E-Commerce Case Study 2022

(Image courtesy: https://www.forbes.com)

MAY 4

upGrad & IIITB(DS_B17_C3)
Authored by: Vijaykumar Rangvani
Sonam Gupta

(Image courtesy: https://en.wikipedia.org/)

Introduction

With online sales gaining popularity, tech companies are exploring ways to improve their sales by analyzing customer behavior and gaining insights about product trends. Furthermore, the websites make it easier for customers to find the products they require without much scavenging.

Problem Statement

One of the most popular use cases of Big Data is in ecommerce companies such as Amazon or Flipkart. So before we get into the details of the dataset, let us understand how ecommerce companies make use of these concepts to give customers product recommendations. This is done by tracking your clicks on their website and searching for patterns within them. This kind of data is called clickstream data.

The clickstream data contains all the logs as to how you navigated through the website. It also contains other details such as time spent on every page, etc. From this, they make use of data ingesting frameworks such as Apache Kafka or AWS Kinesis in order to store it in frameworks such as Hadoop.

Business Objective

Therefore, as part of this case study, as a big data analyst, we will be extracting data and will be gathering insights from a real-life data set of an e-commerce company.

Solution Steps:

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:

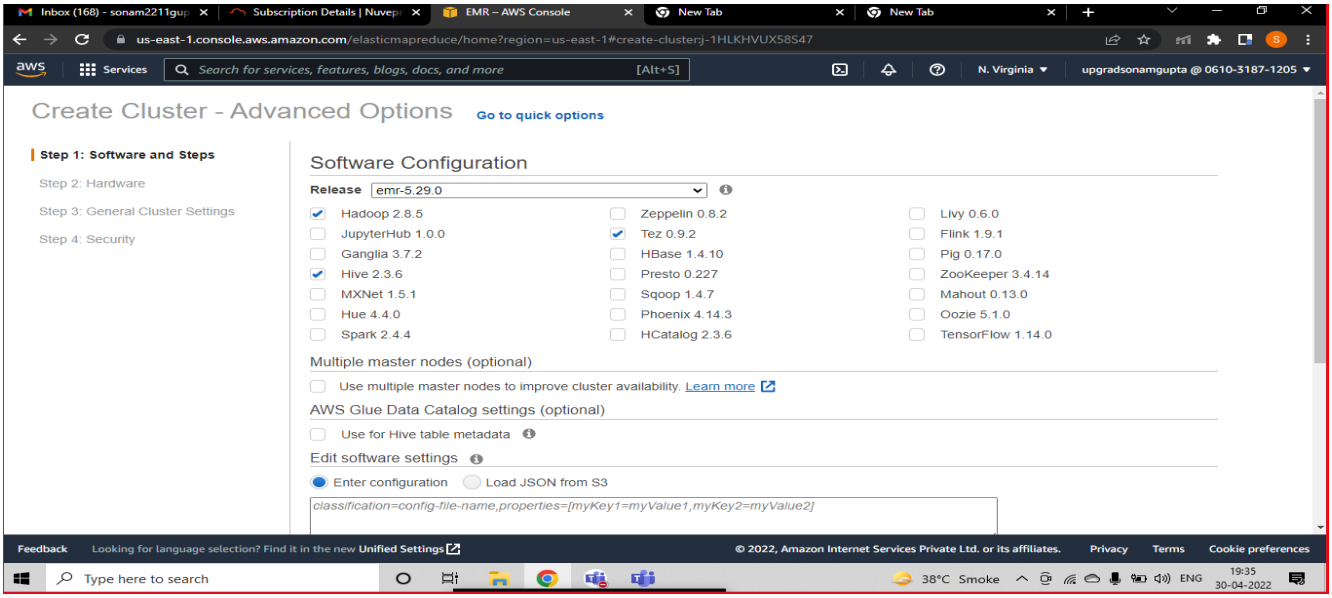
- Create the structure of your database,
- Using optimized techniques to run your queries as efficiently as possible
- Run Hive queries to answer the questions.

Cleaning up

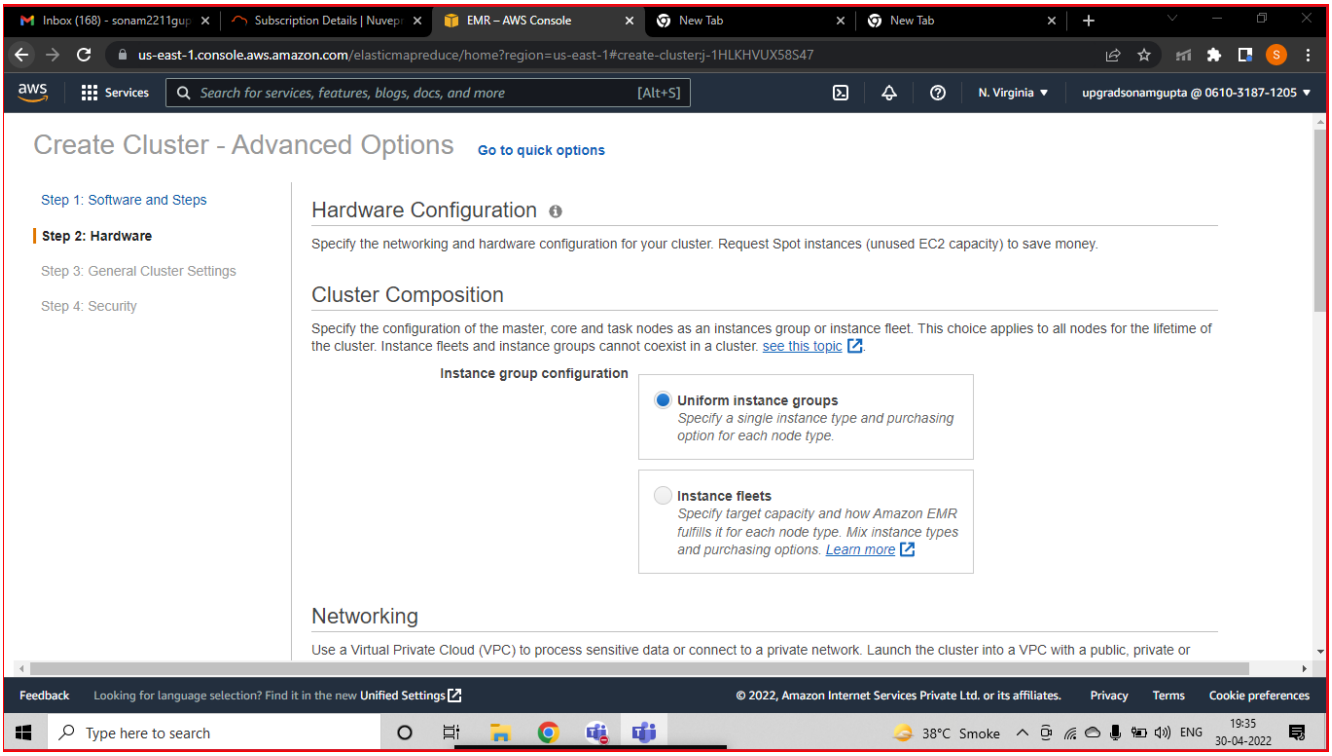
- Drop database, and
- Terminate cluster

Launching an EMR Cluster:

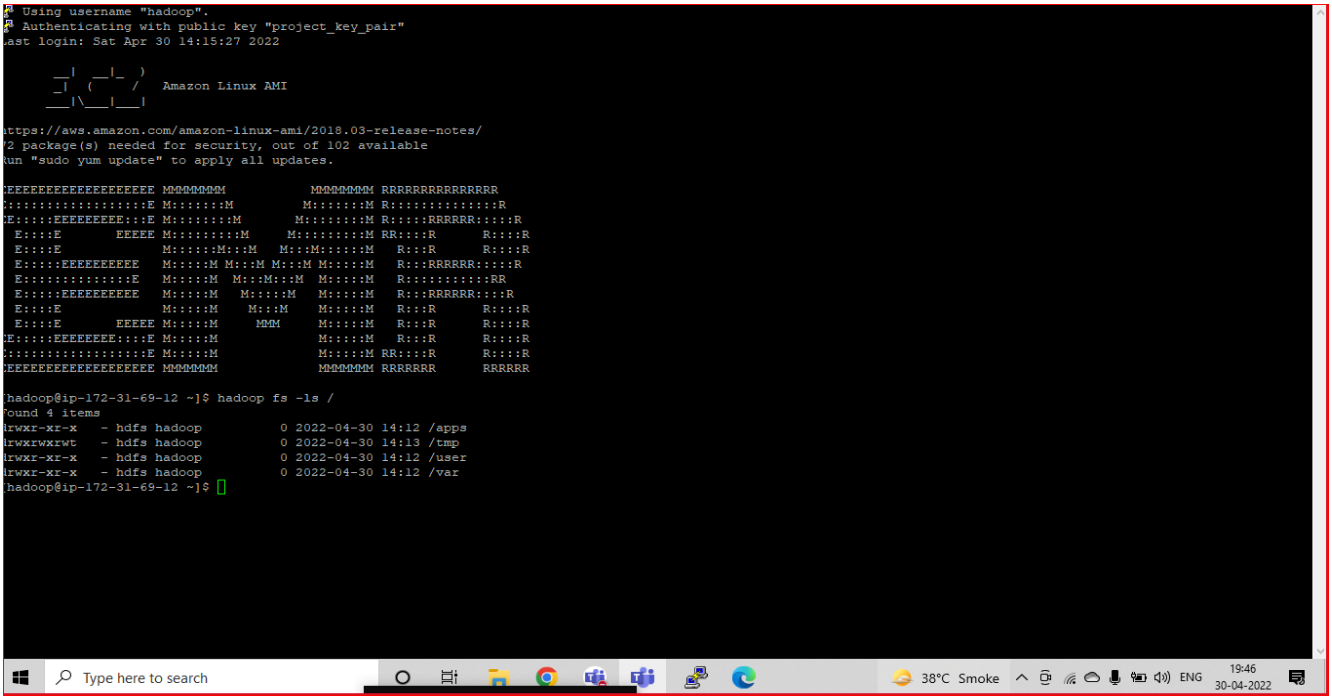
- We have created EMR Cluster having **emr-5.29.0** release along with Hive service installed.



