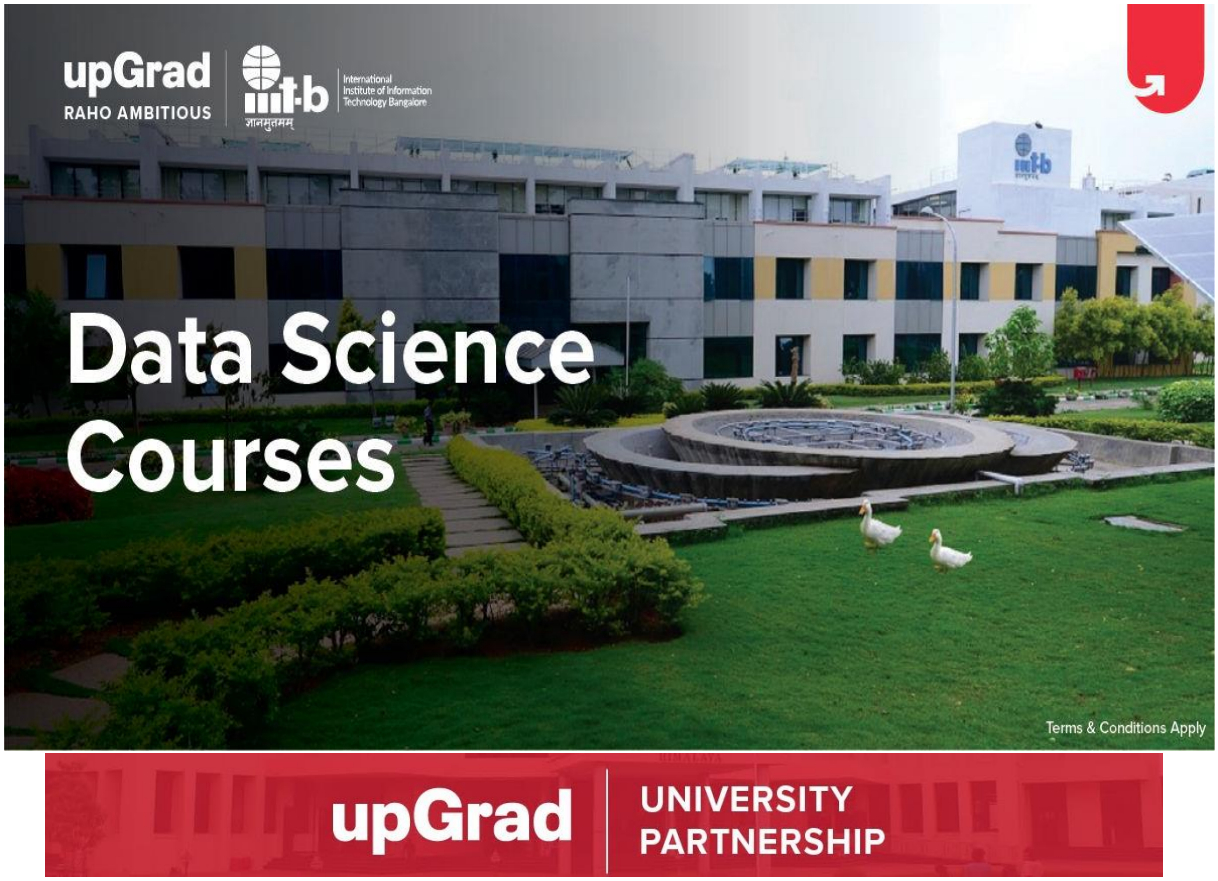


Hive Case study -DA track

Submission by :

1] SIDHESH TONAPE

2] JYOTIRMAYEE SAHOO

PROBLEM STATEMENT:

In this modern era, Tech companies are exploring ways to improve their sales by analyzing customer behavior and gaining insights into product trends. In order to make better business decision, E-commerce websites are finding their way by tracking the number of clicks made by customers and their spending time on websites in searching for patterns within them.

OBJECTIVE:

The aim is to extract the data and gather insights from a real-life data set of an e-commerce company

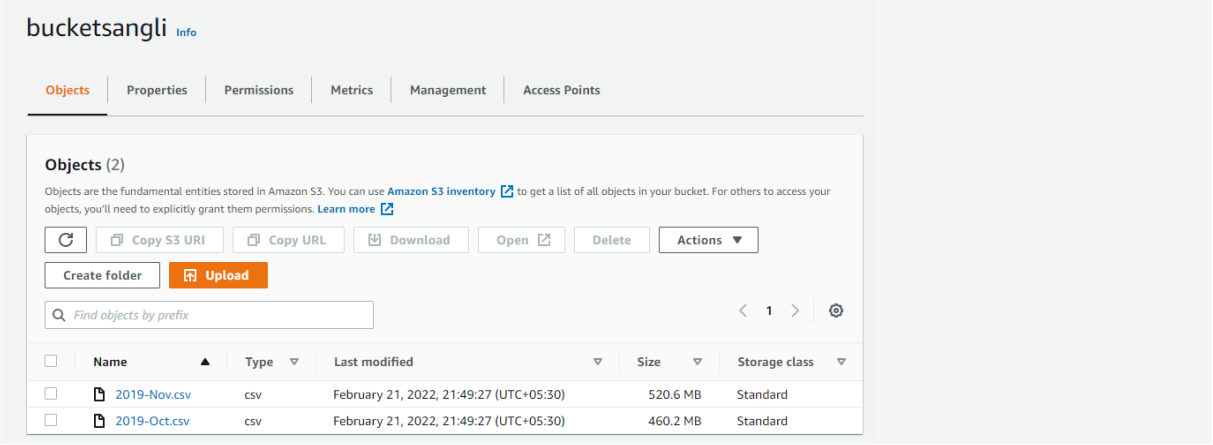
Business Objectives:

Using a public clickstream dataset of a cosmetics store we need to extract valuable insights that can improve their sales.

