HIVE CASE STUDY

Copying the data set into the HDFS:

- To check the Hadoop file system

hadoop fs -ls /

```
[hadoop@ip-172-31-16-69 ~]$ hadoop fs -ls /
Found 4 items
drwxr-xr-x   - hdfs hadoop          0 2021-04-24 05:20 /apps
drwxrwxrwt   - hdfs hadoop          0 2021-04-24 05:21 /tmp
drwxr-xr-x   - hdfs hadoop          0 2021-04-24 05:41 /user
drwxr-xr-x   - hdfs hadoop          0 2021-04-24 05:20 /var
```

hadoop fs –ls /user/

```
[hadoop@ip-172-31-16-69 ~]$ hadoop fs -ls /user/
Found 4 items
drwxrwxrwx   - hadoop hadoop          0 2021-04-24 05:20 /user/hadoop
drwxr-xr-x   - mapred mapred          0 2021-04-24 05:20 /user/history
drwxrwxrwx   - hdfs   hadoop          0 2021-04-24 05:20 /user/hive
drwxrwxrwx   - root   hadoop          0 2021-04-24 05:20 /user/root
[hadoop@ip-172-31-16-69 ~]$ hadoop fs -ls /user/hive/
Found 1 items
drwxrwxrwt   - hdfs hadoop          0 2021-04-24 05:20 /user/hive/warehouse
```

- Creating a directory

hdfs dfs -mkdir /user/clickstream

```
[hadoop@ip-172-31-16-69 ~]$ hadoop fs -mkdir /user/clickstream
[hadoop@ip-172-31-16-69 ~]$ hadoop fs -ls /user/
Found 5 items
drwxr-xr-x   - hadoop hadoop          0 2021-04-24 08:24 /user/clickstream
drwxrwxrwx   - hadoop hadoop          0 2021-04-24 07:21 /user/hadoop
drwxr-xr-x   - mapred mapred          0 2021-04-24 05:20 /user/history
drwxrwxrwx   - hdfs   hadoop          0 2021-04-24 07:58 /user/hive
drwxrwxrwx   - root   hadoop          0 2021-04-24 05:20 /user/root
```

hdfs dfs -mkdir /user/clickstream/salesdata

```
[hadoop@ip-172-31-23-227 ~]$ hdfs dfs -mkdir /user/clickstream/salesdata
[hadoop@ip-172-31-23-227 ~]$ hadoop fs -ls /user/clickstream
Found 1 items
drwxr-xr-x   - hadoop hadoop          0 2021-04-29 14:16 /user/clickstream/salesdata
[hadoop@ip-172-31-23-227 ~]$
```

- Move data from the S3 bucket into the HDFS

hadoop distcp s3n://e-commerce-events-ml/2019-Oct.csv /user/clickstream/salesdata

```
[hadoop@ip-172-31-23-116 ~]$ hadoop distcp s3n://e-commerce-events-ml/2019-Oct.csv /user/clickstream/salesdata
```

hadoop distcp s3n://e-commerce-events-ml/2019-Nov.csv
/user/clickstream/salesdata

```
-rw-r--r--   1 hadoop hadoop  482542278 2021-04-28 15:00 /user/clickstream/salesdata/2019-Oct.csv
[hadoop@ip-172-31-23-116 ~]$ hadoop distcp s3n://e-commerce-events-ml/2019-Nov.csv /user/clickstream/salesdata
```

- To check if the files are imported or not

hadoop fs -ls /user/clickstream/salesdata

```
[hadoop@ip-172-31-23-116 ~]$ hadoop fs -ls /user/clickstream/salesdata
Found 2 items
-rw-r--r--   1 hadoop hadoop  545839412 2021-04-28 15:18 /user/clickstream/salesdata/2019-Nov.csv
-rw-r--r--   1 hadoop hadoop  482542278 2021-04-28 15:00 /user/clickstream/salesdata/2019-Oct.csv
[hadoop@ip-172-31-23-116 ~]$
```

Creating the database and launching Hive queries on EMR cluster:

- Launching the Hive CLI

```
[hadoop@ip-172-31-16-69 ~]$ hive

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

- Creating the database

create database if not exists cs
location '/user/clickstream/salesdata' ;

```
hive> create database if not exists cs
    > location '/user/clickstream/salesdata' ;
OK
Time taken: 0.687 seconds
```

describe database extended cs ;

```
hive> describe database extended cs ;
OK
cs              hdfs://ip-172-31-23-116.ec2.internal:8020/user/clickstream/salesdata    hadoop  USER
Time taken: 0.186 seconds, Fetched: 1 row(s)
hive>
```

use cs ;