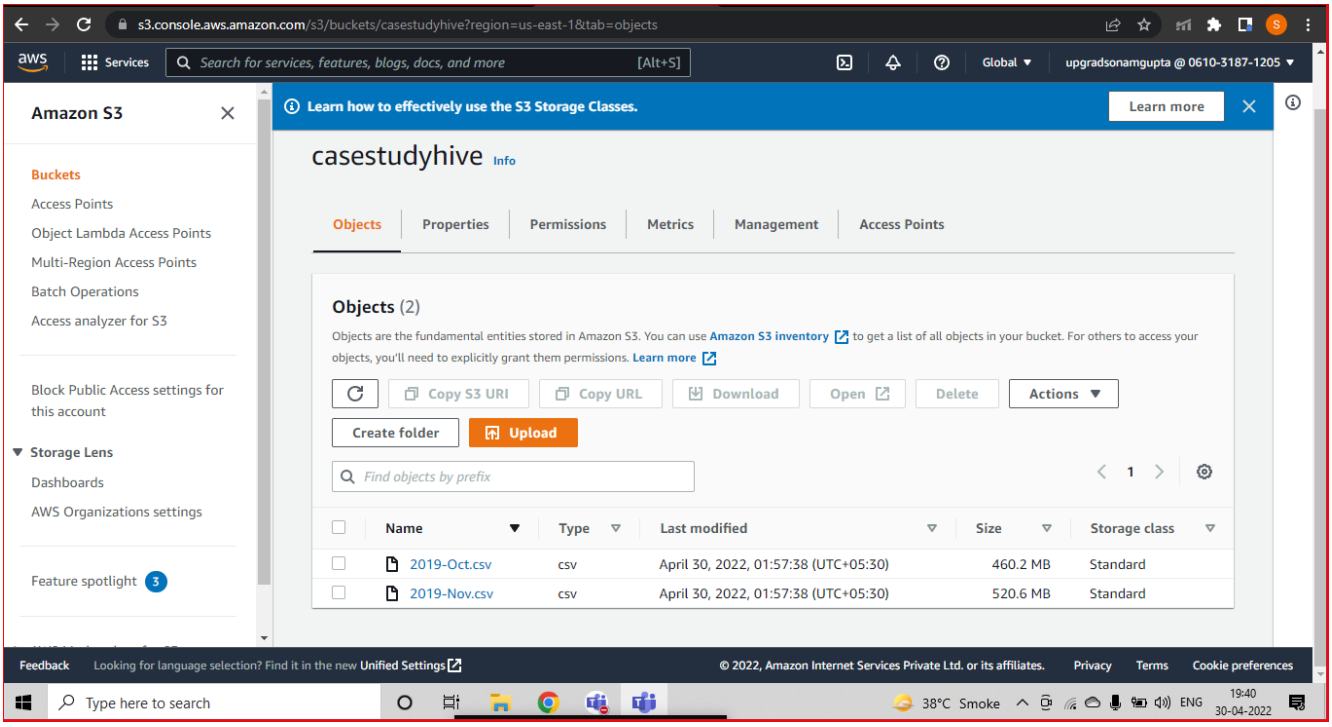
- Next, as a part of Hardware configuration, created 2-Node EMR cluster with Master and Core node as **M4.large.**



- Verifying that EMR Cluster is up and running now and able to login into it using Putty.



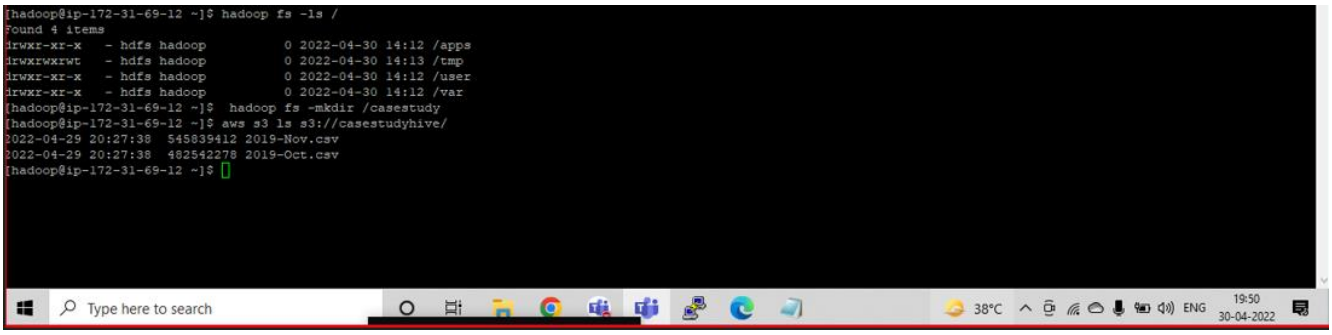
- Uploaded files into S3 bucket under bucket named **casestudyhive**.



Importing data from S3 bucket into HDFS

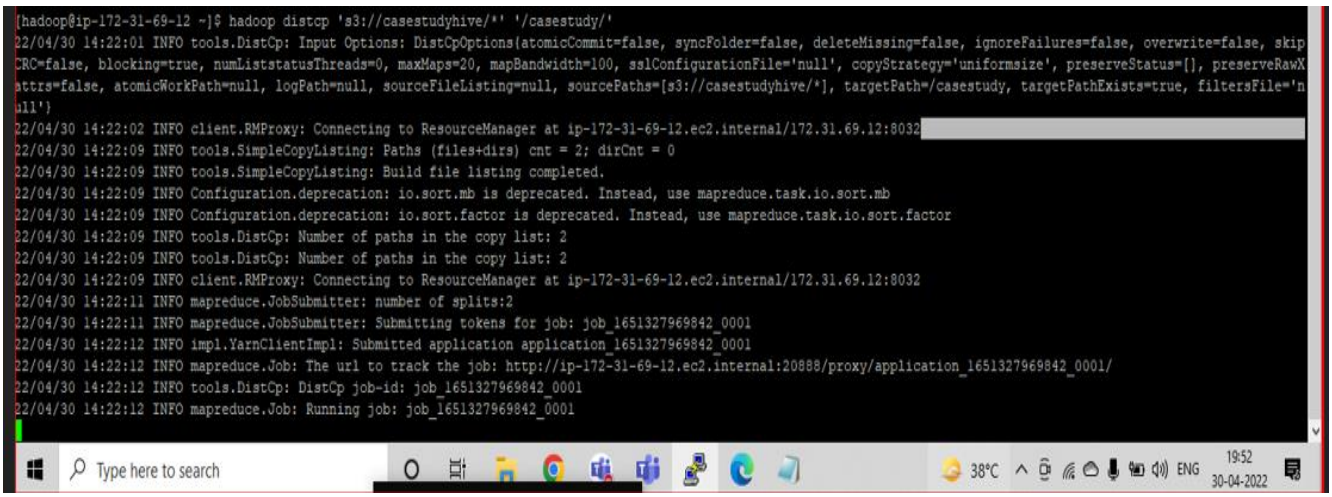
- Creating the Directory named “casestudy”.

