HIVE CASE STUDY

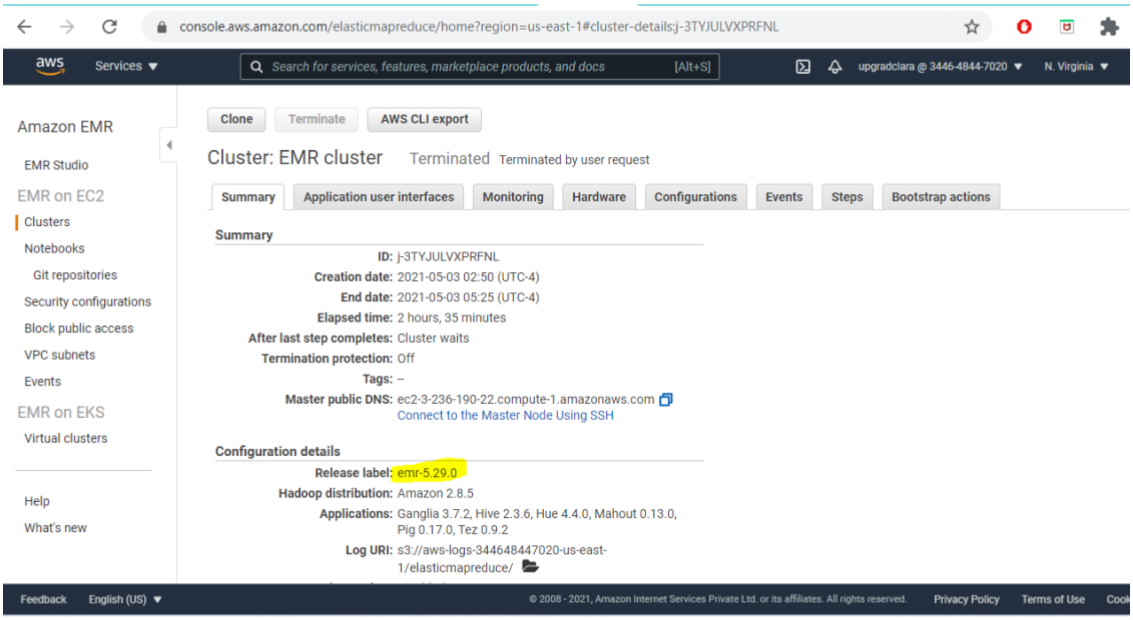
Submission by:
Gunavina Mehta
Clara Rosalind Francisca
Pieyush C Joy

Steps Followed -

- Copying the data set into the HDFS:
  - Launch an EMR cluster that utilizes the Hive services, and
  - Move the data from the S3 bucket into the HDFS
- Creating the database and launching Hive queries on your EMR cluster:
  - Create the structure of your database,
  - Use optimized techniques to run your queries as efficiently as possible
  - Show the improvement of the performance after using optimization on any single query.
  - Run Hive queries to answer the questions given below.
- Cleaning up
  - Drop your database, and
  - Terminate your cluster

Step -1

Creating the EMR Cluster –



Cluster: EMR cluster Terminated Terminated by user request

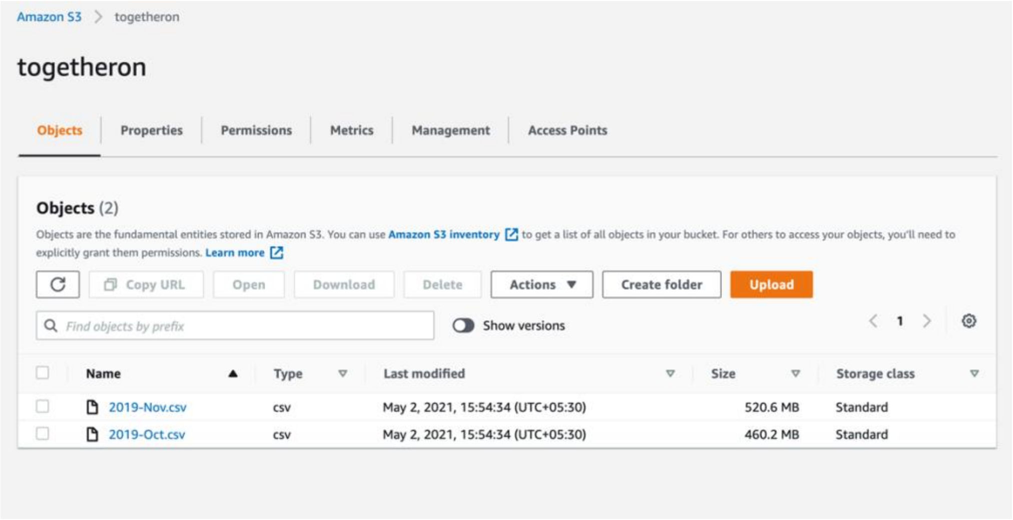
Summary

ID: j-3TYJULVXPRFNL
Creation date: 2021-05-03 02:50 (UTC-4)
End date: 2021-05-03 05:25 (UTC-4)
Elapsed time: 2 hours, 35 minutes
After last step completes: Cluster waits
Termination protection: Off
Tags: –
Master public DNS: ec2-3-236-190-22.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: Ganglia 3.7.2, Hive 2.3.6, Hue 4.4.0, Mahout 0.13.0, Pig 0.17.0, Tez 0.9.2
Log URI: s3://aws-logs-344648447020-us-east-1/elasticmapreduce/

S3 Clusters –



togetheron

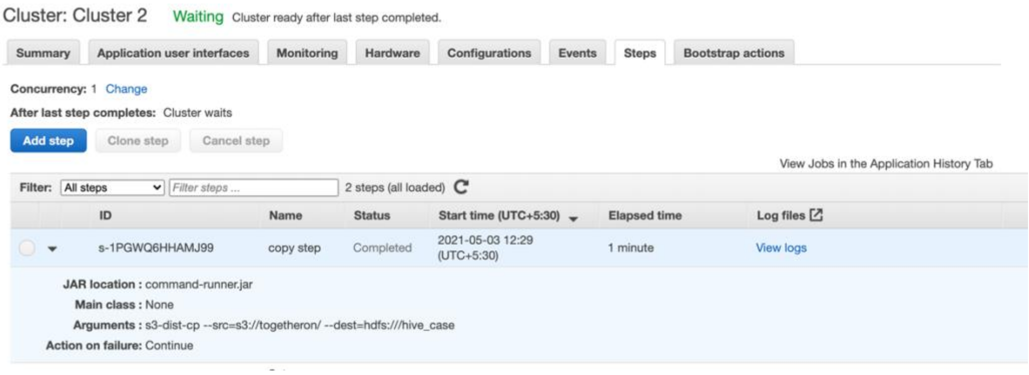
Objects (2)

Objects are the fundamental entities stored in Amazon S3. You can use **Amazon S3 inventory** to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. **Learn more**

□	Name	Type	Last modified	Size	Storage class
□	2019-Nov.csv	csv	May 2, 2021, 15:54:34 (UTC+05:30)	520.6 MB	Standard
□	2019-Oct.csv	csv	May 2, 2021, 15:54:34 (UTC+05:30)	460.2 MB	Standard

Connecting the EMR Master Node

Included a step of loading the data from s3 here –



Cluster: Cluster 2 Waiting Cluster ready after last step completed.

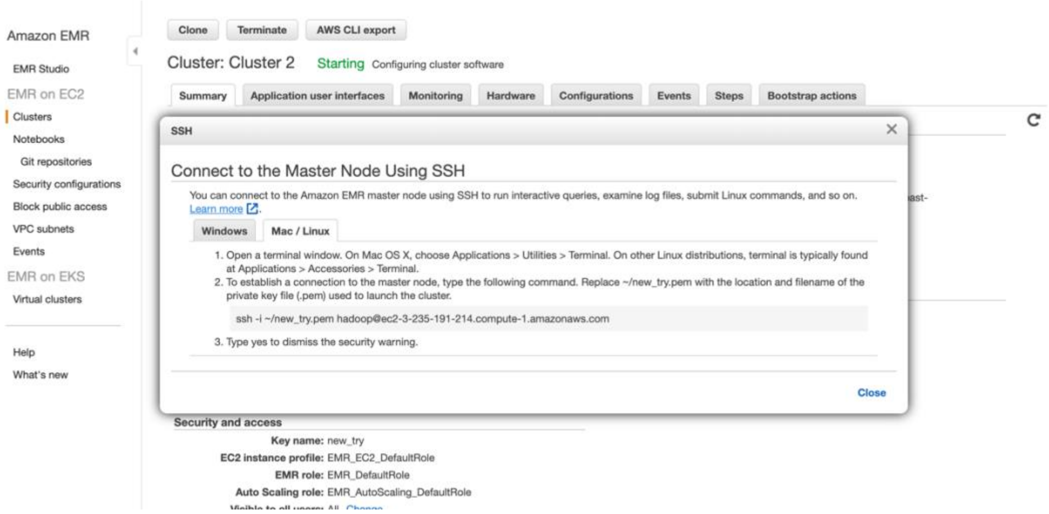
Concurrency: 1 Change

After last step completes: Cluster waits

Filter: All steps | Filter steps ... | 2 steps (all loaded)

