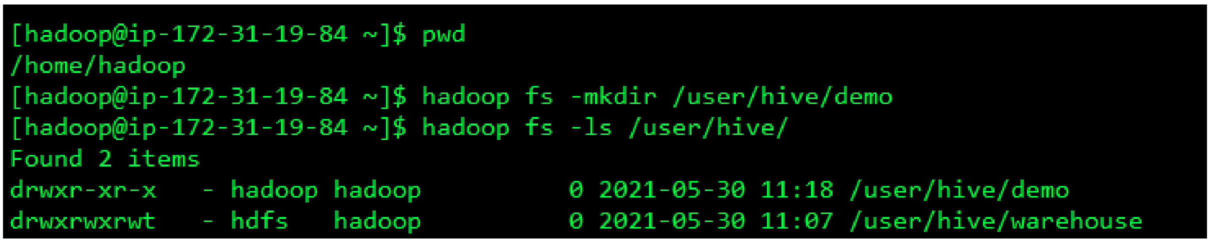
Assignment on Hive Case Study

Name: - Jaydeep Kumar Parida and Sharath Chandra VC

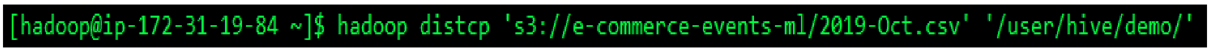
Date: - 31-05-2021

1. Creating folder in the Hadoop to store the files.

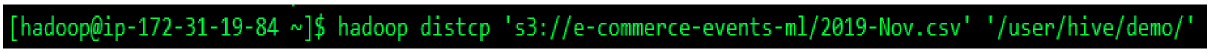


```
[hadoop@ip-172-31-19-84 ~]$ pwd
/home/hadoop
[hadoop@ip-172-31-19-84 ~]$ hadoop fs -mkdir /user/hive/demo
[hadoop@ip-172-31-19-84 ~]$ hadoop fs -ls /user/hive/
Found 2 items
drwxr-xr-x   - hadoop hadoop          0 2021-05-30 11:18 /user/hive/demo
drwxrwxrwt   - hdfs   hadoop          0 2021-05-30 11:07 /user/hive/warehouse
```

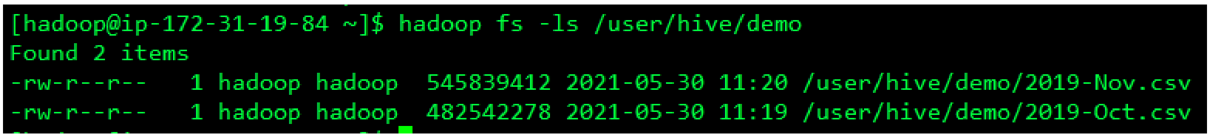
2. Files copied directly to the folder into Hadoop from S3.



```
[hadoop@ip-172-31-19-84 ~]$ hadoop distcp 's3://e-commerce-events-ml/2019-Oct.csv' '/user/hive/demo/'
```

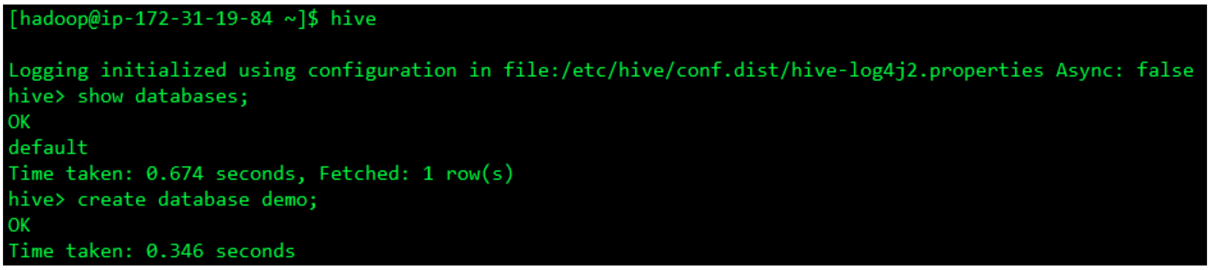


```
[hadoop@ip-172-31-19-84 ~]$ hadoop distcp 's3://e-commerce-events-ml/2019-Nov.csv' '/user/hive/demo/'
```

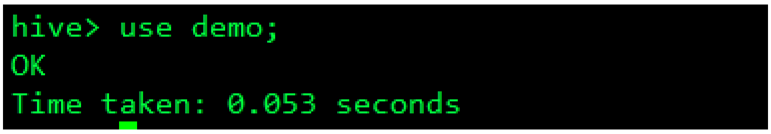


```
[hadoop@ip-172-31-19-84 ~]$ hadoop fs -ls /user/hive/demo
Found 2 items
-rw-r--r--   1 hadoop hadoop  545839412 2021-05-30 11:20 /user/hive/demo/2019-Nov.csv
-rw-r--r--   1 hadoop hadoop  482542278 2021-05-30 11:19 /user/hive/demo/2019-Oct.csv
```