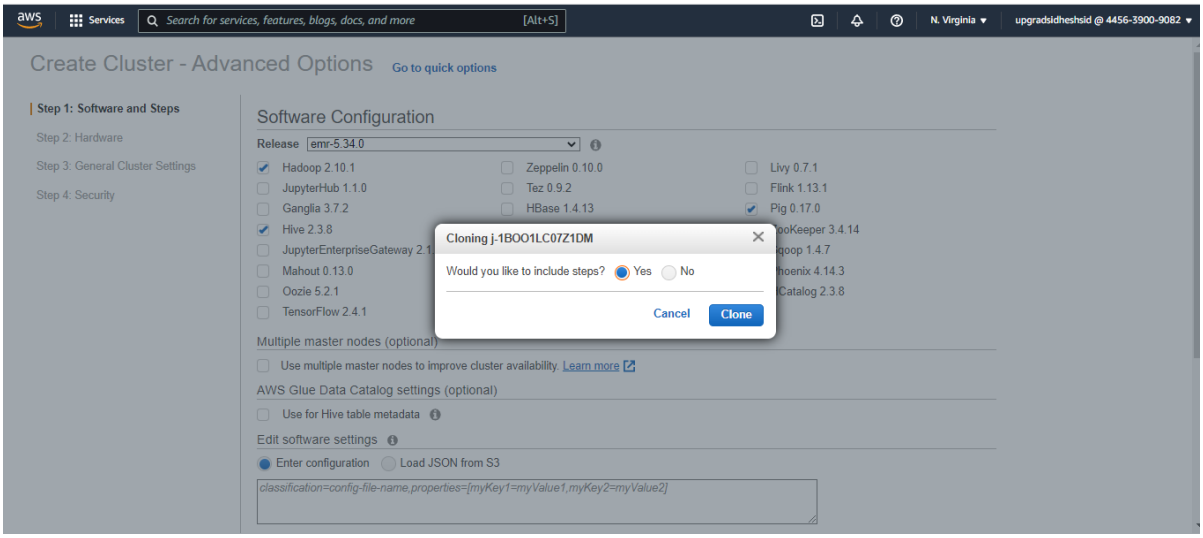
You will find the data in the link given below.

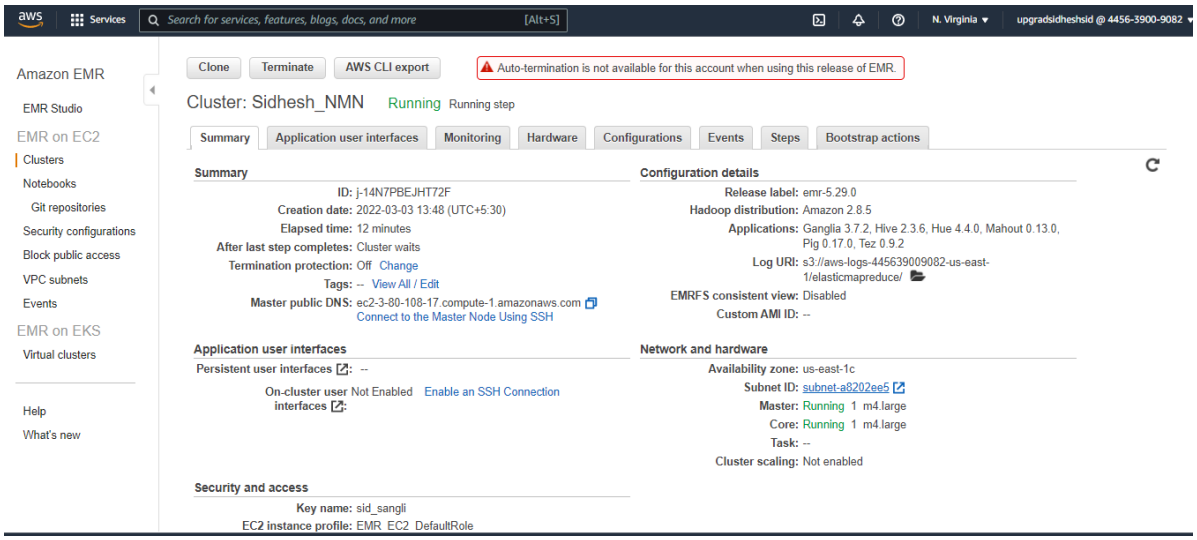
https://e-commerce-events-ml.s3.amazonaws.com/2019-Oct.csv

https://e-commerce-events-ml.s3.amazonaws.com/2019-Nov.csv



Creating an EMR Cluster





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.

Windows | Mac / Linux

1. Download PuTTY.exe to your computer from:
   http://www.chiark.greenend.org.uk/~sgtatham/putty/download.html
2. Start PuTTY.
3. In the Category list, click Session.
4. In the Host Name field, type hadoop@ec2-3-80-108-17.compute-1.amazonaws.com
5. In the Category list, expand Connection > SSH, and then click Auth.
6. For Private key file for authentication, click Browse and select the private key file (**sid_sangli.ppk**) used to launch the cluster.
7. Click Open.
8. Click Yes to dismiss the security alert.

