HIVE CASE STUDY

BY PRADYUMNA DESHPANDE

&

DHARITRI SENAPATI

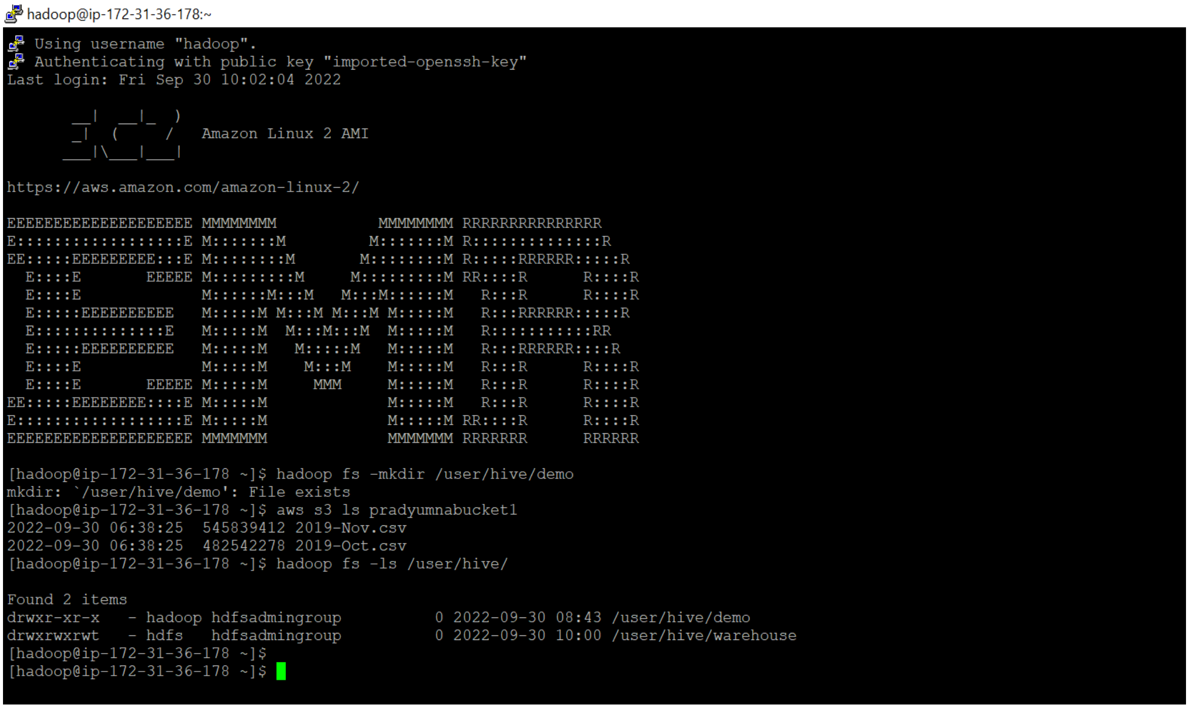
Problem Definition:

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. 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.

For this assignment, you will be working with a public click stream dataset of a cosmetics store. Using this dataset, your job is to extract valuable insights which generally data engineers come up within an e-retail company.

STEP 1 :

Connecting Cluster via Putty :

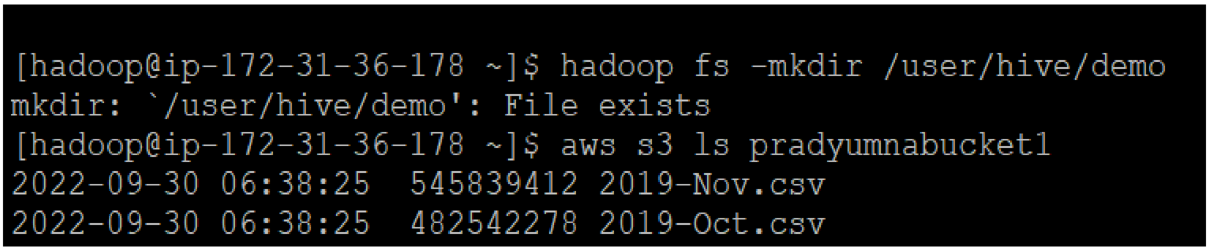


```
hadoop@ip-172-31-36-178:~
 Using username "hadoop".
 Authenticating with public key "imported-openssh-key"
Last login: Fri Sep 30 10:02:04 2022

       __|  __|_  )
       _|  (     /   Amazon Linux 2 AMI
      ___|\___|___|

https://aws.amazon.com/amazon-linux-2/

EEEEEEEEEEEEEEEEEEEE MMMMMMMM           MMMMMMMM RRRRRRRRRRRRRRR
E::::::::::::::::::E M:::::::M         M:::::::M R::::::::::::::R
EE:::::EEEEEEEEE:::E M::::::::M       M::::::::M R:::::RRRRRR:::::R
  E::::E       EEEEE M:::::::::M     M:::::::::M RR::::R      R::::R
  E::::E             M::::::M:::M   M:::M::::::M   R:::R      R::::R
  E:::::EEEEEEEEEE   M:::::M M:::M M:::M M:::::M   R:::RRRRRR:::::R
  E::::::::::::::E   M:::::M  M:::M:::M  M:::::M   R:::::::::::RR
  E:::::EEEEEEEEEE   M:::::M   M:::::M   M:::::M   R:::RRRRRR::::R
  E::::E             M:::::M    M:::M    M:::::M   R:::R      R::::R
  E::::E       EEEEE M:::::M     MMM     M:::::M   R:::R      R::::R
EE:::::EEEEEEEE::::E M:::::M             M:::::M   R:::R      R::::R
E::::::::::::::::::E M:::::M             M:::::M RR::::R      R::::R
EEEEEEEEEEEEEEEEEEEE MMMMMMM             MMMMMMM RRRRRRR      RRRRRR

[hadoop@ip-172-31-36-178 ~]$ hadoop fs -mkdir /user/hive/demo
mkdir: `/user/hive/demo': File exists
[hadoop@ip-172-31-36-178 ~]$ aws s3 ls pradyumnabucket1
2022-09-30 06:38:25  545839412 2019-Nov.csv
2022-09-30 06:38:25  482542278 2019-Oct.csv
[hadoop@ip-172-31-36-178 ~]$ hadoop fs -ls /user/hive/

Found 2 items
drwxr-xr-x   - hadoop hdfsadmingroup          0 2022-09-30 08:43 /user/hive/demo
drwxrwxrwt   - hdfs   hdfsadmingroup          0 2022-09-30 10:00 /user/hive/warehouse
[hadoop@ip-172-31-36-178 ~]$
[hadoop@ip-172-31-36-178 ~]$
```

Copying data into HDFS :

A. Creating directory in HDFS :
   Command : hadoop fs -mkdir /user/hive/demo

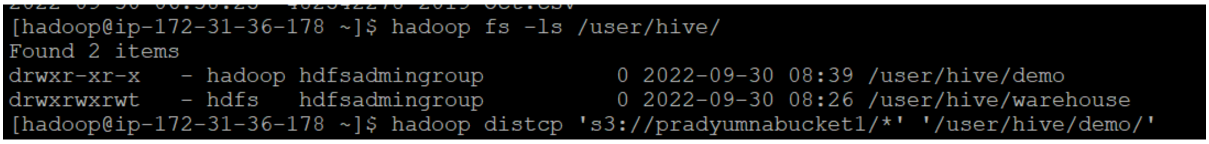
B. To access the public s3 bucket:
   Command : aws s3 ls pradyumnabucket1



```
[hadoop@ip-172-31-36-178 ~]$ hadoop fs -mkdir /user/hive/demo
mkdir: `/user/hive/demo': File exists
[hadoop@ip-172-31-36-178 ~]$ aws s3 ls pradyumnabucket1
2022-09-30 06:38:25  545839412 2019-Nov.csv
2022-09-30 06:38:25  482542278 2019-Oct.csv
```