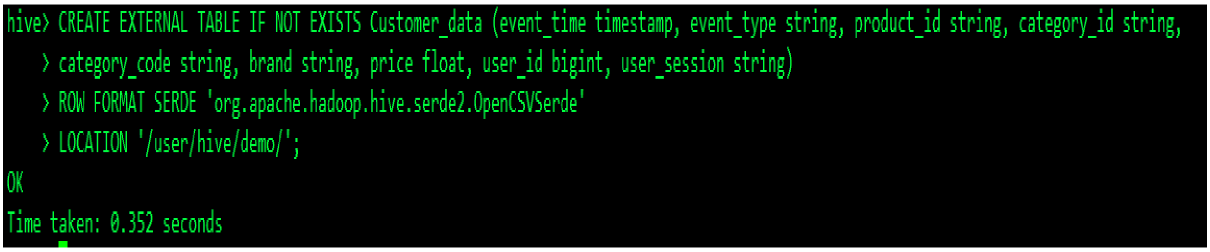
3. Invoking hive session, creating database and using the database for creation of table (Customer_data) into it.



```
[hadoop@ip-172-31-19-84 ~]$ hive

Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j2.properties Async: false
hive> show databases;
OK
default
Time taken: 0.674 seconds, Fetched: 1 row(s)
hive> create database demo;
OK
Time taken: 0.346 seconds
```

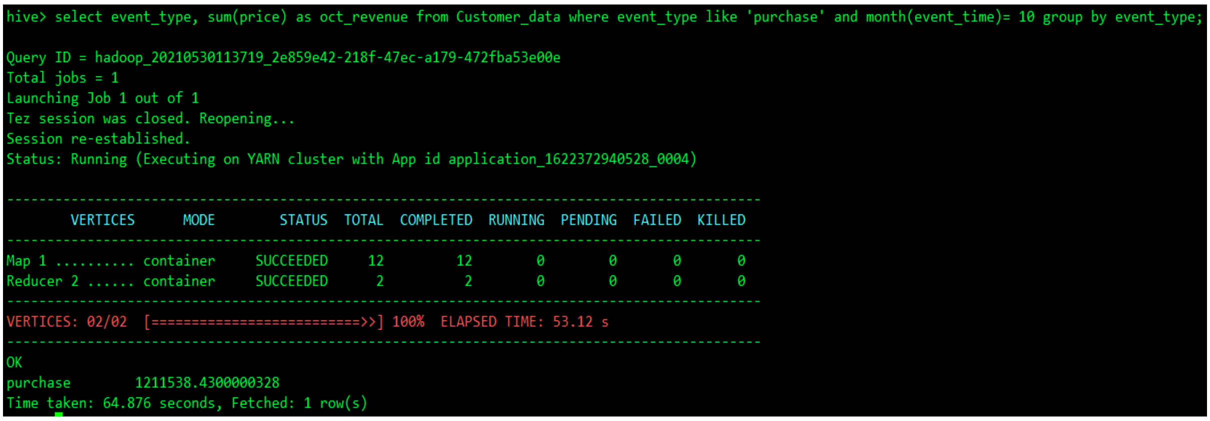


```
hive> use demo;
OK
Time taken: 0.053 seconds
```



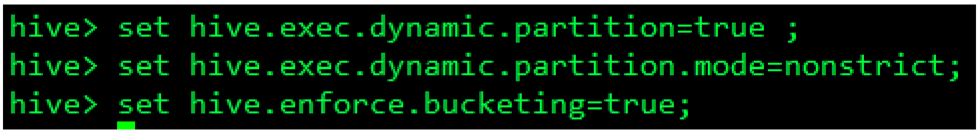
```
hive> CREATE EXTERNAL TABLE IF NOT EXISTS Customer_data (event_time timestamp, event_type string, product_id string, category_id string,
    > category_code string, brand string, price float, user_id bigint, user_session string)
    > ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.OpenCSVSerde'
    > LOCATION '/user/hive/demo/';
OK
Time taken: 0.352 seconds
```

4. Running the 1st assignment query to observe the time taken to run the query without optimized table (Query runtime 64.876 seconds).



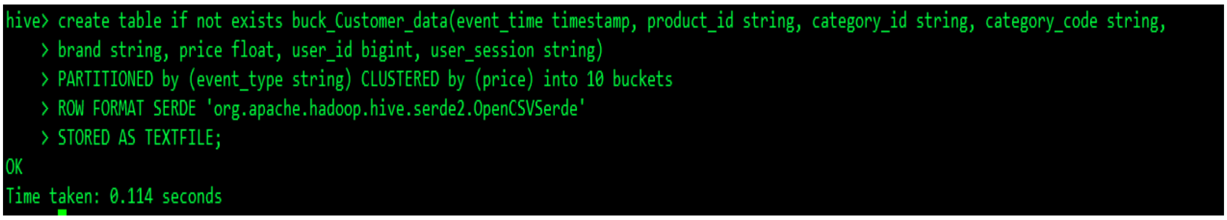
```
hive> select event_type, sum(price) as oct_revenue from Customer_data where event_type like 'purchase' and month(event_time)= 10 group by event_type;

Query ID = hadoop_20210530113719_2e859e42-218f-47ec-a179-472fba53e00e
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_1622372940528_0004)

----------------------------------------------------------------------------------------------
        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: 53.12 s
----------------------------------------------------------------------------------------------
OK
purchase        1211538.4300000328
Time taken: 64.876 seconds, Fetched: 1 row(s)
```