Terminal Opened. Created a new Directory to copy the files into HDFS from S3

```
Using username "hadoop".
Authenticating with public key "sid_sangli"
Last login: Thu Mar  3 08:30:35 2022

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

https://aws.amazon.com/amazon-linux-ami/2018.03-release-notes/
71 package(s) needed for security, out of 106 available
Run "sudo yum update" to apply all updates.

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 MMMMMM             MMMMMM RRRRRRR      RRRRRR

[hadoop@ip-172-31-19-96 ~]$
[hadoop@ip-172-31-19-96 ~]$
[hadoop@ip-172-31-19-96 ~]$ hadoop fs -ls /
Found 4 items
drwxr-xr-x   - hdfs hadoop          0 2022-03-03 08:25 /apps
drwxrwxrwt   - hdfs hadoop          0 2022-03-03 08:27 /tmp
drwxr-xr-x   - hdfs hadoop          0 2022-03-03 08:25 /user
drwxr-xr-x   - hdfs hadoop          0 2022-03-03 08:25 /var
[hadoop@ip-172-31-19-96 ~]$
[hadoop@ip-172-31-19-96 ~]$ hadoop fs -mkdir /hivesid_case_study
[hadoop@ip-172-31-19-96 ~]$ hadoop fs -ls /
Found 5 items
drwxr-xr-x   - hdfs   hadoop          0 2022-03-03 08:25 /apps
drwxr-xr-x   - hadoop hadoop          0 2022-03-03 08:34 /hivesid_case_study
drwxrwxrwt   - hdfs   hadoop          0 2022-03-03 08:27 /tmp
drwxr-xr-x   - hdfs   hadoop          0 2022-03-03 08:25 /user
drwxr-xr-x   - hdfs   hadoop          0 2022-03-03 08:25 /var
[hadoop@ip-172-31-19-96 ~]$
```

Creating a directory and checking the Loaded data , Connecting to HIVE

```
[hadoop@ip-172-31-19-96 ~]$
[hadoop@ip-172-31-19-96 ~]$ hadoop fs -ls /
Found 4 items
drwxr-xr-x   - hdfs hadoop          0 2022-03-03 08:25 /apps
drwxrwxrwt   - hdfs hadoop          0 2022-03-03 08:27 /tmp
drwxr-xr-x   - hdfs hadoop          0 2022-03-03 08:25 /user
drwxr-xr-x   - hdfs hadoop          0 2022-03-03 08:25 /var
[hadoop@ip-172-31-19-96 ~]$
[hadoop@ip-172-31-19-96 ~]$ hadoop fs -mkdir /hivesid_case_study
[hadoop@ip-172-31-19-96 ~]$ hadoop fs -ls /
Found 5 items
drwxr-xr-x   - hdfs   hadoop          0 2022-03-03 08:25 /apps
drwxr-xr-x   - hadoop hadoop          0 2022-03-03 08:34 /hivesid_case_study
drwxrwxrwt   - hdfs   hadoop          0 2022-03-03 08:27 /tmp
drwxr-xr-x   - hdfs   hadoop          0 2022-03-03 08:25 /user
drwxr-xr-x   - hdfs   hadoop          0 2022-03-03 08:25 /var
[hadoop@ip-172-31-19-96 ~]$
```

```
[hadoop@ip-172-31-19-96 ~]$
[hadoop@ip-172-31-19-96 ~]$ hadoop fs -ls /hivesid_case_study
Found 2 items
-rw-r--r--   1 hadoop hadoop  545839412 2022-03-03 08:38 /hivesid_case_study/2019-Nov.csv
-rw-r--r--   1 hadoop hadoop  482542278 2022-03-03 08:36 /hivesid_case_study/2019-Oct.csv
[hadoop@ip-172-31-19-96 ~]$
[hadoop@ip-172-31-19-96 ~]$ hive

Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j2.properties Async: false
```

Launched Hive Session and created database .

```
[hadoop@ip-172-31-19-96 ~]$ hive

Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j2.properties Async: false
hive> SHOW DATABASES;
OK
default
Time taken: 1.252 seconds, Fetched: 1 row(s)
hive> CREATE DATABASE IF NOT EXISTS clicksid_stream_data;
OK
Time taken: 0.559 seconds
hive>
    > set hive.cli.print.header=true;
hive>
    >
    > USE clicksid_stream_data;
OK
Time taken: 0.074 seconds
hive>
```

Creating new table “retail”

```
hive>
    > CREATE EXTERNAL TABLE IF NOT EXISTS ecomsid_events (event_time timestamp, event_type string, product_id string, category_id string, category_code string, brand st
ring, price float, user_id bigint, user_session string) ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.OpenCSVSerde' STORED AS TEXTFILE LOCATION  '/hivesid_case_study'
 TBLPROPERTIES ('serialization.null.format'='', 'skip.header.line.count' = '1');
OK
Time taken: 0.558 seconds
hive>
```