C. Checking the available directory :

Command: hadoop fs -ls /user/hive/

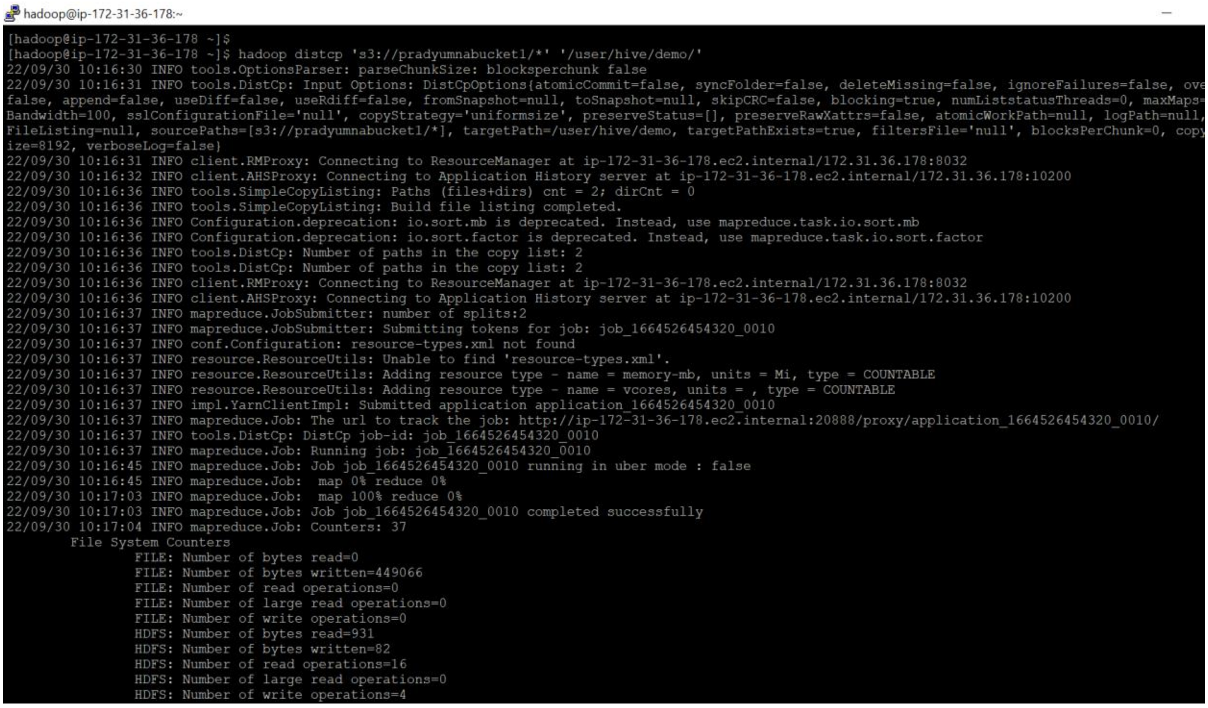


```
[hadoop@ip-172-31-36-178 ~]$ hadoop fs -ls /user/hive/
Found 2 items
drwxr-xr-x   - hadoop hdfsadmingroup          0 2022-09-30 08:39 /user/hive/demo
drwxrwxrwt   - hdfs   hdfsadmingroup          0 2022-09-30 08:26 /user/hive/warehouse
[hadoop@ip-172-31-36-178 ~]$ hadoop distcp 's3://pradyumnabucket1/*' '/user/hive/demo/'
```

D. Loading the s3 dataset to created directory 'demo' in hadoop

Command: hadoop distcp 's3://pradyumnabucket1/* '

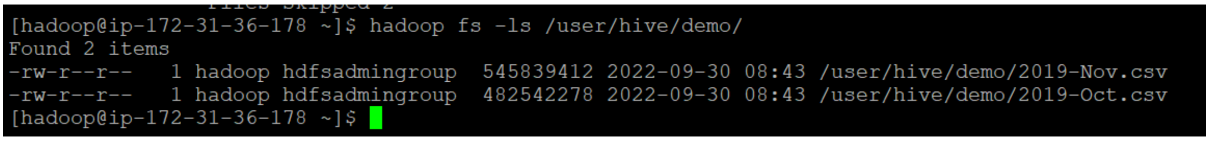
'/user/hive/demo/'



```
[hadoop@ip-172-31-36-178 ~]$
[hadoop@ip-172-31-36-178 ~]$ hadoop distcp 's3://pradyumnabucket1/*' '/user/hive/demo/'
22/09/30 10:16:30 INFO tools.OptionsParser: parseChunkSize: blocksperchunk false
22/09/30 10:16:31 INFO tools.DistCp: Input Options: DistCpOptions{atomicCommit=false, syncFolder=false, deleteMissing=false, ignoreFailures=false, ove
false, append=false, useDiff=false, useRdiff=false, fromSnapshot=null, toSnapshot=null, skipCRC=false, blocking=true, numListstatusThreads=0, maxMaps=
Bandwidth=100, sslConfigurationFile='null', copyStrategy='uniformsize', preserveStatus=[], preserveRawXattrs=false, atomicWorkPath=null, logPath=null,
FileListing=null, sourcePaths=[s3://pradyumnabucket1/*], targetPath=/user/hive/demo, targetPathExists=true, filtersFile='null', blocksPerChunk=0, copy
ize=8192, verboseLog=false}
22/09/30 10:16:31 INFO client.RMProxy: Connecting to ResourceManager at ip-172-31-36-178.ec2.internal/172.31.36.178:8032
22/09/30 10:16:32 INFO client.AHSProxy: Connecting to Application History server at ip-172-31-36-178.ec2.internal/172.31.36.178:10200
22/09/30 10:16:36 INFO tools.SimpleCopyListing: Paths (files+dirs) cnt = 2; dirCnt = 0
22/09/30 10:16:36 INFO tools.SimpleCopyListing: Build file listing completed.
22/09/30 10:16:36 INFO Configuration.deprecation: io.sort.mb is deprecated. Instead, use mapreduce.task.io.sort.mb
22/09/30 10:16:36 INFO Configuration.deprecation: io.sort.factor is deprecated. Instead, use mapreduce.task.io.sort.factor
22/09/30 10:16:36 INFO tools.DistCp: Number of paths in the copy list: 2
22/09/30 10:16:36 INFO tools.DistCp: Number of paths in the copy list: 2
22/09/30 10:16:36 INFO client.RMProxy: Connecting to ResourceManager at ip-172-31-36-178.ec2.internal/172.31.36.178:8032
22/09/30 10:16:36 INFO client.AHSProxy: Connecting to Application History server at ip-172-31-36-178.ec2.internal/172.31.36.178:10200
22/09/30 10:16:37 INFO mapreduce.JobSubmitter: number of splits:2
22/09/30 10:16:37 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1664526454320_0010
22/09/30 10:16:37 INFO conf.Configuration: resource-types.xml not found
22/09/30 10:16:37 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
22/09/30 10:16:37 INFO resource.ResourceUtils: Adding resource type - name = memory-mb, units = Mi, type = COUNTABLE
22/09/30 10:16:37 INFO resource.ResourceUtils: Adding resource type - name = vcores, units = , type = COUNTABLE
22/09/30 10:16:37 INFO impl.YarnClientImpl: Submitted application application_1664526454320_0010
22/09/30 10:16:37 INFO mapreduce.Job: The url to track the job: http://ip-172-31-36-178.ec2.internal:20888/proxy/application_1664526454320_0010/
22/09/30 10:16:37 INFO tools.DistCp: DistCp job-id: job_1664526454320_0010
22/09/30 10:16:37 INFO mapreduce.Job: Running job: job_1664526454320_0010
22/09/30 10:16:45 INFO mapreduce.Job: Job job_1664526454320_0010 running in uber mode : false
22/09/30 10:16:45 INFO mapreduce.Job:  map 0% reduce 0%
22/09/30 10:17:03 INFO mapreduce.Job:  map 100% reduce 0%
22/09/30 10:17:03 INFO mapreduce.Job: Job job_1664526454320_0010 completed successfully
22/09/30 10:17:04 INFO mapreduce.Job: Counters: 37
        File System Counters
                FILE: Number of bytes read=0
                FILE: Number of bytes written=449066
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=931
                HDFS: Number of bytes written=82
                HDFS: Number of read operations=16
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=4
```