ID	Name	Status	Start time (UTC+5:30)	Elapsed time	Log files
s-1PGWQ6HHAMJ99	copy step	Completed	2021-05-03 12:29 (UTC+5:30)	1 minute	View logs

JAR location : command-runner.jar
Main class : None
Arguments : s3-dist-cp --src=s3://togetheron/ --dest=hdfs:///hive_case
Action on failure: Continue



Cluster: Cluster 2 Starting Configuring cluster software

Connect to the Master Node Using SSH

You can connect to the Amazon EMR master node using SSH to run interactive queries, examine log files, submit Linux commands, and so on. Learn more

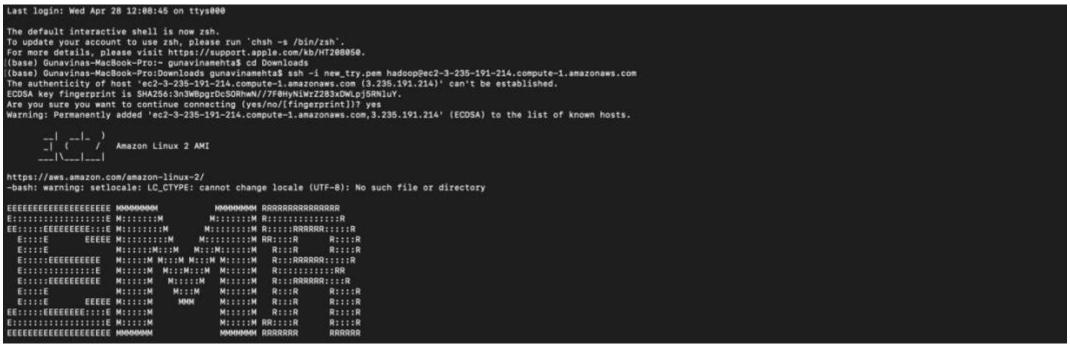
1. Open a terminal window. On Mac OS X, choose Applications > Utilities > Terminal. On other Linux distributions, terminal is typically found at Applications > Accessories > Terminal.
2. To establish a connection to the master node, type the following command. Replace ~/new_try.pem with the location and filename of the private key file (.pem) used to launch the cluster.

```
ssh -i ~/new_try.pem hadoop@ec2-3-235-191-214.compute-1.amazonaws.com
```

3. Type yes to dismiss the security warning.

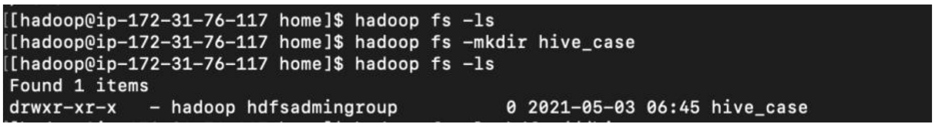
Security and access

Key name: new_try
EC2 instance profile: EMR_EC2_DefaultRole
EMR role: EMR_DefaultRole
Auto Scaling role: EMR_AutoScaling_DefaultRole

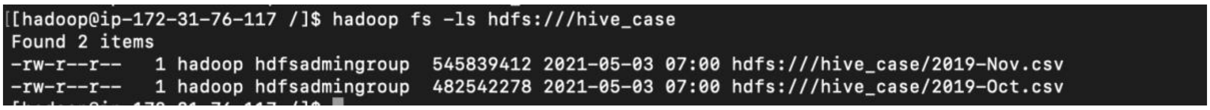


```
Last login: Wed Apr 28 12:08:45 on ttys000

The default interactive shell is now zsh.
To update your account to use zsh, please run `chsh -s /bin/zsh`.
For more details, please visit https://support.apple.com/kb/HT208050.
(base) Gunavinas-MacBook-Pro:~ gunavinamehta$ cd Downloads
(base) Gunavinas-MacBook-Pro:Downloads gunavinamehta$ ssh -i new_try.pem hadoop@ec2-3-235-191-214.compute-1.amazonaws.com
The authenticity of host 'ec2-3-235-191-214.compute-1.amazonaws.com (3.235.191.214)' can't be established.
ECDSA key fingerprint is SHA256:3n3WBpgrDcSORhwN//7F0HyNiWrZ283xDWLpj5RNluY.
Are you sure you want to continue connecting (yes/no/[fingerprint])? yes
Warning: Permanently added 'ec2-3-235-191-214.compute-1.amazonaws.com,3.235.191.214' (ECDSA) to the list of known hosts.

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

https://aws.amazon.com/amazon-linux-2/
-bash: warning: setlocale: LC_CTYPE: cannot change locale (UTF-8): No such file or directory

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
```

Creating a directory and checking the loaded data



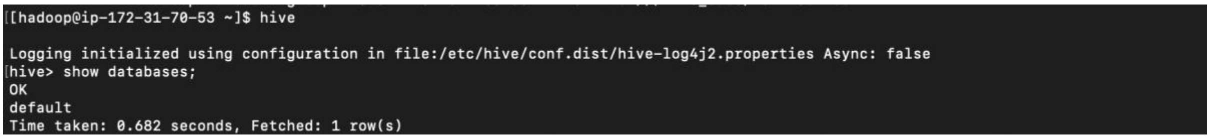
```
[[hadoop@ip-172-31-76-117 home]$ hadoop fs -ls
[[hadoop@ip-172-31-76-117 home]$ hadoop fs -mkdir hive_case
[[hadoop@ip-172-31-76-117 home]$ hadoop fs -ls
Found 1 items
drwxr-xr-x   - hadoop hdfsadmingroup          0 2021-05-03 06:45 hive_case
```



```
[[hadoop@ip-172-31-76-117 /]$ hadoop fs -ls hdfs:///hive_case
Found 2 items
-rw-r--r--   1 hadoop hdfsadmingroup  545839412 2021-05-03 07:00 hdfs:///hive_case/2019-Nov.csv
-rw-r--r--   1 hadoop hdfsadmingroup  482542278 2021-05-03 07:00 hdfs:///hive_case/2019-Oct.csv
```

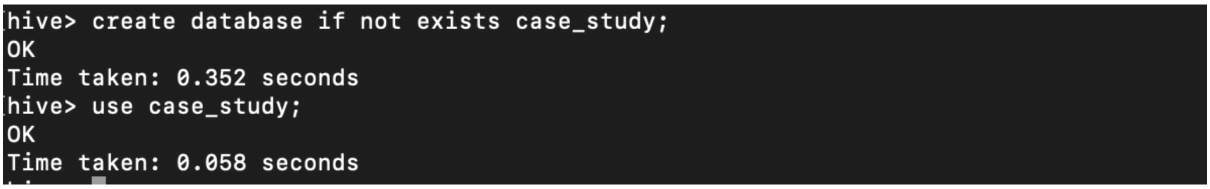
STEP – 2

Connecting to HIVE



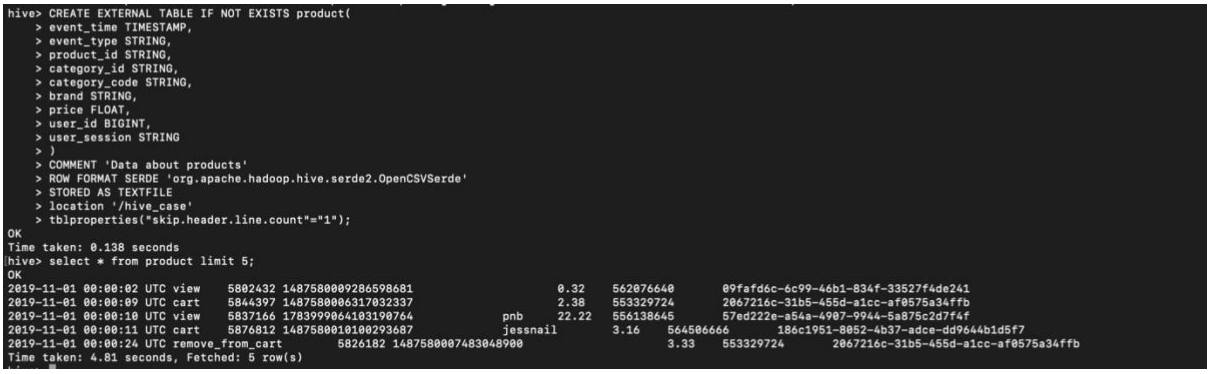
```
[[hadoop@ip-172-31-70-53 ~]$ hive

Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j2.properties Async: false
[hive> show databases;
OK
default
Time taken: 0.682 seconds, Fetched: 1 row(s)
```

