Submitted by

Shubhang singh and Sandeep Challas

Problem statement:

With online sales gaining popularity, tech companies are exploring ways to improve their sales by analysing customer behaviour and gaining insights about product trends. Furthermore, the websites make it easier for customers to find the products they require without much scavenging. Needless to say, the role of big data analysts is among the most sought-after job profiles of this decade. Therefore, as part of this assignment, we will be challenging you, as a big data analyst, to extract data and gather insights from a real-life data set of an e-commerce company.

Case Study:

Before the given steps we created an EMR cluster of version 5.29.0 with two nodes of M4.large type instances after which we linked the dataset to the HDFS via the CLI. Finally launching the Hive tool to perform query analysis on it.

1. Launching terminal with Hadoop

   create a directory in HDFS to collect the input data:

   >> Hadoop fs -mkdir /user/hive/online_sales

2. Reading the file from S3

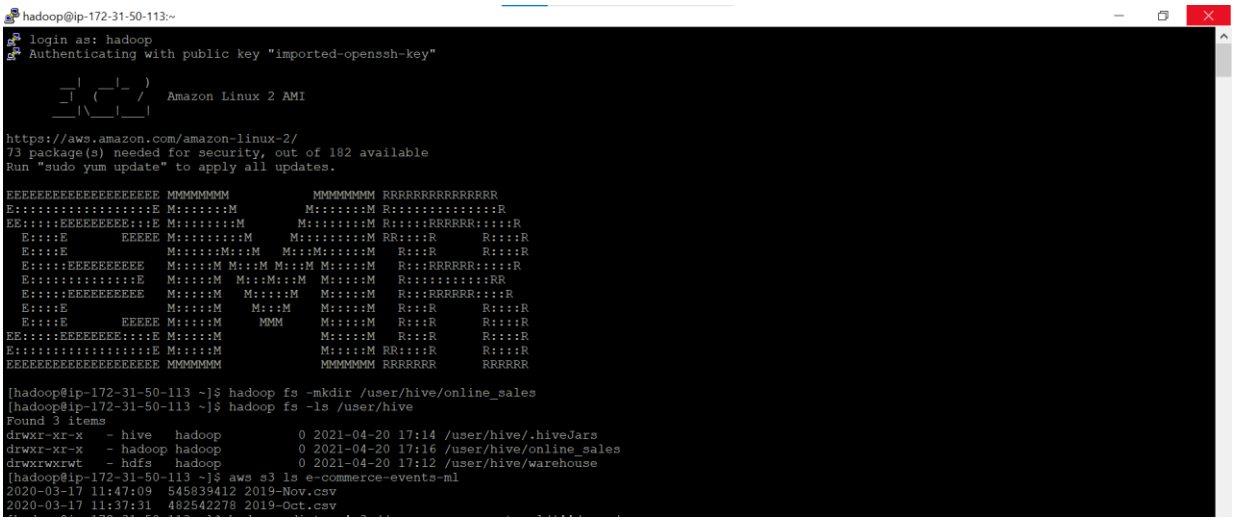
   Import file from S3

   >> aws s3 ls e-commerce-events-ml

   Check the directory

   >> hadoop fs -ls /user/hive

   Screenshot
   #1

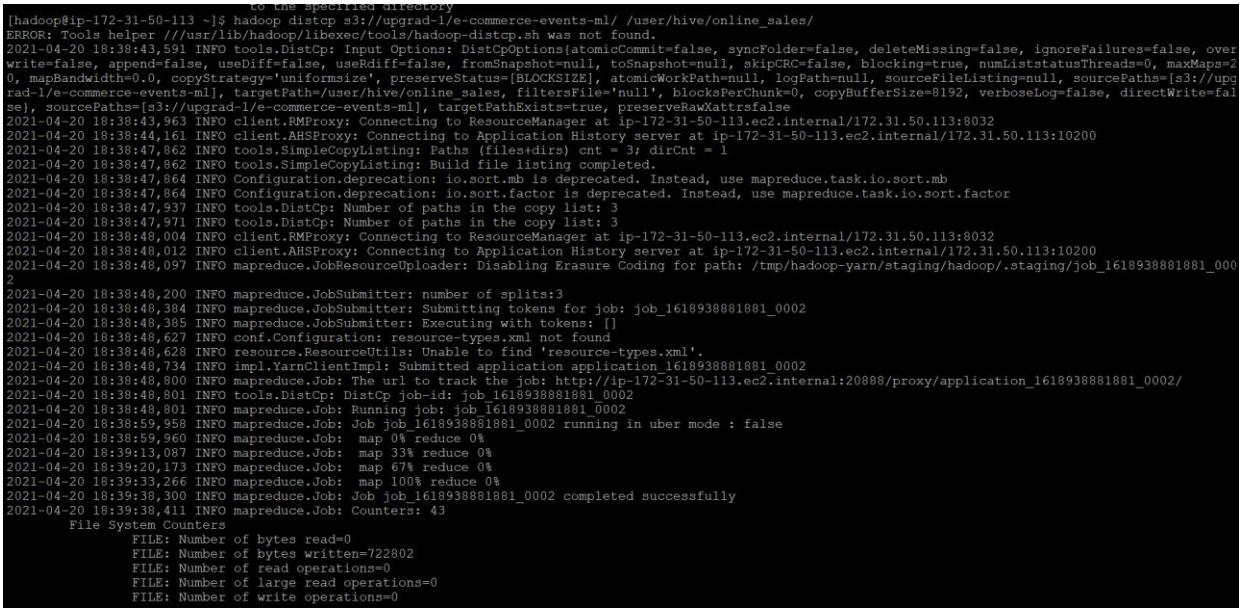


3. Copying the dataset from S3 to HDFS:

