**HIVE CASE STUDY ASSIGNMENT**

**[DS C41 – 2022]**

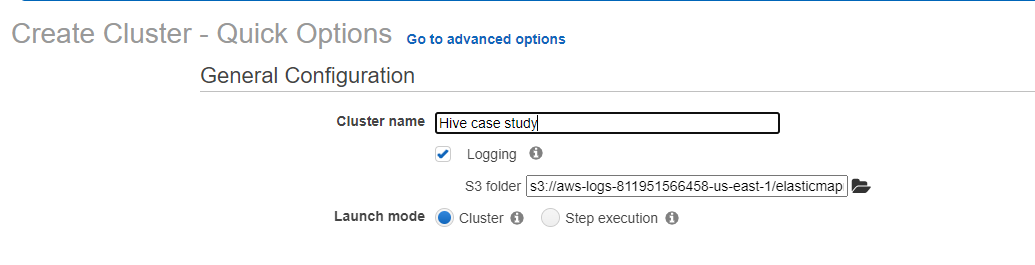
**E-COMMERCE SALES DATA ANALYSIS**



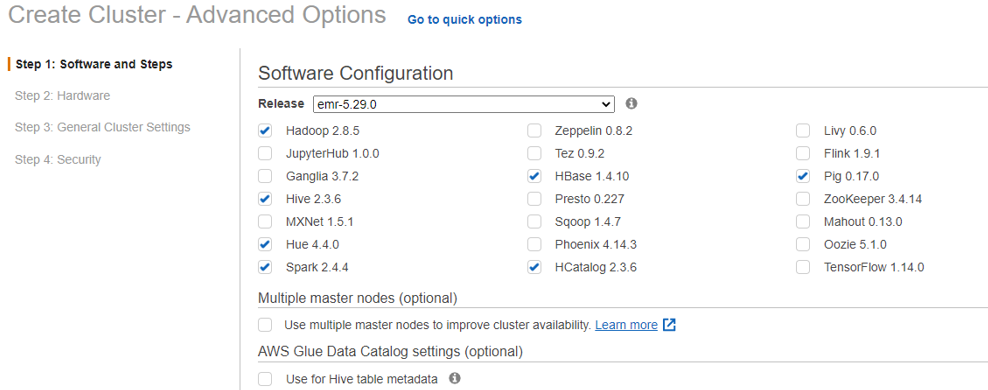
**Submitted by :-**

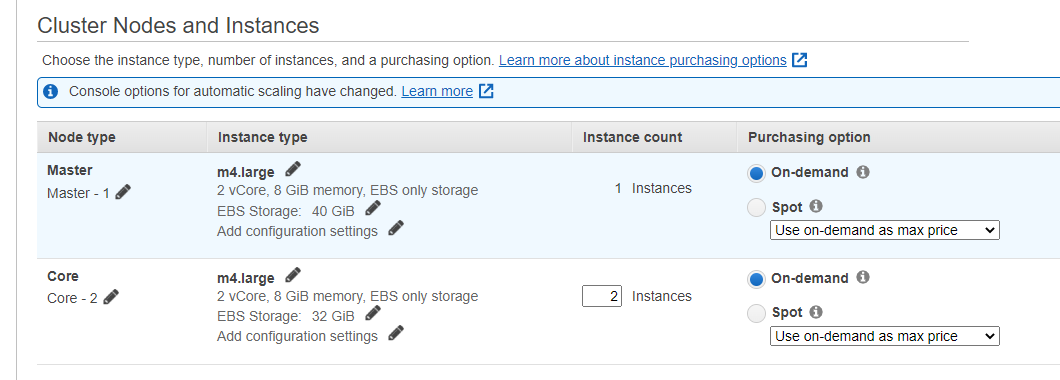
HARIOM AWASTHI & KARTIKEY SHUKLA

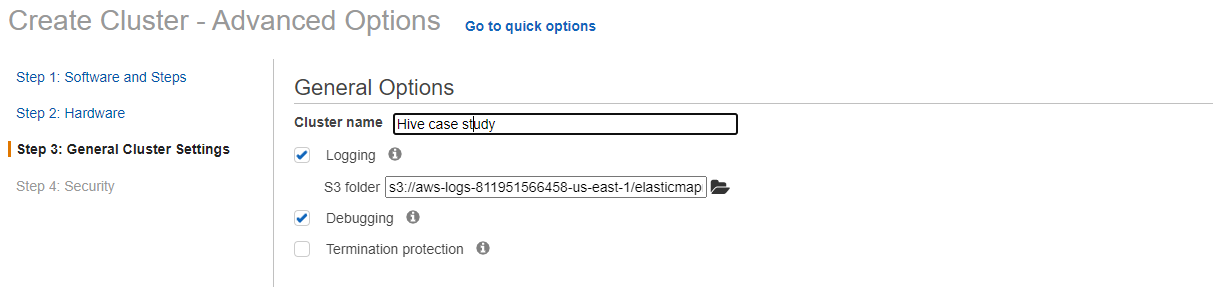
Creating the EMR cluster that utilizes the Hive services and Move the data from the S3 bucket into the HDFS.