5. To turn ON the partitioning and bucketing on hive.



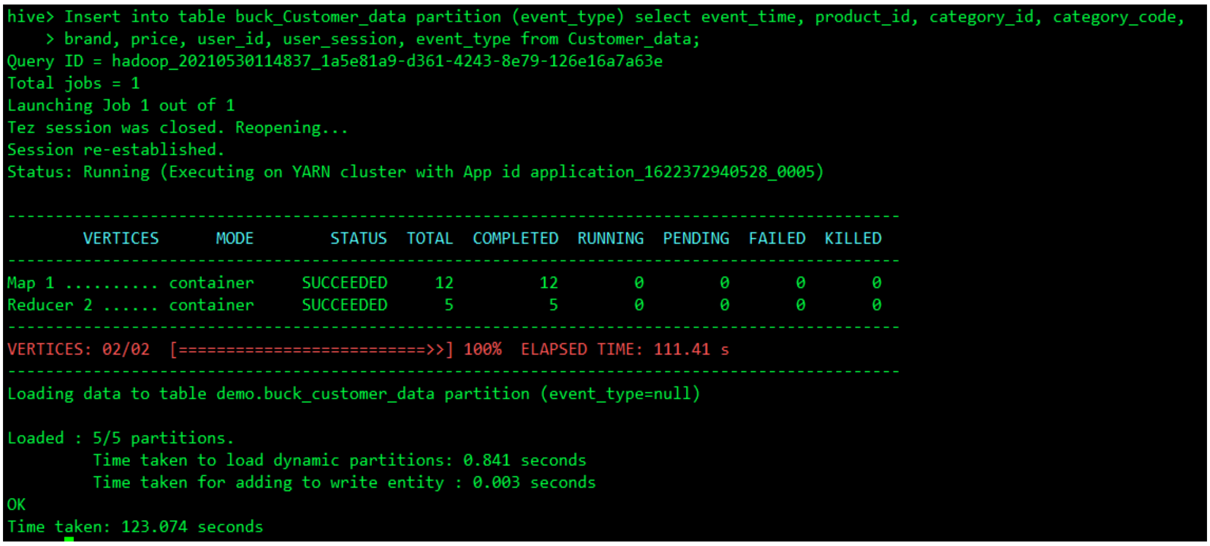
```
hive> set hive.exec.dynamic.partition=true ;
hive> set hive.exec.dynamic.partition.mode=nonstrict;
hive> set hive.enforce.bucketing=true;
```

6. Creating a table with Partitioning and bucketing for optimizing the query time (buck_Customer_data).



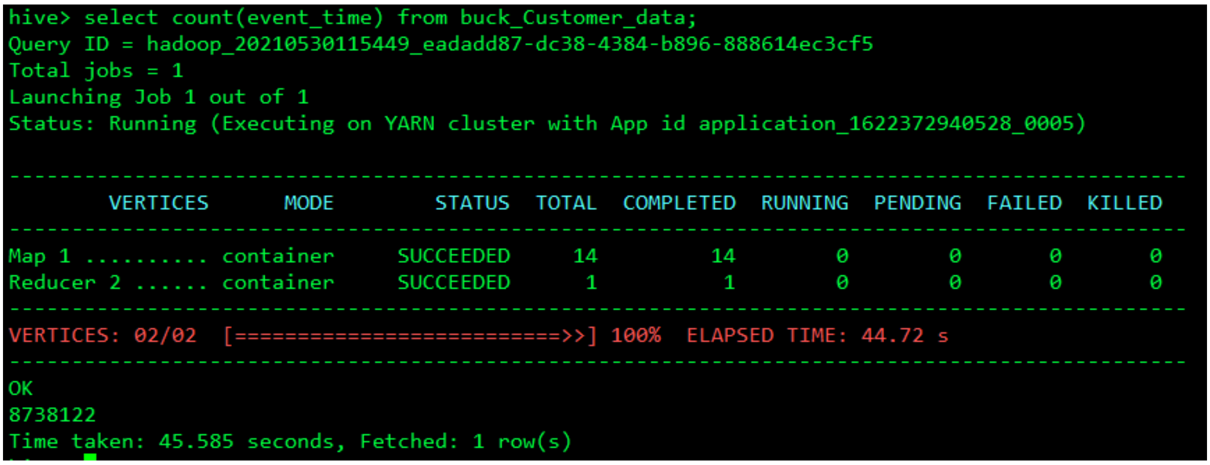
```
hive> create table if not exists buck_Customer_data(event_time timestamp, product_id string, category_id string, category_code string,
    > brand string, price float, user_id bigint, user_session string)
    > PARTITIONED by (event_type string) CLUSTERED by (price) into 10 buckets
    > ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.OpenCSVSerde'
    > STORED AS TEXTFILE;
OK
Time taken: 0.114 seconds
```