Creating a database named – **CASE_STUDY**



```
[hive> create database if not exists case_study;
OK
Time taken: 0.352 seconds
[hive> use case_study;
OK
Time taken: 0.058 seconds
```

Creating a table named - **product** and loading the data from hdfs to hive



```
hive> CREATE EXTERNAL TABLE IF NOT EXISTS product(
    > 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
    > )
    > COMMENT 'Data about products'
    > ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.OpenCSVSerde'
    > STORED AS TEXTFILE
    > location '/hive_case'
    > tblproperties("skip.header.line.count"="1");
OK
Time taken: 0.138 seconds
[hive> select * from product 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: 4.81 seconds, Fetched: 5 row(s)
```

STEP - 3

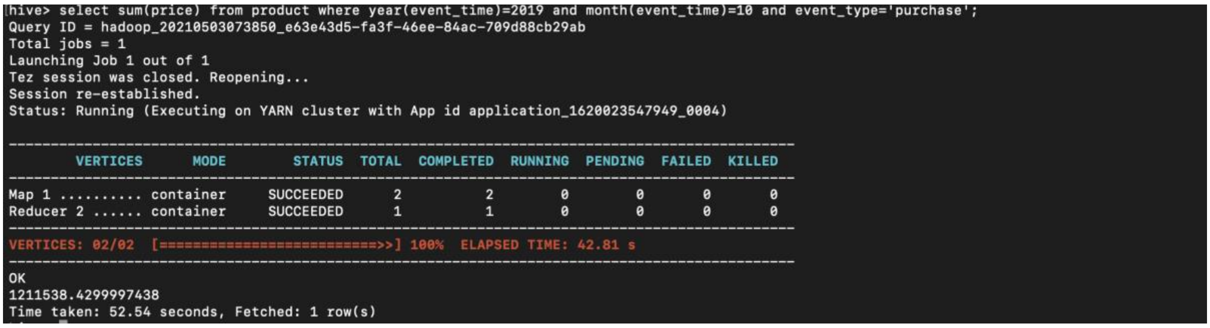
Starting with Querying –

Note: They are first performed without any partitioning

Query – 1

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

Answer - select sum(price) from product where year(event_time)=2019 and month(event_time)=10 and event_type='purchase';

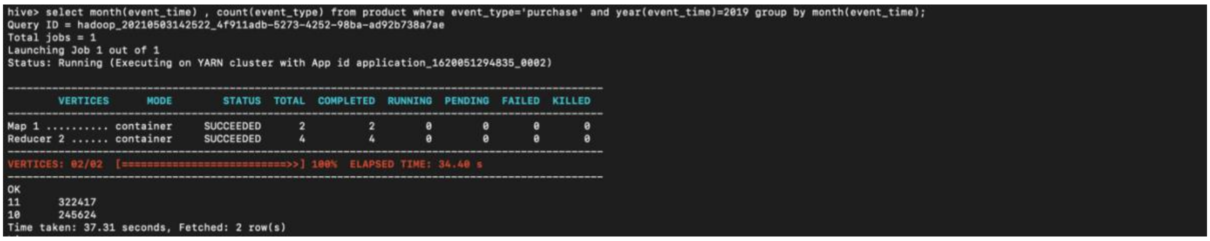


```
hive> select sum(price) from product where year(event_time)=2019 and month(event_time)=10 and event_type='purchase';
Query ID = hadoop_20210503073850_e63e43d5-fa3f-46ee-84ac-709d88cb29ab
Total jobs = 1
Launching Job 1 out of 1
Tez session was closed. Reopening...
Session re-established.
Status: Running (Executing on YARN cluster with App id application_1620023547949_0004)

----------------------------------------------------------------------------------------------
        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: 42.81 s
----------------------------------------------------------------------------------------------
OK
1211538.4299997438
Time taken: 52.54 seconds, Fetched: 1 row(s)
```

Query – 2

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

Answer - select month(event_time) , count(event_type) from product where event_type='purchase' and year(event_time)=2019 group by month(event_time);

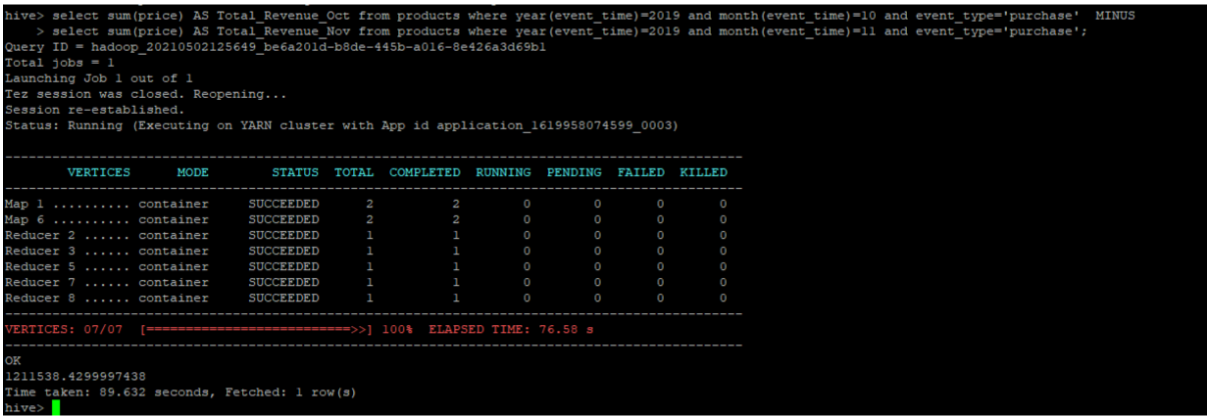


```
hive> select month(event_time) , count(event_type) from product where event_type='purchase' and year(event_time)=2019 group by month(event_time);
Query ID = hadoop_20210503142522_4f911adb-5273-4252-98ba-ad92b738a7ae
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1620051294835_0002)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      2          2        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      4          4        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 34.40 s
----------------------------------------------------------------------------------------------
OK
11      322417
10      245624
Time taken: 37.31 seconds, Fetched: 2 row(s)
```

Query – 3

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

Answer - select sum(price) AS Total_Revenue_Oct from products where year(event_time)=2019 and month(event_time)=10 and event_type='purchase' MINUS select sum(price) AS Total_Revenue_Nov from products where year(event_time)=2019 and month(event_time)=11 and event_type='purchase' ;

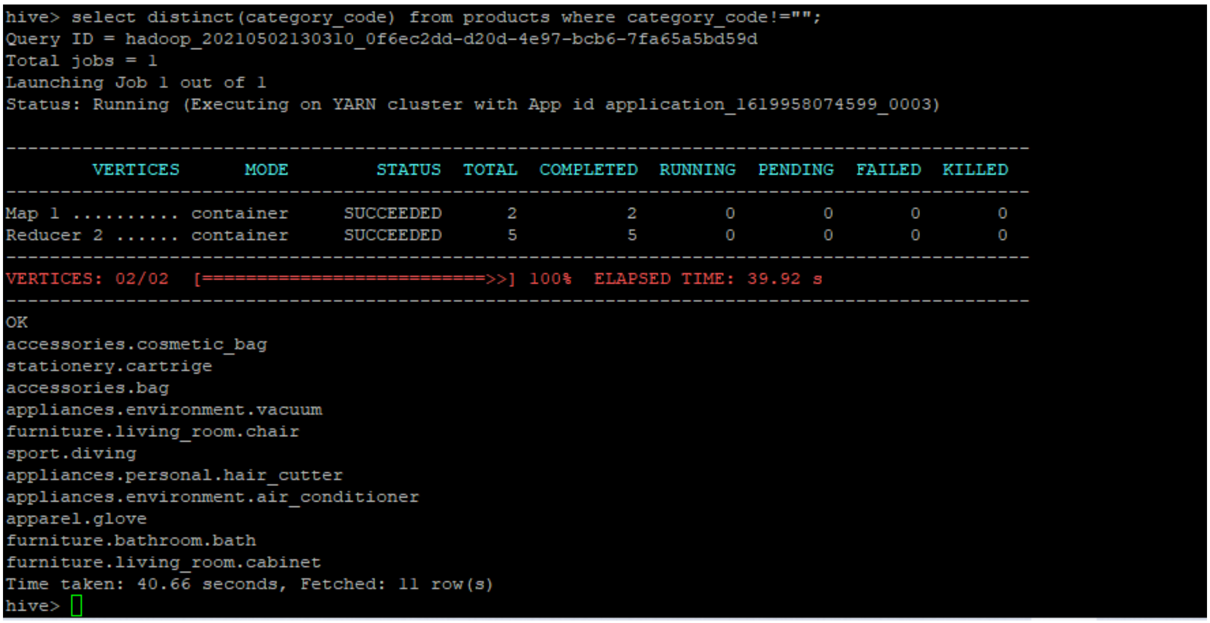


```
hive> select sum(price) AS Total_Revenue_Oct from products where year(event_time)=2019 and month(event_time)=10 and event_type='purchase'  MINUS
    > select sum(price) AS Total_Revenue_Nov from products where year(event_time)=2019 and month(event_time)=11 and event_type='purchase';
Query ID = hadoop_20210502125649_be6a201d-b8de-445b-a016-8e426a3d69b1
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_1619958074599_0003)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      2          2        0        0       0       0
Map 6 .......... container     SUCCEEDED      2          2        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 5 ...... container     SUCCEEDED      1          1        0        0       0       0
Reducer 7 ...... container     SUCCEEDED      1          1        0        0       0       0
Reducer 8 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 07/07  [==========================>>] 100%  ELAPSED TIME: 76.58 s
----------------------------------------------------------------------------------------------
OK
1211538.4299997438
Time taken: 89.632 seconds, Fetched: 1 row(s)
hive>
```

