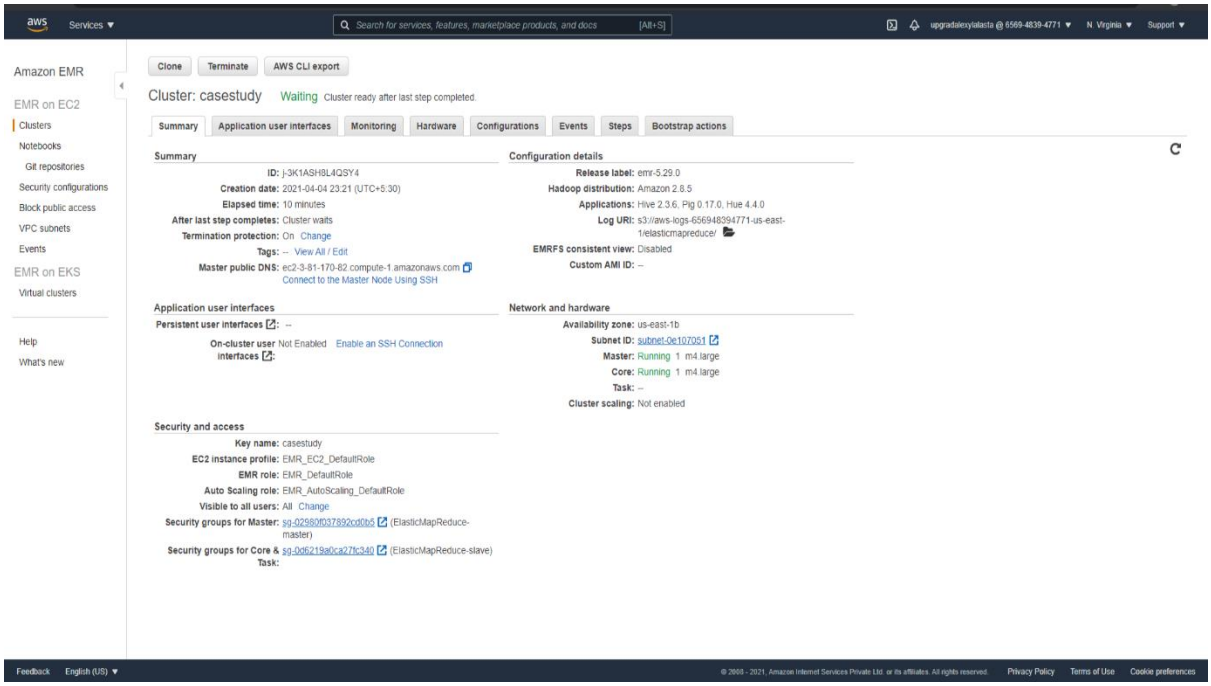
HIVE CASE STUDY

Submitted by

ALEXY LALAS T A

MAYURI SIRCAR

PART 1: Cluster details and starting of putty

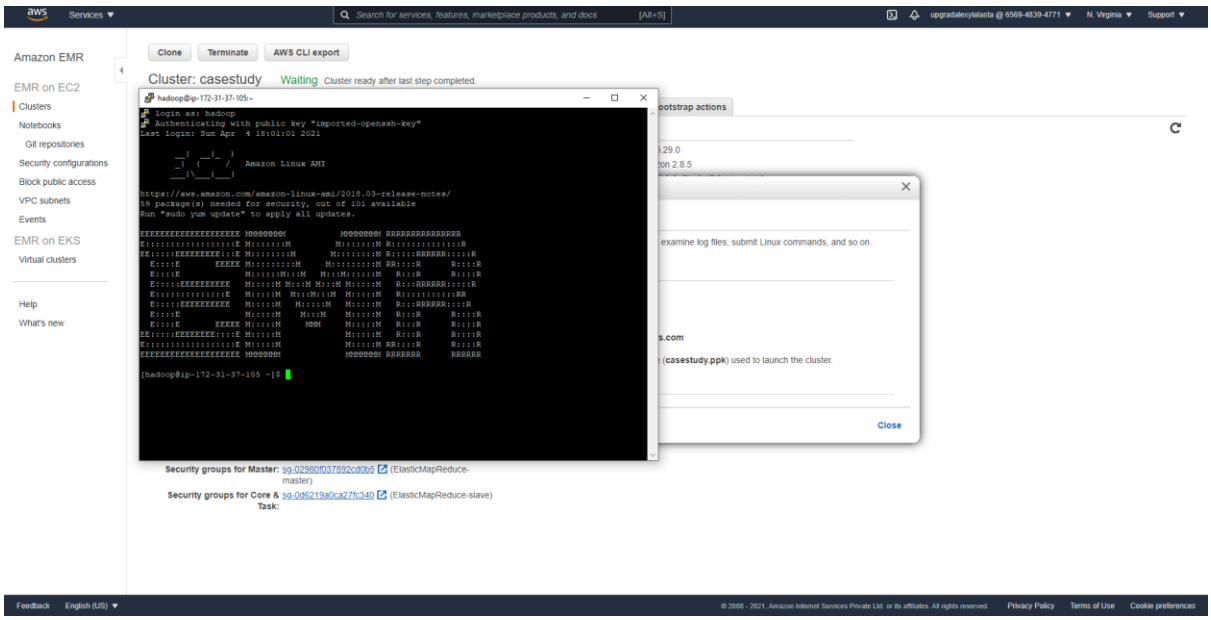


The above screen shot tells us about the cluster that has been created

Name of cluster : '**casestudy'**