   To move datasets from S3 to HDFS:

   >> Hadoop distcp s3://upgrad-1/e-commerce-events-ml/ /user/hive/online_sales/

   #Screenshot 2



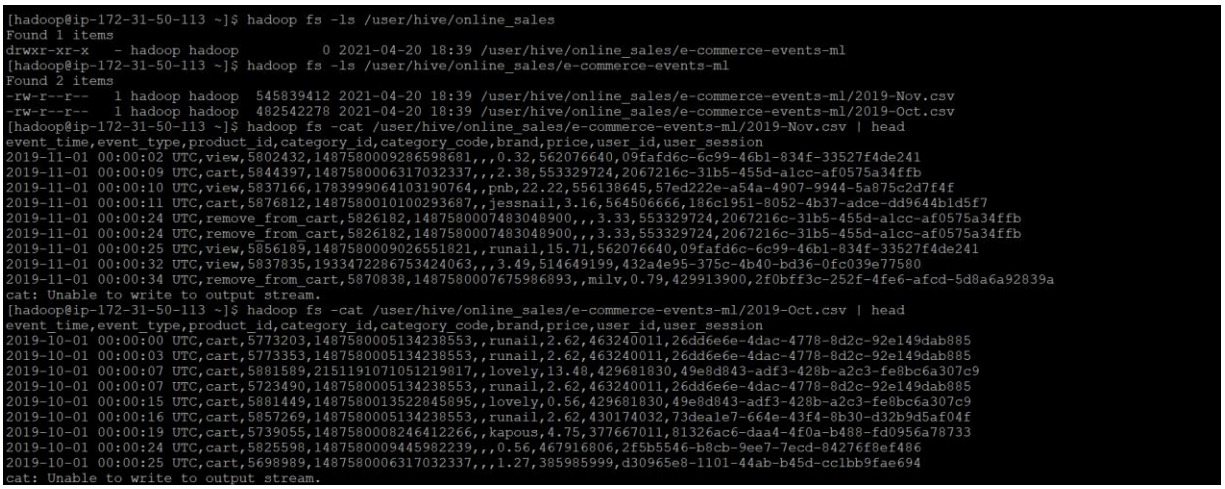
4. Reading the dataset

   To check the files

   >> Hadoop fs -cat /user/hive/online_sales/2019-Nov.csv|head

   >> Hadoop fs -cat /user/hive/online_sales/2019-Oct.csv|head

   #Screenshot 3



5. Loading Hive and creating initial tables

   Use hive command

   >> hive

   Create and use database

   >> create database online_sales;

   >> use online_sales;

   >> create external table if not exists sales_input(event_time timestamp,event_type string,product_id string, category_id string,category_code string,brand string,price float,user_id nigint,user_session string)

   ROW FORMAT SERDE ‘org.apache.hadoop.hive.serde2.OpenCSVSerde’

   STORED AS TEXT FILE

   LOCATION ‘/user/hive/online_sales/’

   Tblproperties(“skip.header.line.count”=”1”);

   #Screenshot 4

```
hive> create external table if not exists sales_input(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/hive/online_sales/'
    > tblproperties("skip.header.line.count"="1");
OK
Time taken: 0.354 seconds
hive> select * from sales_input 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: 2.664 seconds, Fetched: 5 row(s)
hive>
```