Query – 4

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

Answer - select distinct(category_code) from products where category_code!="";

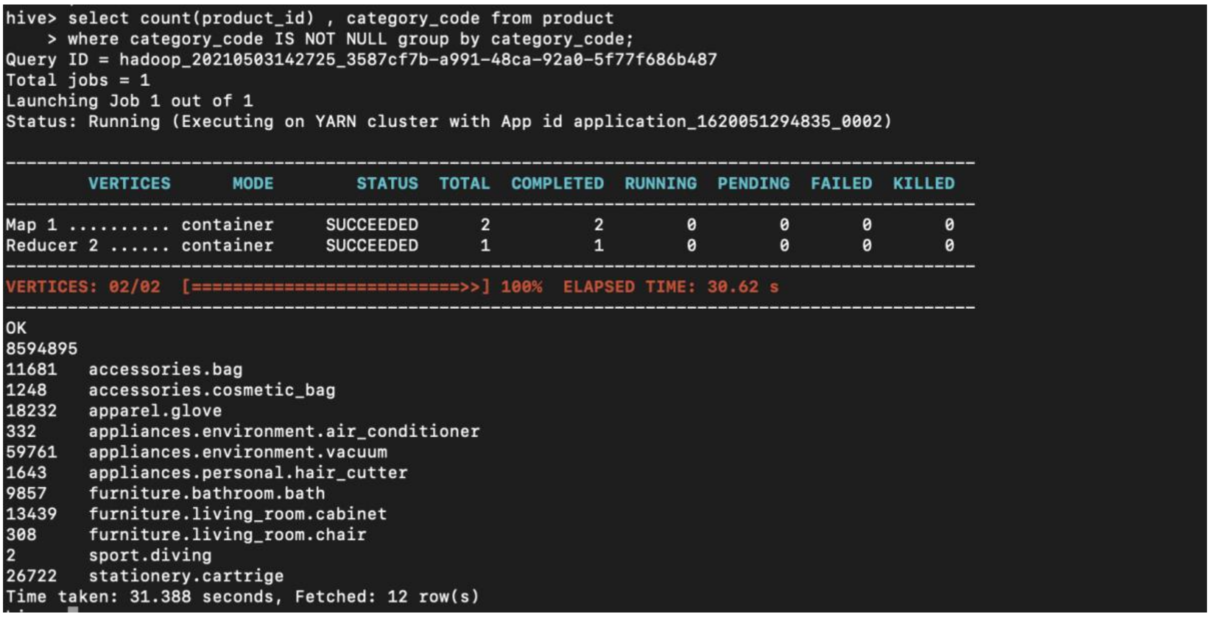


```
hive> select distinct(category_code) from products where category_code!="";
Query ID = hadoop_20210502130310_0f6ec2dd-d20d-4e97-bcb6-7fa65a5bd59d
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1619958074599_0003)

----------------------------------------------------------------------------------------------
        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: 39.92 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: 40.66 seconds, Fetched: 11 row(s)
hive>
```

Query – 5

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

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

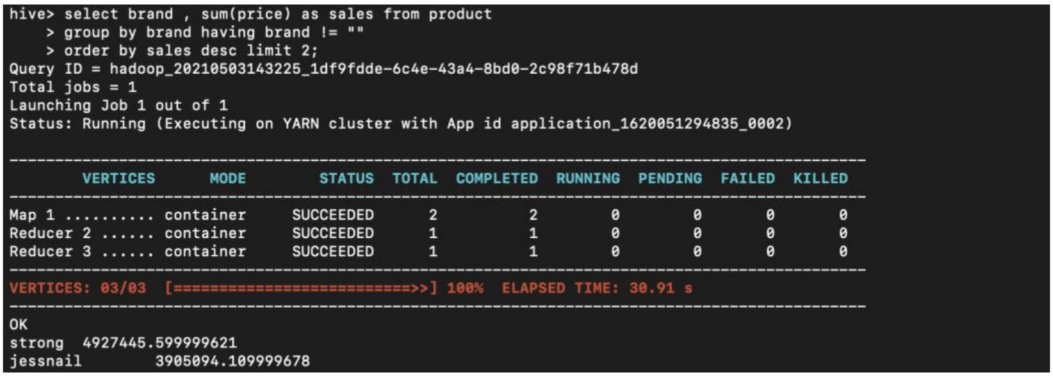


```
hive> select count(product_id) , category_code from product
    > where category_code IS NOT NULL group by category_code;
Query ID = hadoop_20210503142725_3587cf7b-a991-48ca-92a0-5f77f686b487
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1620051294835_0002)

----------------------------------------------------------------------------------------------
        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: 30.62 s
----------------------------------------------------------------------------------------------
OK
8594895
11681   accessories.bag
1248    accessories.cosmetic_bag
18232   apparel.glove
332     appliances.environment.air_conditioner
59761   appliances.environment.vacuum
1643    appliances.personal.hair_cutter
9857    furniture.bathroom.bath
13439   furniture.living_room.cabinet
308     furniture.living_room.chair
2       sport.diving
26722   stationery.cartrige
Time taken: 31.388 seconds, Fetched: 12 row(s)
```

Query – 6

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

Answer - select brand , sum(price) as sales from product group by brand having brand != "" order by sales desc limit 2;

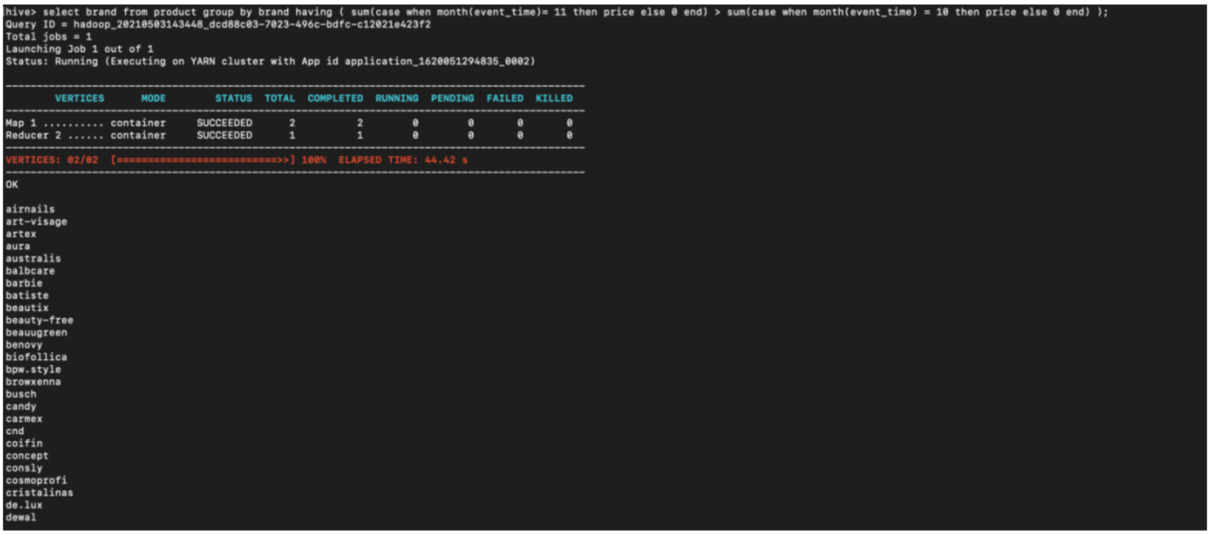


```
hive> select brand , sum(price) as sales from product
    > group by brand having brand != ""
    > order by sales desc limit 2;
Query ID = hadoop_20210503143225_1df9fdde-6c4e-43a4-8bd0-2c98f71b478d
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1620051294835_0002)

----------------------------------------------------------------------------------------------
        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
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 30.91 s
----------------------------------------------------------------------------------------------
OK
strong  4927445.599999621
jessnail        3905094.109999678
```

