_type=cart
event_type=purchase
event_type=remove_from_cart
Time taken: 1.899 seconds, Fetched: 4 row(s)
hive> show partitions bucket_part_cosmetic_OctNov2019_data;
OK
month=11/event_type=remove_from_cart
month=10/event_type=cart
month=11/event_type=view
month=10/event_type=purchase
month=10/event_type=view
month=11/event_type=purchase
month=10/event_type=remove_from_cart
month=11/event_type=cart
Time taken: 1.612 seconds, Fetched: 8 row(s)
hive>
```

HDFS location of partitioned table

There are 4 directories, one for each partition and each partition has a single file.

```
hadoop@ip-172-31-18-44 ~]$ hadoop fs -ls hdfs://ip-172-31-18-44.ec2.internal:8020/user/hive/warehouse/cosmetics_clickstream_data.db/part_cosmetic_octnov2019_data
Found 4 items
-rwxrwxrwt  - hadoop hadoop          0 2021-09-05 13:27 hdfs://ip-172-31-18-44.ec2.internal:8020/user/hive/warehouse/cosmetics_clickstream_data.db/part_cosmetic_octnov2019_data/event_type=cart
-rwxrwxrwt  - hadoop hadoop          0 2021-09-05 13:27 hdfs://ip-172-31-18-44.ec2.internal:8020/user/hive/warehouse/cosmetics_clickstream_data.db/part_cosmetic_octnov2019_data/event_type=purchase
-rwxrwxrwt  - hadoop hadoop          0 2021-09-05 13:27 hdfs://ip-172-31-18-44.ec2.internal:8020/user/hive/warehouse/cosmetics_clickstream_data.db/part_cosmetic_octnov2019_data/event_type=remove_from_cart
-rwxrwxrwt  - hadoop hadoop          0 2021-09-05 13:27 hdfs://ip-172-31-18-44.ec2.internal:8020/user/hive/warehouse/cosmetics_clickstream_data.db/part_cosmetic_octnov2019_data/event_type=view
hadoop@ip-172-31-18-44 ~]$
hadoop@ip-172-31-18-44 ~]$ hadoop fs -ls hdfs://ip-172-31-18-44.ec2.internal:8020/user/hive/warehouse/cosmetics_clickstream_data.db/part_cosmetic_octnov2019_data/event_type=purchase
Found 1 items
-rwxrwxrwt  1 hadoop hadoop 61814679 2021-09-05 13:27 hdfs://ip-172-31-18-44.ec2.internal:8020/user/hive/warehouse/cosmetics_clickstream_data.db/part_cosmetic_octnov2019_data/event_type=purchase/000001.0
```

HDFS location of bucketed table

There are 2 directories, one for each month. Inside each directory, there are 4 directories for each event_type. The directory has 20 files within since the rows are bucketed into 20 files.