No.of nodes : **1 master node and 1 slave node**

EMR version used is '**emr-5.29.0'**



- The putty has been started

i) Working with datasets and warehouses

Find the dataset file present in the hadoop.

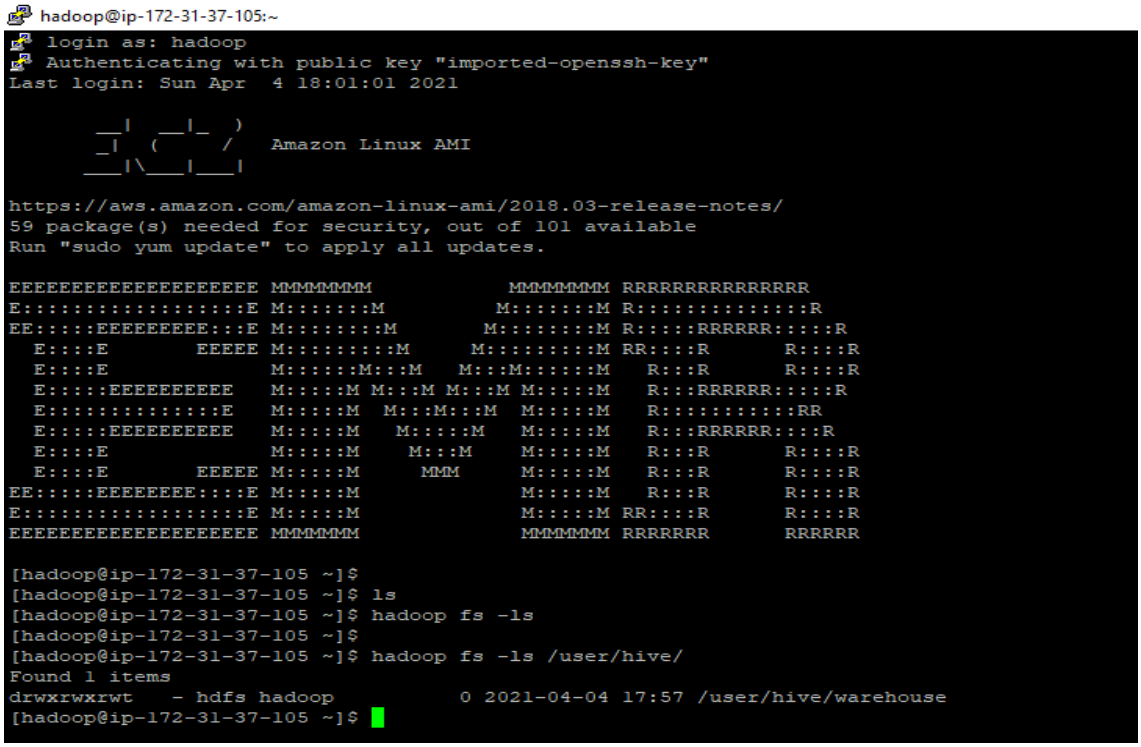
Code : **Hadoop fs -ls**

- This code helps in finding the dataset present in the Hadoop.