Data got loaded into the table correctly

```
hive> SELECT * FROM ecomsid_events limit 7;
OK
ecomsid_events.event_time       ecomsid_events.event_type       ecomsid_events.product_id       ecomsid_events.category_id      ecomsid_events.category_code    ecomsid_
events.brand    ecomsid_events.price    ecomsid_events.user_id  ecomsid_events.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
Time taken: 2.966 seconds, Fetched: 7 row(s)
hive>
```

```
hive>
    > INSERT INTO TABLE ecomsid_events_part1 PARTITION (event_type) SELECT event_time, product_id, category_id, category_code, brand, price, user_id, user_session, even
t_type FROM ecomsid_events;
Query ID = hadoop_20220303084249_e7e4806e-911f-4b8c-8041-aa6d2df88dec
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1646296003361_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: 142.22 s
--------------------------------------------------------------------------------
Loading data to table clicksid_stream_data.ecomsid_events_part1 partition (event_type=null)

Loaded : 4/4 partitions.
         Time taken to load dynamic partitions: 0.597 seconds
         Time taken for adding to write entity : 0.005 seconds
OK
_col0   _col1   _col2   _col3   _col4   _col5   _col6   _col7   _col8
Time taken: 146.709 seconds
```

```
hive>
    > -- Checking Data of Partitioned event_type column table ecomsid_events_part1
    >
    > SELECT * FROM ecomsid_events_part1 LIMIT 7;
OK
ecomsid_events_part1.event_time ecomsid_events_part1.product_id ecomsid_events_part1.category_id        ecomsid_events_part1.category_code      ecomsid_events_part1.bra
nd      ecomsid_events_part1.price      ecomsid_events_part1.user_id    ecomsid_events_part1.user_session       ecomsid_events_part1.event_type
2019-10-09 19:28:14 UTC 5773158 1487580012969197740     NULL    irisk   2.79    558203129       b1e0b3d3-d60b-bd68-d7e5-500a7af8920f    cart
2019-10-07 20:53:09 UTC 5663062 1487580009622143014     NULL    NULL    1.43    251478914       a99a5589-0f7a-40a5-9748-b19961fc4d30    cart
2019-10-07 20:53:11 UTC 5796751 1487580004916134735     NULL    NULL    4.29    533966373       a834973c-30a5-4268-9fbc-597ea0865ff4    cart
2019-10-07 20:53:11 UTC 5835333 1926797403503985079     NULL    NULL    4.76    552795963       24632cad-25f2-d02c-cfe7-6a59a096565a    cart
2019-10-07 20:53:14 UTC 5854897 1487580009445982239     NULL    irisk   0.32    540100212       b445a30f-1eef-4275-9d0e-911e7bbdfbc2    cart
2019-10-07 20:53:21 UTC 5860194 1487580007634043851     NULL    NULL    0.63    438698700       3f107f57-4776-1988-bf33-03430f9d9160    cart
2019-10-07 20:53:23 UTC 5824194 1487580006551913373     NULL    domix   5.95    533267875       5d5364da-6082-4301-971c-be9ed27fd75e    cart
Time taken: 0.268 seconds, Fetched: 7 row(s)
```

HDFS Buckets Creating :

```
[hadoop@ip-172-31-19-96 ~]$ Buckets created in HDFS
-bash: Buckets: command not found
[hadoop@ip-172-31-19-96 ~]$ hadoop fs -ls /user/hive/warehouse/clicksid_stream_data.db/ecomsid_events_optimize
Found 4 items
drwxrwxrwt   - hadoop hadoop          0 2022-03-03 08:49 /user/hive/warehouse/clicksid_stream_data.db/ecomsid_events_optimize/event_type=cart
drwxrwxrwt   - hadoop hadoop          0 2022-03-03 08:49 /user/hive/warehouse/clicksid_stream_data.db/ecomsid_events_optimize/event_type=purchase
drwxrwxrwt   - hadoop hadoop          0 2022-03-03 08:49 /user/hive/warehouse/clicksid_stream_data.db/ecomsid_events_optimize/event_type=remove_from_cart
drwxrwxrwt   - hadoop hadoop          0 2022-03-03 08:49 /user/hive/warehouse/clicksid_stream_data.db/ecomsid_events_optimize/event_type=view
[hadoop@ip-172-31-19-96 ~]$
[hadoop@ip-172-31-19-96 ~]$
[hadoop@ip-172-31-19-96 ~]$ hadoop fs -ls /user/hive/warehouse/clicksid_stream_data.db/ecomsid_events_optimize/event_type=purchase
Found 10 items
-rwxrwxrwt   1 hadoop hadoop    8848048 2022-03-03 08:48 /user/hive/warehouse/clicksid_stream_data.db/ecomsid_events_optimize/event_type=purchase/000000_0
-rwxrwxrwt   1 hadoop hadoop    6082854 2022-03-03 08:49 /user/hive/warehouse/clicksid_stream_data.db/ecomsid_events_optimize/event_type=purchase/000001_0
-rwxrwxrwt   1 hadoop hadoop    5771952 2022-03-03 08:49 /user/hive/warehouse/clicksid_stream_data.db/ecomsid_events_optimize/event_type=purchase/000002_0
-rwxrwxrwt   1 hadoop hadoop    5810623 2022-03-03 08:48 /user/hive/warehouse/clicksid_stream_data.db/ecomsid_events_optimize/event_type=purchase/000003_0
-rwxrwxrwt   1 hadoop hadoop    3127215 2022-03-03 08:48 /user/hive/warehouse/clicksid_stream_data.db/ecomsid_events_optimize/event_type=purchase/000004_0
-rwxrwxrwt   1 hadoop hadoop    6592136 2022-03-03 08:48 /user/hive/warehouse/clicksid_stream_data.db/ecomsid_events_optimize/event_type=purchase/000005_0
-rwxrwxrwt   1 hadoop hadoop    6760201 2022-03-03 08:49 /user/hive/warehouse/clicksid_stream_data.db/ecomsid_events_optimize/event_type=purchase/000006_0
-rwxrwxrwt   1 hadoop hadoop    5320537 2022-03-03 08:49 /user/hive/warehouse/clicksid_stream_data.db/ecomsid_events_optimize/event_type=purchase/000007_0
-rwxrwxrwt   1 hadoop hadoop    6034152 2022-03-03 08:48 /user/hive/warehouse/clicksid_stream_data.db/ecomsid_events_optimize/event_type=purchase/000008_0
-rwxrwxrwt   1 hadoop hadoop    8035002 2022-03-03 08:48 /user/hive/warehouse/clicksid_stream_data.db/ecomsid_events_optimize/event_type=purchase/000009_0
[hadoop@ip-172-31-19-96 ~]$
```

Optimized Technique:

Scenario Used: The company wants to reward the top 10 users of its website with a Golden Customer plan. We need to write a query to generate a list of the top 10 users who spend the most.