E. After loading the dataset we have used following command to check the dataset file in the hadoop directory

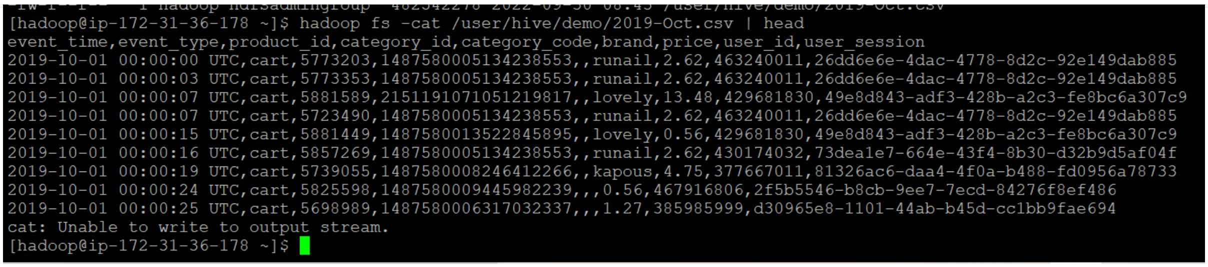
Command: hadoop fs -ls /user/hive/demo/



```
[hadoop@ip-172-31-36-178 ~]$ hadoop fs -ls /user/hive/demo/
Found 2 items
-rw-r--r--   1 hadoop hdfsadmingroup  545839412 2022-09-30 08:43 /user/hive/demo/2019-Nov.csv
-rw-r--r--   1 hadoop hdfsadmingroup  482542278 2022-09-30 08:43 /user/hive/demo/2019-Oct.csv
[hadoop@ip-172-31-36-178 ~]$
```

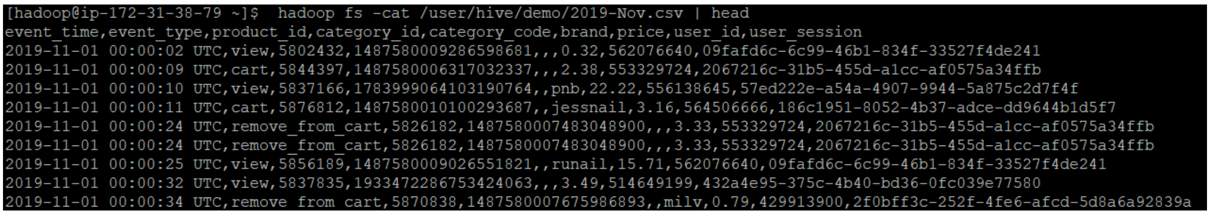
F. We have used below command to check the saved data set in the hadoop directory.

Command: hadoop fs -cat /user/hive/demo/2019-Oct.csv | head



```
[hadoop@ip-172-31-36-178 ~]$ hadoop fs -cat /user/hive/demo/2019-Oct.csv | head
event_time,event_type,product_id,category_id,category_code,brand,price,user_id,user_session
2019-10-01 00:00:00 UTC,cart,5773203,1487580005134238553,,runail,2.62,463240011,26dd6e6e-4dac-4778-8d2c-92e149dab885
2019-10-01 00:00:03 UTC,cart,5773353,1487580005134238553,,runail,2.62,463240011,26dd6e6e-4dac-4778-8d2c-92e149dab885
2019-10-01 00:00:07 UTC,cart,5881589,2151191071051219817,,lovely,13.48,429681830,49e8d843-adf3-428b-a2c3-fe8bc6a307c9
2019-10-01 00:00:07 UTC,cart,5723490,1487580005134238553,,runail,2.62,463240011,26dd6e6e-4dac-4778-8d2c-92e149dab885
2019-10-01 00:00:15 UTC,cart,5881449,1487580013522845895,,lovely,0.56,429681830,49e8d843-adf3-428b-a2c3-fe8bc6a307c9
2019-10-01 00:00:16 UTC,cart,5857269,1487580005134238553,,runail,2.62,430174032,73dea1e7-664e-43f4-8b30-d32b9d5af04f
2019-10-01 00:00:19 UTC,cart,5739055,1487580008246412266,,kapous,4.75,377667011,81326ac6-daa4-4f0a-b488-fd0956a78733
2019-10-01 00:00:24 UTC,cart,5825598,1487580009445982239,,,0.56,467916806,2f5b5546-b8cb-9ee7-7ecd-84276f8ef486
2019-10-01 00:00:25 UTC,cart,5698989,1487580006317032337,,,1.27,385985999,d30965e8-1101-44ab-b45d-cc1bb9fae694
cat: Unable to write to output stream.
[hadoop@ip-172-31-36-178 ~]$
```

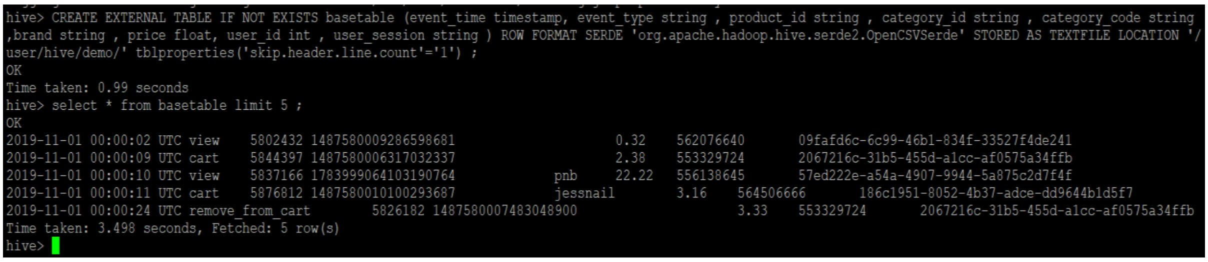
Command: hadoop fs -cat /user/hive/demo/2019-Nov.csv | head



```
[hadoop@ip-172-31-38-79 ~]$  hadoop fs -cat /user/hive/demo/2019-Nov.csv | head
event_time,event_type,product_id,category_id,category_code,brand,price,user_id,user_session
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
2019-11-01 00:00:24 UTC,remove_from_cart,5826182,1487580007483048900,,,3.33,553329724,2067216c-31b5-455d-a1cc-af0575a34ffb
2019-11-01 00:00:25 UTC,view,5856189,1487580009026551821,,runail,15.71,562076640,09fafd6c-6c99-46b1-834f-33527f4de241
2019-11-01 00:00:32 UTC,view,5837835,1933472286753424063,,,3.49,514649199,432a4e95-375c-4b40-bd36-0fc039e77580
2019-11-01 00:00:34 UTC,remove_from_cart,5870838,1487580007675986893,,milv,0.79,429913900,2f0bff3c-252f-4fe6-afcd-5d8a6a92839a
```

G. After moving the data to the directory we create the base table and check for the data inthe table.

Command: CREATE EXTERNAL TABLE IF NOT EXISTS basetable (event_time timestamp, event_type string , product_id string , category_id string , category_code string ,brand string , price float, user_id int , user_session string) ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.OpenCSVSerde' STORED AS TEXTFILE LOCATION '/user/hive/demo/' tblproperties('skip.header.line.count'='1') ;



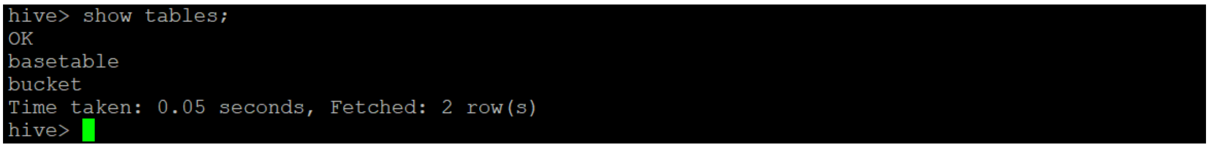
```
hive> CREATE EXTERNAL TABLE IF NOT EXISTS basetable (event_time timestamp, event_type string , product_id string , category_id string , category_code string
,brand string , price float, user_id int , user_session string ) ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.OpenCSVSerde' STORED AS TEXTFILE LOCATION '/
user/hive/demo/' tblproperties('skip.header.line.count'='1') ;
OK
Time taken: 0.99 seconds
hive> select * from basetable 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: 3.498 seconds, Fetched: 5 row(s)
hive>
```