```
hive> use cs;
OK
Time taken: 0.014 seconds
```

- Creating external table

  create external table if not exists clickstreamtab (event_time timestamp , event_type string , product_id string , category_id string , category_code string , brand string , price float, user_id bigint , user_session string)
  row format serde 'org.apache.hadoop.hive.serde2.OpenCSVSerde'
  stored as textfile
  location '/user/clickstream/salesdata'
  tblproperties("skip.header.line.count"="1");

```
hive> create external table if not exists clickstreamtab (event_time timestamp , event_type string , product_id string , category_id string , category_code s
tring , brand string , price float, user_id bigint , user_session string)
    > row format serde 'org.apache.hadoop.hive.serde2.OpenCSVSerde'
    > stored as textfile
    > location '/user/clickstream/salesdata'
    > tblproperties("skip.header.line.count"="1");
OK
Time taken: 0.333 seconds
hive> select * from clickstreamtab limit 5;
OK
2019-11-01 00:00:02 UTC view    5802432 1487580009286598681                     0.32    562076640       09fafd6c-6c99-46b1-834f-33527f4de241
2019-11-01 00:00:09 UTC cart    5844397 1487580006317032337                     2.38    553329724       2067216c-31b5-455d-a1cc-af0575a34ffb
2019-11-01 00:00:10 UTC view    5837166 1783999064103190764             pnb     22.22   556138645       57ed222e-a54a-4907-9944-5a875c2d7f4f
2019-11-01 00:00:11 UTC cart    5876812 1487580010100293687             jessnail        3.16    564506666       186c1951-8052-4b37-adce-dd9644b1d5f7
2019-11-01 00:00:24 UTC remove_from_cart        5826182 1487580007483048900                     3.33    553329724       2067216c-31b5-455d-a1cc-af0575a34ffb
Time taken: 2.051 seconds, Fetched: 5 row(s)
hive>
```

- Checking the table

  select * from clickstreamtab limit 5;

```
hive> select * from clickstreamtab limit 5;
OK
2019-11-01 00:00:02 UTC view    5802432 1487580009286598681                     0.32    562076640       09fafd6c-6c99-46b1-834f-33527f4de241
2019-11-01 00:00:09 UTC cart    5844397 1487580006317032337                     2.38    553329724       2067216c-31b5-455d-a1cc-af0575a34ffb
2019-11-01 00:00:10 UTC view    5837166 1783999064103190764             pnb     22.22   556138645       57ed222e-a54a-4907-9944-5a875c2d7f4f
2019-11-01 00:00:11 UTC cart    5876812 1487580010100293687             jessnail        3.16    564506666       186c1951-8052-4b37-adce-dd9644b1d5f7
2019-11-01 00:00:24 UTC remove_from_cart        5826182 1487580007483048900                     3.33    553329724       2067216c-31b5-455d-a1cc-af0575a34ffb
Time taken: 2.051 seconds, Fetched: 5 row(s)
hive>
```

- Creating another table ‘monthdata’ by extracting month separately from timestamp column and adding extra column ‘month’

  create external table if not exists monthdata (month int , event_time timestamp , event_type string , product_id string , category_id string , category_code string , brand string , price float, user_id bigint , user_session string)
  row format serde 'org.apache.hadoop.hive.serde2.OpenCSVSerde'
  stored as textfile;

```
hive>  create external table if not exists monthdata (month int , event_time timestamp , event_type string , product_id string , category_id string , categor
y_code string , brand string , price float, user_id bigint , user_session string)
    > row format serde 'org.apache.hadoop.hive.serde2.OpenCSVSerde'
    > stored as textfile;
OK
Time taken: 0.265 seconds
```

  insert into table monthdata select month(event_time) as month, event_time , event_type , product_id , category_id , category_code , brand , price , user_id , user_session from clickstreamtab ;

```
hive> insert into table monthdata select month(event_time) as month, event_time , event_type , product_id , category_id , category_code , brand , price , use
r_id , user_session from clickstreamtab ;
Query ID = hadoop_20210428153854_45764171-d2e0-42f7-bfa5-a13d6b81c708
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_1619621564682_0007)

--------------------------------------------------------------------------------
        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: 98.25 s
--------------------------------------------------------------------------------
Loading data to table cs.monthdata
OK
Time taken: 108.742 seconds
```

Hive queries to answer the questions:

Question 1, 2 & 3

- Partition by event_type and bucketing on month:

  create table if not exists dyn_month1 (month int, price float)
  PARTITIONED BY (event_type string) CLUSTERED BY (month) into 2 buckets
  ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.OpenCSVSerde'
  WITH SERDEPROPERTIES
  ("separatorChar" = ",", "quoteChar" = "\"", "escapeChar" = "\\")
  STORED AS TEXTFILE;