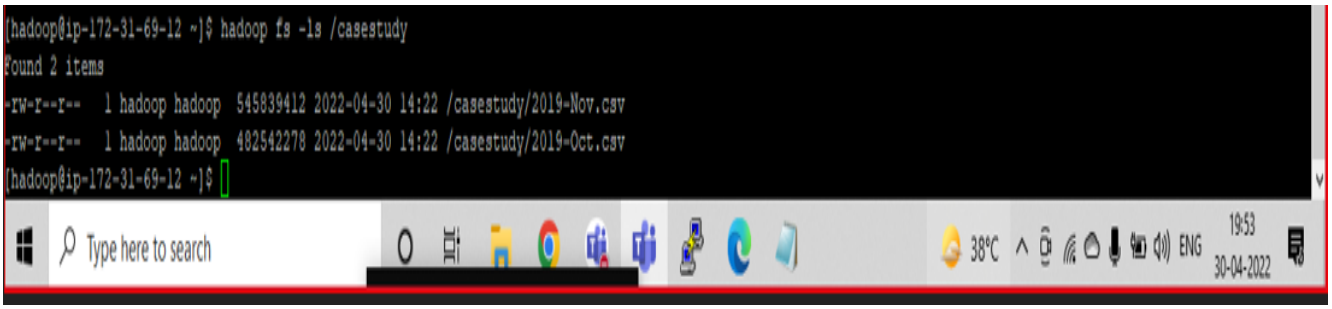
hadoop fs -mkdir /casestudy



- Copying files from S3 bucket to HDFS.

hadoop distcp 's3://casestudyhive/*' '/casestudy/'

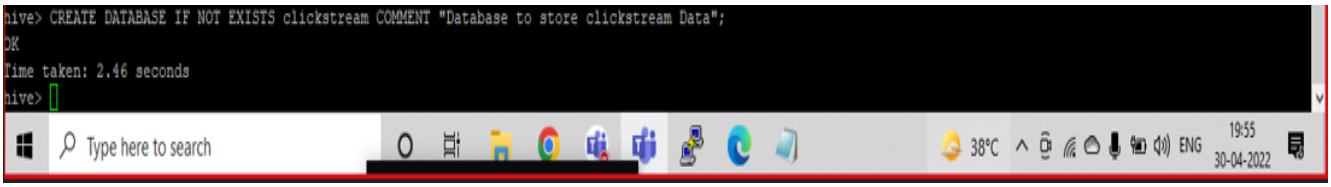




Creating Hive Tables:

- Creating Database named “clickstream”.

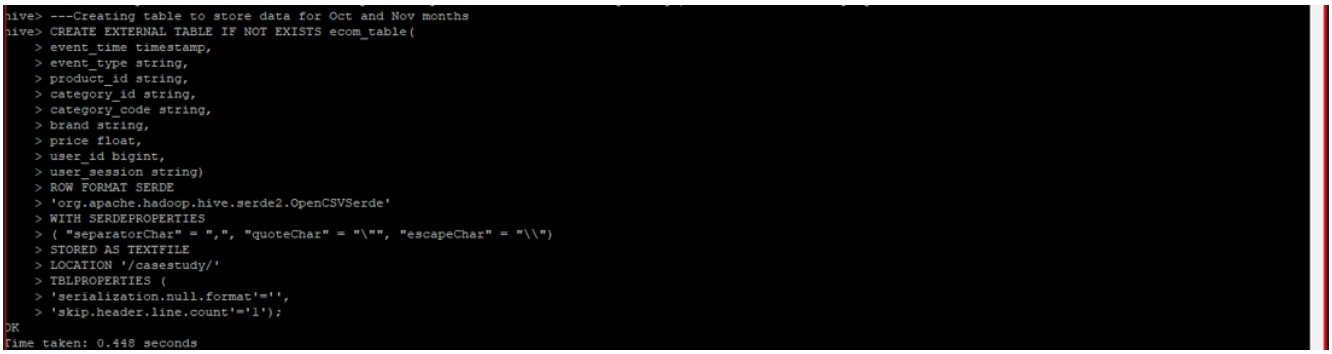
CREATE DATABASE IF NOT EXISTS clickstream COMMENT "Database to store clickstream Data";



- ---Setting up hive paramter to display the header
  set hive.cli.print.header=True;
- ---Using clickstream DB
  USE clickstream;

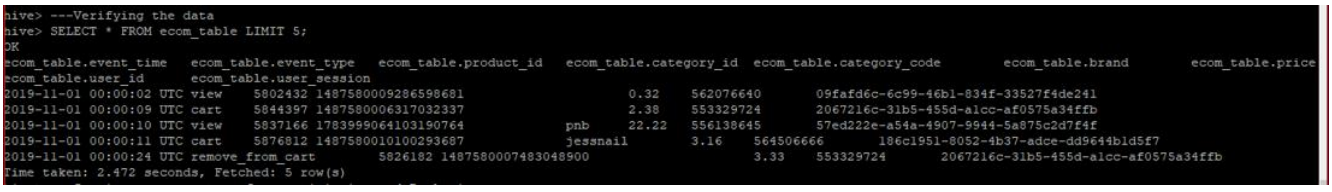
BASE TABLE (ecom_tab)

- Creating table named “ecom_table”.



- Verifying data under table “ecom_table”.

SELECT * FROM ecom_table LIMIT 5;



- Setting up parameters for creating tables with bucketing and partitioning.

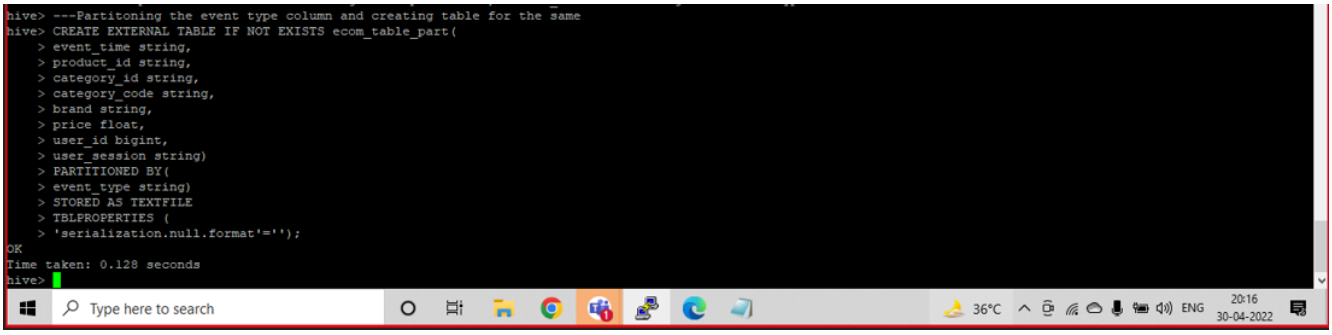
set hive.exec.dynamic.partition=true;

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

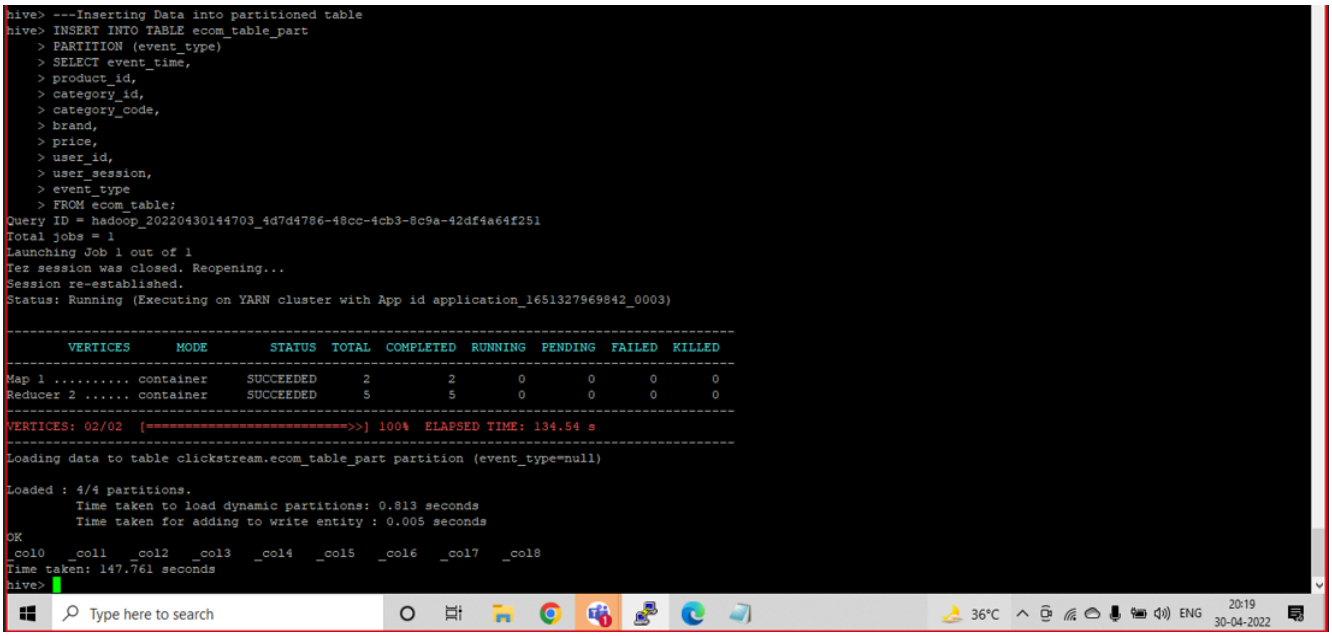
set hive.enforce.bucketing=true;

PARTITIONED TABLE (ecom_table_part)

- Partitioning the event type column and creating table for the same named “ecom_table_part”.