Checking the data directory in the hadoop .

Code: **hadoop fs -ls /user/hive/**

- This command helps in finding the directories.Only a default directory known as '**warehouse'** is present



ii) Creating a new directory:

Creating a directory in HDFS.

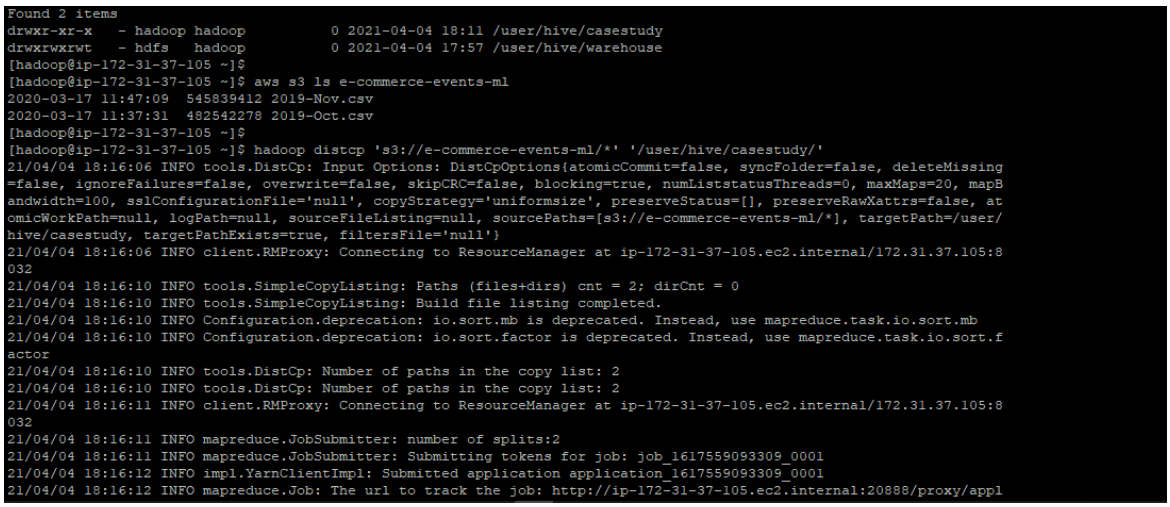
Code : **hadoop fs -mkdir /user/hive/casestudy**

- Mkdir : this command helps in the making of a new directory called "**casestudy**"

Checking the whether the directory has been created

Code : **hadoop fs -ls /user/hive/**

- Result : The output shows that a directory called "**casestudy**" has been created



Loading the s3 public data set to directory "**Casestudy**" in hadoop .

Code : **hadoop distcp 's3://e-commerce-events-ml/*' '/user/hive/casestudy/'**

- This command loads the data from the public dataset to the casestudy directory

```
                S3: Number of bytes written=0
                S3: Number of read operations=0
                S3: Number of large read operations=0
                S3: Number of write operations=0
        Job Counters
                Launched map tasks=2
                Other local map tasks=2
                Total time spent by all maps in occupied slots (ms)=2036800
                Total time spent by all reduces in occupied slots (ms)=0
                Total time spent by all map tasks (ms)=63650
                Total vcore-milliseconds taken by all map tasks=63650
                Total megabyte-milliseconds taken by all map tasks=65177600
        Map-Reduce Framework
                Map input records=2
                Map output records=0
                Input split bytes=272
                Spilled Records=0
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=1273
                CPU time spent (ms)=43780
                Physical memory (bytes) snapshot=977768448
                Virtual memory (bytes) snapshot=6593753088
                Total committed heap usage (bytes)=786432000
        File Input Format Counters
                Bytes Read=626
        File Output Format Counters
                Bytes Written=0
        DistCp Counters
                Bytes Copied=1028381690
                Bytes Expected=1028381690
                Files Copied=2
[hadoop@ip-172-31-37-105 ~]$
[hadoop@ip-172-31-37-105 ~]$
```