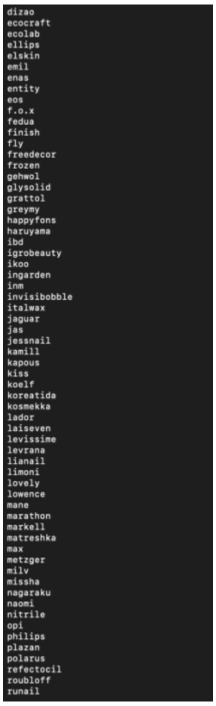
Query – 7

- Which brands increased their sales from October to November?

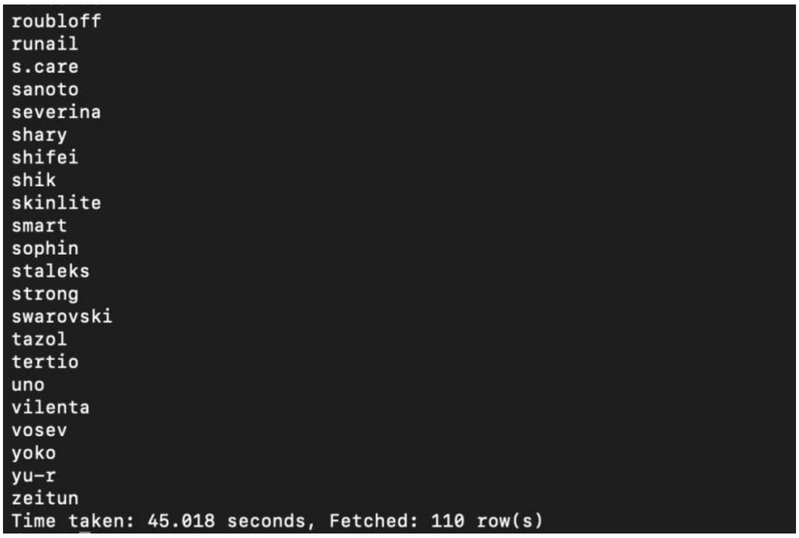
Answer - select brand from product group by brand having (sum(case when month(event_time)= 11 then price else 0 end) > sum(case when month(event_time) = 10 then price else 0 end));



```
hive> select brand from product group by brand having ( sum(case when month(event_time)= 11 then price else 0 end) > sum(case when month(event_time) = 10 then price else 0 end) );
Query ID = hadoop_20210503143448_dcd88c03-7023-496c-bdfc-c12021e423f2
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1620051294835_0002)

----------------------------------------------------------------------------------------------
        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: 44.42 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
cosmoprofi
cristalinas
de.lux
dewal
```



```
dizao
ecocraft
ecolab
ellips
elskin
emil
enas
entity
eos
f.o.x
fedua
finish
fly
freedecor
frozen
gehwol
glysolid
grattol
greymy
happyfons
haruyama
ibd
igrobeauty
ikoo
ingarden
inm
invisibobble
italwax
jaguar
jas
jessnail
kamill
kapous
kiss
koelf
koreatida
kosmekka
lador
laiseven
levissime
levrana
lianail
limoni
lovely
lowence
mane
marathon
markell
matreshka
max
metzger
milv
missha
nagaraku
naomi
nitrile
opi
philips
plazan
polarus
refectocil
roubloff
runail
```

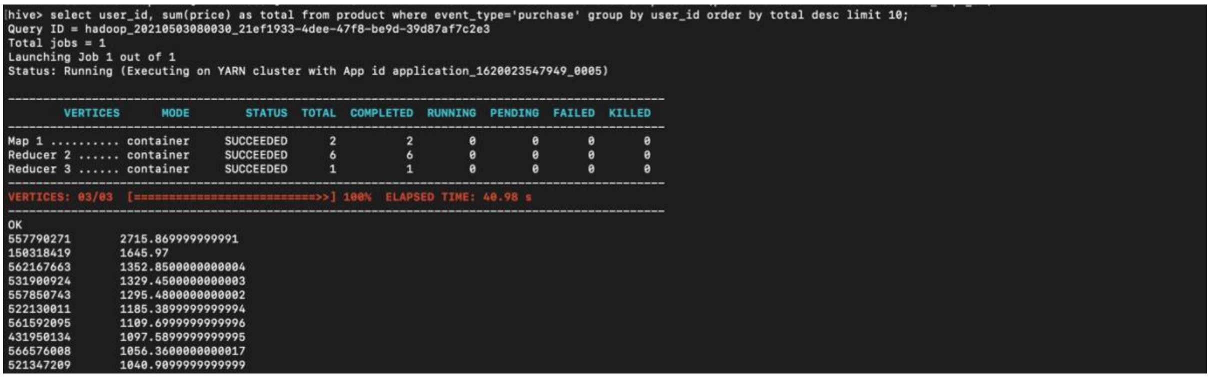


```
roubloff
runail
s.care
sanoto
severina
shary
shifei
shik
skinlite
smart
sophin
staleks
strong
swarovski
tazol
tertio
uno
vilenta
vosev
yoko
yu-r
zeitun
Time taken: 45.018 seconds, Fetched: 110 row(s)
```

Query – 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.

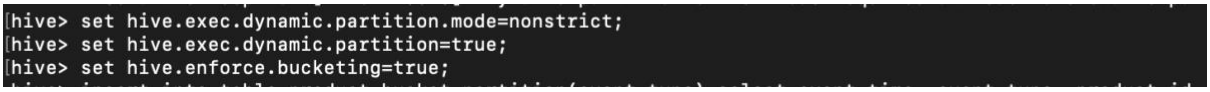
Answer - select user_id, sum(price) as total from product where event_type='purchase' group by user_id order by total desc limit 10;



```
hive> select user_id, sum(price) as total from product where event_type='purchase' group by user_id order by total desc limit 10;
Query ID = hadoop_20210503080030_21ef1933-4dee-47f8-be9d-39d87af7c2e3
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1620023547949_0005)

----------------------------------------------------------------------------------------------
        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
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 40.98 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
```

Bucketing & Partitioning

Performing all the above queries by partitioning and bucketing

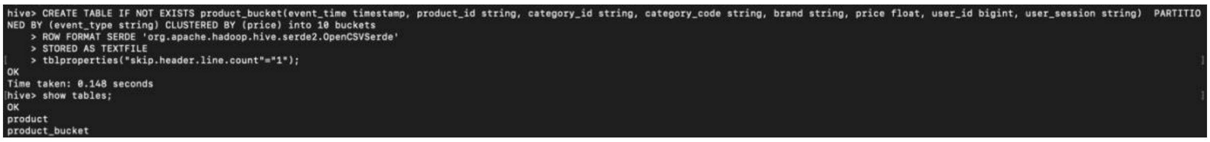


```
[hive> set hive.exec.dynamic.partition.mode=nonstrict;
[hive> set hive.exec.dynamic.partition=true;
[hive> set hive.enforce.bucketing=true;
```

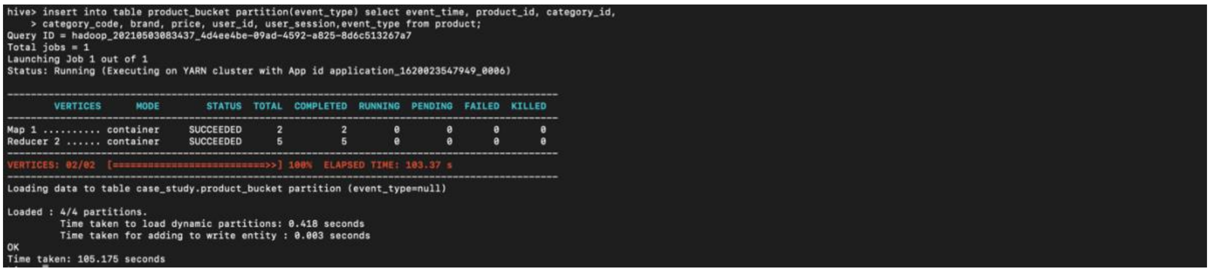
Product_Bucket1 is used for solving query – 1,2,3,7,8