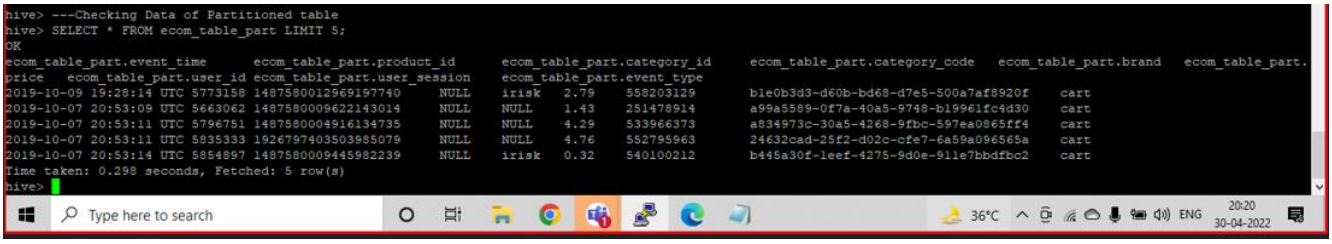


- Inserting Data into partitioned table



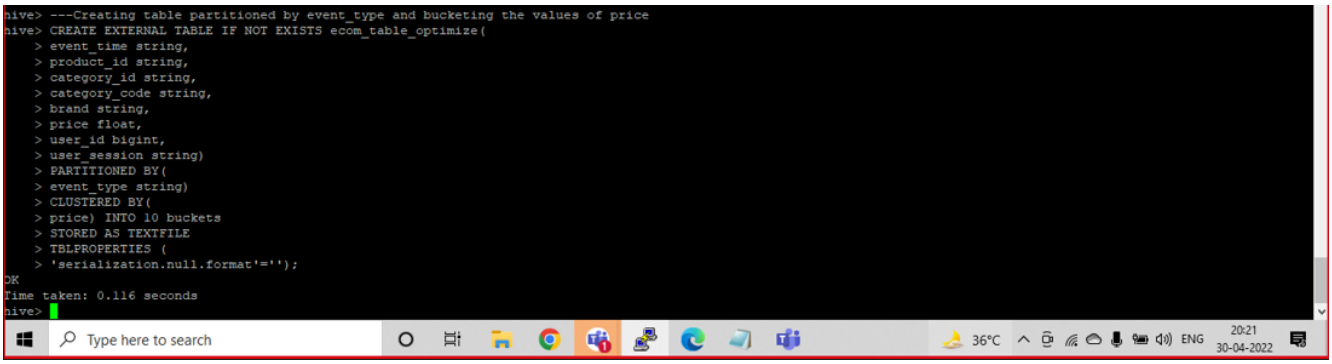
- Verifying data under table “ecom_table_part”.

SELECT * FROM ecom_table_part LIMIT 5;

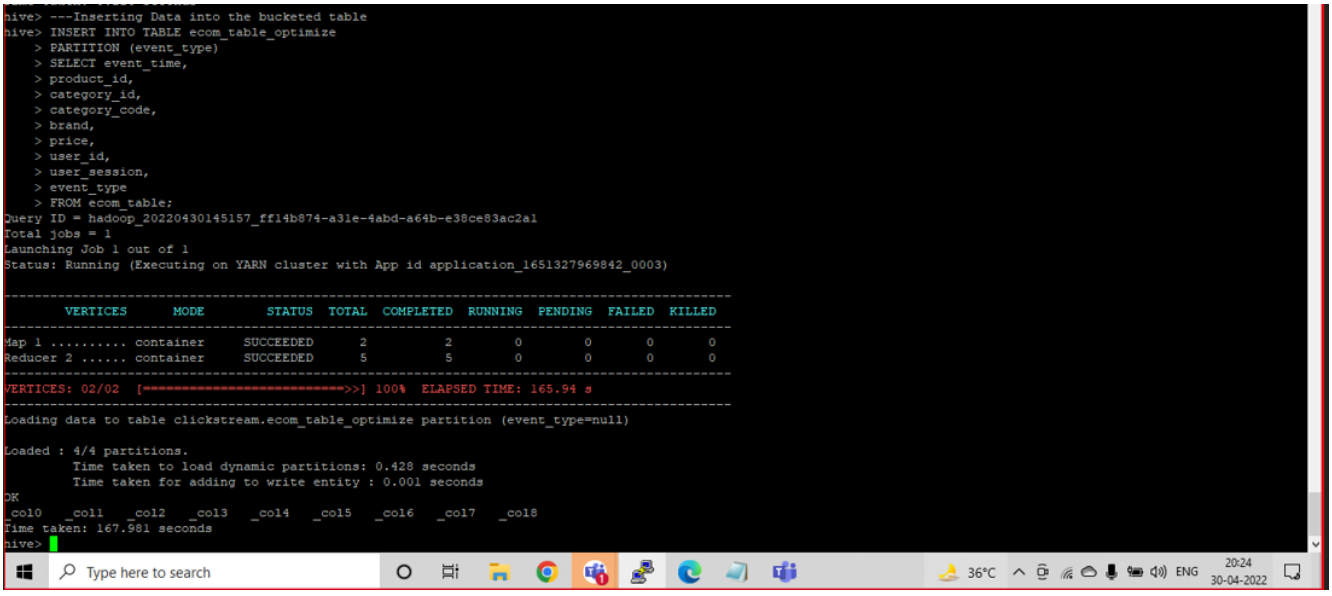


BUCKETING TABLE (ecom_table_optimize)

- Creating table partitioned by event_type and bucketing the values of price.

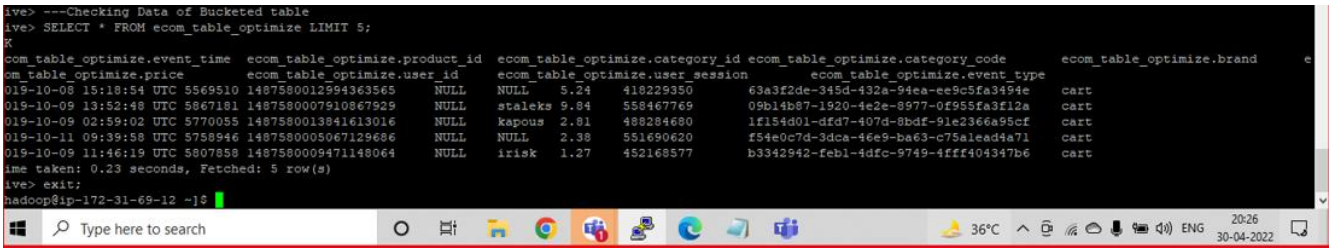


- Inserting Data into the bucketed table



- Verifying data under table “ecom_table_optimize”.

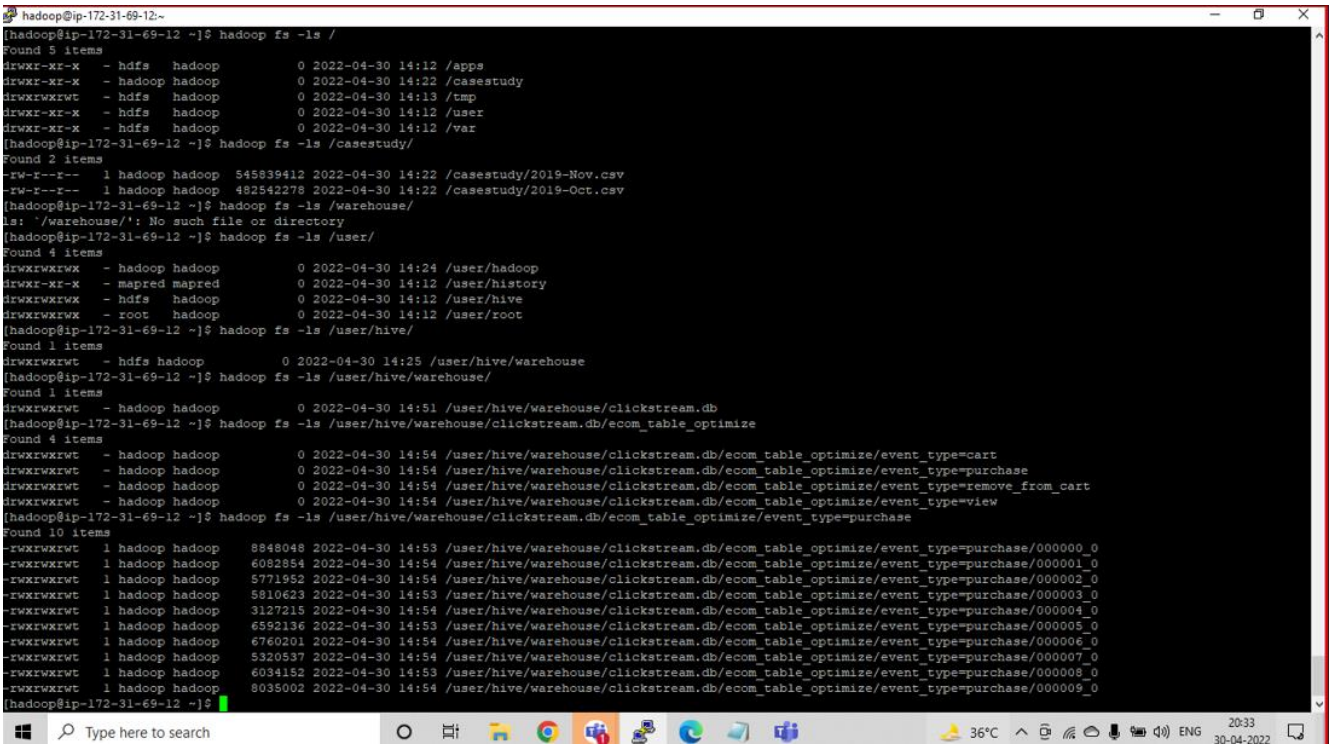
SELECT * FROM ecom_table_optimize LIMIT 5;