- This screen shots show that datas **“2019-Oct.csv”** and **“2019-Nov.csv”** has been copied to the casestudy directory

```
[hadoop@ip-172-31-37-105 ~]$
[hadoop@ip-172-31-37-105 ~]$ hadoop fs -ls /user/hive/casestudy/
Found 2 items
-rw-r--r--   1 hadoop hadoop  545839412 2021-04-04 18:16 /user/hive/casestudy/2019-Nov.csv
-rw-r--r--   1 hadoop hadoop  482542278 2021-04-04 18:16 /user/hive/casestudy/2019-Oct.csv
[hadoop@ip-172-31-37-105 ~]$ hadoop fs -cat /user/hive/casestudy/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,2.62,430174032,73deale7-664e-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-172-31-37-105 ~]$ hadoop fs -cat /user/hive/casestudy/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-dd9644b1d5f
7
2019-11-01 00:00:24 UTC,remove_from_cart,5826182,1487580007483048900,,,3.33,553329724,2067216c-31b5-455d-a1cc-af0575a
34ffb
2019-11-01 00:00:24 UTC,remove_from_cart,5826182,1487580007483048900,,,3.33,553329724,2067216c-31b5-455d-a1cc-af0575a
34ffb
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-5d8
a6a92839a
cat: Unable to write to output stream.
[hadoop@ip-172-31-37-105 ~]$
```

Checking of the files in the directory

Code : **Hadoop fs -ls /user/hive/casestudy/**

- The output shows that the data has a copied to the casestudy directory

Checking whether the data copied is correct

Code: **hadoop fs -cat /user/hive/casestudy/2019-Oct.csv | head**

Code: **hadoop fs -cat /user/hive/casestudy/2019-Nov.csv | head**

Part 2 : Starting of hive

```
hadoop@ip-172-31-37-105:~
[hadoop@ip-172-31-37-105 ~]$ hive

Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j2.properties Async: false
hive> show databases ;
OK
default
Time taken: 1.269 seconds, Fetched: 1 row(s)
hive> create database if not exists hivecasestudy;
OK
Time taken: 0.058 seconds
hive> show databases ;
OK
default
hivecasestudy
Time taken: 0.016 seconds, Fetched: 2 row(s)
hive> use hivecasestudy;
OK
Time taken: 0.029 seconds
hive>
```

Checking the databases present

Code :**Show databases;**

Creating a new database

Code : **create database if not exists hivecasestudy;**

Use the right database

Code : **use hivecasestudy;**

```
hive> CREATE EXTERNAL TABLE IF NOT EXISTS basedata (event_time timestamp, event_type string ,product_id string ,
    > category_id string , category_code string ,brand string , price float, user_id bigint , user_session string )
    > ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.OpenCSVSerde'
    > STORED AS TEXTFILE
    > LOCATION '/user/hive/casestudy/'
    > tblproperties('skip.header.line.count'='1');
OK
Time taken: 0.058 seconds
hive> select  * from basedata 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: 0.189 seconds, Fetched: 5 row(s)
hive>
```

Creating a table for the working with data

Code : **CREATE EXTERNAL TABLE IF NOT EXISTS basedata (event_time timestamp, event_type string ,product_id string , category_id string , category_code string ,brand string , price float, user_id bigint , user_session string)**

ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.OpenCSVSerde'

STORED AS TEXTFILE

LOCATION '/user/hive/casestudy/'

tblproperties('skip.header.line.count'='1');

- The above code helps in the creation of the table with just the help of location

partitioning and clustering

```
hive> set hive.exec.dynamic.partition.mode=nonstrict;
hive> set hive.exec.dynamic.partition=true;
hive> set hive.enforce.bucketing=true;
hive> create table if not exists data_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 20 buckets
    > ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.OpenCSVSerde'
    > STORED AS TEXTFILE;
OK
Time taken: 0.018 seconds
hive>
    > insert into table data_bucket partition(event_type) select event_time,product_id,category_id,category_code,brand,price,user_id,user_session,event_type from basedata;
Query ID = hadoop_20210405164446_5293894b-d593-4b15-a78a-c127b3932c3e
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_1617632176718_0010)

--------------------------------------------------------------------------------
        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: 157.04 s
--------------------------------------------------------------------------------
Loading data to table hivecasestudy.data_bucket partition (event_type=null)

Loaded : 4/4 partitions.
        Time taken to load dynamic partitions: 0.742 seconds
        Time taken for adding to write entity : 0.009 seconds
OK
Time taken: 166.858 seconds
hive>
```

Code :

commands for partitioning

---

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

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

hive>set hive.enforce.bucketing=true;