7. Inserting the values into the buck_Customer_data from the Customer_data.



```
hive> Insert into table buck_Customer_data partition (event_type) select event_time, product_id, category_id, category_code,
    > brand, price, user_id, user_session, event_type from Customer_data;
Query ID = hadoop_20210530114837_1a5e81a9-d361-4243-8e79-126e16a7a63e
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_1622372940528_0005)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED     12         12        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      5          5        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 111.41 s
----------------------------------------------------------------------------------------------
Loading data to table demo.buck_customer_data partition (event_type=null)

Loaded : 5/5 partitions.
         Time taken to load dynamic partitions: 0.841 seconds
         Time taken for adding to write entity : 0.003 seconds
OK
Time taken: 123.074 seconds
```

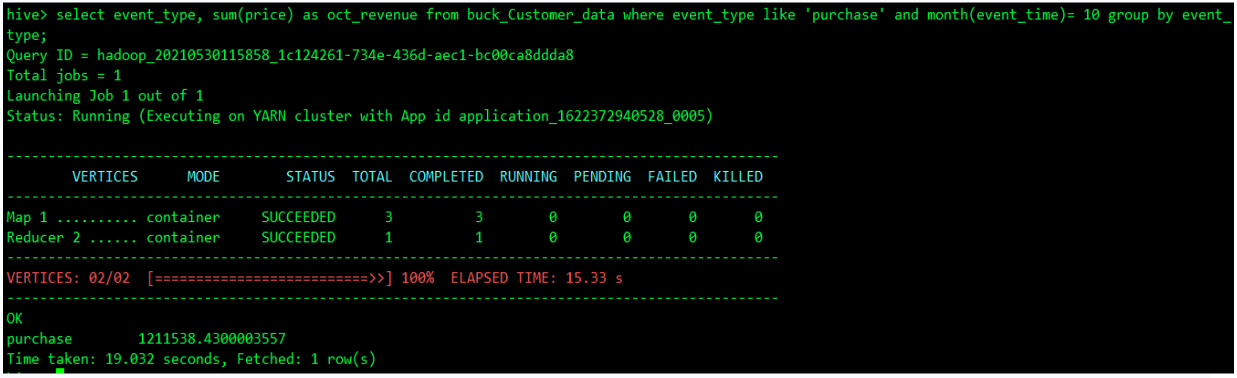
8. Checking the number of rows on the table.



```
hive> select count(event_time) from buck_Customer_data;
Query ID = hadoop_20210530115449_eadadd87-dc38-4384-b896-888614ec3cf5
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1622372940528_0005)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED     14         14        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 44.72 s
----------------------------------------------------------------------------------------------
OK
8738122
Time taken: 45.585 seconds, Fetched: 1 row(s)
```

9.1. Running the 1st assignment query on the optimised table (Query runtime 19.032 seconds)

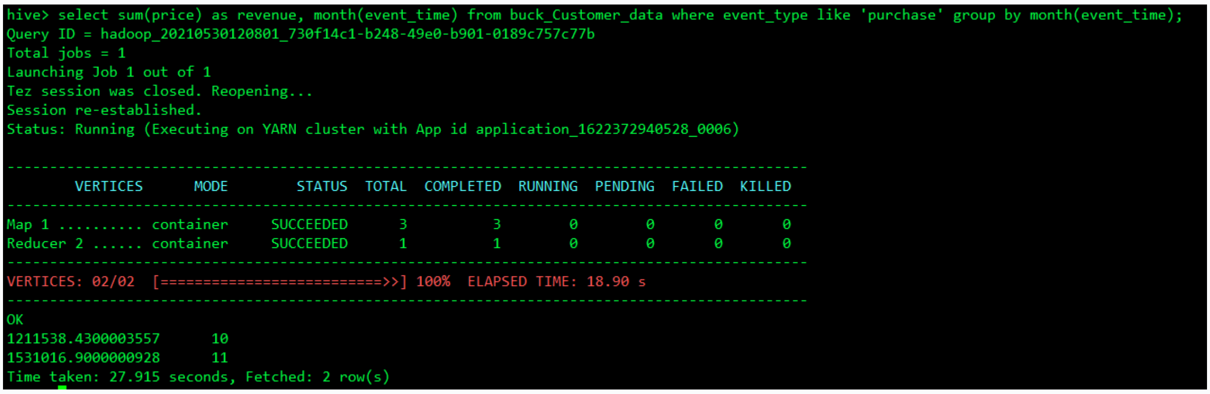
Find the total revenue generated due to purchases made in October.



```
hive> select event_type, sum(price) as oct_revenue from buck_Customer_data where event_type like 'purchase' and month(event_time)= 10 group by event_
type;
Query ID = hadoop_20210530115858_1c124261-734e-436d-aec1-bc00ca8ddda8
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1622372940528_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: 15.33 s
----------------------------------------------------------------------------------------------
OK
purchase        1211538.4300003557
Time taken: 19.032 seconds, Fetched: 1 row(s)
```

From the above query 1211538.43 is the revenue generated due to purchase in October.