H. We will be using optimization on one of the queries by partitioning and bucketing.

Command: create table if not exists bucket (event_time string, 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 (category_code) into 13 buckets ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.OpenCSVSerde' STORED AS TEXTFILE LOCATION '/user/hive/demo/' tblproperties('skip.header.line.count'='1') ;

I. Once the table is created check for the created tables.

Command: show tables;



```
hive> show tables;
OK
basetable
bucket
Time taken: 0.05 seconds, Fetched: 2 row(s)
hive>
```

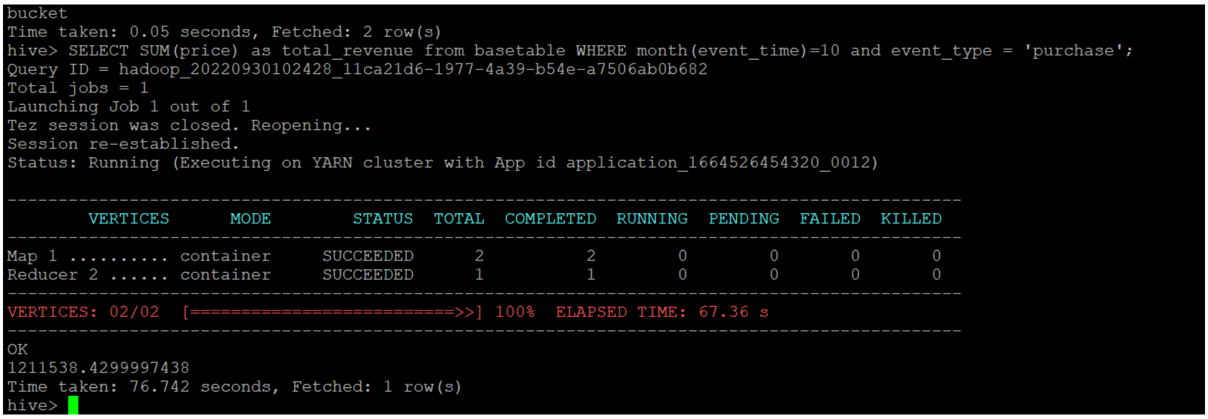
Query Analysis:

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

SELECT SUM(price) as total_revenue

from basetable

WHERE month(event_time)=10 and event_type = 'purchase';

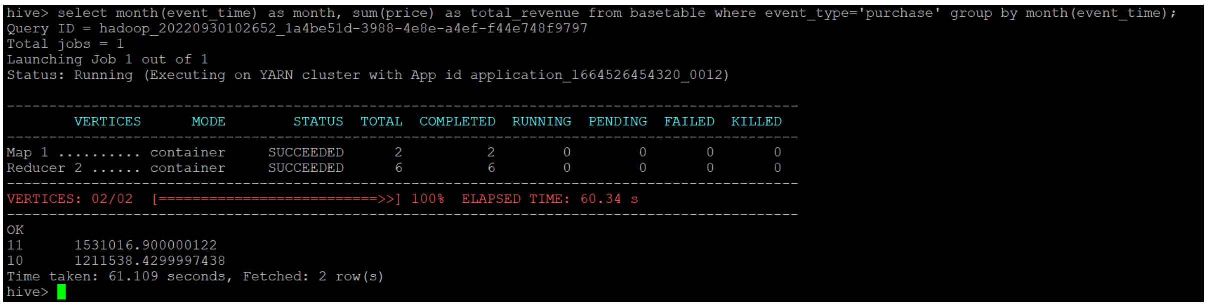


```
bucket
Time taken: 0.05 seconds, Fetched: 2 row(s)
hive> SELECT SUM(price) as total_revenue from basetable WHERE month(event_time)=10 and event_type = 'purchase';
Query ID = hadoop_20220930102428_11ca21d6-1977-4a39-b54e-a7506ab0b682
Total jobs = 1
Launching Job 1 out of 1
Tez session was closed. Reopening...
Session re-established.
Status: Running (Executing on YARN cluster with App id application_1664526454320_0012)

----------------------------------------------------------------------------------------------
        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: 67.36 s
----------------------------------------------------------------------------------------------
OK
1211538.4299997438
Time taken: 76.742 seconds, Fetched: 1 row(s)
hive>
```

Insight: Total Revenue is 1211538.4299997438

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

select month(event_time) as month, sum(price) as total_revenue from basetable where event_type='purchase' group by month(event_time) ;

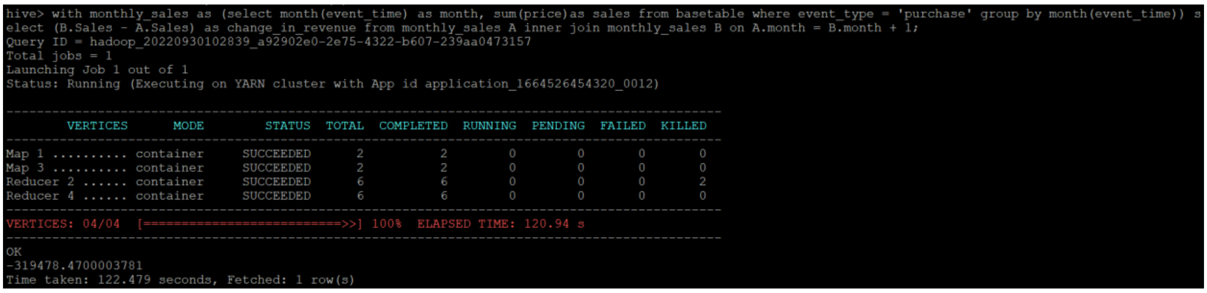


```
hive> select month(event_time) as month, sum(price) as total_revenue from basetable where event_type='purchase' group by month(event_time);
Query ID = hadoop_20220930102652_1a4be51d-3988-4e8e-a4ef-f44e748f9797
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1664526454320_0012)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      2          2        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      6          6        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 60.34 s
----------------------------------------------------------------------------------------------
OK
11      1531016.900000122
10      1211538.4299997438
Time taken: 61.109 seconds, Fetched: 2 row(s)
hive>
```

Insight: Revenue of November is higher than October.

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

with monthly_sales as (select month(event_time) as month, sum(price)as sales from basetable where event_type = 'purchase' group by month(event_time)) select (B.Sales - A.Sales) as change_in_revenue from monthly_sales A inner join monthly_sales B on A.month = B.month + 1;