- These above commands enables us to work on the dynamic partitioning

Creation of table for partitioning and clustering

Code :

create table if not exists data_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 20 buckets

ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.OpenCSVSerde'

STORED AS TEXTFILE;

Loading into optimized table

code :

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

- The partition has been done on the basis of the event_type and clustered on category_code into 20 buckets

QUERIES

Question 1:

Find the revenue generated due to purchases made in October

Answer

```
hive> SELECT SUM(price) as total_revenue from basedata WHERE month(event_time)=10 and
    > event_type = 'purchase';
Query ID = hadoop_20210404190621_a697c2d9-64f4-410c-a36c-918b4641e661
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1617559093309_0010)

--------------------------------------------------------------------------------
        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: 73.20 s
--------------------------------------------------------------------------------
OK
1211538.429999982
Time taken: 73.972 seconds, Fetched: 1 row(s)
hive>
```

Code : **SELECT SUM(price) as total_revenue**

from basedata

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

- The query helps in the finding the total revenue of the month october

Question 2

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

Answer

```
hive> SELECT SUM( CASE WHEN MONTH(event_time) = '10'THEN price else 0 end) AS Oct_purchase,
    > SUM( CASE WHEN MONTH(event_time) = '11'THEN price else 0 end) AS Nov_purchase
    > FROM basedata
    > WHERE event_type = 'purchase';
Query ID = hadoop_20210404190903_5bf6424b-d1bf-4865-aa2f-388578ab6aa1
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1617559093309_0010)

--------------------------------------------------------------------------------
        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: 63.88 s
--------------------------------------------------------------------------------
OK
1211538.429999982       1531016.8999999657
Time taken: 64.515 seconds, Fetched: 1 row(s)
hive>
```

Code :**SELECT SUM(CASE WHEN MONTH(event_time) = '10'THEN price else 0 end) AS Oct_purchase,**

SUM(CASE WHEN MONTH(event_time) = '11'THEN price else 0 end) AS Nov_purchase

FROM data_bucket

WHERE event_type = 'purchase';

- The query helps in finding the revenue of both October and November separately

Question 3

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

```
hive> WITH revenue_diff AS
    > (SELECT
    >  SUM(case when MONTH(event_time) = '10' then price else 0 end) AS Oct_purchase,
    >  SUM(case when MONTH(event_time) = '11' then price else 0 end) AS Nov_purchase
    >  FROM basedata
    >  WHERE event_type= 'purchase'
    >  ) SELECT (Nov_purchase - Oct_purchase) as revenue_diff FROM revenue_diff ;
Query ID = hadoop_20210404191509_19bcdc98-2b3d-4cbb-a27a-15422e1ea6fe
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1617559093309_0010)

----------------------------------------------------------------------------------------------
        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: 64.25 s
----------------------------------------------------------------------------------------------
OK
319478.4699999837
Time taken: 64.747 seconds, Fetched: 1 row(s)
hive>
```

Code :

WITH revenue_diff AS

(SELECT

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

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

FROM basedata

WHERE event_type= 'purchase'

) SELECT (Nov_purchase - Oct_purchase) as revenue_diff FROM revenue_diff ;

- The revenue difference between both the month of October and the November has been calculated with this code

Question 4

Find the distinct categories of products. Categories with null value can be ignored.

Answer :

```
hive> SELECT distinct(category_code) as Category_codes FROM basedata WHERE category_code !='' ;
Query ID = hadoop_20210404191802_9cc19f17-3b07-4641-a21f-43ba4d4ca79c
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1617559093309_0010)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      8          8        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      5          5        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 61.66 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: 62.333 seconds, Fetched: 12 row(s)
hive>
```

Code:

SELECT distinct(category_code) as Category_codes

FROM basedata

WHERE category_code !='' ;

Distinct category_code has been retrieved from the data set

Question 5:

Find the total number of products available under each category.

Answer:

```
hive> SELECT category_code, count(product_id) as num_of_products FROM basedata WHERE category_code !='' GROUP BY category_code ;
Query ID = hadoop_20210404192103_922b31d1-925b-4967-abc7-8dcbaf7a3d1f
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1617559093309_0010)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      8          8        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      5          5        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 61.04 s
--------------------------------------------------------------------------------
OK
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: 61.622 seconds, Fetched: 12 row(s)
hive>
```

Code:

SELECT category_code, count(product_id) as num_of_products

FROM basedata

WHERE category_code !=''

GROUP BY category_code ;

- This query helps in the total number in each category products.