- Checking created Buckets in HDFS.

hadoop fs -ls /user/hive/warehouse/ClickStream.db/ecom_table_optimize

hadoop fs ls /user/hive/warehouse/ClickStream.db/ecom_table_optimize/event_type=purchase

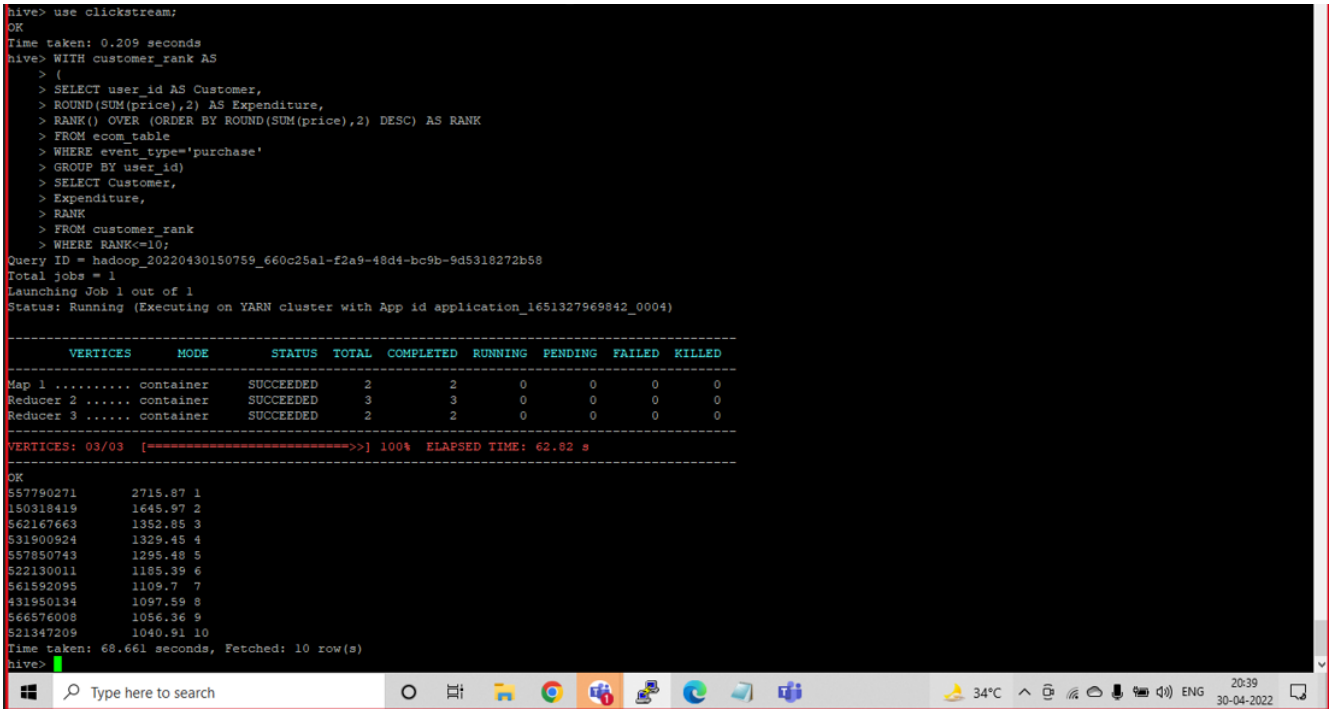


Optimization Techniques:

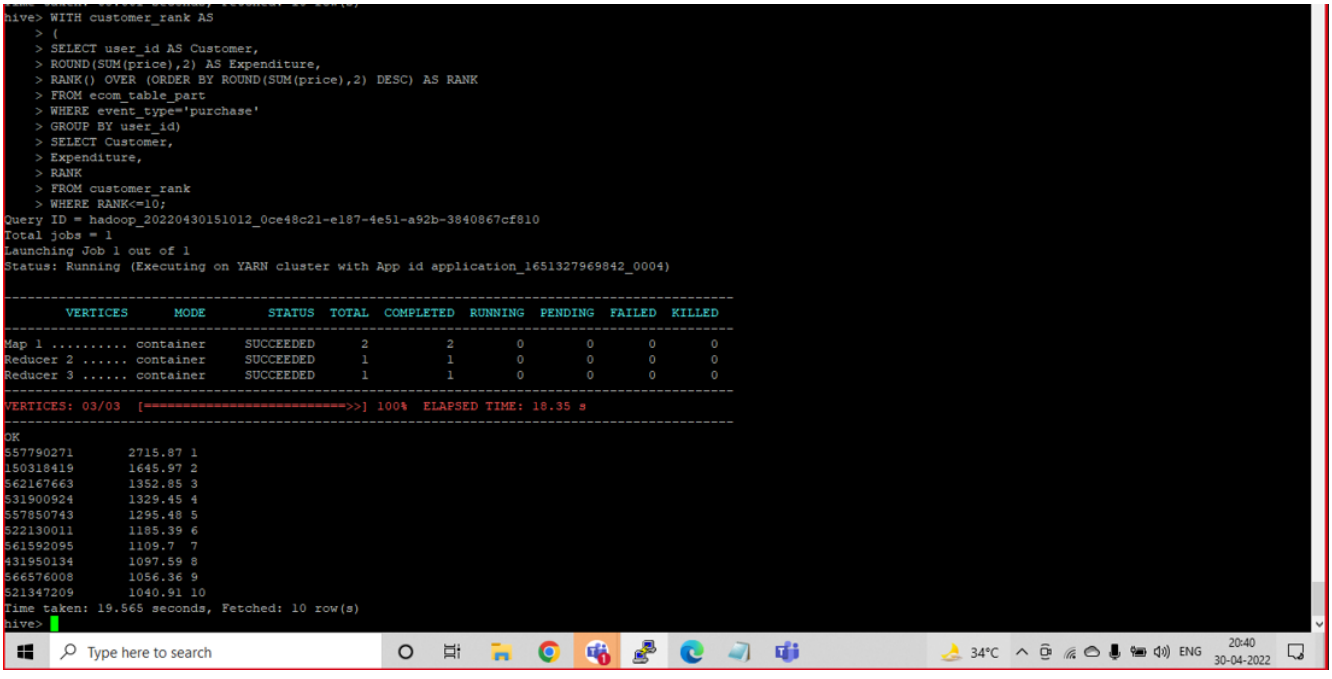
Comparing the performance:

Query: Finding top 10 customers who are making most of purchases.

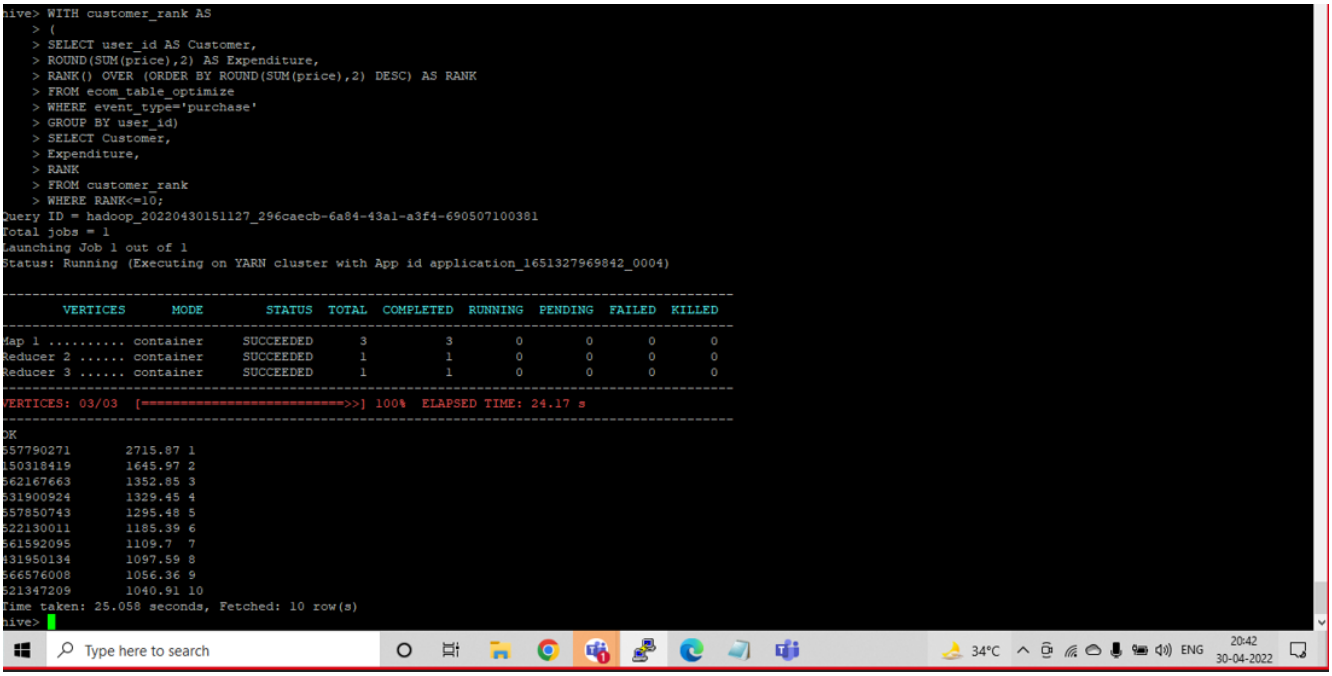
1. **ecom_table:** table without partitioning and bucketing