6. Enable partitioning and bucketing

   To enable portioning and dynamic partitioning

   >> set hive.exec.dynamic.partition.mode=nonstrict;

   >> set hive.exec.dynamic.partition=true;

   >> set hive.enforce.bucketing=true;

   >> create table if not exists sales_bucket(event_time timestamp,product_id string,category_id string,category_code string,brand string,price float,user_id bigint,user_session string)

   >> PARTITIONED BY (event_type string) CLUSTERED BY (price) into 10 buckets

   >> ROW FORMAT SERDE ‘org.apache.hadoop.hive.serde2.OpenCSVSerde’

   >> STORED AS TEXTFILE;

   To load the optimised hive table:

   >> insert into table sales_bucket partition(event_type) select event_time,product_id,category_id,category_code,brand,price,user_id,user_session,event_type from sales_input

   >> distributed by event_type;

   #Screenshot 5

```
hive> create table if not exists sales_bucket(event_time timestamp,product_id string,category_id string,category_code string,brand string,price float,user_id
 bigint,user_session string)partitioned by (event_type string) clustered by (price) into 10 buckets row format serde 'org.apache.hadoop.hive.serde2.OpenCSVSe
rde' stored as textfile;
OK
Time taken: 0.074 seconds
hive> insert into table sales_bucket partition(event_type)select event_time,product_id,category_id,category_code,brand,price,user_id,user_session,event_type
from sales_input distribute by event_type  ;
Query ID = hadoop_20210423113744_0cd38add-774e-485e-b2ea-97e68c687344
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_1619173061468_0005)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 ..........   container     SUCCEEDED      2          2        0        0       0       0
Reducer 2 ......   container     SUCCEEDED      5          5        0        0       0       0
Reducer 3 ......   container     SUCCEEDED      5          5        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 196.17 s
--------------------------------------------------------------------------------
Loading data to table online_sales.sales_bucket partition (event_type=null)

Loaded : 4/4 partitions.
        Time taken to load dynamic partitions: 0.882 seconds
        Time taken for adding to write entity : 0.003 seconds
OK
Time taken: 205.938 seconds
```

7. Solutions to the questions asked

- Find the total revenue generated due to purchase made in October

  >> select sum(price) from sales_bucket where event_type=”purchase” and month(event_time)=10;

  Output:

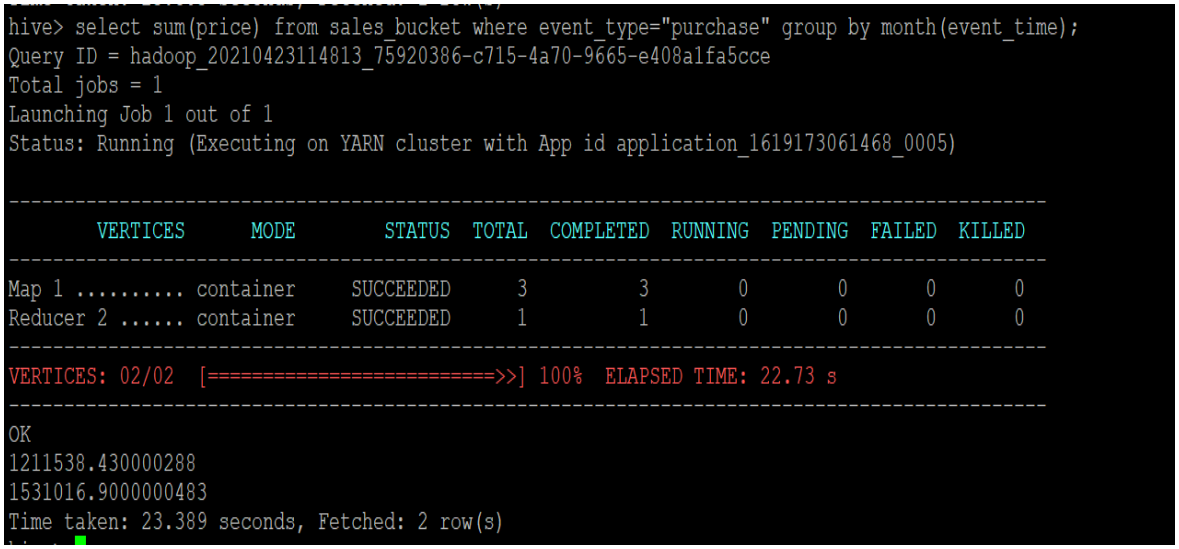
```
hive> select sum(price) from sales_bucket where event_type="purchase" and month(event_time)=10;
Query ID = hadoop_20210423114514_716b46f4-6fe9-4cdd-a85f-c6ce1400c4c2
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1619173061468_0005)

--------------------------------------------------------------------------------
        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: 22.42 s
--------------------------------------------------------------------------------
OK
1211538.430000288
Time taken: 23.503 seconds, Fetched: 1 row(s)
```