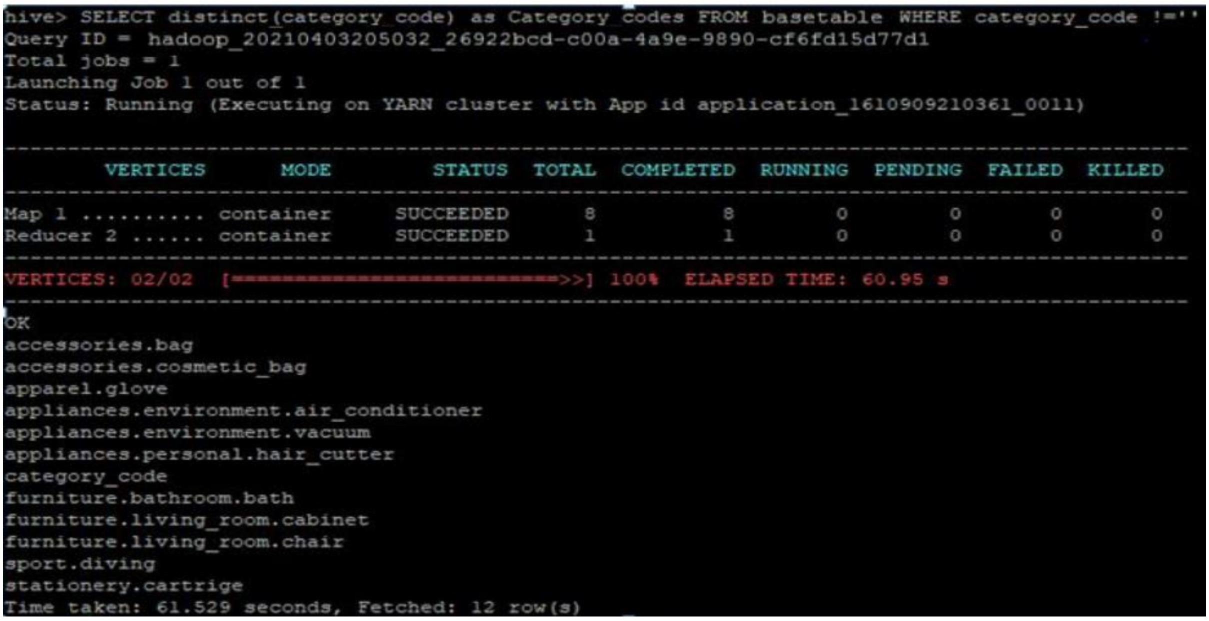
```
hive> with monthly_sales as (select month(event_time) as month, sum(price)as sales from basetable where event_type = 'purchase' group by month(event_time)) s
elect (B.Sales - A.Sales) as change_in_revenue from monthly_sales A inner join monthly_sales B on A.month = B.month + 1;
Query ID = hadoop_20220930102839_a92902e0-2e75-4322-b607-239aa0473157
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1664526454320_0012)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      2          2        0        0       0       0
Map 3 .......... container     SUCCEEDED      2          2        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      6          6        0        0       0       2
Reducer 4 ...... container     SUCCEEDED      6          6        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 04/04  [==========================>>] 100%  ELAPSED TIME: 120.94 s
----------------------------------------------------------------------------------------------
OK
-319478.4700003781
Time taken: 122.479 seconds, Fetched: 1 row(s)
```

Insight - The change in revenue is 319478.47

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

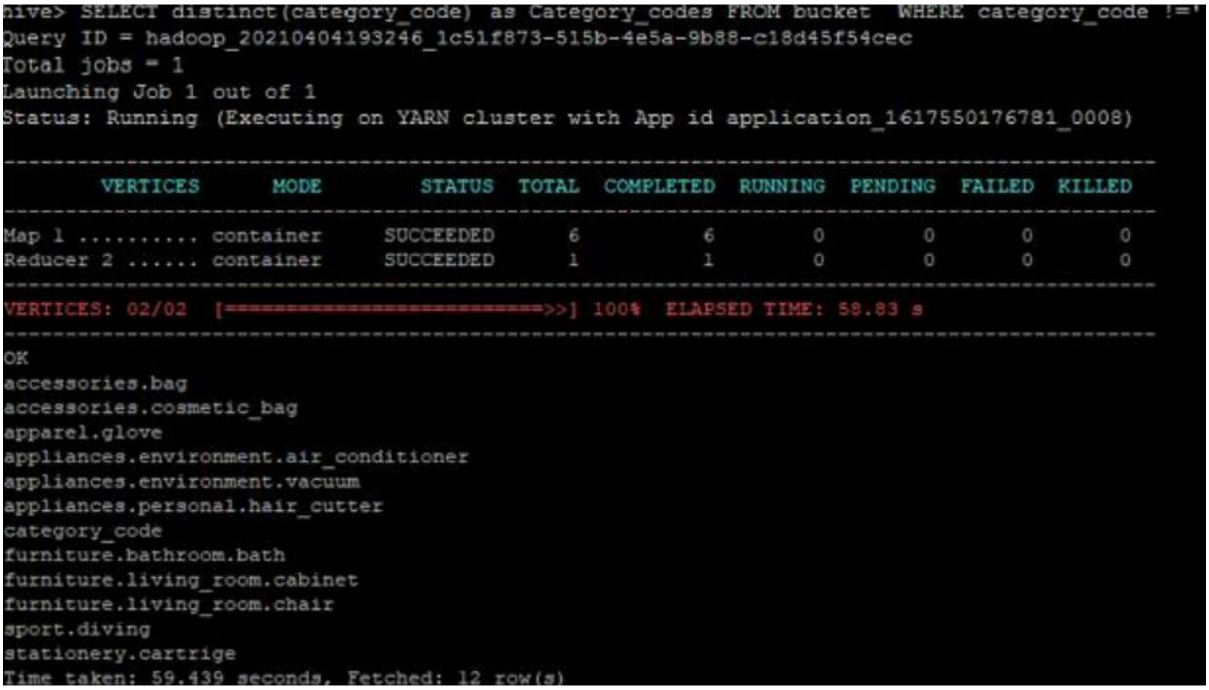
We have used bucketing over here and the optimization time is less in the second query

SELECT distinct(category_code) as Category_codes FROM basetable WHERE category_code !='' ;



```
hive> SELECT distinct(category_code) as Category_codes FROM basetable WHERE category_code !=''
Query ID = hadoop_20210403205032_26922bcd-c00a-4a9e-9890-cf6fd15d77d1
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1610909210361_0011)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      8          8        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 60.95 s
----------------------------------------------------------------------------------------------
OK
accessories.bag
accessories.cosmetic_bag
apparel.glove
appliances.environment.air_conditioner
appliances.environment.vacuum
appliances.personal.hair_cutter
category_code
furniture.bathroom.bath
furniture.living_room.cabinet
furniture.living_room.chair
sport.diving
stationery.cartrige
Time taken: 61.529 seconds, Fetched: 12 row(s)
```

SELECT distinct(category_code) as Category_codes FROM bucket WHERE category_code !='' ;



```
hive> SELECT distinct(category_code) as Category_codes FROM bucket  WHERE category_code !='
Query ID = hadoop_20210404193246_1c51f873-515b-4e5a-9b88-c18d45f54cec
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1617550176781_0008)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      6          6        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 58.83 s
----------------------------------------------------------------------------------------------
OK
accessories.bag
accessories.cosmetic_bag
apparel.glove
appliances.environment.air_conditioner
appliances.environment.vacuum
appliances.personal.hair_cutter
category_code
furniture.bathroom.bath
furniture.living_room.cabinet
furniture.living_room.chair
sport.diving
stationery.cartrige
Time taken: 59.439 seconds, Fetched: 12 row(s)
```

Insight : The Distinct categories of products are

Bags , Cosmetic_bag , Glove , Air Conditioner , Vacuum , hair_cutter , bath (furniture) , cabinet , chair , sports.diving , Cartrige

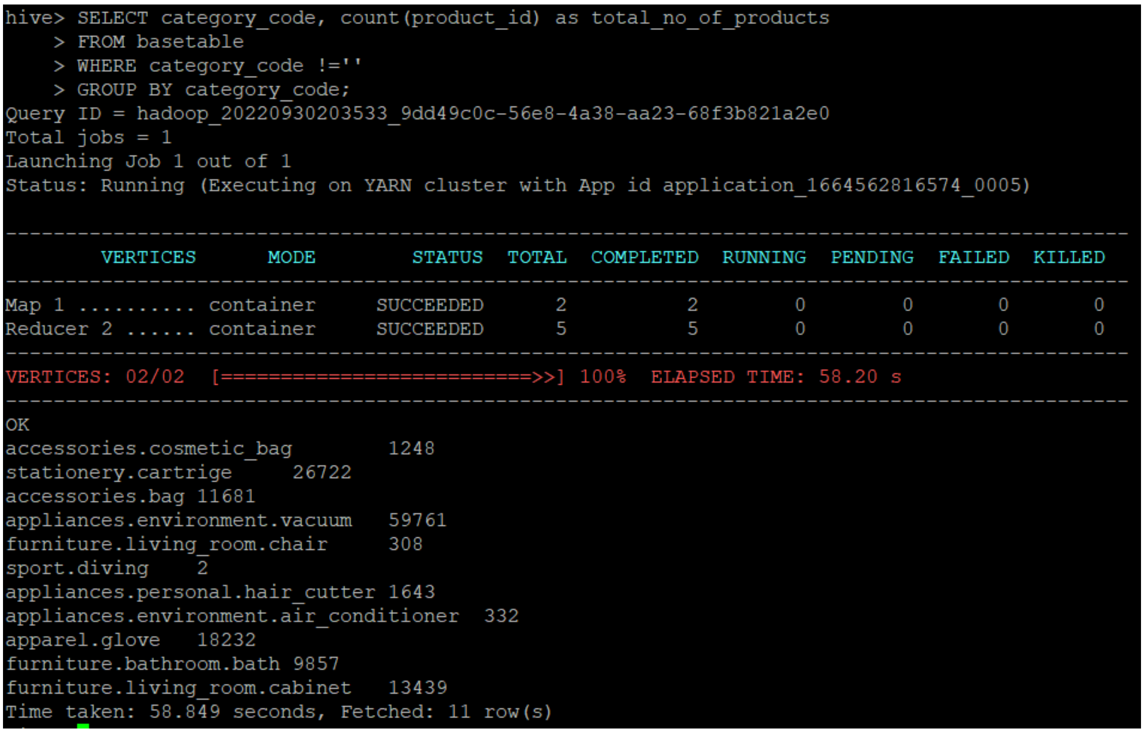
Q.5. Find the total number of products available under each category.