Question 6:

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

Answer:

```
hive> SELECT brand,sum(price) as total_price
    > from basedata
    > where brand !='' and event_type ='purchase'
    > group by brand
    > order by total_price desc limit 1;
Query ID = hadoop_20210404192520_6e04452f-ec67-4854-9e16-08b25cd564d2
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1617559093309_0010)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      8          8        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: 62.18 s
--------------------------------------------------------------------------------
OK
runail  148297.93999999977
Time taken: 62.898 seconds, Fetched: 1 row(s)
hive>
```

Code: SELECT brand,sum(price) as total_price

from basedata

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

group by brand

order by total_price desc limit 1;

- The query helps in finding the maximum sales from the month of October and November .

Question 7 :

Which brands increased their sales from October to November?

```
hive> WITH revenue_diff AS
    > (SELECT brand,SUM(case when MONTH(event_time) = '10' then price else 0 end) AS Oct_purchase,
    > SUM(case when MONTH(event_time) = '11' then price else 0 end) AS Nov_purchase
    > FROM basedata
    > WHERE event_type= 'purchase' group by brand)
    > SELECT brand FROM revenue_diff WHERE (Nov_purchase - Oct_purchase) > 0 ;
Query ID = hadoop_20210404192919_9a0bd1ee-1835-403f-b1dc-0562069d0efb
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1617559093309_0010)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      8          8        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      3          3        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 66.29 s
--------------------------------------------------------------------------------
OK
airnails
artex
binacil
bioaqua
blixz
bluesky
bpw.style
carmex
chi
concept
cosima
```

```
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: 66.769 seconds, Fetched: 161 row(s)
hive>
    >
```

Code:

WITH revenue_diff AS

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

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

FROM basedata

WHERE event_type= 'purchase' group by brand)

SELECT brand FROM revenue_diff WHERE (Nov_purchase - Oct_purchase) > 0 ;

Same query when done with partitioning

```
hive> WITH revenue_diff AS
    > (SELECT brand,SUM(case when MONTH(event_time) = '10' then price else 0 end) AS Oct_purchase,
    > SUM(case when MONTH(event_time) = '11' then price else 0 end) AS Nov_purchase
    > FROM data_bucket
    > WHERE event_type= 'purchase' group by brand)
    > SELECT brand FROM revenue_diff WHERE (Nov_purchase - Oct_purchase) > 0 ;
Query ID = hadoop_20210405170333_626d1011-e822-4281-819a-ef19985d77f2
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1617632176718_0010)

--------------------------------------------------------------------------------
        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: 17.99 s
--------------------------------------------------------------------------------
OK

airnails
art-visage
skinity
skinlite
smart
soleo
solomeya
sophin
staleks
strong
supertan
swarovski
tertio
treaclemoon
trind
uno
uskusi
veraclara
vilenta
yoko
yu-r
zeitun
Time taken: 18.477 seconds, Fetched: 161 row(s)
hive>
```

A total of 161 brands has been selected where the brand have a revenue of october greater than november.

- The optimization plays an important role in the large datas as the optimized data takes just **18 sec** to complete the code while a normal format required **66 sec** to complete the code

- So partitioning and clustering acts as an important factor in the quick access of the data

Question 8:

Your Company wants to reward the top 10 users in its website with a Golden Customer plan.

Write a query to generate a list of top 10 users who spent the most.

```
hive>
    > With golden_customer AS
    > (SELECT user_id,SUM(price) AS total_price
    > FROM basedata
    > WHERE event_type = "purchase"
    > GROUP BY user_id
    > ORDER BY total_price DESC LIMIT 10)
    > SELECT user_id from golden_customer ;
Query ID = hadoop_20210404193443_30c1febc-41db-4fc1-b71c-0947dd3f44d7
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1617559093309_0010)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      8          8        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: 64.36 s
----------------------------------------------------------------------------------------------
OK
557790271
150318419
562167663
531900924
557850743
522130011
561592095
431950134
566576008
521347209
Time taken: 64.995 seconds, Fetched: 10 row(s)
hive>
```

Code:

With golden_customer AS

(SELECT user_id,SUM(price) AS total_price

FROM basedata

WHERE event_type = "purchase"

GROUP BY user_id

ORDER BY total_price DESC LIMIT 10)

SELECT user_id from golden_customer ;

- The above considered 10 users is the ones that are to be rewarded as ,they are the users who spent the most.