```

hadoop@ip-172-31-18-44 ~]$ hadoop fs -ls hdfs://ip-172-31-18-44.ec2.internal:8020/user/hive/warehouse/cosmetics_clickstream_data.db/bucket_part_cosmetic_octnov2019_data
Found 2 items
-rwxrwxrwt - hadoop hadoop 0 2021-09-05 16:43 hdfs://ip-172-31-18-44.ec2.internal:8020/user/hive/warehouse/cosmetics_clickstream_data.db/bucket_part_cosmetic_octnov2019_data/month=10
-rwxrwxrwt - hadoop hadoop 0 2021-09-05 16:43 hdfs://ip-172-31-18-44.ec2.internal:8020/user/hive/warehouse/cosmetics_clickstream_data.db/bucket_part_cosmetic_octnov2019_data/month=11
hadoop@ip-172-31-18-44 ~]$ hadoop fs -ls hdfs://ip-172-31-18-44.ec2.internal:8020/user/hive/warehouse/cosmetics_clickstream_data.db/bucket_part_cosmetic_octnov2019_data/month=10
Found 4 items
-rwxrwxrwt - hadoop hadoop 0 2021-09-05 16:43 hdfs://ip-172-31-18-44.ec2.internal:8020/user/hive/warehouse/cosmetics_clickstream_data.db/bucket_part_cosmetic_octnov2019_data/month=10/event_type=cart
-rwxrwxrwt - hadoop hadoop 0 2021-09-05 16:43 hdfs://ip-172-31-18-44.ec2.internal:8020/user/hive/warehouse/cosmetics_clickstream_data.db/bucket_part_cosmetic_octnov2019_data/month=10/event_type=purchase
-rwxrwxrwt - hadoop hadoop 0 2021-09-05 16:43 hdfs://ip-172-31-18-44.ec2.internal:8020/user/hive/warehouse/cosmetics_clickstream_data.db/bucket_part_cosmetic_octnov2019_data/month=10/event_type=remove_from_cart
-rwxrwxrwt - hadoop hadoop 0 2021-09-05 16:43 hdfs://ip-172-31-18-44.ec2.internal:8020/user/hive/warehouse/cosmetics_clickstream_data.db/bucket_part_cosmetic_octnov2019_data/month=10/event_type=view
hadoop@ip-172-31-18-44 ~]$ hadoop fs -ls hdfs://ip-172-31-18-44.ec2.internal:8020/user/hive/warehouse/cosmetics_clickstream_data.db/bucket_part_cosmetic_octnov2019_data/month=10/event_type=purchase
Found 20 items
-rwxrwxrwt 1 hadoop hadoop 10842044 2021-09-05 16:42 hdfs://ip-172-31-18-44.ec2.internal:8020/user/hive/warehouse/cosmetics_clickstream_data.db/bucket_part_cosmetic_octnov2019_data/month=10/event_type=purchase/0000000_0
-rwxrwxrwt 1 hadoop hadoop 1744915 2021-09-05 16:42 hdfs://ip-172-31-18-44.ec2.internal:8020/user/hive/warehouse/cosmetics_clickstream_data.db/bucket_part_cosmetic_octnov2019_data/month=10/event_type=purchase/0000001_0
-rwxrwxrwt 1 hadoop hadoop 345213 2021-09-05 16:42 hdfs://ip-172-31-18-44.ec2.internal:8020/user/hive/warehouse/cosmetics_clickstream_data.db/bucket_part_cosmetic_octnov2019_data/month=10/event_type=purchase/0000002_0
-rwxrwxrwt 1 hadoop hadoop 1034579 2021-09-05 16:42 hdfs://ip-172-31-18-44.ec2.internal:8020/user/hive/warehouse/cosmetics_clickstream_data.db/bucket_part_cosmetic_octnov2019_data/month=10/event_type=purchase/0000003_0
-rwxrwxrwt 1 hadoop hadoop 330076 2021-09-05 16:42 hdfs://ip-172-31-18-44.ec2.internal:8020/user/hive/warehouse/cosmetics_clickstream_data.db/bucket_part_cosmetic_octnov2019_data/month=10/event_type=purchase/0000004_0

```

Task 3: Query Optimization

Used optimization techniques such as partitioning and bucketing and compared the query execution time on the three tables created.

Query: Check the total revenue generated due to purchases made in October

Set the below properties to execute the queries in “tez” mode and to display column headers in the queries being executed.

```

hive>
> set hive.cli.print.header=true;
hive> set hive.execution.engine=tez;
hive>

```

Non-bucketed table

```
hive>
> select ROUND(sum(price),2) as TotalRevenue, month as purchaseMonth
> from cosmetic_OctNov2019_data
> where month=10 and event_type="purchase"
> group by month;
Query ID = hadoop_20210905095321_9a2e8993-1fcd-43bf-9fc2-257510d09583
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1630501538063_0096)

-----
VERTICES      MODE      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
-----
Map 1 ..... container  SUCCEEDED  61      61          0         0         0         0
Reducer 2 ..... container  SUCCEEDED   2        2          0         0         0         0
-----
VERTICES: 02/02  [=====>>>] 100%  ELAPSED TIME: 25.67 s
-----
OK
totalrevenue    purchasemonth
1211538.43      10
Time taken: 27.687 seconds, Fetched: 1 row(s)
```

Partitioned table

```
hive>
> select ROUND(sum(price),2) as TotalRevenue, month as purchaseMonth
> from part_cosmetic_OctNov2019_data
> where month=10 and event_type="purchase"
> group by month;
Query ID = hadoop_20210905095408_b7a70144-94e9-46e6-880c-ac6ee58886ad
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1630501538063_0096)