```
hive> create table if not exists dyn_month1 (month int, price float)
    > PARTITIONED BY (event_type string) CLUSTERED BY (month) into 2 buckets
    > ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.OpenCSVSerde'
    > WITH SERDEPROPERTIES
    > ( "separatorChar" = ",", "quoteChar" = "\"", "escapeChar" = "\\")
    > STORED AS TEXTFILE;
OK
Time taken: 0.089 seconds
```

  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;
```

  insert into table dyn_month1 partition(event_type) select month, price ,
  event_type from monthdata ;

```
hive> insert into table dyn_month1 partition( event_type ) select month, price , event_type  from monthdata ;
Query ID = hadoop_20210428155543_dbd10566-bd02-4f8e-8c6d-708522de92b6
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1619621564682_0007)

--------------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      9          9        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      5          5        0        0       0       0
--------------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 133.51 s
--------------------------------------------------------------------------------------------------
Loading data to table cs.dyn_month1 partition (event_type=null)

Loaded : 4/4 partitions.
         Time taken to load dynamic partitions: 0.343 seconds
         Time taken for adding to write entity : 0.001 seconds
OK
Time taken: 135.015 seconds
```

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

- select sum(price) as total_revenue from dyn_month1
  where month = 10 and event_type = 'purchase' ;

```
hive> select sum(price) as total_revenue from dyn_month1
    > where month = 10 and event_type = 'purchase' ;
Query ID = hadoop_20210428160240_556c9df0-316c-4952-b654-ab44525a5c73
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1619621564682_0007)

--------------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 10.17 s
--------------------------------------------------------------------------------------------------
OK
1211538.4299997666
Time taken: 12.021 seconds, Fetched: 1 row(s)
```

Answer: 1211538.4299

<u>Without optimization:</u>

select sum(price) as total_revenue from monthdata
where month = 10 and event_type = 'purchase' ;

```
hive>  select sum(price) as total_revenue from monthdata
    > where month = 10 and event_type = 'purchase' ;
Query ID = hadoop_20210428171357_14c4f562-0a59-48a8-8e21-93855763144d
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_1619621564682_0010)

--------------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      9          9        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 69.25 s
--------------------------------------------------------------------------------------------------
OK
1211538.4300000046
Time taken: 78.414 seconds, Fetched: 1 row(s)
hive>
```

Same query is done with and without partitioning, job is done in 10.17 s with bucketing whereas time taken to complete the job without optimization is 69.25 s

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

- select month, count(event_type) as purchase_count
  from dyn_month1
  where event_type = 'purchase'
  group by month ;

```
hive> select month, count(event_type) as purchase_count
    >    from dyn_month1
    >    where event_type = 'purchase'
    >    group by month ;
Query ID = hadoop_20210428160851_cb38f6f3-e778-4b48-bc15-a99d89324e70
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1619621564682_0007)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 9.84 s
--------------------------------------------------------------------------------
OK
10      245624
11      322417
Time taken: 10.496 seconds, Fetched: 2 row(s)
```

Oct – 245624
Nov – 322417

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

- With revenue_change as
  (
  select sum(case when month = 10 then price else 0 end) AS October , sum(case when month = 11 then price else 0 end) AS November
  from dyn_month1
  where event_type = 'purchase'
)
  select October , November , November - October as Difference
  from revenue_change;

```
hive>
    > With revenue_change as
    > (
    > select sum(case when month = 10 then price else 0 end) AS October , sum(case when month = 11 then price else 0 end) AS November
    > from dyn_month1
    > where event_type = 'purchase'
    > )
    > select October , November , November - October as Difference
    > from revenue_change;
Query ID = hadoop_20210429183208_0daa6728-5c59-4054-b4c8-a2943d5d9a20
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1619718961631_0005)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reduce 2 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 10.92 s
--------------------------------------------------------------------------------
OK
1211538.4299997666      1531016.900000061       319478.47000029427
Time taken: 11.703 seconds, Fetched: 1 row(s)
```