Find the Result with Base Table without any Optimization Technique

```
hive> --Finding top 10 customers from the base table without any Optimization Technique
hive>
    > WITH Customer_Rank AS(SELECT user_id AS Customer, ROUND(SUM(price),2) AS Expenditure, RANK() OVER(ORDER BY ROUND(SUM(price),2) DESC) AS Rank FROM ecomsid_events W
HERE event_type = 'purchase' GROUP BY user_id) SELECT Customer, Expenditure, Rank FROM Customer_Rank WHERE Rank <=10;
Query ID = hadoop_20220303090106_e427320d-006a-4da7-b4b3-e70b7c0f595e
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_1646296003361_0004)

--------------------------------------------------------------------------------
        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      2          2        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 66.66 s
--------------------------------------------------------------------------------
OK
customer        expenditure     rank
557790271       2715.87 1
150318419       1645.97 2
562167663       1352.85 3
531900924       1329.45 4
557850743       1295.48 5
522130011       1185.39 6
561592095       1109.7  7
431950134       1097.59 8
566576008       1056.36 9
521347209       1040.91 10
Time taken: 79.448 seconds, Fetched: 10 row(s)
```

Partitioned Table using column event_type

```
hive> --Finding top 10 customers from the partitioned table
hive> WITH Customer_Rank AS (SELECT user_id AS Customer, ROUND(SUM(price),2) AS Expenditure, RANK() OVER(ORDER BY ROUND(SUM(price),2) DESC) AS Rank FROM ecomsid_events_
part1 WHERE event_type = 'purchase' GROUP BY user_id) SELECT Customer, Expenditure, Rank FROM Customer_Rank WHERE Rank <=10;
Query ID = hadoop_20220303090335_a336e2b6-6689-434e-a872-7972bab394dd
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1646296003361_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
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 19.70 s
--------------------------------------------------------------------------------
OK
customer        expenditure     rank
557790271       2715.87 1
150318419       1645.97 2
562167663       1352.85 3
531900924       1329.45 4
557850743       1295.48 5
522130011       1185.39 6
561592095       1109.7  7
431950134       1097.59 8
566576008       1056.36 9
521347209       1040.91 10
Time taken: 20.775 seconds, Fetched: 10 row(s)
```

Run Hive queries to answer the Case study questions.

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

Ans.-

SELECT ROUND(SUM(price),2) AS total_revenue_october FROM ecomsid_events_optimize WHERE event_type = 'purchase' GROUP BY month(event_time) HAVING month(event_time) = 10;

```
    > ;
hive> --The total revenue generated due to purchases made in October.
hive> SELECT ROUND(SUM(price),2) AS total_revenue_october FROM ecomsid_events_optimize WHERE event_type = 'purchase' GROUP BY month(event_time) HAVING month(event_time)
 = 10;
Query ID = hadoop_20220303090542_387479aa-101c-4af7-b7fb-d23ce66eae82
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1646296003361_0004)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 ......... container     SUCCEEDED      3          3        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 25.26 s
--------------------------------------------------------------------------------
OK
total_revenue_october
1211538.43
Time taken: 25.88 seconds, Fetched: 1 row(s)
hive>
```

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

Ans.- SELECT CASE WHEN (month(event_time) ==10) THEN 'Oct' ELSE 'Nov' END AS Month, ROUND(SUM(price),2) AS total_purchases FROM ecomsid_events_optimize WHERE event_type = 'purchase' GROUP BY month(event_time);

```
hive> --Total purchases per month
hive> SELECT CASE WHEN (month(event_time) ==10) THEN 'Oct' ELSE 'Nov' END AS Month, ROUND(SUM(price),2) AS total_purchases FROM ecomsid_events_optimize WHERE event_type
 = 'purchase' GROUP BY month(event_time);
Query ID = hadoop_20220303090856_8963b8a4-e4c9-4b02-9b29-af2fca3961db
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1646296003361_0004)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 ......... container     SUCCEEDED      3          3        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 23.62 s
--------------------------------------------------------------------------------
OK
month   total_purchases
Oct     1211538.43
Nov     1531016.9
Time taken: 24.256 seconds, Fetched: 2 row(s)
```

3] Write a query to find the change in revenue generated due to purchasesfrom October to November.

Ans.- SELECT (Rev_November - Rev_October) AS change_in_revenue FROM (SELECT ROUND(SUM(CASE WHEN month(event_time)=10 THEN price ELSE 0 END),2) AS Rev_October, ROUND(SUM(CASE WHEN month(event_time)=11 THEN price ELSE 0 END),2) AS Rev_November FROM ecomsid_events_optimize WHERE event_type='purchase') AS Rev_Dtls;

```
hive> --change in revenue generated due to purchases from October to November.
hive> SELECT (Rev_November - Rev_October) AS change_in_revenue FROM (SELECT ROUND(SUM(CASE WHEN month(event_time)=10 THEN price ELSE 0 END),2) AS Rev_October, ROUND(SUM
(CASE WHEN month(event_time)=11 THEN price ELSE 0 END),2) AS Rev_November FROM ecomsid_events_optimize WHERE event_type='purchase') AS Rev_Dtls;
Query ID = hadoop_20220303091011_206bbd4c-8386-4ef3-80dc-e27cb6d5254e
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1646296003361_0004)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 ..........  container     SUCCEEDED      3          3        0        0       0       0
Reducer 2 ......  container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 25.78 s
--------------------------------------------------------------------------------
OK
change_in_revenue
319478.47
Time taken: 26.526 seconds, Fetched: 1 row(s)
```