SELECT category_code, count(product_id) as total_no_of_products

FROM basetable

WHERE category_code !=''

GROUP BY category_code ;



```
hive> SELECT category_code, count(product_id) as total_no_of_products
    > FROM basetable
    > WHERE category_code !=''
    > GROUP BY category_code;
Query ID = hadoop_20220930203533_9dd49c0c-56e8-4a38-aa23-68f3b821a2e0
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1664562816574_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
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 58.20 s
----------------------------------------------------------------------------------------------
OK
accessories.cosmetic_bag        1248
stationery.cartrige     26722
accessories.bag 11681
appliances.environment.vacuum   59761
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: 58.849 seconds, Fetched: 11 row(s)
```

Insight: Vacuum has the maximum products whereas sport.diving has the least number of products.

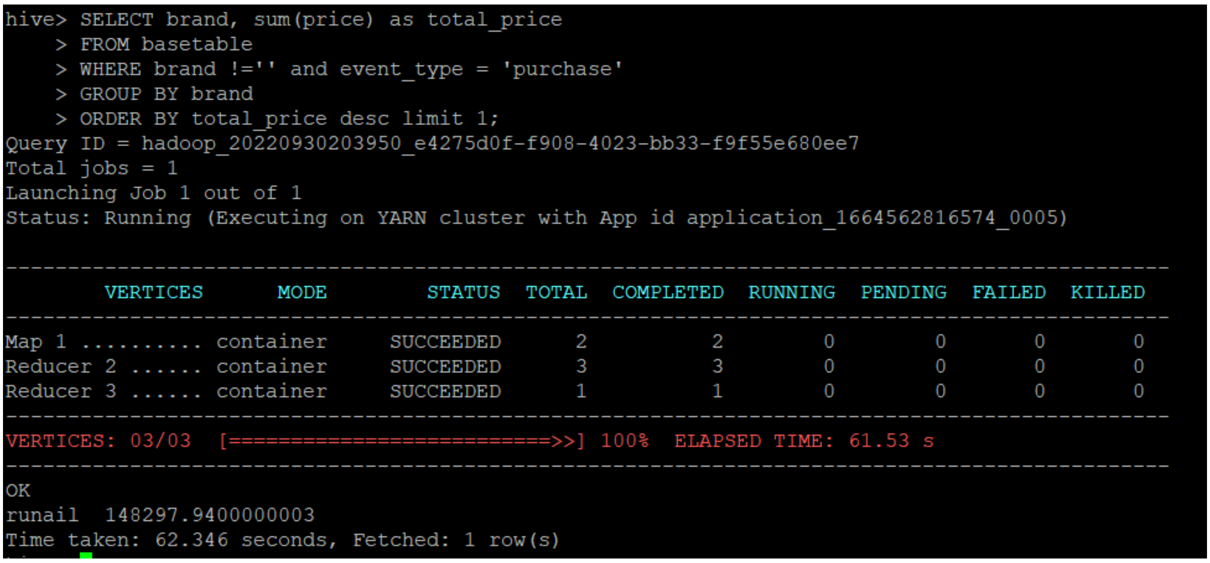
Q.6. Which brand had the maximum sales in October and November combined?

SELECT brand,sum(price) as total_price from basetable

where brand !='' and event_type ='purchase'

group by brand

order by total_price desc limit 1 ;



```
hive> SELECT brand, sum(price) as total_price
    > FROM basetable
    > WHERE brand !='' and event_type = 'purchase'
    > GROUP BY brand
    > ORDER BY total_price desc limit 1;
Query ID = hadoop_20220930203950_e4275d0f-f908-4023-bb33-f9f55e680ee7
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1664562816574_0005)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      2          2        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      3          3        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 61.53 s
----------------------------------------------------------------------------------------------
OK
runail  148297.9400000003
Time taken: 62.346 seconds, Fetched: 1 row(s)
```

Insight: Runail has highest sales with 148297.93999

Q.7 Which brands increased their sales from October to November?

with Revenue_difference AS

(

SELECT brand,SUM(case when MONTH(event_time) = '10' then price else 0 end) AS Revenue_in_Oct,

SUM(case when MONTH(event_time) = '11' then price else 0 end) AS Revenue_in_Nov

FROM basetable

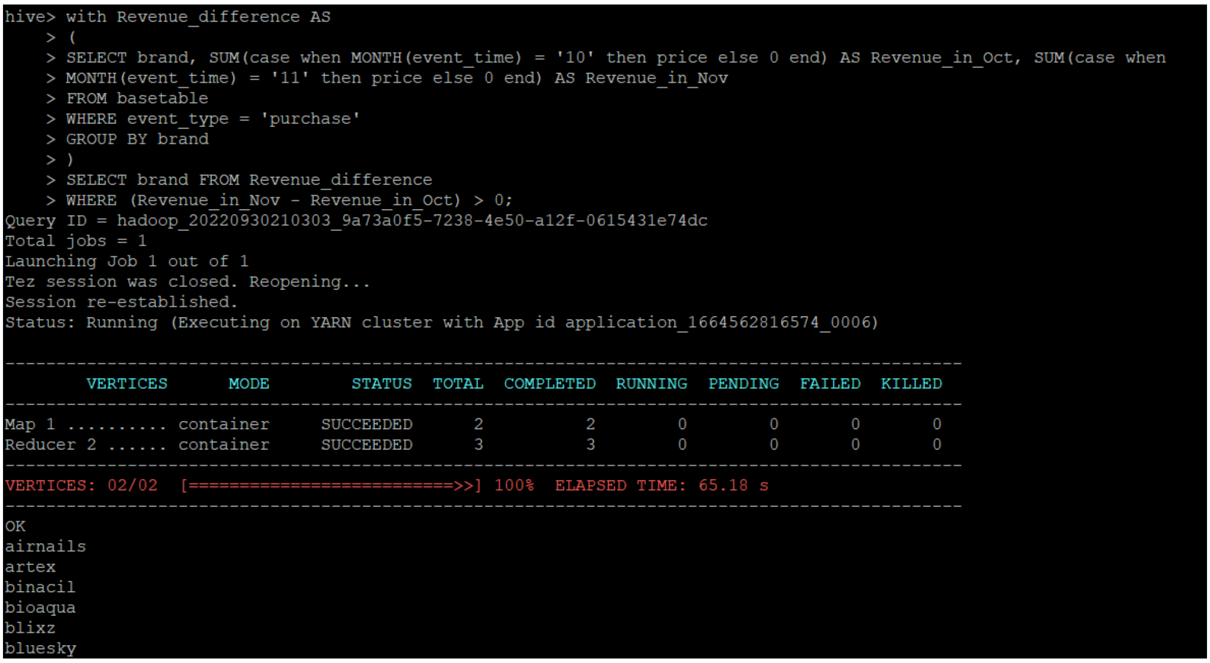
WHERE event_type = 'purchase'

group by brand

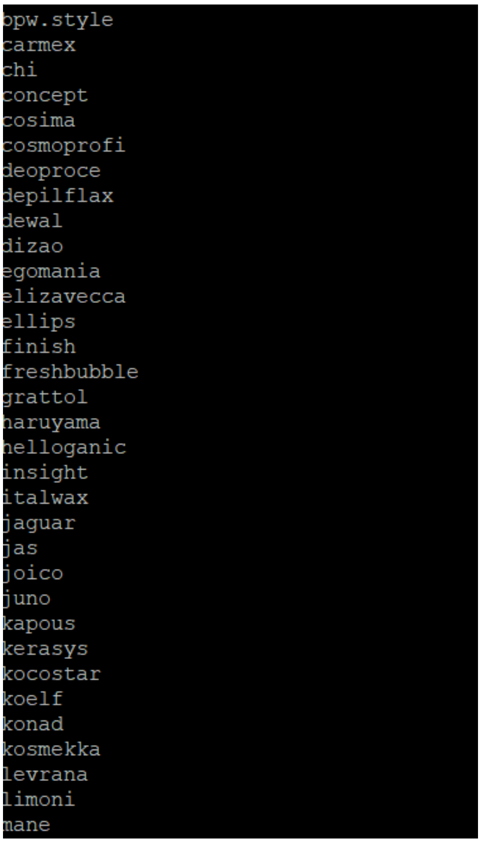
)