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

>> select sum(price) from sales_bucket where event_type=”purchase” group by month(event_type);

Output:



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

  >>Select Oct,Nov,Nov-Oct Difference from

  >>(select sum(case when date_format(event_time,’MM’)=10 then price else 0 end) as Oct, sum(case when date_format(event_time,’MM’)=11 then price else 0 end)as Nov from sales_bucket where date_format(event_time,’MM’)in(10,11)and event_type=”purchase”)s;

  Output:

```
hive> select Oct,Nov, Nov-Oct Difference
    > from (select sum(case when date_format(event_time,'MM')=10 then price else 0 end)as Oct,sum(case when date_format(event_time,'MM')=11 then price else 0
 end)as Nov from sales_bucket where date_format(event_time,'MM')in (10,11)and event_type="purchase")s;
Query ID = hadoop_20210423115645_3b0097d9-3c84-4a2e-9052-7c4b7ea11fda
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1619173061468_0006)

--------------------------------------------------------------------------------
        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: 38.12 s
--------------------------------------------------------------------------------
OK
1211538.430000288       1531016.9000000483      319478.4699997604
Time taken: 38.769 seconds, Fetched: 1 row(s)
```

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

  >> select distinct category_code from sales_bucket;

  Output:

```
hive> select distinct category_code from sales_bucket;
Query ID = hadoop_20210423120133_e2ac0e5e-b17d-4f34-9204-97578396e6c7
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1619173061468_0006)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      6          6        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      5          5        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 63.50 s
--------------------------------------------------------------------------------
OK

accessories.cosmetic_bag
stationery.cartrige
accessories.bag
appliances.environment.vacuum
furniture.living_room.chair
sport.diving
appliances.personal.hair_cutter
appliances.environment.air_conditioner
apparel.glove
furniture.bathroom.bath
furniture.living_room.cabinet
Time taken: 64.192 seconds, Fetched: 12 row(s)
```

- Find the total number of products available under each category

>>select category_id, count(product_id) from sales_bucket group by category_id;

Output:

```
hive> select category_id, count(product_id)from sales_bucket group by category_id;
Query ID = hadoop_20210423121227_44fdd961-ea4a-44cc-bb62-8705f2033d59
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_1619173061468_0007)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      6          6        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      5          5        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 65.16 s
--------------------------------------------------------------------------------
OK
1487580004966466385     16
1487580005050352469     83278
1487580005176181595     127
1487580005369119587     3
1487580005570446188     24
1487580005671109489     300570
1487580005687886706     14
1487580005922767741     640
1487580006056985476     1794
1487580006073762693     7556
1487580006174425994     466
1487580006216369036     3
1487580006585467807     17
1487580007281722301     34854
1487580007432717250     64400
1487580007508214725     36
1487580007592100809     12450
1487580007852147670     42694
1487580007894090712     4957
1487580007910867929     51488
1487580007952810971     24742
1487580008053474272     2629
1487580008070251489     1643
1487580008087028706     778
1487580008221246441     46
1487580008472904691     65
```

Output part 2:

```
1487580012574933146     10450
1487580012616876188     1146
1487580012876923048     1608
1487580013011140782     29205
1487580013388628160     20437
1487580013472514244     7463
1487580013522845895     70148
1487580013539623112     17687
1487580013640286413     18309
1487580013799669974     3316
1487580013933887709     376
1487580014093271270     523
1495705810662064688     10994
1495705810729173554     5433
1542195323827388674     19035
1604427094756950459     2397
1658462125284131265     49543
1715102773747384334     3045
1725504706412807026     1819
1752742617696698537     25
1783999063314661546     39932
1783999067181810204     286
1783999071199952917     19733
1783999071325782053     5830
1810470908326838736     225
1897124478404526487     22374
1911999884991397970     394
1924049110428549877     20300
1958278551207674674     4957
1982860263572898112     6507
1998040849203594085     5995
2007399943458784057     18070
2035665444290953519     7792
2069171133327868014     2028
2093602042093240877     3188
2106514244487873093     1472
2134354356349173879     257
2141560642253881670     12861
2166295400451933025     11
2193074740493550411     1749
2193074740552270669     13772
2195085258339123402     25
Time taken: 73.298 seconds, Fetched: 500 row(s)
```