Part 3

Dropping of database

```
OK
Time taken: 5.433 seconds
hive> drop table data_bucket ;
OK
Time taken: 0.088 seconds
hive> drop table basedata ;
OK
Time taken: 0.015 seconds
```

Code : **drop table data_bucket ;**

- This code drops the partition table that was created for the study

Code : **drop table basedata;**

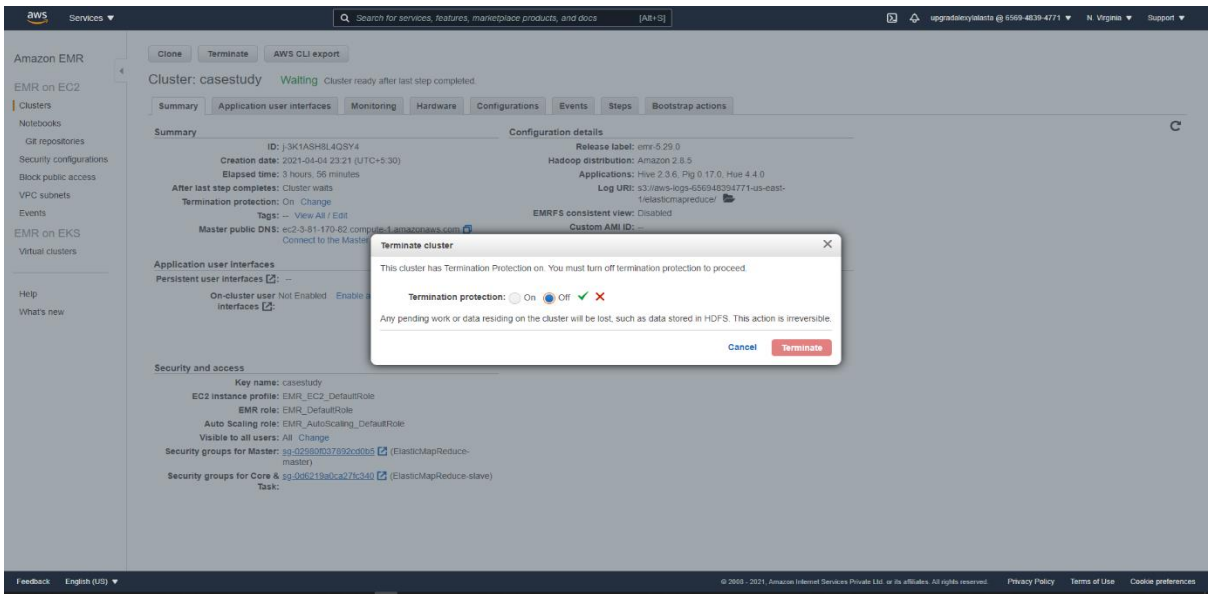
- This code drops the external table that we created for the study