SELECT brand FROM Revenue_difference

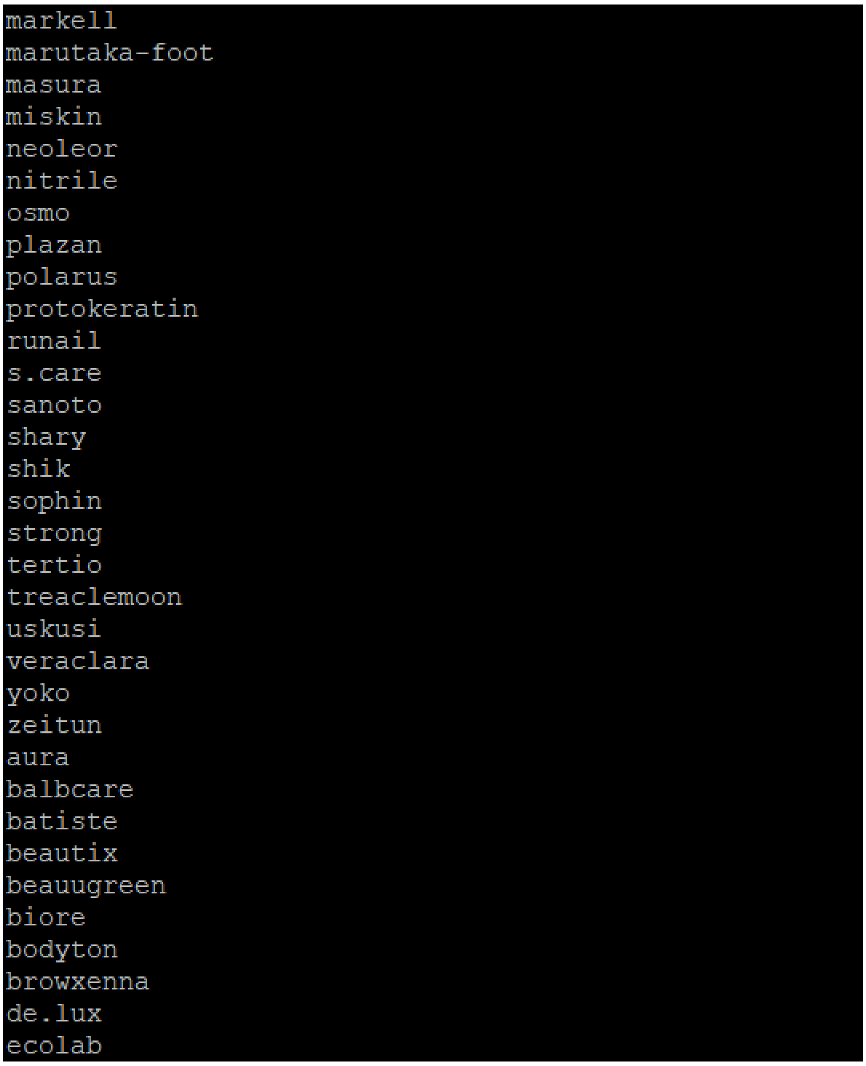
WHERE (Revenue_in_Nov - Revenue_in_Oct) > 0;



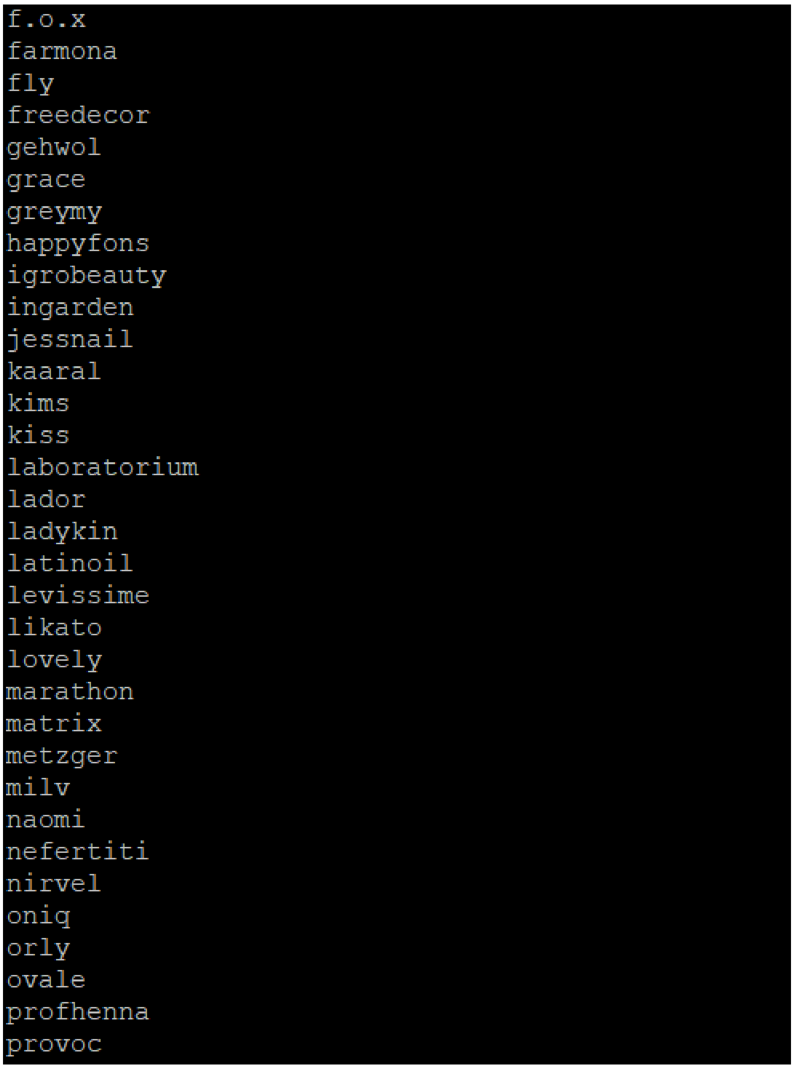
```
hive> with Revenue_difference AS
    > (
    > SELECT brand, SUM(case when MONTH(event_time) = '10' then price else 0 end) AS Revenue_in_Oct, SUM(case when
    > MONTH(event_time) = '11' then price else 0 end) AS Revenue_in_Nov
    > FROM basetable
    > WHERE event_type = 'purchase'
    > GROUP BY brand
    > )
    > SELECT brand FROM Revenue_difference
    > WHERE (Revenue_in_Nov - Revenue_in_Oct) > 0;
Query ID = hadoop_20220930210303_9a73a0f5-7238-4e50-a12f-0615431e74dc
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_1664562816574_0006)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      2          2        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      3          3        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 65.18 s
----------------------------------------------------------------------------------------------
OK
airnails
artex
binacil
bioaqua
blixz
bluesky
```