-----
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
-----
VERTICES: 02/02  [=====>>>] 100%  ELAPSED TIME: 7.16 s
-----
OK
totalrevenue    purchasemonth
1211538.43      10
Time taken: 9.818 seconds, Fetched: 1 row(s)
```

Bucketed+ partitioned table

```
hive>
> select ROUND(sum(price),2) as TotalRevenue, month as purchaseMonth
> from bucket_part_cosmetic_OctNov2019_data
> where month=10 and event_type="purchase"
> group by month;
Query ID = hadoop_20210905100250_725eef50-9ad2-40de-98a5-59514cc4b4eb
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1630501538063_0099)

-----
VERTICES      MODE        STATUS      TOTAL   COMPLETED   RUNNING   PENDING   FAILED   KILLED
-----
Map 1 ..... container  SUCCEEDED    3         3           0         0         0         0
Reducer 2 ..... container  SUCCEEDED    2         2           0         0         0         0
-----
VERTICES: 02/02 [=====] 100% ELAPSED TIME: 2.60 s
-----
OK
totalrevenue    purchasemonth
1211538.43      10
Time taken: 4.805 seconds, Fetched: 1 row(s)
hive>
```

As seen from the above screenshots, the query executed faster for bucketed+ partitioned table followed by partitioned table and non-bucketed table.

Hence, the queries are executed on bucket_part_cosmetic_OctNov2019_data and the other tables created to check and compare the query optimization are deleted.

Task 4: Analysis using hive queries

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

```
SELECT Round(Sum(price), 2) AS TotalRevenue,
        month                AS purchaseMonth
FROM    bucket_part_cosmetic_octnov2019_data
WHERE   month = 10
        AND event_type = "purchase"
GROUP BY month;
```

```
hive>
> select ROUND(sum(price),2) as TotalRevenue, month as purchaseMonth
> from bucket_part_cosmetic_OctNov2019_data
> where month=10 and event_type="purchase"
> group by month;
Query ID = hadoop_20210905100250_725eef50-9ad2-40de-98a5-59514cc4b4eb
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1630501538063_0099)

-----
VERTICES      MODE      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
-----
Map 1 ..... container  SUCCEEDED    3         3         0         0         0         0
Reducer 2 ..... container  SUCCEEDED    2         2         0         0         0         0
-----
VERTICES: 02/02 [=====] 100% ELAPSED TIME: 2.60 s
-----
OK
totalrevenue  purchasemonth
1211538.43    10
Time taken: 4.805 seconds, Fetched: 1 row(s)
hive>
```

Total revenue generated due to purchases made in October is 1211538.43

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

```
SELECT Round(Sum(price), 2) AS TotalRevenue,
        month                AS purchaseMonth
FROM    bucket_part_cosmetic_octnov2019_data
WHERE   event_type = "purchase"
GROUP BY month;
```



```

> select ROUND(sum(price),2) as TotalRevenue, month as purchaseMonth
> from bucket_part_cosmetic_OctNov2019_data
> where event_type="purchase"
> group by month;
Query ID = hadoop_20210905102722_d6fe036c-6ec2-498e-bec0-e2fe0df4cda5
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1630501538063_0100)

-----
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: 2.86 s
-----
K
TotalRevenue  purchaseMonth
211538.43      10
1531016.9      11
Time taken: 5.058 seconds, Fetched: 2 row(s)
hive>

```

Total revenue generated due to purchases made in October is 1211538.43 and that in November is 1531016.9.

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

```

WITH octrevenue
  AS (SELECT Round(Sum(price), 2) AS OctRevenue,
            month                  AS purchaseMonth
        FROM bucket_part_cosmetic_octnov2019_data
        WHERE month = 10
            AND event_type = "purchase"
        GROUP BY month),
    novrevenue
  AS (SELECT Round(Sum(price), 2) AS NovRevenue,
            month                  AS purchaseMonth
        FROM bucket_part_cosmetic_octnov2019_data
        WHERE month = 11
            AND event_type = "purchase"
        GROUP BY month)
SELECT novrevenue - octrevenue AS ChangeInRevenue
FROM   octrevenue, novrevenue;

```