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

>> select brand, sum(price) as pr from sales_bucket where event_type=”purchase” group by brand order by pr desc limit 2;

Output:

```
hive> select brand, sum(price) as pr from sales_bucket where event_type="purchase" group by brand order by pr desc limit 2;
Query ID = hadoop_20210423121628_c87d4eb4-75b8-446c-abfb-bed1bde52b2c
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1619173061468_0007)

----------------------------------------------------------------------------------------------
        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: 22.18 s
----------------------------------------------------------------------------------------------
OK
        1094188.300000215
runail  148297.93999997302
Time taken: 23.234 seconds, Fetched: 2 row(s)
```

- Which brand increased their sales from October to November?

  >> select brand, sum(price) from sales_bucket where event_type=”purchase” and (month(event_time)=10)<(month(event_time)=11) group by brand;

  Output:

```
hive> select brand,sum(price) from sales_bucket where event_type="purchase" and (month(event_time)=10)<(month(event_time)=11)group by brand;
Query ID = hadoop_20210423122012_e477337e-1c62-4949-a0e0-5b69204aa12d
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1619173061468_0007)

----------------------------------------------------------------------------------------------
        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: 23.20 s
----------------------------------------------------------------------------------------------
OK
        619509.2400000421
airnails        5691.520000000035
almea   973.8699999999999
ardell  843.6500000000003
art-visage      2997.8000000000056
artex   4327.249999999996
aura    177.50999999999996
balbcare        212.37999999999997
barbie  12.39
batiste 874.1699999999998
beautix 12222.950000000003
beauty-free     1782.8599999999983
beautyblender   109.40999999999998
beauugreen      768.3499999999999
benovy  3259.970000000003
bergamo 144.3
bespecial       70.5
binacil 24.259999999999998
bioaqua 1398.1200000000001
biofollica      257.93999999999994
biore   90.31
```

Output part 2:

```
rocknailstar    1.9
rosi    3841.5600000000018
roubloff        4913.770000000011
runail  76758.66000000475
s.care  913.07
sanoto  1209.6799999999998
sawa    45.5
severina        6120.479999999961
shary   1176.4899999999996
shik    4839.719999999999
siberina        337.6500000000001
skinity 12.440000000000001
skinlite        890.4499999999999
skipofit        8.49
smart   5902.140000000005
soleo   212.52999999999986
solomeya        2685.7999999999947
sophin  1515.5200000000018
staleks 11875.61000000002
strong  38671.269999999975
sun     65.9
sunuv   8042.149999999995
supertan        66.51000000000003
swarovski       3043.159999999979
tannymaxx       171.28000000000003
tazol   7.18
tertio  245.80000000000007
thuya   2604.939999999999
tosowoong       27.3
treaclemoon     181.49
trind   542.96
uno     51039.75000000041
uskusi  5690.310000000033
veraclara       71.21000000000001
vilenta 231.20999999999998
vosev   316.7
weaver  6.48
yoko    11707.879999999988
ypsed   436.31999999999994
yu-r    673.7099999999999
zeitun  2009.63
zinger  6684.860000000028
Time taken: 23.916 seconds, Fetched: 214 row(s)
```

- Your company wants to reward the top 10 users of its websites 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 pr from sales_bucket where event_type=”purchase” group by user_id order by pr desc limit 10;

  Output:

```
hive> select user_id,sum(price)as pr from sales_bucket where event_type="purchase" group by user_id order by pr desc limit 10;
Query ID = hadoop_20210423122353_78bc6d77-c4e7-487b-874c-9c955d01342e
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1619173061468_0007)

--------------------------------------------------------------------------------
        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: 25.88 s
--------------------------------------------------------------------------------
OK
557790271       2715.8699999999953
150318419       1645.9699999999996
562167663       1352.8500000000001
531900924       1329.4499999999996
557850743       1295.4800000000005
522130011       1185.3899999999999
561592095       1109.7000000000005
431950134       1097.5899999999997
566576008       1056.3599999999997
521347209       1040.9100000000003
Time taken: 26.635 seconds, Fetched: 10 row(s)
```