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

Ans.- SELECT DISTINCT category_code FROM ecomsid_events_optimize WHERE category_code IS NOT NULL;

```
hive> -- Distinct categories of products
hive> SELECT DISTINCT category_code FROM ecomsid_events_optimize WHERE category_code IS NOT NULL;
Query ID = hadoop_20220303091109_6e7b918a-d45e-4646-924a-4ff6aeee6993
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1646296003361_0004)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 ..........  container     SUCCEEDED      6          6        0        0       0       0
Reducer 2 ......  container     SUCCEEDED      4          4        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 29.72 s
--------------------------------------------------------------------------------
OK
category_code
accessories.bag
appliances.environment.vacuum
appliances.personal.hair_cutter
sport.diving
apparel.glove
furniture.bathroom.bath
furniture.living_room.cabinet
stationery.cartrige
accessories.cosmetic_bag
appliances.environment.air_conditioner
furniture.living_room.chair
Time taken: 30.494 seconds, Fetched: 11 row(s)
```

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

Ans.- SELECT Category_code, COUNT(product_id) AS no_of_products FROM ecomsid_events_optimize WHERE category_code IS NOT NULL GROUP BY category_code ORDER BY no_of_products DESC;

```
hive> --Total no. of products in each category
hive> SELECT Category_code, COUNT(product_id) AS no_of_products FROM ecomsid_events_optimize WHERE category_code IS NOT NULL GROUP BY category_code ORDER BY no_of_produ
cts DESC;
Query ID = hadoop_20220303091243_3ec8f263-d4d9-4f61-ab5e-be6653ad0cc3
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1646296003361_0004)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 ..........  container     SUCCEEDED      6          6        0        0       0       0
Reducer 2 ......  container     SUCCEEDED      4          4        0        0       0       0
Reducer 3 ......  container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 30.30 s
--------------------------------------------------------------------------------
OK
category_code   no_of_products
appliances.environment.vacuum   59761
stationery.cartrige     26722
apparel.glove   18232
furniture.living_room.cabinet   13439
accessories.bag 11681
furniture.bathroom.bath 9857
appliances.personal.hair_cutter 1643
accessories.cosmetic_bag        1248
appliances.environment.air_conditioner  332
furniture.living_room.chair     308
sport.diving    2
Time taken: 31.144 seconds, Fetched: 11 row(s)
```

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

Ans.- SELECT brand, ROUND(SUM(price),2) AS total_sales FROM ecomsid_events_optimize WHERE brand IS NOT NULL AND event_type = 'purchase' GROUP BY brand ORDER BY total_sales DESC LIMIT 1;

```
hive> --Brand having maximum sales in October & November
hive> SELECT brand, ROUND(SUM(price),2) AS total_sales FROM ecomsid_events_optimize WHERE brand IS NOT NULL AND event_type = 'purchase' GROUP BY brand ORDER BY total_sa
les DESC LIMIT 1;
Query ID = hadoop_20220303091343_09059870-d0a3-49c0-901c-993bba2e4c95
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1646296003361_0004)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      3          3        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 22.35 s
--------------------------------------------------------------------------------
OK
brand   total_sales
runail  148297.94
Time taken: 22.93 seconds, Fetched: 1 row(s)
```

7] Which brands increased their sales from October to November?