```

ive> with OctRevenue as
> (
> select ROUND(sum(price),2) as OctRevenue, month as purchaseMonth
> from bucket_part_cosmetic_OctNov2019_data
> where month =10 and event_type="purchase"
> group by month
> ),
> NovRevenue as
> (
> select ROUND(sum(price),2) as NovRevenue, month as purchaseMonth
> from bucket_part_cosmetic_OctNov2019_data
> where month =11 and event_type="purchase"
> group by month
> )
> select NovRevenue-OctRevenue as ChangeInRevenue
> from OctRevenue, NovRevenue;
Warning: Map Join MAPJOIN[23][bigTable=?] in task 'Reducer 4' is a cross product
Query ID = hadoop_20210906115307_ff0efb83-f7b9-4822-9f7b-e90c1761175b
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1630501538063_0188)

-----
VERTICES      MODE      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
-----
map 3 ..... container  SUCCEEDED   3         3         0         0         0         0
map 1 ..... container  SUCCEEDED   3         3         0         0         0         0
reducer 2 ..... container  SUCCEEDED   2         2         0         0         0         0
reducer 4 ..... container  SUCCEEDED   2         2         0         0         0         0
-----
VERTICES: 04/04  [=====>>>] 100%  ELAPSED TIME: 2.36 s
-----
K
19478.47
Time taken: 5.351 seconds, Fetched: 1 row(s)

```

The change in revenue generated due to purchases from October to November is 319478.47.

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

```

SELECT DISTINCT( category_code )
FROM          bucket_part_cosmetic_octnov2019_data
WHERE         category_code IS NOT NULL
AND          category_code != "";
Query: find the total number
of products available under EACH category.
SELECT      count(*),
            category_code
FROM        bucket_part_cosmetic_octnov2019_data
GROUP BY   category_code;

```

```

hive>
> select DISTINCT(category_code)
> from bucket_part_cosmetic_OctNov2019_data
> where category_code IS NOT NULL AND category_code!= "";
Query ID = hadoop_20210905105724_b7aeff7a-2e8c-48be-92fb-92459dffd1c
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1630501538063_0101)

```

| | VERTICES | MODE | STATUS | TOTAL | COMPLETED | RUNNING | PENDING | FAILED | KILLED |
|-----------------|-----------|-----------|--------|-------|-----------|---------|---------|--------|--------|
| Map 1 | container | SUCCEEDED | 66 | 66 | 0 | 0 | 0 | 0 | |
| Reducer 2 | container | SUCCEEDED | 1 | 1 | 0 | 0 | 0 | 0 | |

```

VERTICES: 02/02 [=====>] 100% ELAPSED TIME: 8.18 s
OK
category_code
accessories.bag
accessories.cosmetic_bag
apparel.glove
appliances.environment.air_conditioner
appliances.environment.vacuum
appliances.personal.hair_cutter
furniture.bathroom.bath
furniture.living_room.cabinet
furniture.living_room.chair
sport.diving
stationery.cartridge
Time taken: 10.35 seconds, Fetched: 11 row(s)
hive>

```

There are 11 distinct categories of products.

Query: Find the total number of products available under each category.

```

SELECT Count(*) ,
       category_code
FROM   bucket_part_cosmetic_octnov2019_data
GROUP BY category_code;

```

```

> select count(*),category_code
> from bucket_part_cosmetic_OctNov2019_data
> group by category_code;
Query ID = hadoop_20210905111947_fffe5618f-7aa3-4b5d-913c-b3f28486b722
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1630501538063_0102)

-----
      VERTICES      MODE      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
-----
Map 1 ..... container  SUCCEEDED    66         66         0         0         0         0
Reducer 2 ..... container  SUCCEEDED     1          1         0         0         0         0
-----
VERTICES: 02/02 [=====>>] 100% ELAPSED TIME: 7.46 s
-----
OK
_c0      category_code
8594895
11681    accessories.bag
1248     accessories.cosmetic_bag
18232    apparel.glove
332      appliances.environment.air_conditioner
59761    appliances.environment.vacuum
1643     appliances.personal.hair_cutter
9857     furniture.bathroom.bath
13439    furniture.living_room.cabinet
308      furniture.living_room.chair
2        sport.diving
26722    stationery.cartridge
Time taken: 9.562 seconds, Fetched: 12 row(s)
hive>

```

There are 8594895 entries with blank category code in the data.

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