CREATE TABLE IF NOT EXISTS **product_bucket1**(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
tblproperties("skip.header.line.count"="1");

insert into table product_bucket1 partition(event_type) select event_time, product_id, category_id,
category_code, brand, price, user_id, user_session,event_type from product;



```
hive> CREATE TABLE IF NOT EXISTS product_bucket(event_time timestamp, product_id string, category_id string, category_code string, brand string, price float, user_id bigint, user_session string)  PARTITIO
NED BY (event_type string) CLUSTERED BY (price) into 10 buckets
    > ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.OpenCSVSerde'
    > STORED AS TEXTFILE
    > tblproperties("skip.header.line.count"="1");
OK
Time taken: 0.148 seconds
hive> show tables;
OK
product
product_bucket
```

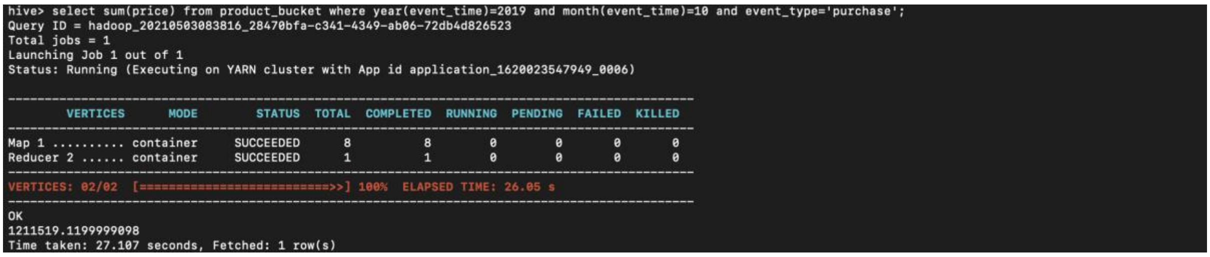


```
hive> insert into table product_bucket partition(event_type) select event_time, product_id, category_id,
    > category_code, brand, price, user_id, user_session,event_type from product;
Query ID = hadoop_20210503083437_4d4ee4be-09ad-4592-a825-8d6c513267a7
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1620023547949_0006)

----------------------------------------------------------------------------------------------
        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: 103.37 s
----------------------------------------------------------------------------------------------
Loading data to table case_study.product_bucket partition (event_type=null)

Loaded : 4/4 partitions.
         Time taken to load dynamic partitions: 0.418 seconds
         Time taken for adding to write entity : 0.003 seconds
OK
Time taken: 105.175 seconds
```

Query 1 – Partitioned

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

select sum(price) from product_bucket1 where year(event_time)=2019 and month(event_time)=10 and event_type='purchase';

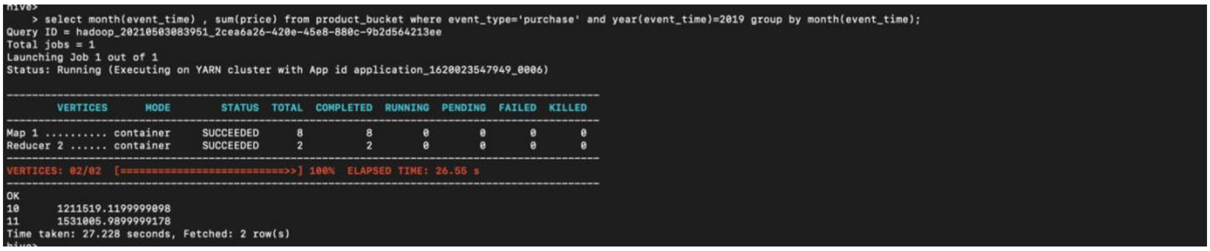


```
hive> select sum(price) from product_bucket where year(event_time)=2019 and month(event_time)=10 and event_type='purchase';
Query ID = hadoop_20210503083816_28470bfa-c341-4349-ab06-72db4d826523
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1620023547949_0006)

----------------------------------------------------------------------------------------------
        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: 26.05 s
----------------------------------------------------------------------------------------------
OK
1211519.1199999098
Time taken: 27.107 seconds, Fetched: 1 row(s)
```

Query -2 Partitioned

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

Answer - select month(event_time) , count(event_type) from product_bucket1 where event_type='purchase' and year(event_time)=2019 group by month(event_time);

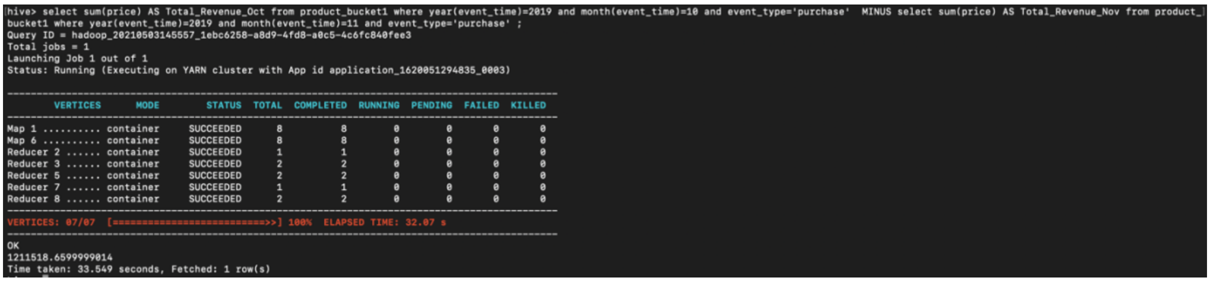


```
hive>
    > select month(event_time) , sum(price) from product_bucket where event_type='purchase' and year(event_time)=2019 group by month(event_time);
Query ID = hadoop_20210503083951_2cea6a26-420e-45e8-880c-9b2d564213ee
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1620023547949_0006)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      8          8        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      2          2        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 26.55 s
----------------------------------------------------------------------------------------------
OK
10      1211519.1199999098
11      1531005.9899999178
Time taken: 27.228 seconds, Fetched: 2 row(s)
```

Query – 3 Partitioned

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

Answer - select sum(price) AS Total_Revenue_Oct from product_bucket1 where year(event_time)=2019 and month(event_time)=10 and event_type='purchase' MINUS select sum(price) AS Total_Revenue_Nov from product_bucket1 where year(event_time)=2019 and month(event_time)=11 and event_type='purchase' ;



```
hive> select sum(price) AS Total_Revenue_Oct from product_bucket1 where year(event_time)=2019 and month(event_time)=10 and event_type='purchase'  MINUS select sum(price) AS Total_Revenue_Nov from product_
bucket1 where year(event_time)=2019 and month(event_time)=11 and event_type='purchase' ;
Query ID = hadoop_20210503145557_1ebc6258-a8d9-4fd8-a0c5-4c6fc840fee3
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1620051294835_0003)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      8          8        0        0       0       0
Map 6 .......... container     SUCCEEDED      8          8        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      2          2        0        0       0       0
Reducer 5 ...... container     SUCCEEDED      2          2        0        0       0       0
Reducer 7 ...... container     SUCCEEDED      1          1        0        0       0       0
Reducer 8 ...... container     SUCCEEDED      2          2        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 07/07  [==========================>>] 100%  ELAPSED TIME: 32.07 s
----------------------------------------------------------------------------------------------
OK
1211518.6599999014
Time taken: 33.549 seconds, Fetched: 1 row(s)
```

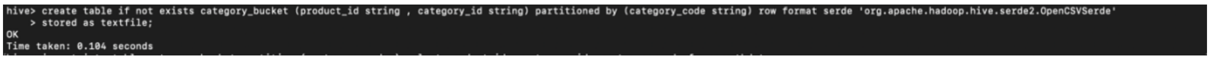
Time_Taken parameter was reduced by a significant amount If we compare the partitioned vs non-partitioned.