Ans.- SELECT oct.brand as Brand, ROUND((nov.Nov_sale - oct.Oct_sale),2) AS increase_in_sale FROM (SELECT brand, SUM(price) as Oct_sale FROM ecomsid_events_optimize WHERE event_type = 'purchase' AND brand IS NOT NULL AND month(event_time) = 10 GROUP BY brand) oct JOIN (SELECT brand, SUM(price) as Nov_sale FROM ecomsid_events_optimize WHERE event_type = 'purchase' AND brand IS NOT NULL AND month(event_time) = 11 GROUP BY brand) nov ON oct.brand = nov.brand WHERE nov.Nov_sale > oct.Oct_sale ORDER BY increase_in_sale DESC;

```
hive> --Brands showing increase in sales from october to november
hive> SELECT oct.brand as Brand, ROUND((nov.Nov_sale - oct.Oct_sale),2) AS increase_in_sale FROM (SELECT brand, SUM(price) as Oct_sale FROM ecomsid_events_optimize WHER
E event_type = 'purchase' AND brand IS NOT NULL AND month(event_time) = 10 GROUP BY brand) oct JOIN (SELECT brand, SUM(price) as Nov_sale FROM ecomsid_events_optimize W
HERE event_type = 'purchase' AND brand IS NOT NULL AND month(event_time) = 11 GROUP BY brand) nov ON oct.brand = nov.brand WHERE nov.Nov_sale > oct.Oct_sale ORDER BY in
crease_in_sale DESC;
Query ID = hadoop_20220303091515_7e60840e-428d-4d9a-9648-ab5d7409af3b
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1646296003361_0004)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      3          3        0        0       0       0
Map 5 .......... container     SUCCEEDED      3          3        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
Reducer 4 ...... container     SUCCEEDED      1          1        0        0       0       0
Reducer 6 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 06/06  [==========================>>] 100%  ELAPSED TIME: 31.56 s
--------------------------------------------------------------------------------
OK
```

```
OK
brand   increase_in_sale
grattol 36027.17
uno     15737.72
lianail 10501.4
ingarden        10404.82
strong  9474.64
jessnail        7057.39
cosmoprofi      6214.18
polarus 5358.21
runail  5219.38
freedecor       4250.02
staleks 3355.88
bpw.style       3265.29
lovely  3234.68
marathon        2992.35
haruyama        2962.22
yoko    2950.97
italwax 2859.13
benovy  2850.35
kaypro  2387.36
estel   2385.92
concept 2348.26
kapous  2165.92
f.o.x   1953.05
masura  1792.39
milv    1737.07
beautix 1729.0
artex   1596.61
domix   1537.12
shik    1498.52
smart   1444.88
roubloff        1422.41
levrana 1420.54
oniq    1416.24
irisk   1354.08
severina        1344.6
joico   1309.58
```

```
joico   1309.58
zeitun  1300.97
beauty-free     1228.69
swarovski       1155.23
de.lux  1115.81
metzger 1083.71
markell 1065.68
sanoto  1052.54
nagaraku        957.94
ecolab  951.45
art-visage      905.09
levissime       857.81
missha  856.45
solomeya        786.1
rosi    764.52
refectocil      759.4
kaaral  673.64
kosmekka        631.93
kinetics        611.01
browxenna       585.36
airnails        572.62
uskusi  548.04
coifin  525.49
s.care  500.39
limoni  487.7
matrix  483.49
gehwol  468.61
greymy  460.28
bioaqua 455.23
farmavita       454.6
sophin  447.66
yu-r    402.3
kiss    395.78
lador   387.92
ellips  360.19
jas     338.47
lowence 324.91
nitrile 315.4
shary   304.53
kims    302.0
happyfons       289.67
kocostar        284.08
insight 278.26
```

```
kocostar        284.08
insight 278.26
candy   264.42
bluesky 258.29
beauugreen      256.84
protokeratin    255.54
trind   244.89
entity  239.55
skinlite        238.51
provoc  235.83
fedua   211.43
ecocraft        200.79
keen    199.27
mane    193.47
freshbubble     183.64
chi     179.67
cristalinas     157.32
farmona 150.97
latinoil        135.07
miskin  135.03
elizavecca      133.77
nefertiti       133.12
finish  132.0
igrobeauty      131.41
dizao   126.38
osmo    116.73
batiste 101.77
carmex  98.28
eos     98.27
depilflax       96.71
enjoy   95.22
kerasys 94.29
aura    93.56
plazan  92.64
koelf   84.56
nirvel  71.29
konad   70.84
egomania        68.57
cutrin  68.25
laboratorium    66.02
inm     63.19
marutaka-foot   60.11
profhenna       57.62
koelcia 57.25
```

```
inm     63.19
marutaka-foot   60.11
profhenna       57.62
koelcia 57.25
balbcare        57.05
elskin  56.56
foamie  45.45
ladykin 44.92
likato  44.91
mavala  37.28
vilenta 33.61
beautyblender   30.67
biore   29.66
orly    28.71
estelare        27.06
profepil        24.66
blixz   24.45
godefroy        23.9
glysolid        21.86
veraclara       21.1
kamill  18.48
treaclemoon     18.12
supertan        16.14
deoproce        12.33
rasyan  10.14
fly     10.03
tertio  9.64
jaguar  8.54
soleo   8.33
neoleor 8.29
moyou   4.57
bodyton 4.3
skinity 3.56
grace   1.69
cosima  0.7
ovale   0.56
Time taken: 32.461 seconds, Fetched: 152 row(s)
hive>
```

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

Ans.- WITH Customer_Rank AS(SELECT user_id AS Customer, ROUND(SUM(price),2) AS Expenditure, RANK() OVER(ORDER BY ROUND(SUM(price),2) DESC) AS Rank FROM ecomsid_events_optimize WHERE event_type = 'purchase' GROUP BY user_id) SELECT Customer, Expenditure, Rank FROM Customer_Rank WHERE Rank <=10;

```
    > ;
hive> --Top 10 users eligible for Golden Customer Plan
hive> WITH Customer_Rank AS( SELECT user_id AS Customer, ROUND(SUM(price),2) AS Expenditure, RANK() OVER(ORDER BY ROUND(SUM(price),2) DESC) AS Rank FROM ecomsid_events_
optimize WHERE event_type = 'purchase' GROUP BY user_id) SELECT Customer, Expenditure, Rank FROM Customer_Rank WHERE Rank <=10;
Query ID = hadoop_20220303091657_975a6872-76e4-4077-9486-97a7d9cbd6c6
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1646296003361_0004)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 ......... container     SUCCEEDED      3          3        0        0       0       0
Reduce 2 ...... container     SUCCEEDED      1          1        0        0       0       0
Reduce 3 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 27.55 s
--------------------------------------------------------------------------------
OK
customer        expenditure     rank
557790271       2715.87 1
150318419       1645.97 2
562167663       1352.85 3
531900924       1329.45 4
557850743       1295.48 5
522130011       1185.39 6
561592095       1109.7  7
431950134       1097.59 8
566576008       1056.36 9
521347209       1040.91 10
Time taken: 28.378 seconds, Fetched: 10 row(s)
hive>
    >
```