```

WITH maxsalesoct
AS
(
    SELECT    brand,
             sum(price) AS total_sales
    FROM      bucket_part_cosmetic_octnov2019_data
    WHERE     month=10
    AND       event_type="purchase"
    GROUP BY  brand
    ORDER BY  (total_sales) DESC
    LIMIT     10),
maxsalesnov
AS
(
    SELECT    brand,

```

```

        sum(price) AS total_sales
FROM      bucket_part_cosmetic_octnov2019_data
WHERE     month=11
AND       event_type="purchase"
GROUP BY  brand
ORDER BY  (total_sales) DESC
LIMIT     10)

SELECT      maxsalesnov.brand
FROM        maxsalesoct
LEFT OUTER JOIN maxsalesnov
ON          maxsalesoct.brand =maxsalesnov.brand
WHERE       maxsalesnov.brand!= ""
ORDER BY    maxsalesnov.total_sales+maxsalesoct.total_sales
DESC
LIMIT      1;

```

```

> with maxSalesOct as
> (
> select brand, sum(price) as total_sales
> from bucket_part_cosmetic_OctNov2019_data
> where month=10
> and event_type="purchase"
> group by brand
> order by (total_sales) DESC
> limit 10),
> maxSalesNov as
> (
> select brand, sum(price) as total_sales
> from bucket_part_cosmetic_OctNov2019_data
> where month=11
> and event_type="purchase"
> group by brand
> order by (total_sales) DESC
> limit 10)
> select maxSalesNov.brand
> from maxSalesOct left outer join maxSalesNov
> on maxSalesOct.brand =maxSalesNov.brand
> where maxSalesNov.brand!= ""
> order by maxSalesNov.total_sales+maxSalesOct.total_sales DESC
> limit 1;
Query ID = hadoop_20210905115003_c200b8dc-75b8-4b6f-ae80-70da4980a004
total jobs = 1
Launching Job 1 out of 1
status: Running (Executing on YARN cluster with App id application_1630501538063_0105)

-----
      VERTICES      MODE      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
-----
Map 4 ..... container  SUCCEEDED    3         3         0         0         0         0
reducer 5 ..... container  SUCCEEDED    2         2         0         0         0         0
Map 1 ..... container  SUCCEEDED    3         3         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 6 ..... container  SUCCEEDED    1         1         0         0         0         0
reducer 7 ..... container  SUCCEEDED    1         1         0         0         0         0
-----
VERTICES: 07/07 [=====>] 100% ELAPSED TIME: 3.21 s
-----
K
maxsalesnov.brand
unail
time taken: 6.241 seconds. Fetched: 1 row(s)

```

Runail had the maximum sales in October and November combined

Query: Which brands increased their sales from October to November?