Partition by Category Code –

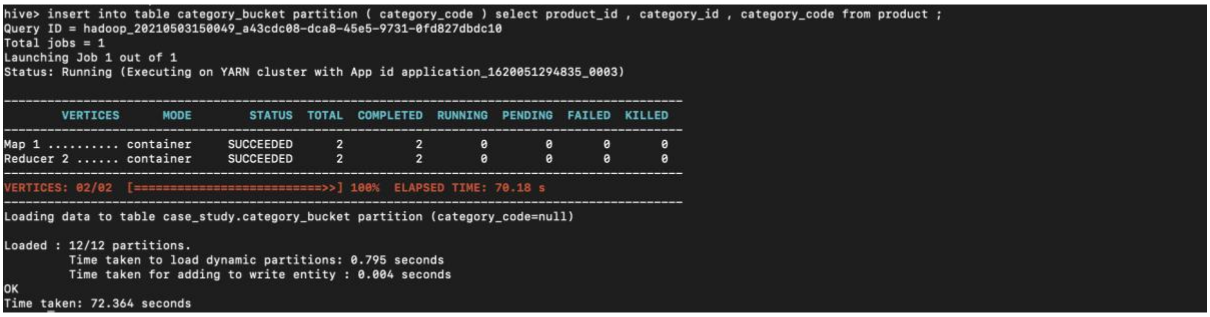
Category_Bucket is used in query for Q – 4,5,6

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

insert into table category_bucket partition (category_code) select product_id , category_id , category_code from product ;



```
hive> create table if not exists category_bucket (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: 0.104 seconds
```

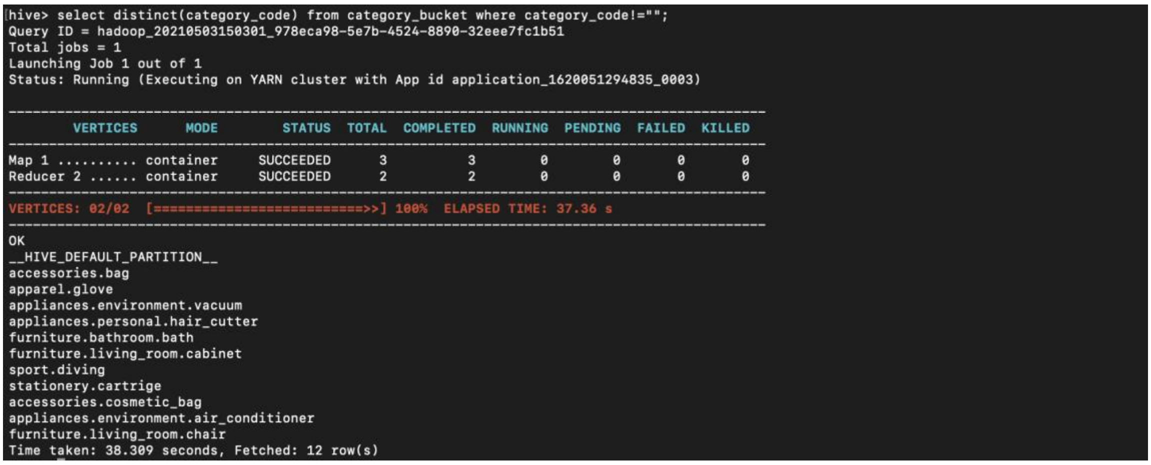


```
hive> insert into table category_bucket partition ( category_code ) select product_id , category_id , category_code from product ;
Query ID = hadoop_20210503150049_a43cdc08-dca8-45e5-9731-0fd827dbdc10
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1620051294835_0003)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      2          2        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      2          2        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 70.18 s
----------------------------------------------------------------------------------------------
Loading data to table case_study.category_bucket partition (category_code=null)

Loaded : 12/12 partitions.
         Time taken to load dynamic partitions: 0.795 seconds
         Time taken for adding to write entity : 0.004 seconds
OK
Time taken: 72.364 seconds
```

Query – 4

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

select distinct(category_code) from category_bucket where category_code!="";

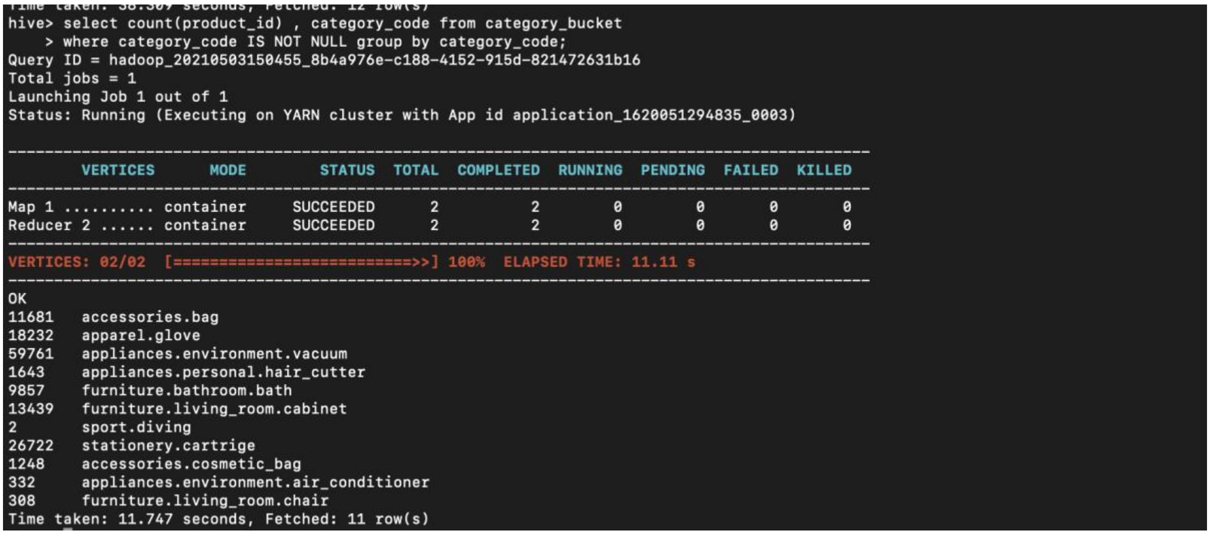


```
hive> select distinct(category_code) from category_bucket where category_code!="";
Query ID = hadoop_20210503150301_978eca98-5e7b-4524-8890-32eee7fc1b51
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1620051294835_0003)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      3          3        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      2          2        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 37.36 s
----------------------------------------------------------------------------------------------
OK
__HIVE_DEFAULT_PARTITION__
accessories.bag
apparel.glove
appliances.environment.vacuum
appliances.personal.hair_cutter
furniture.bathroom.bath
furniture.living_room.cabinet
sport.diving
stationery.cartrige
accessories.cosmetic_bag
appliances.environment.air_conditioner
furniture.living_room.chair
Time taken: 38.309 seconds, Fetched: 12 row(s)
```

Query – 5

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

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

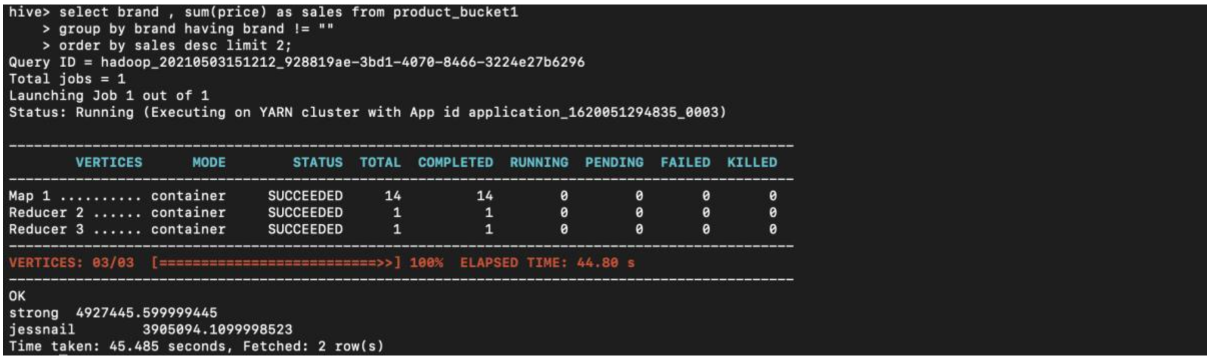


```
Time taken: 38.309 seconds, Fetched: 12 row(s)
hive> select count(product_id) , category_code from category_bucket
    > where category_code IS NOT NULL group by category_code;
Query ID = hadoop_20210503150455_8b4a976e-c188-4152-915d-821472631b16
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1620051294835_0003)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      2          2        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      2          2        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 11.11 s
----------------------------------------------------------------------------------------------
OK
11681   accessories.bag
18232   apparel.glove
59761   appliances.environment.vacuum
1643    appliances.personal.hair_cutter
9857    furniture.bathroom.bath
13439   furniture.living_room.cabinet
2       sport.diving
26722   stationery.cartrige
1248    accessories.cosmetic_bag
332     appliances.environment.air_conditioner
308     furniture.living_room.chair
Time taken: 11.747 seconds, Fetched: 11 row(s)
```