2. **ecom_table_part:** table with partitioning



3. **ecom_table_optimize** – table with partitioning and bucketing



Performance Comparison:

Below table shows the time taken to execute the query:

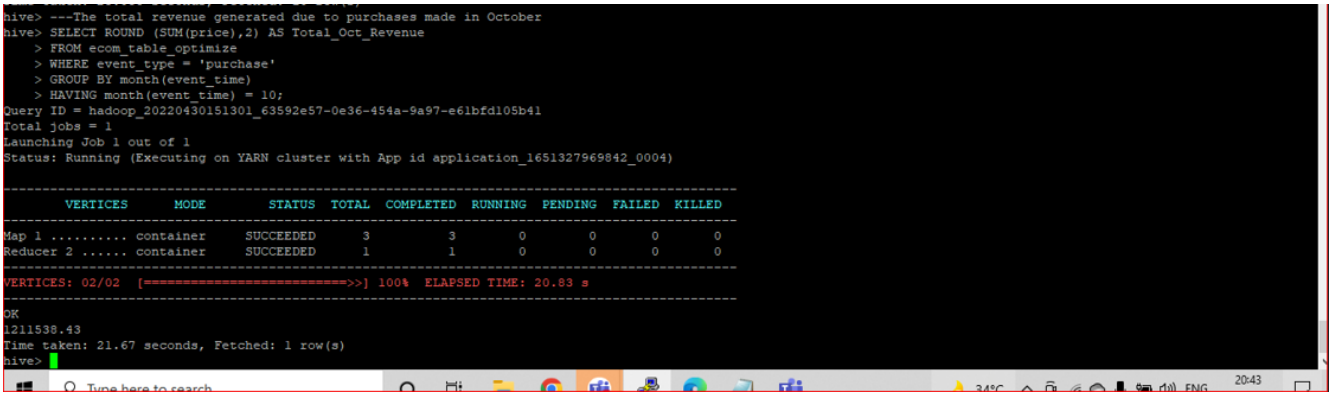
Without Partition and Bucketing (ecom_table)	**With Partitioning** (ecom_table_part)	**With Partitioning and Bucketing** (ecom_table_optimize)
68.66 seconds	19.56 seconds	25.05 seconds

Conclusion:

Hence, it can be clearly observed that partitioned tables and Bucketed tables definitely has better performance in terms of query execution time.

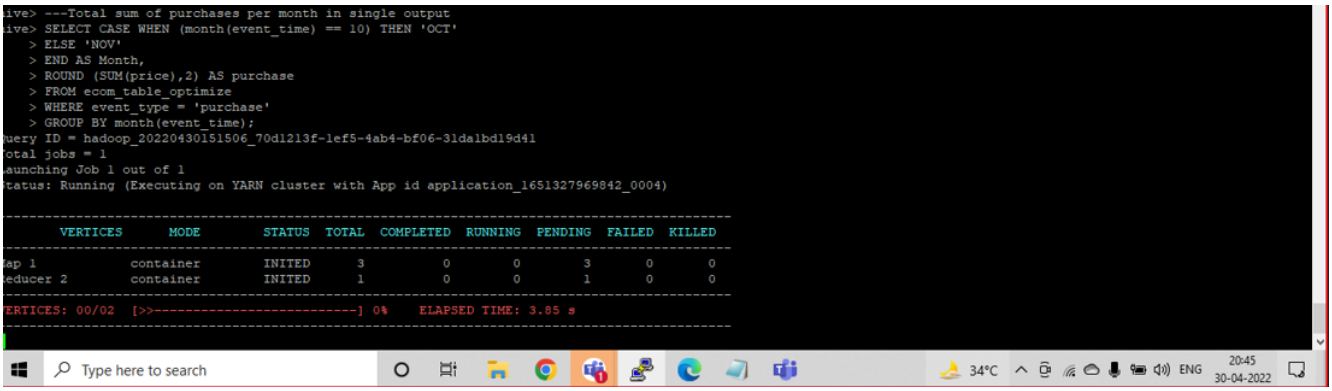
Analysis using Hive Queries:

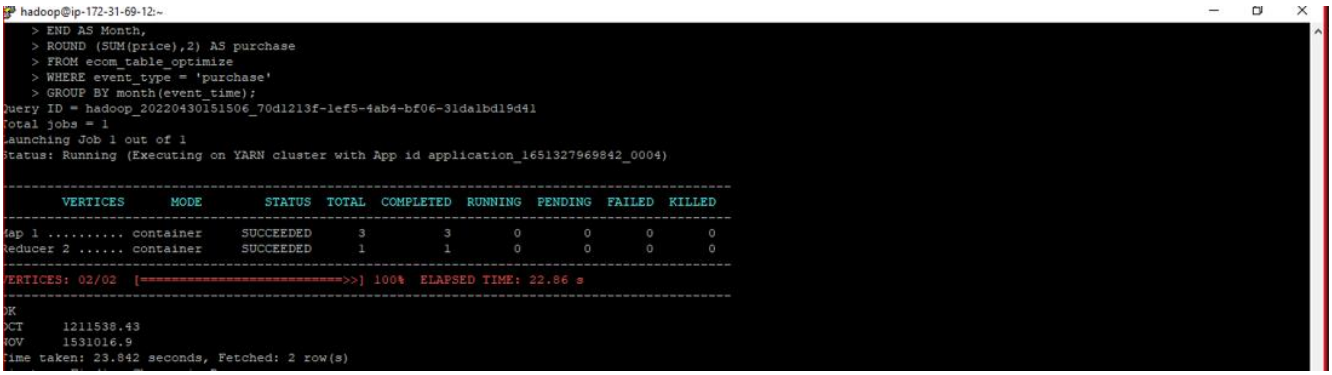
Question-1 Find the total revenue generated due to purchases made in October.



The total revenue generated due to purchases made in October is 1211538.43.

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





Sum of purchases in October is 1211538.43 and November is 1531016.9

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

```
hive> ---Finding Change in Revenue
hive> SELECT (Nov_Rev-Oct_Rev) AS Change_In_Revenue
    > FROM (
    > SELECT ROUND(SUM(CASE WHEN month(event_time)=10 THEN price ELSE 0 END),2) AS Oct_rev,
    > ROUND(SUM(CASE WHEN month(event_time)=11 THEN price ELSE 0 END),2) AS Nov_rev
    > FROM ecom_table_optimize
    > WHERE event_type = 'purchase'
    > ) AS rev_cal;
Query ID = hadoop_20220430151547_9b7f6916-b224-42de-8f82-5da9d87bee5c
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1651327969842_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: 22.67 s
----------------------------------------------------------------------------------------------
OK
319478.47
Time taken: 23.506 seconds, Fetched: 1 row(s)
```

Change in revenue from October to November is 319478.47.

This indicates that there is strong positive increase in business.

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

```
hive> ---Finding Distinct categories of products
hive> SELECT DISTINCT category_code
    > FROM ecom_table_optimize
    > WHERE category_code IS NOT NULL;
Query ID = hadoop_20220430151650_0fcdad5e-e95e-4389-9667-53020a6148dd
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1651327969842_0004)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1            container     RUNNING      6          0        3        3       0       0
Reducer 2        container      INITED      4          0        0        4       0       0
----------------------------------------------------------------------------------------------
VERTICES: 00/02  [>>--------------------------] 0%    ELAPSED TIME: 8.71 s
----------------------------------------------------------------------------------------------
```

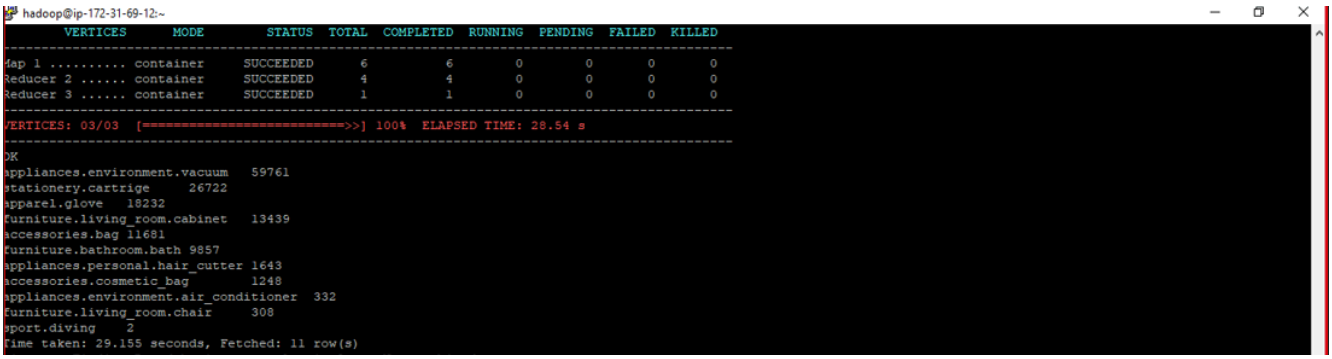


```
hadoop@ip-172-31-69-12:~

----------------------------------------------------------------------------------------------
        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: 28.51 s
----------------------------------------------------------------------------------------------
OK
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: 29.349 seconds, Fetched: 11 row(s)
```

There are total 11 distinct categories of products.