```
WITH maxsalesoct
  AS (SELECT brand,
             Sum(price) AS total_sales
       FROM bucket_part_cosmetic_octnov2019_data
      WHERE month = 10
             AND event_type = "purchase"
      GROUP BY brand),
maxsalesnov
  AS (SELECT brand,
             Sum(price) AS total_sales
       FROM bucket_part_cosmetic_octnov2019_data
      WHERE month = 11
             AND event_type = "purchase"
      GROUP BY brand)
SELECT maxsalesnov.brand
FROM   maxsalesoct
      JOIN maxsalesnov
        ON maxsalesoct.brand = maxsalesnov.brand
WHERE  maxsalesnov.total_sales - maxsalesoct.total_sales > 0
      AND maxsalesnov.brand != "";
```

There are 152 brands which have increased their sales from October to November. First few and last few brands of the list are displayed in the screenshot



| | | | |
|-------------------|--------------|--------------|---------------|
| OK | lianail | blizz | marutaka-foot |
| maxsalesnov.brand | limoni | bluesky | matrix |
| matiste | lovely | browxenna | mavala |
| beautyblender | marathon | candy | metzger |
| bodyton | markell | carmex | maskin |
| lpa.style | masura | chi | neoleor |
| coffin | milv | cosima | oniq |
| concept | missha | cosmoprofi | plazan |
| cristalinas | moyou | cutrin | rasyan |
| deoproce | nagaraku | de.lux | refectocil |
| domix | nefertiti | depilflax | s.care |
| ecolab | nirvel | dizao | severina |
| elizavecca | nitrile | ecocraft | shary |
| ellips | orly | egomania | shik |
| enjoy | osmo | elskin | skinlite |
| entity | ovale | estel | solomeya |
| eos | polarus | estelare | sophin |
| f.o.x | profepil | farmavita | staleks |
| fedua | profhenna | farmona | strong |
| finish | protokeratin | foamie | swarovski |
| fly | provoc | freshbubble | treaclemoon |
| freedecor | rosi | greymy | trind |
| gehwoi | roubloff | happyfons | uno |
| glysolid | runail | haruyoma | uskusi |
| godefroy | sanoto | insight | veraclara |
| grace | skinity | irisk | vilenta |
| grattol | smart | jessnail | yoko |
| agrobeauty | soleo | joico | zeitun |
| ingarden | supertan | kaaral | |
| inn | tertio | kamill | |
| italwax | yu-r | kaypro | |
| jaguar | airnails | keen | |
| jas | art-visage | kerasys | |
| kapous | artex | kims | |
| kinetics | aura | kocostar | |
| kiss | balbcare | koelcia | |
| koelf | beautix | konad | |
| kosmekka | beauty-free | laboratorium | |
| lador | beauugreen | latinoil | |
| ladykin | benovy | likato | |
| levissime | bioaqua | lowence | |
| levrana | biore | mane | |


```

find
uno
uskusi
veraclara
vilenta
yoko
zeitun
Time taken: 6.544 seconds, Fetched: 152 row(s)
hive>
> With maxSalesOct as
> (
> select brand, sum(price) as total_sales
> from bucket_part_cosmetic_OctNov2019_data
> where month=10
> and event_type="purchase"
> group by brand),
> maxSalesNov as
> (
> select brand, sum(price) as total_sales
> from bucket_part_cosmetic_OctNov2019_data
> where month=11
> and event_type="purchase"
> group by brand)
> select maxSalesNov.brand
> from maxSalesOct join maxSalesNov
> on maxSalesOct.brand = maxSalesNov.brand
> where maxSalesNov.total_sales-maxSalesOct.total_sales >0
> and maxSalesNov.brand!="";
Query ID = hadoop_20210905115252_f6a154a1-3ed6-40b6-a03a-c07a5de58760
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1630501538063_0105)

-----
VERTICES    MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
-----
Map 3 ..... container  SUCCEEDED  3      3           0         0         0         0
Map 1 ..... container  SUCCEEDED  3      3           0         0         0         0
Reducer 2 ..... container  SUCCEEDED  2      2           0         0         0         0
Reducer 4 ..... container  SUCCEEDED  2      2           0         0         0         0
-----
VERTICES: 04/04  [=====>>>] 100%  ELAPSED TIME: 4.20 s
-----
OK
maxsalesnov.brand
batiste
beautyblender
bodyton
opw.style
toifin
concept
cristalinas
deoproce
domix
ecolab
elizavecca

```

Query: 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.

```

WITH user_summary
AS
(
    SELECT      user_id ,
               sum(price)
               AS total,
               rank() over (partition BY user_id ORDER BY sum
m(price) DESC ) AS user_rank
    FROM        bucket_part_cosmetic_octnov2019_data
    WHERE       event_type='purchase'
    GROUP BY   user_id )

SELECT  user_id
FROM    user_summary
ORDER BY total DESC
LIMIT   10;

```

```

hive>
> With user_summary as
> (
> select user_id , sum(price) as total, rank() over (partition by user_id order by sum(price) DESC ) as user_rank
> from bucket_part_cosmetic_OctNov2019_data
> where event_type='purchase'
> group by user_id
> )
>
> select user_id from user_summary
> order by total DESC
> limit 10;
Query ID = hadoop_20210905115915_c283e483-6f65-4fc9-9219-50a1fb9b3c39
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1630501538063_0105)

-----
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: 2.20 s
-----
K
user_id
57790271
50318419
62167663
31900924
57850743
22130011
61592095
31950134
66576008
21347209
Time taken: 4.519 seconds, Fetched: 10 row(s)
hive>

```

The user ids of top 10 users of the website who spend the most is displayed.