Revenue generated due to purchases

Oct – 1211538.43

Nov – 1531016.90

Change in revenue – 319478.47

Question 4 & 5

- Partition by Category_code

create table if not exists dyn_category (product_id string , category_id string) partitioned by (category_code string)
row format serde 'org.apache.hadoop.hive.serde2.OpenCSVSerde'
stored as textfile;

```
hive> create table if not exists dyn_category (product_id string , category_id string) partitioned by (category_code string)
    > row format serde 'org.apache.hadoop.hive.serde2.OpenCSVSerde'
    > stored as textfile;
OK
Time taken: 0.084 seconds
```

insert into table dyn_category partition (category_code) select product_id , category_id , category_code from monthdata ;

```
hive> insert into table dyn_category partition ( category_code ) select product_id , category_id , category_code from monthdata ;
Query ID = hadoop_20210429144855_57153614-de9e-47c6-8a6b-b2f883a83fe0
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1619705042474_0005)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      9          9        0        0       0       0
Reduce 2 ...... container     SUCCEEDED      5          5        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 106.67 s
--------------------------------------------------------------------------------
Loading data to table cs.dyn_category partition (category_code=null)

Loaded : 12/12 partitions.
         Time taken to load dynamic partitions: 0.895 seconds
         Time taken for adding to write entity : 0.006 seconds
OK
Time taken: 109.196 seconds
```

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

- SELECT distinct(category_code)
  FROM dyn_category
  WHERE category_code IS NOT NULL;

```
hive> SELECT distinct(category_code)
    > FROM dyn_category
    > WHERE category_code IS NOT NULL;
Query ID = hadoop_20210429152239_f3db4f84-ca56-4603-8e38-10ad4bed1ee6
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1619705042474_0005)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reduce 2 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 8.03 s
--------------------------------------------------------------------------------
OK
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.cartrige
Time taken: 8.619 seconds, Fetched: 11 row(s)
```

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

- select count(product_id) , category_code
  from dyn_category
  where category_code IS NOT NULL
  group by category_code;

```
hive> select count(product_id) , category_code
    > from dyn_category
    > where category_code IS NOT NULL
    > group by category_code;
Query ID = hadoop_20210429151954_fac49d78-3589-4fb8-aef1-c77dca891c4d
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1619705042474_0005)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 7.97 s
--------------------------------------------------------------------------------
OK
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.cartrige
Time taken: 8.754 seconds, Fetched: 11 row(s)
```

Question 6 & 7

- Partition based on month and bucketing on brand

select count(distinct(brand)) from monthdata ;

```
hive> select count(distinct(brand)) from monthdata ;
Query ID = hadoop_20210428162802_d6ea53fe-6643-4a47-9ca1-d8c2ecdc8664
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1619621564682_0008)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      9          9        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      5          5        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 64.61 s
--------------------------------------------------------------------------------
OK
245
```

Distinct brands (including ‘blanks’) = 245

create table if not exists dyn_brand (brand string, price float)
PARTITIONED BY (month int) CLUSTERED BY (brand) into 245 buckets
ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.OpenCSVSerde'
WITH SERDEPROPERTIES ("separatorChar" = ",", "quoteChar" = "\"", "escapeChar" = "\\")
STORED AS TEXTFILE;

```
hive> create table if not exists dyn_brand (brand string, price float)
    > PARTITIONED BY (month int) CLUSTERED BY (brand) into 245 buckets
    > ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.OpenCSVSerde'
    > WITH SERDEPROPERTIES ( "separatorChar" = ",", "quoteChar" = "\"", "escapeChar" = "\\")
    > STORED AS TEXTFILE;
OK
Time taken: 0.099 seconds
```

insert into table dyn_brand partition(month) select brand , price , month from monthdata ;

```
hive> insert into table dyn_brand partition( month ) select brand , price , month  from monthdata ;
Query ID = hadoop_20210429155209_85d65bed-1095-4bed-b507-6276bdaf2e22
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1619705042474_0006)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      9          9        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      5          5        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 140.49 s
--------------------------------------------------------------------------------
Loading data to table cs.dyn_brand partition (month=null)

Loaded : 2/2 partitions.
         Time taken to load dynamic partitions: 0.785 seconds
         Time taken for adding to write entity : 0.003 seconds
OK
Time taken: 145.738 seconds
```

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

- select brand , sum(price) as sales
  from dyn_brand
  group by brand
  order by sales desc
  limit 2;

```
hive> select brand , sum(price) as sales
    > from dyn_brand
    > group by brand
    > order by sales desc
    > limit 2;
Query ID = hadoop_20210429162230_cee03fed-b70a-40f5-bad9-aaf51acd8b2c
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1619705042474_0007)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      4          4        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: 54.08 s
--------------------------------------------------------------------------------
OK
        2.6194508599906653E7
strong  4927445.599999651
Time taken: 54.743 seconds, Fetched: 2 row(s)
```

Ignoring the ‘blanks’ in brand column, the maximum sales are of brand ‘strong’ which is = 4927445.599999651