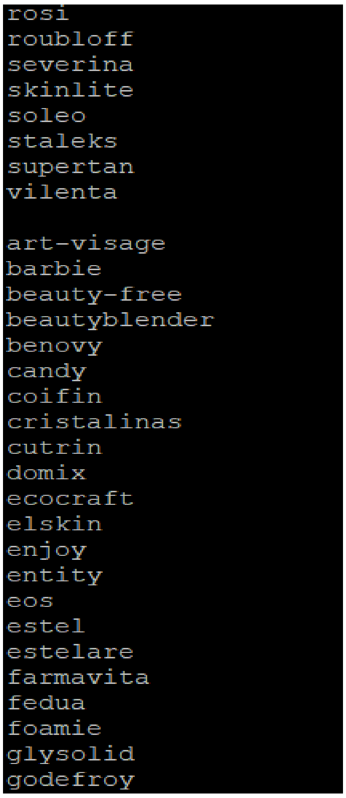
```
bpw.style
carmex
chi
concept
cosima
cosmoprofi
deoproce
depilflax
dewal
dizao
egomania
elizavecca
ellips
finish
freshbubble
grattol
haruyama
helloganic
insight
italwax
jaguar
jas
joico
juno
kapous
kerasys
kocostar
koelf
konad
kosmekka
levrana
limoni
mane
```



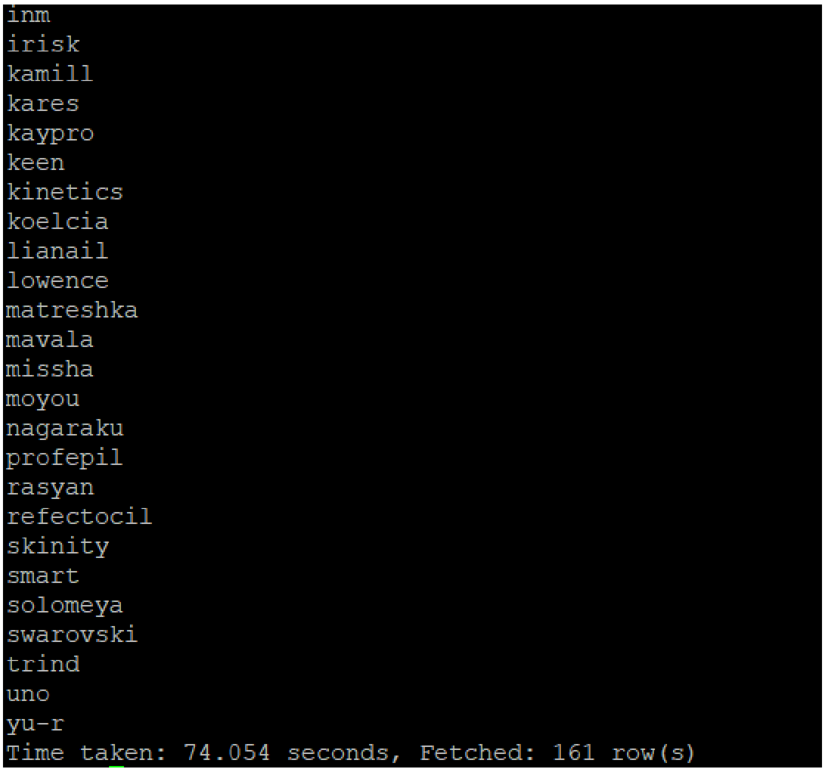
```
markell
marutaka-foot
masura
miskin
neoleor
nitrile
osmo
plazan
polarus
protokeratin
runail
s.care
sanoto
shary
shik
sophin
strong
tertio
treaclemoon
uskusi
veraclara
yoko
zeitun
aura
balbcare
batiste
beautix
beauugreen
biore
bodyton
browxenna
de.lux
ecolab
```



```
f.o.x
farmona
fly
freedecor
gehwol
grace
greymy
happyfons
igrobeauty
ingarden
jessnail
kaaral
kims
kiss
laboratorium
lador
ladykin
latinoil
levissime
likato
lovely
marathon
matrix
metzger
milv
naomi
nefertiti
nirvel
oniq
orly
ovale
profhenna
provoc
```



```
rosi
roubloff
severina
skinlite
soleo
staleks
supertan
vilenta

art-visage
barbie
beauty-free
beautyblender
benovy
candy
coifin
cristalinas
cutrin
domix
ecocraft
elskin
enjoy
entity
eos
estel
estelare
farmavita
fedua
foamie
glysolid
godefroy
```

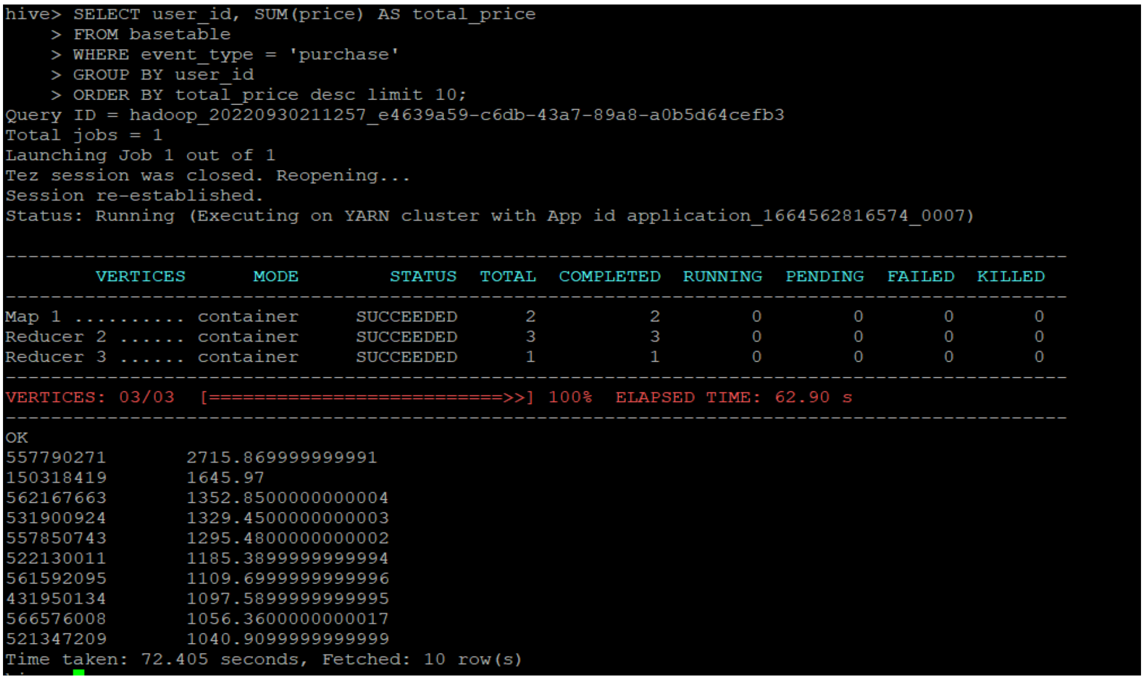


```
inm
irisk
kamill
kares
kaypro
keen
kinetics
koelcia
lianail
lowence
matreshka
mavala
missha
moyou
nagaraku
profepil
rasyan
refectocil
skinity
smart
solomeya
swarovski
trind
uno
yu-r
Time taken: 74.054 seconds, Fetched: 161 row(s)
```

Insight : Around 161 brands had a increased sales in Oct and Nov combined.

Q.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.

SELECT user_id,SUM(price) AS total_price FROM basetable WHERE event_type = 'purchase' GROUP BY user_id ORDER BY total_price desc limit 10;



```
hive> SELECT user_id, SUM(price) AS total_price
    > FROM basetable
    > WHERE event_type = 'purchase'
    > GROUP BY user_id
    > ORDER BY total_price desc limit 10;
Query ID = hadoop_20220930211257_e4639a59-c6db-43a7-89a8-a0b5d64cefb3
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_1664562816574_0007)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      2          2        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      3          3        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 62.90 s
----------------------------------------------------------------------------------------------
OK
557790271       2715.869999999991
150318419       1645.97
562167663       1352.8500000000004
531900924       1329.4500000000003
557850743       1295.4800000000002
522130011       1185.3899999999994
561592095       1109.6999999999996
431950134       1097.5899999999995
566576008       1056.3600000000017
521347209       1040.9099999999999
Time taken: 72.405 seconds, Fetched: 10 row(s)
```

Insight: Here's a list of the top 10 users who have spent the most on purchasing the goods in the ecommerce website.