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

```
hive> ---Finding total No. of products under each category
hive> SELECT category_code,
    > COUNT(product_id) AS number_of_products
    > FROM ecom_table_optimize
    > WHERE category_code IS NOT NULL
    > GROUP BY category_code
    > ORDER BY number_of_products DESC;
Query ID = hadoop_20220430151744_71baa324-56c6-4b2f-95ba-dfb3637588bc
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1651327969842_0004)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1           container       RUNNING      6          0        3        3       0       0
Reducer 2       container        INITED      4          0        0        4       0       0
Reducer 3       container        INITED      1          0        0        1       0       0
--------------------------------------------------------------------------------
VERTICES: 00/03  [>>--------------------------] 0%    ELAPSED TIME: 19.09 s
--------------------------------------------------------------------------------
```



As we already know that there are 11 distinct categories of products, the above query result shows the total number of available under each category.

Category having highest products available is appliances environment vacuum.

Category having lowest products available is sport diving.

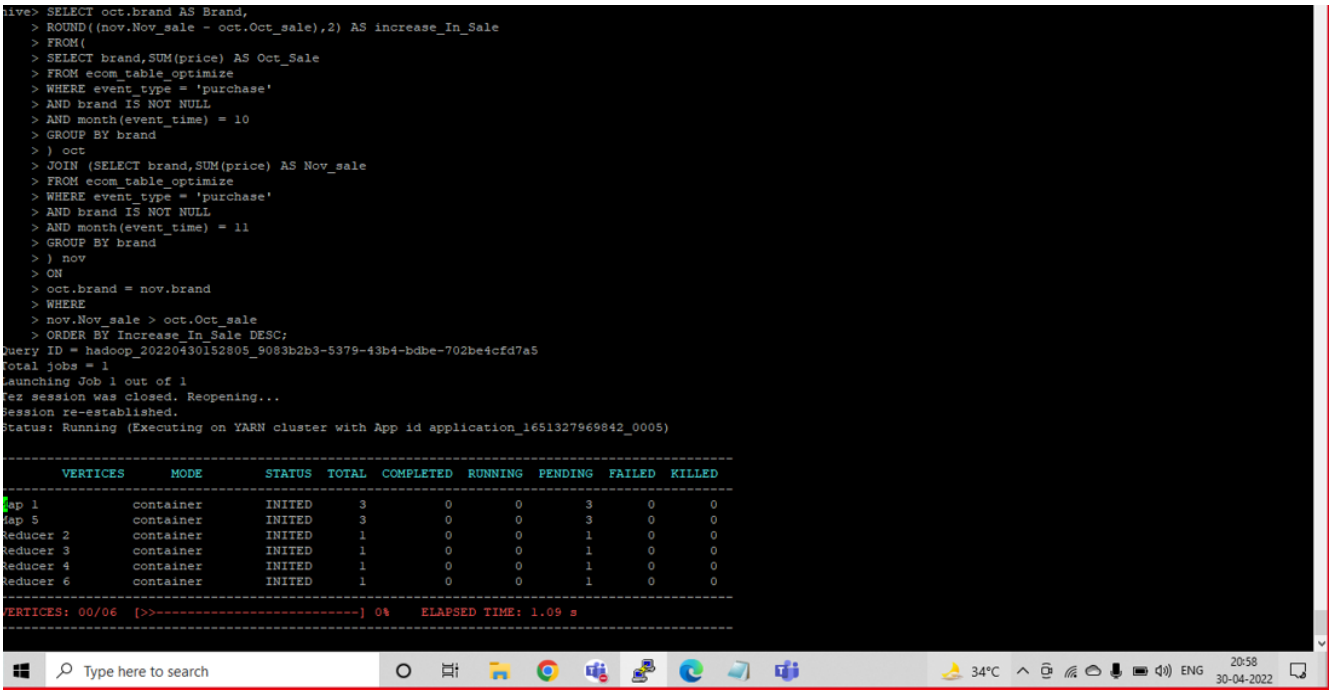
Question:6 Which brand had the maximum sales in October and November combined?

```
hive> ---Finding Brand having max sales in Oct & Nov combined
hive> SELECT brand,
    > ROUND(SUM(price),2) AS Sales
    > FROM ecom_table_optimize
    > WHERE brand IS NOT NULL AND
    > event_type = 'purchase'
    > GROUP BY brand
    > ORDER BY Sales DESC LIMIT 1;
Query ID = hadoop_20220430151859_c262c14f-5ce0-4b5d-b43a-853636233a87
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1651327969842_0004)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1           container        INITED      3          0        0        3       0       0
Reducer 2       container        INITED      1          0        0        1       0       0
Reducer 3       container        INITED      1          0        0        1       0       0
--------------------------------------------------------------------------------
VERTICES: 00/03  [>>--------------------------] 0%    ELAPSED TIME: 3.85 s
--------------------------------------------------------------------------------
```

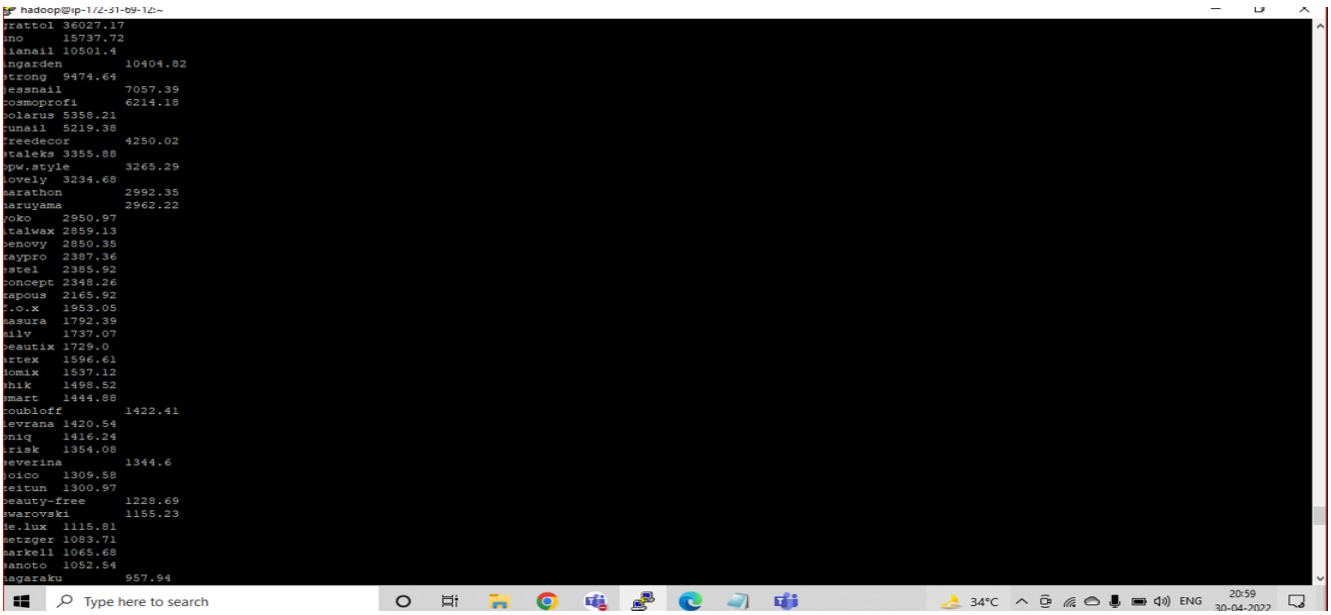
```
runail  148297.94
```

Runail is the brand that made the maximum sales in October and November combined with sales amount of 148297.94