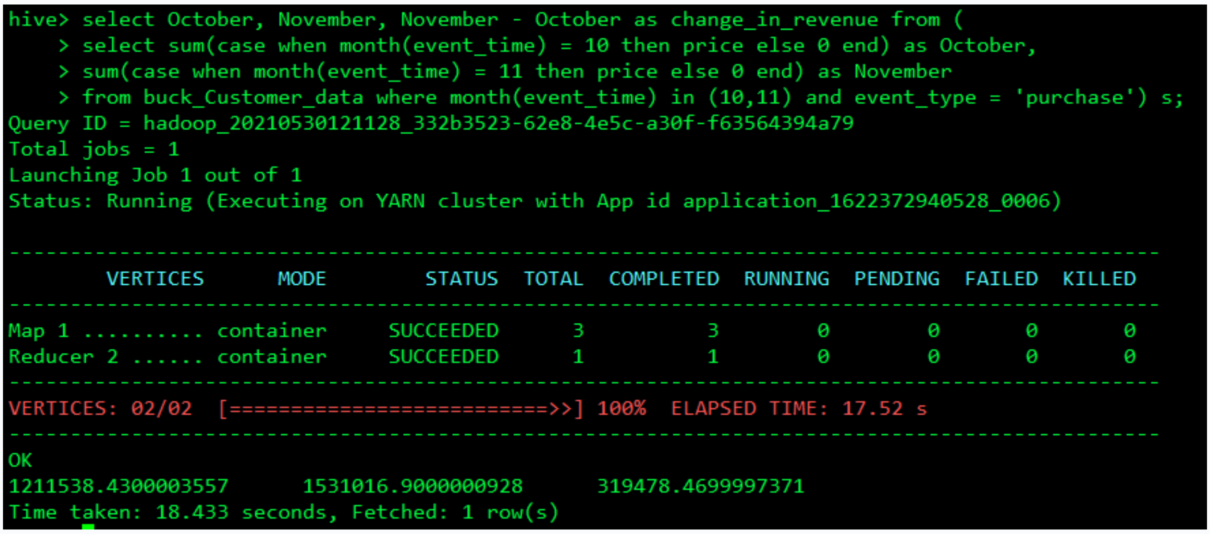
9.2. Write a query to yield the total sum of purchases per month in a single output.



```
hive> select sum(price) as revenue, month(event_time) from buck_Customer_data where event_type like 'purchase' group by month(event_time);
Query ID = hadoop_20210530120801_730f14c1-b248-49e0-b901-0189c757c77b
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_1622372940528_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: 18.90 s
----------------------------------------------------------------------------------------------
OK
1211538.4300003557      10
1531016.9000000928      11
Time taken: 27.915 seconds, Fetched: 2 row(s)
```

From the above query 1211538.43 for the month of October and 1531016.90 for month of November the total revenue was generated.

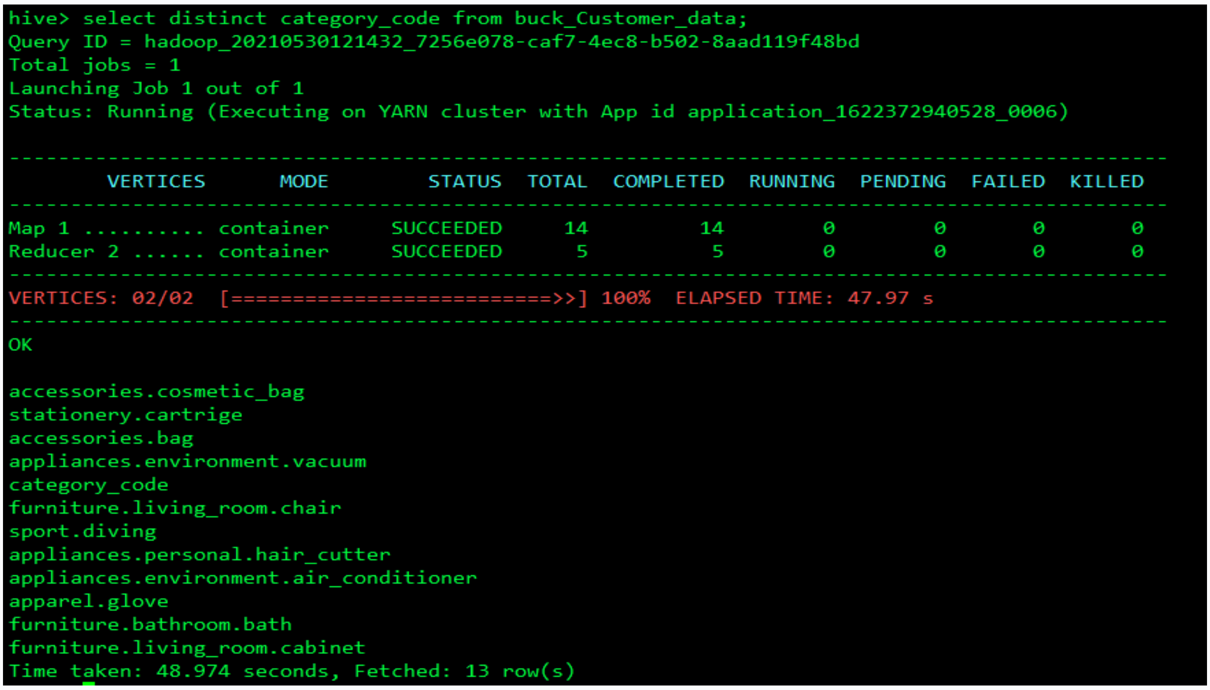
9.3. Write a query to find the change in revenue generated due to purchases from October to November.



```
hive> select October, November, November - October as change_in_revenue from (
    > select sum(case when month(event_time) = 10 then price else 0 end) as October,
    > sum(case when month(event_time) = 11 then price else 0 end) as November
    > from buck_Customer_data where month(event_time) in (10,11) and event_type = 'purchase') s;
Query ID = hadoop_20210530121128_332b3523-62e8-4e5c-a30f-f63564394a79
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1622372940528_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: 17.52 s
----------------------------------------------------------------------------------------------
OK
1211538.4300003557      1531016.9000000928      319478.4699997371
Time taken: 18.433 seconds, Fetched: 1 row(s)
```