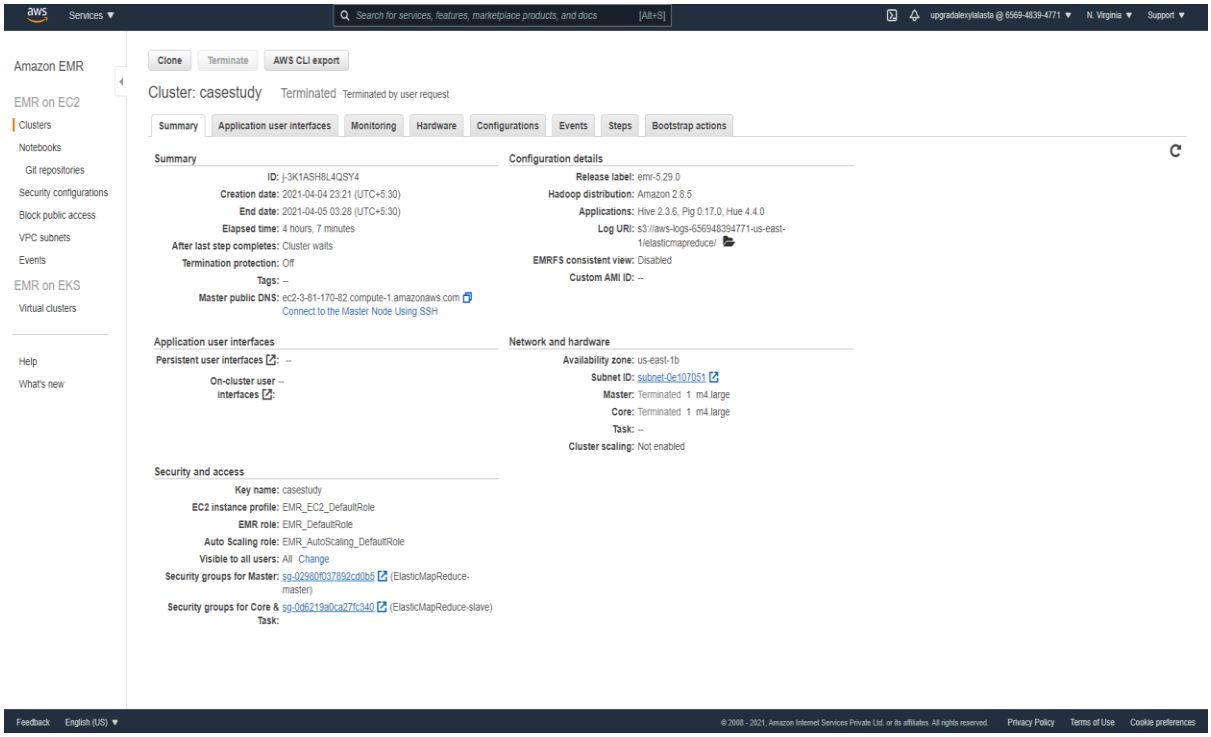
```
hive> drop database hivecasestudy ;
OK
Time taken: 0.048 seconds
hive>
```

Code : **drop database hive;**

- This code drops the database that was created for the study

- Termination of cluster




Services
Search for services, features, marketplace products, and docs [Alt+S]
upgradalexylalasta @ 6569-4939-4771
N. Virginia
Support
Amazon EMR
EMR on EC2
Clusters
Notebooks
Git repositories
Security configurations
Block public access
VPC subnets
Events
EMR on EKS
Virtual clusters
Help
What's new
Clone
Terminate
AWS CLI export
Cluster: casestudy
Terminated
Terminated by user request
Summary
Application user interfaces
Monitoring
Hardware
Configurations
Events
Steps
Bootstrap actions
Summary
ID: j-3K1ASH8L4QSY4
Creation date: 2021-04-04 23:21 (UTC+5:30)
End date: 2021-04-05 03:28 (UTC+5:30)
Elapsed time: 4 hours, 7 minutes
After last step completes: Cluster waits
Termination protection: Off
Tags: --
Master public DNS: ec2-3-81-170-82.compute-1.amazonaws.com
Connect to the Master Node Using SSH
Configuration details
Release label: emr-5.29.0
Hadoop distribution: Amazon 2.8.5
Applications: Hive 2.3.6, Pig 0.17.0, Hue 4.4.0
Log URI: s3://aws-logs-656948394771-us-east-1/elasticmapreduce/
EMRFS consistent view: Disabled
Custom AMI ID: --
Application user interfaces
Persistent user interfaces: --
On-cluster user interfaces: --
Network and hardware
Availability zone: us-east-1b
Subnet ID: subnet-0e107051
Master: Terminated 1 m4.large
Core: Terminated 1 m4.large
Task: --
Cluster scaling: Not enabled
Security and access
Key name: casestudy
EC2 instance profile: EMR_EC2_DefaultRole
EMR role: EMR_DefaultRole
Auto Scaling role: EMR_AutoScaling_DefaultRole
Visible to all users: All Change
Security groups for Master: sg-02980f037892cd0b5 (ElasticMapReduce-master)
Security groups for Core & Task: sg-0d6219a0ca27fc340 (ElasticMapReduce-slave)
Feedback
English (US)
© 2008 - 2021, Amazon Internet Services Private Ltd. or its affiliates. All rights reserved.
Privacy Policy
Terms of Use
Cookie preferences

Thank you