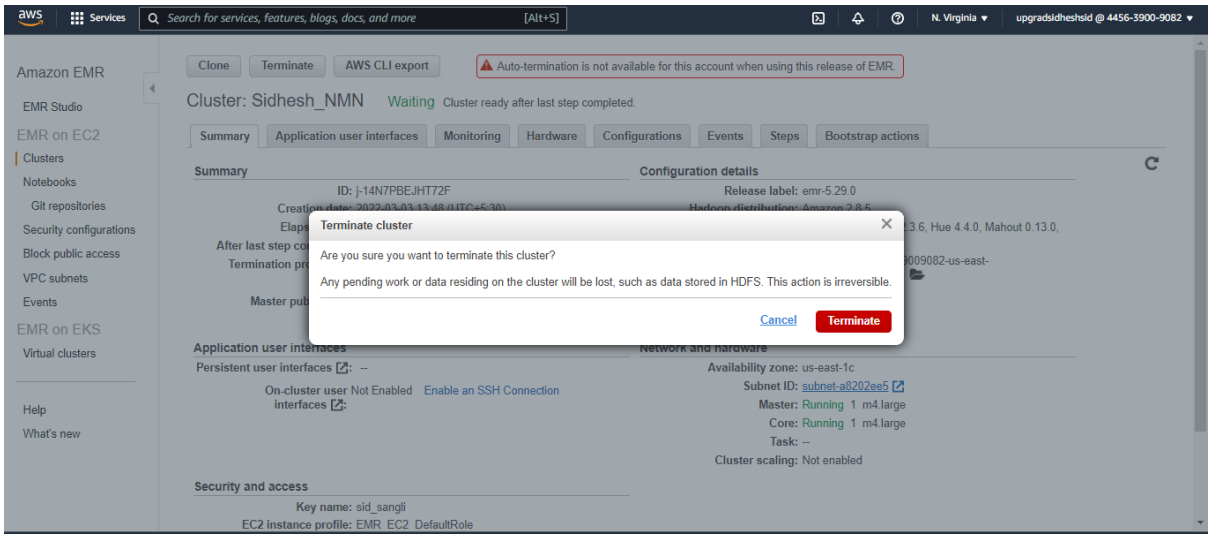
Cleaning up:

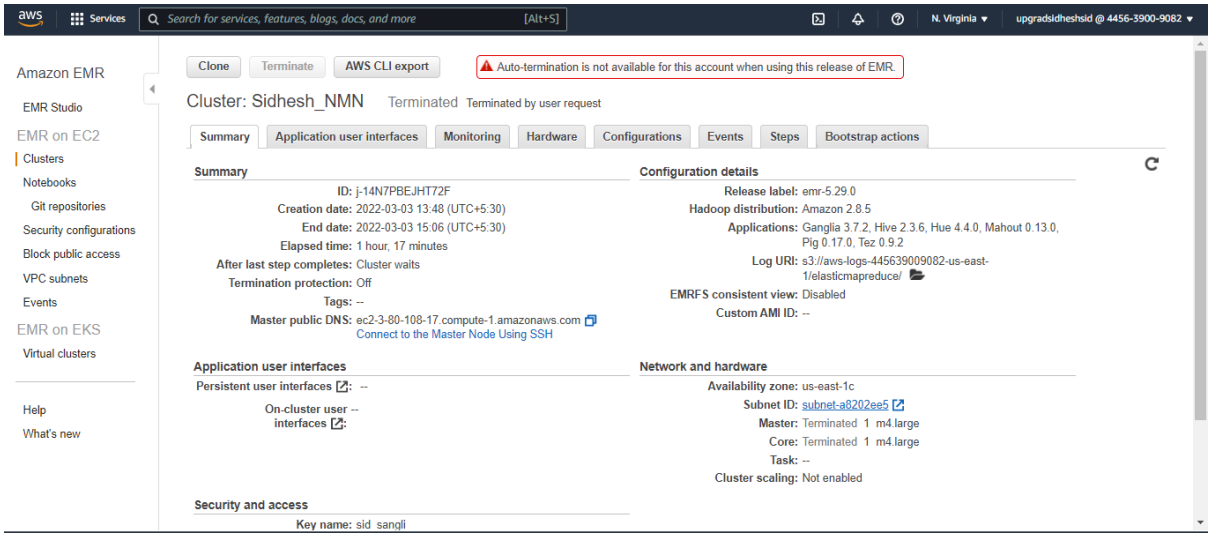
A] Dropping Database The below query Drops off the Database.

```
[hadoop@ip-172-31-19-96 ~]$ hive

Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j2.properties Async: false
hive>
    >
    > SHOW TABLES;
OK
Time taken: 1.222 seconds
hive> DROP TABLE ecomsid_events;
OK
Time taken: 0.101 seconds
hive> DROP TABLE ecomsid_events_part1;
OK
Time taken: 0.026 seconds
hive> DROP TABLE ecomsid_events_optimize;
OK
Time taken: 0.021 seconds
hive> SHOW TABLES;
OK
Time taken: 0.042 seconds
hive> SHOW DATABASES;
OK
clicksid_stream_data
default
Time taken: 0.025 seconds, Fetched: 2 row(s)
hive> DROP DATABASE clicksid_stream_data;
FAILED: Execution Error, return code 1 from org.apache.hadoop.hive.ql.exec.DDLTask. InvalidOperationException(message:Database clicksid_stream_data is not empty. One or
 more tables exist.)
hive>
    >
```

B] Terminating the Cluster The below query terminates the Database




aws
Services
Search for services, features, blogs, docs, and more
[Alt+S]
N. Virginia
upgradsidheshsid @ 4456-3900-9082
Amazon EMR
EMR Studio
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
Auto-termination is not available for this account when using this release of EMR.
Cluster: Sidhesh_NMN
Terminated
Terminated by user request
Summary
Application user interfaces
Monitoring
Hardware
Configurations
Events
Steps
Bootstrap actions
Summary
ID: j-14N7PBEJHT72F
Creation date: 2022-03-03 13:48 (UTC+5:30)
End date: 2022-03-03 15:06 (UTC+5:30)
Elapsed time: 1 hour, 17 minutes
After last step completes: Cluster waits
Termination protection: Off
Tags: --
Master public DNS: ec2-3-80-108-17.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-445639009082-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-1c
Subnet ID: subnet-a8202ee5
Master: Terminated 1 m4.large
Core: Terminated 1 m4.large
Task: --
Cluster scaling: Not enabled
Security and access
Key name: sid_sangli


THANK YOU

==========