From the above query the change in revenue generated was 319478.46 from October to November.

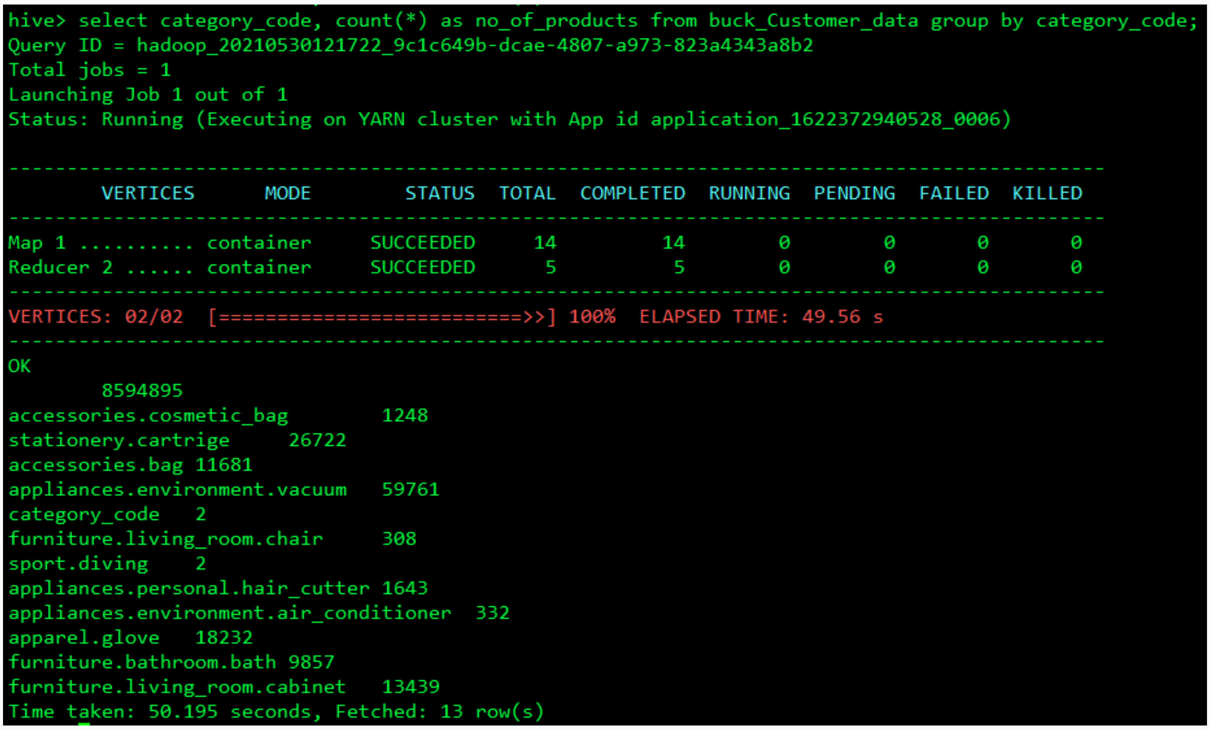
9.4. Find distinct categories of products. Categories with null category code can be ignored.



```
hive> select distinct category_code from buck_Customer_data;
Query ID = hadoop_20210530121432_7256e078-caf7-4ec8-b502-8aad119f48bd
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1622372940528_0006)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED     14         14        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      5          5        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 47.97 s
----------------------------------------------------------------------------------------------
OK

accessories.cosmetic_bag
stationery.cartrige
accessories.bag
appliances.environment.vacuum
category_code
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: 48.974 seconds, Fetched: 13 row(s)
```

From the above query it is observed that 13 categories were generated including the null category value.