7. Which brands increased their sales from October to November

- select brand
  from dyn_brand
  group by brand
  having (
  sum(case when month = 11 then price else 0 end) >
  sum(case when month = 10 then price else 0 end)
  );

```
hive> select brand
    >  from dyn_brand
    >  group by brand
    >  having (
    >                sum(case when month = 11 then price else 0 end) >
    >                sum(case when month = 10 then price else 0 end)
    > );
Query ID = hadoop_20210429162426_a5e99294-c17b-49ee-bf8c-9688ab19e26a
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1619705042474_0007)

--------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      4          4        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 59.78 s
--------------------------------------------------------------------------------------------
OK

airnails
art-visage
artex
aura
australis
balbcare
barbie
batiste
beautix
beauty-free
beauugreen
benovy
biofollica
bpw.style
browxenna
busch
candy
carmex
cnd
coifin
concept
consly
```

Got 110 rows out of 245, attached only the last part of the result as the output is very long. Ignoring the ‘blanks’, there are 109 brands which increased their sales from October to November.

```
s.care
sanoto
severina
shary
shifei
shik
skinlite
smart
sophin
staleks
strong
swarovski
tazol
tertio
uno
vilenta
vosev
yoko
yu-r
zeitun
Time taken: 60.357 seconds, Fetched: 110 row(s)
```

Question 8

8. Top 10 users who spend the most.

- select user_id , sum(price) as total_price from monthdata
  group by user_id order by total_price desc limit 10;

```
hive> select user_id , sum(price) as total_price from monthdata
    > group by user_id order by total_price desc limit 10;
Query ID = hadoop_20210429150622_bbb5d925-6e6d-490b-aed2-bb386aabf4ae
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1619705042474_0005)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      9          9        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      5          5        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 83.77 s
--------------------------------------------------------------------------------
OK
557616099       63266.96999999997
557956487       52370.21999999999
550388516       46264.27999999965
531900924       43504.71000000002
352394658       28205.910000000033
550353491       25317.260000000002
443045778       23742.679999999997
479928991       23540.6
554848397       23359.429999999986
526213023       22983.28
Time taken: 84.72 seconds, Fetched: 10 row(s)
```

Cleaning Up:

- Dropping all tables in database

```
hive> DROP TABLE IF EXISTS monthdata;
OK
Time taken: 0.112 seconds
hive> DROP TABLE IF EXISTS dyn_category;
OK
Time taken: 0.252 seconds
hive> DROP TABLE IF EXISTS dyn_brand;
OK
Time taken: 0.261 seconds
```

```
hive> DROP TABLE IF EXISTS clickstreamtab;
OK
Time taken: 0.068 seconds
```

```
hive> DROP TABLE IF EXISTS dyn_month1 ;
OK
Time taken: 0.081 seconds
```

- Drop Database

```
hive> DROP DATABASE IF EXISTS cs;
OK
Time taken: 0.043 seconds
hive>
```

- Details of the Cluster:

  - 2-node EMR cluster with both the master and core nodes as M4.large.
  - emr-5.29.0

Configuration details

Release label:	emr-5.29.0
Hadoop distribution:	Amazon 2.8.5
Applications:	Hive 2.3.6
Log URI:	s3://aws-logs-425495063791-us-east-1/elasticmapreduce/
EMRFS consistent view:	Disabled
Custom AMI ID:	--

Application user interfaces

Persistent user interfaces:	--
On-cluster user interfaces:	Not Enabled Enable an SSH Connection

Network and hardware

Availability zone:	us-east-1b
Subnet ID:	subnet-0fb88242
Master:	Running 1 m4.large
Core:	Running 1 m4.large
Task:	--
Cluster scaling:	Not enabled

Security and access

Key name:	demo_key_pair
EC2 instance profile:	EMR_EC2_DefaultRole

Terminate Cluster:

Cluster: Mycluster Terminated Terminated by user request

Summary | Application user interfaces | Monitoring | Hardware | C

Summary

ID: j-2GV8BDT51WYNC
Creation date: 2021-04-29 19:29 (UTC+5:30)
End date: 2021-04-29 22:09 (UTC+5:30)
Elapsed time: 2 hours, 39 minutes
After last step completes: Cluster waits
Termination protection: Off
Tags: --
Master public DNS: ec2-54-91-6-225.compute-1.amazonaws.com
Connect to the Master Node Using SSH

Configuration details

Release label: emr-5.29.0
Hadoop distribution: Amazon 2.8.5
Applications: Hive 2.3.6
Log URI: s3://aws-logs-425495063791-us-east-1/elasticmapreduce/