Query – 6

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

```
select brand , sum(price) as sales from product_bucket1
group by brand having brand != ""
order by sales desc limit 2;
```



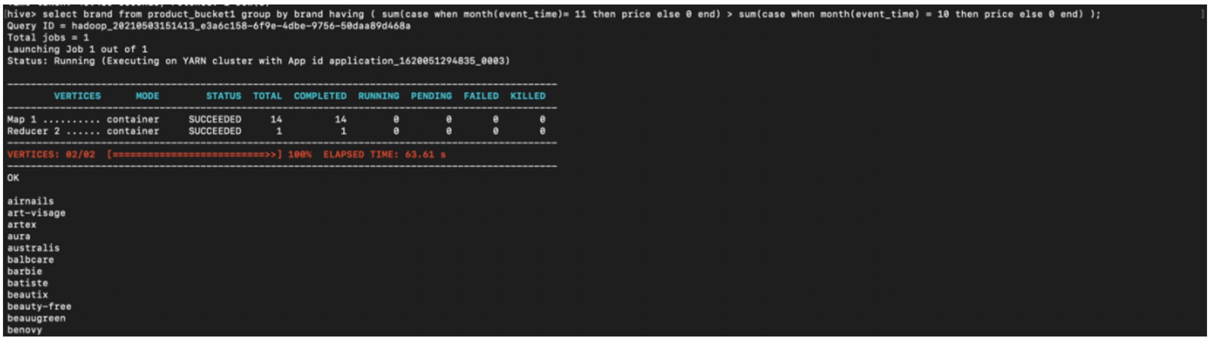
```
hive> select brand , sum(price) as sales from product_bucket1
    > group by brand having brand != ""
    > order by sales desc limit 2;
Query ID = hadoop_20210503151212_928819ae-3bd1-4070-8466-3224e27b6296
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1620051294835_0003)

----------------------------------------------------------------------------------------------
        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
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 44.80 s
----------------------------------------------------------------------------------------------
OK
strong  4927445.599999445
jessnail        3905094.1099998523
Time taken: 45.485 seconds, Fetched: 2 row(s)
```

Query – 7

Note: product_bucket1 is used here.

- Which brands increased their sales from October to November?

```
select brand from product_bucket1 group by brand having ( sum(case when
month(event_time)= 11 then price else 0 end) > sum(case when month(event_time) = 10 then
price else 0 end) );
```

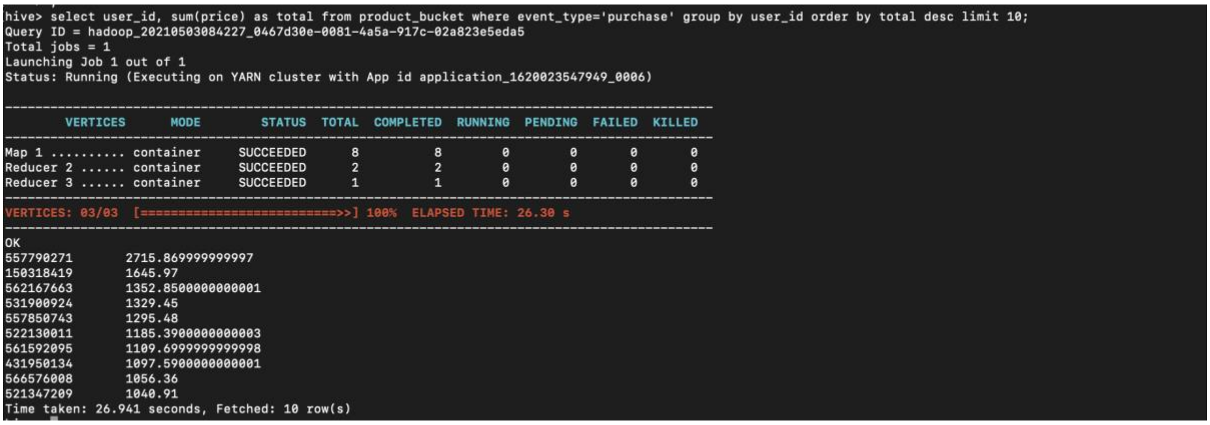


```
hive> select brand from product_bucket1 group by brand having ( sum(case when month(event_time)= 11 then price else 0 end) > sum(case when month(event_time) = 10 then price else 0 end) );
Query ID = hadoop_20210503151413_e3a6c158-6f9e-4dbe-9756-50daa89d468a
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1620051294835_0003)

----------------------------------------------------------------------------------------------
        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: 63.61 s
----------------------------------------------------------------------------------------------
OK

airnails
art-visage
artex
aura
australis
balbcare
barbie
batiste
beautix
beauty-free
beauugreen
benovy
```

Query – 8 Partitioned

- 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.

Answer - select user_id, sum(price) as total from product_bucket where event_type='purchase' group by user_id order by total desc limit 10;

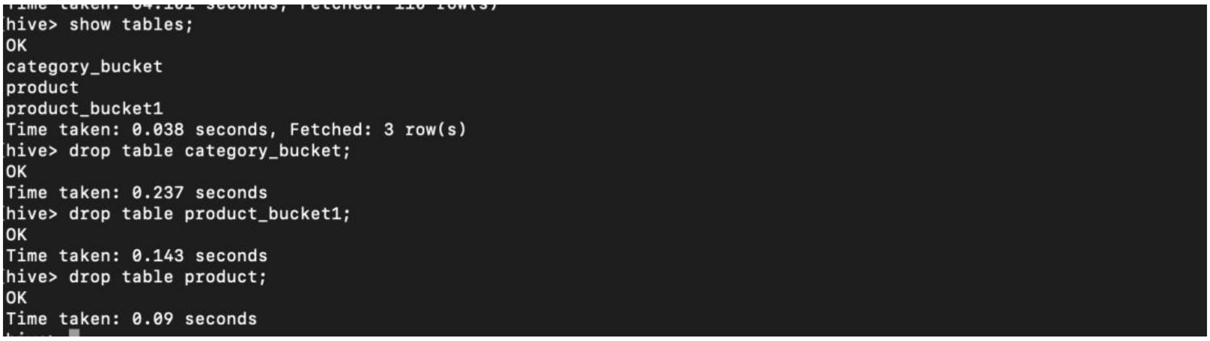


```
hive> select user_id, sum(price) as total from product_bucket where event_type='purchase' group by user_id order by total desc limit 10;
Query ID = hadoop_20210503084227_0467d30e-0081-4a5a-917c-02a823e5eda5
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1620023547949_0006)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      8          8        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      2          2        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 26.30 s
----------------------------------------------------------------------------------------------
OK
557790271       2715.869999999997
150318419       1645.97
562167663       1352.8500000000001
531900924       1329.45
557850743       1295.48
522130011       1185.3900000000003
561592095       1109.6999999999998
431950134       1097.5900000000001
566576008       1056.36
521347209       1040.91
Time taken: 26.941 seconds, Fetched: 10 row(s)
```

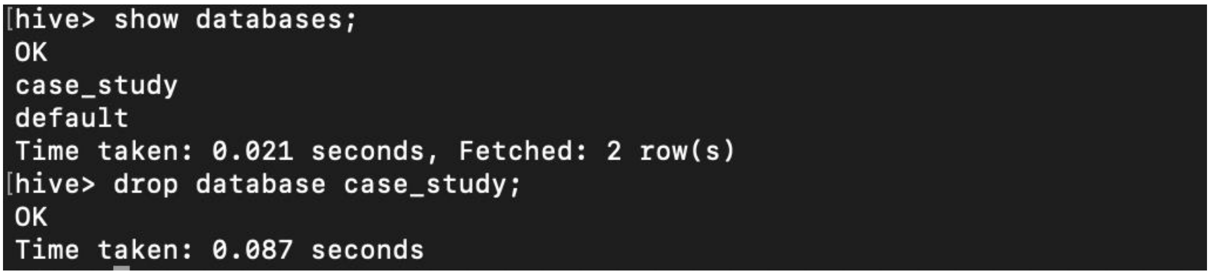
Time_Taken Parameter is changed by a major %.

Cleaning Up -

Dropping all the tables and db



```
hive> show tables;
OK
category_bucket
product
product_bucket1
Time taken: 0.038 seconds, Fetched: 3 row(s)
hive> drop table category_bucket;
OK
Time taken: 0.237 seconds
hive> drop table product_bucket1;
OK
Time taken: 0.143 seconds
hive> drop table product;
OK
Time taken: 0.09 seconds
```



```
[hive> show databases;
 OK
 case_study
 default
 Time taken: 0.021 seconds, Fetched: 2 row(s)
[hive> drop database case_study;
 OK
 Time taken: 0.087 seconds
```