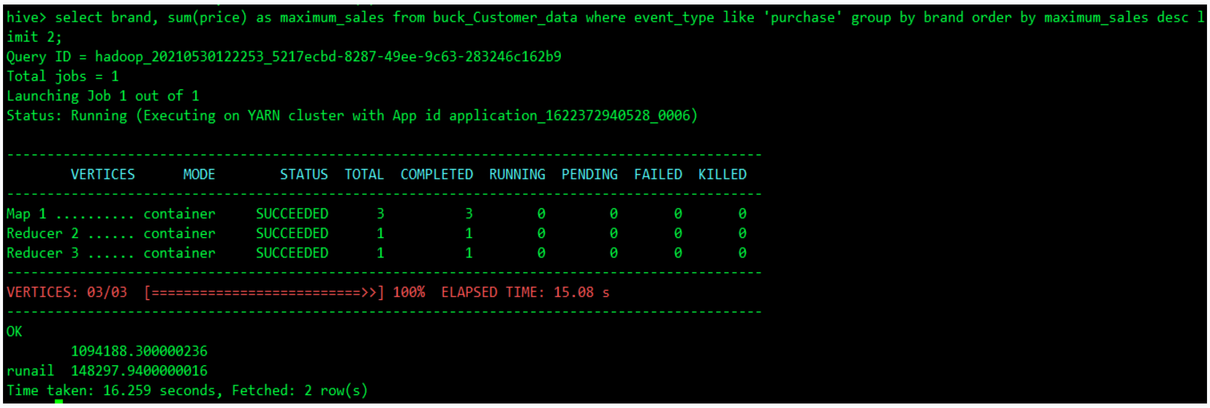
9.5. Find the total number of products available under each category.



```
hive> select category_code, count(*) as no_of_products from buck_Customer_data group by category_code;
Query ID = hadoop_20210530121722_9c1c649b-dcae-4807-a973-823a4343a8b2
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1622372940528_0006)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED     14         14        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      5          5        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 49.56 s
----------------------------------------------------------------------------------------------
OK
        8594895
accessories.cosmetic_bag        1248
stationery.cartrige     26722
accessories.bag 11681
appliances.environment.vacuum   59761
category_code   2
furniture.living_room.chair     308
sport.diving    2
appliances.personal.hair_cutter 1643
appliances.environment.air_conditioner  332
apparel.glove   18232
furniture.bathroom.bath 9857
furniture.living_room.cabinet   13439
Time taken: 50.195 seconds, Fetched: 13 row(s)
```

From the above it is observed that for the null category value 8594894 products were registered.

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

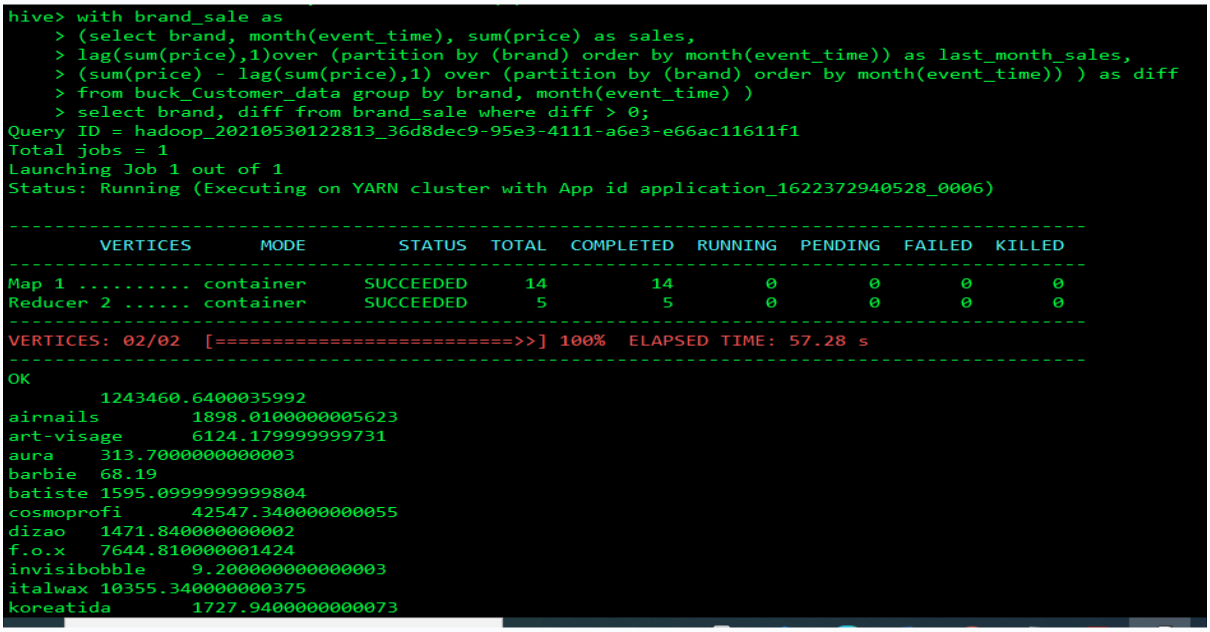


```
hive> select brand, sum(price) as maximum_sales from buck_Customer_data where event_type like 'purchase' group by brand order by maximum_sales desc l
imit 2;
Query ID = hadoop_20210530122253_5217ecbd-8287-49ee-9c63-283246c162b9
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1622372940528_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
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 15.08 s
----------------------------------------------------------------------------------------------
OK
        1094188.300000236
runail  148297.9400000016
Time taken: 16.259 seconds, Fetched: 2 row(s)
```

Hence from the query it is observed that maximum sales brand has the null value and the second most brand is “runail”.