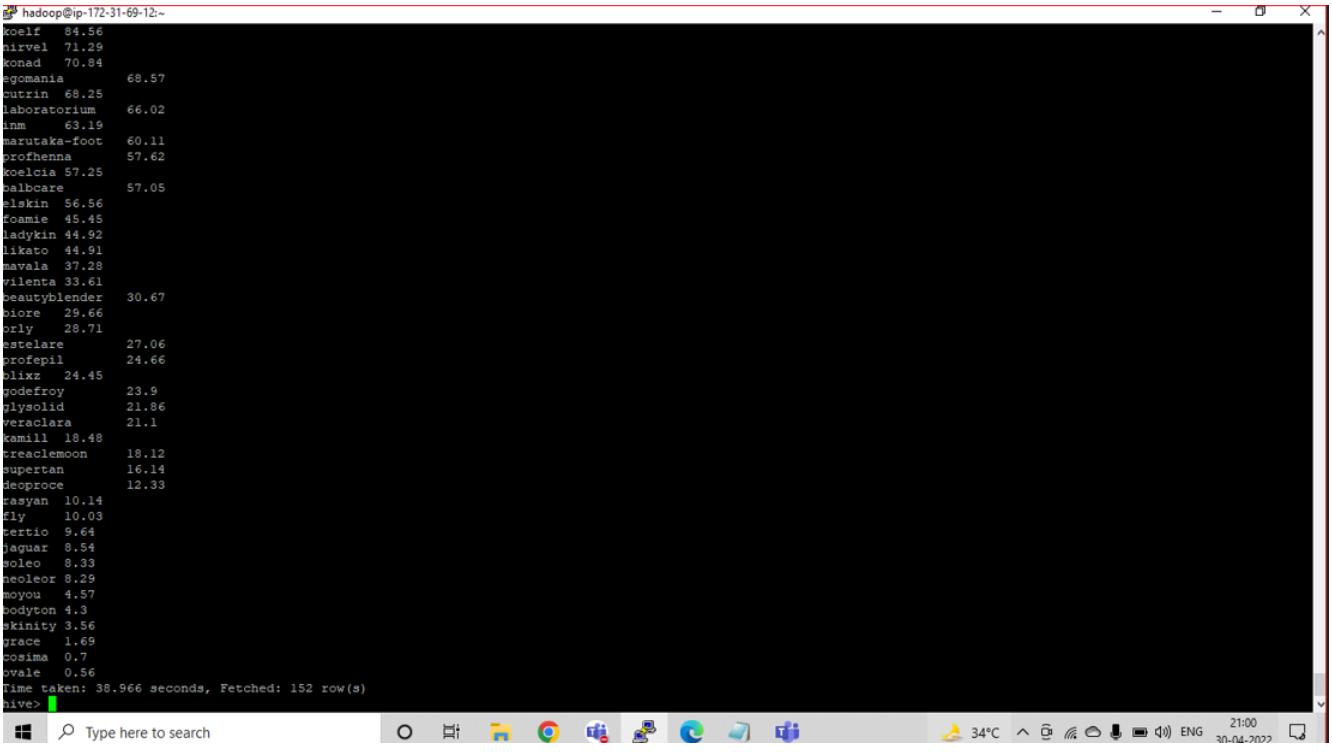
Question:7 Which brands increased their sales from October to November?



```
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 ecom_table_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 ecom_table_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;
Query ID = hadoop_20220430152805_9083b2b3-5379-43b4-bdbe-702be4cfd7a5
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_1651327969842_0005)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1           container     INITED      3          0        0        3       0       0
Map 5           container     INITED      3          0        0        3       0       0
Reducer 2       container     INITED      1          0        0        1       0       0
Reducer 3       container     INITED      1          0        0        1       0       0
Reducer 4       container     INITED      1          0        0        1       0       0
Reducer 6       container     INITED      1          0        0        1       0       0
--------------------------------------------------------------------------------
VERTICES: 00/06  [>>--------------------------] 0%    ELAPSED TIME: 1.09 s
--------------------------------------------------------------------------------
```



```
hadoop@ip-172-31-89-12:~
grattol 36027.17
ano     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
opw.style       3265.29
lovely  3234.68
narathon        2992.35
naruyama        2962.22
yoko    2950.97
italwax 2859.13
penovy  2850.35
kaypro  2387.36
estel   2385.92
concept 2348.26
kapous  2165.92
f.o.x   1953.05
nasura  1792.39
nilv    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
loico   1309.58
reitun  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
```



Total 152 brands have increased their sales value from October to November.

<u>Question – 8</u>: Your Company wants to reward the top 10 users of its website with a Golden Customer plan. Write a query to generate a list of top 10 users who spend the most.

```
hive> 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 ecom_table_optimize
    > WHERE event_type = 'purchase'
    > GROUP BY user_id
    > )
    > SELECT Customer,
    > Expenditure,
    > Rank
    > FROM Customer_Rank
    > WHERE Rank <= 10;
Query ID = hadoop_20220430153303_a2928b81-4527-41b9-9a4d-3bae36151f21
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1651327969842_0005)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 ..........   container     SUCCEEDED      3          3        0        0       0       0
Reducer 2 ......   container     SUCCEEDED      1          1        0        0       0       0
Reducer 3 ......   container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 23.21 s
--------------------------------------------------------------------------------
OK
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: 23.995 seconds, Fetched: 10 row(s)
hive>
```

Query output provides list of top 10 customers who are eligible for Golden plan.

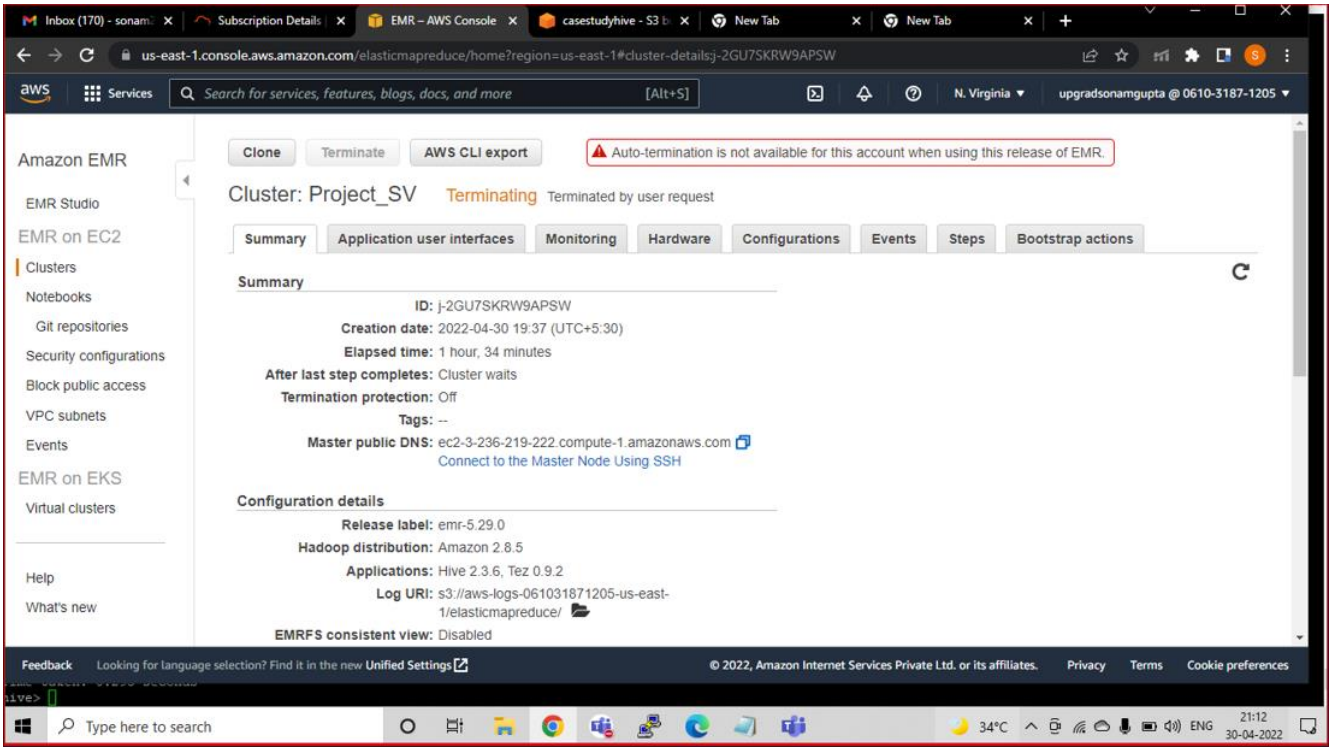
Dropping Database

- At completion, we dropped all created tables and database.

```
hive> SHOW TABLES;
OK
ecom_table
ecom_table_optimize
ecom_table_part
Time taken: 0.042 seconds, Fetched: 3 row(s)
hive> DROP TABLE ecom_table;
OK
Time taken: 0.412 seconds
hive> DROP TABLE ecom_table_optimize;
OK
Time taken: 0.145 seconds
hive> DROP TABLE ecom_table_part;
OK
Time taken: 0.187 seconds
hive> SHOW TABLES;
OK
Time taken: 0.017 seconds
hive> SHOW DATABASES;
OK
clickstream
default
Time taken: 0.015 seconds, Fetched: 2 row(s)
hive> DROP DATABASE clicstream;
FAILED: SemanticException [Error 10072]: Database does not exist: clicstream
hive> DROP DATABASE clickstream;
OK
Time taken: 0.295 seconds
hive>
```

Cluster Termination:

- At completion, we terminated the created cluster named **Project_SV**.



Insights:

- Change in revenue from October to November is 319478.47.This indicates that there is strong positive increase in business.
- Runail is the brand that made the maximum sales in October and November combined with sales amount of 148297.94
- Total 152 brands have increased their sales value from October to November.
- There are 11 distinct categories of products
- Category having highest products available is appliances environment vacuum and lowest products available is sport diving