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

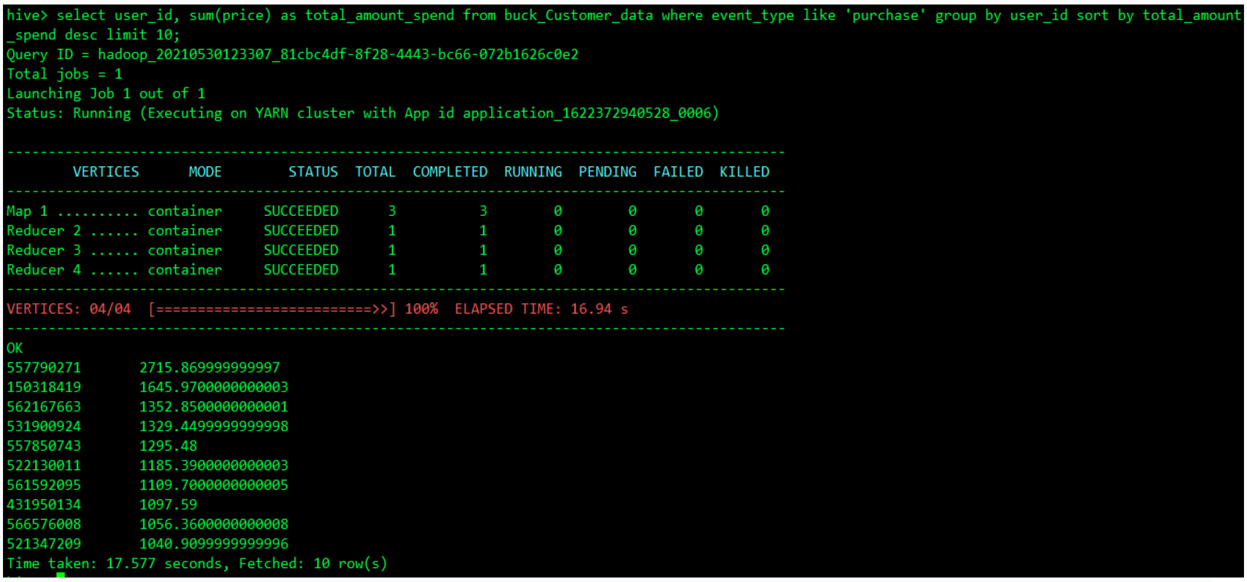
For the below query the record fetched are large hence, updating the 1st part of the screenshot.



```
hive> with brand_sale as
    > (select brand, month(event_time), sum(price) as sales,
    > lag(sum(price),1)over (partition by (brand) order by month(event_time)) as last_month_sales,
    > (sum(price) - lag(sum(price),1) over (partition by (brand) order by month(event_time)) ) as diff
    > from buck_Customer_data group by brand, month(event_time) )
    > select brand, diff from brand_sale where diff > 0;
Query ID = hadoop_20210530122813_36d8dec9-95e3-4111-a6e3-e66ac11611f1
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1622372940528_0006)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED     14         14        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      5          5        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 57.28 s
----------------------------------------------------------------------------------------------
OK
        1243460.6400035992
airnails        1898.0100000005623
art-visage      6124.179999999731
aura    313.70000000000003
barbie  68.19
batiste 1595.0999999999804
cosmoprofi      42547.340000000055
dizao   1471.840000000002
f.o.x   7644.810000001424
invisibobble    9.200000000000003
italwax 10355.340000000375
koreatida       1727.9400000000073
```

From the above it observed that null value brand acquire the top sales with 1243460.64 following the “airnails” with 1898.01.

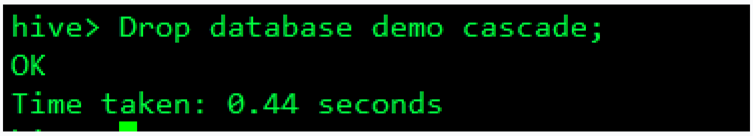
9.8. Your company wants to reward the top 10 users of its website with a Golden Customer plan. Write a query to generate a list of top 10 users who spend the most.



```
hive> select user_id, sum(price) as total_amount_spend from buck_Customer_data where event_type like 'purchase' group by user_id sort by total_amount
_spend desc limit 10;
Query ID = hadoop_20210530123307_81cbc4df-8f28-4443-bc66-072b1626c0e2
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1622372940528_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
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: 16.94 s
----------------------------------------------------------------------------------------------
OK
557790271       2715.869999999997
150318419       1645.9700000000003
562167663       1352.8500000000001
531900924       1329.4499999999998
557850743       1295.48
522130011       1185.3900000000003
561592095       1109.7000000000005
431950134       1097.59
566576008       1056.3600000000008
521347209       1040.9099999999996
Time taken: 17.577 seconds, Fetched: 10 row(s)
```

Top 10 user Id has been generated who spend the most amount in the month of October and November.

10. To drop the database (Cascade is used to drop the database as the database is not empty).



```
hive> Drop database demo cascade;
OK
Time taken: 0.44 seconds
```

From the above case study it is observed that hive plays a vital role in fetching the selected records from total records of 8738122, when the table is properly partitioned and bucked the query time is optimised.