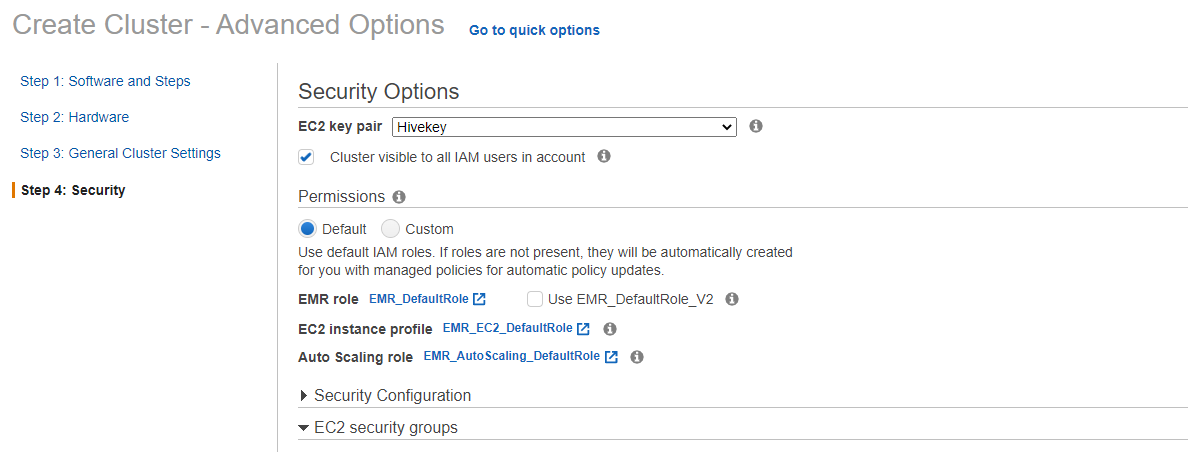


As suggested, we are using the emr-5.29.0 version for case study

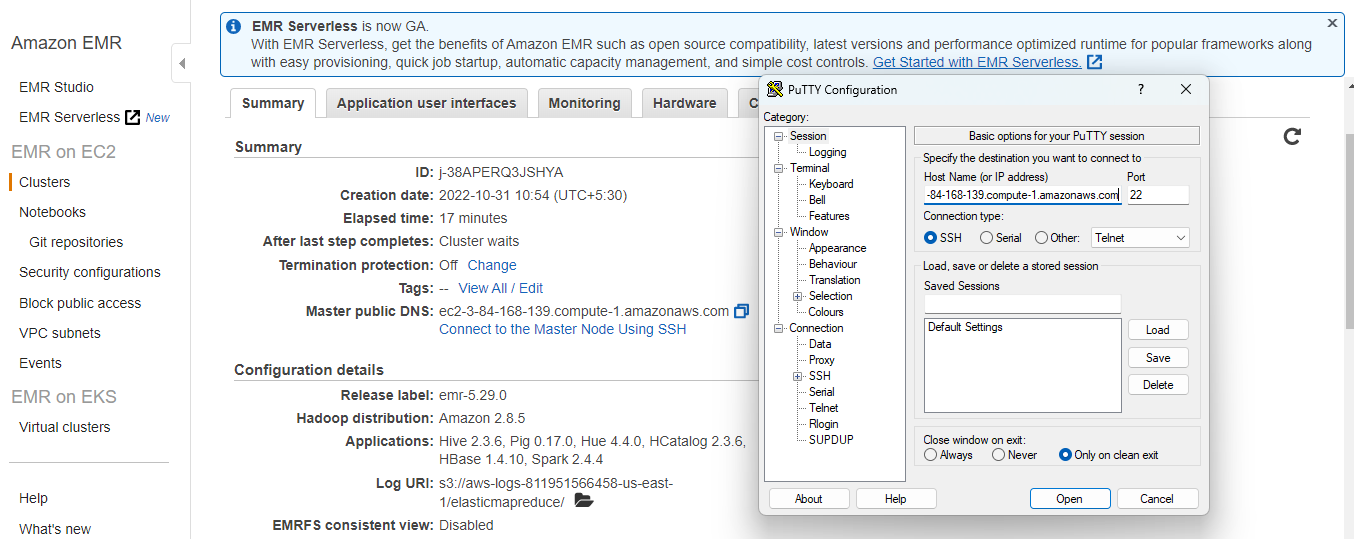


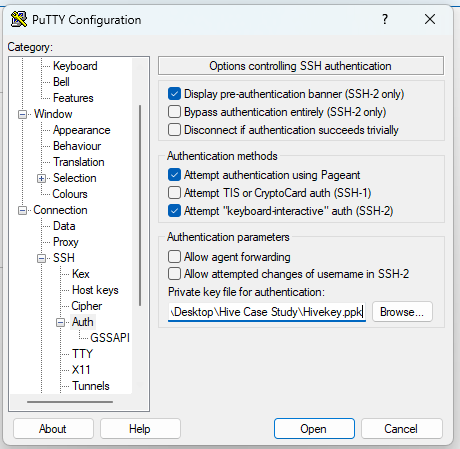
Using a 2-node EMR cluster with both the master and core nodes as M4.large.



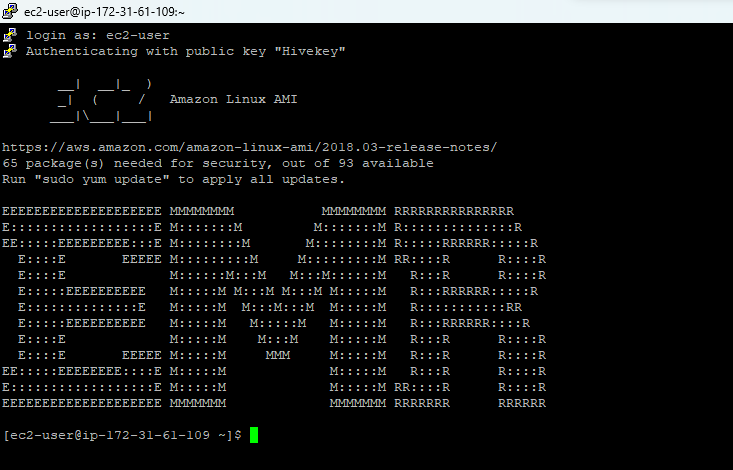


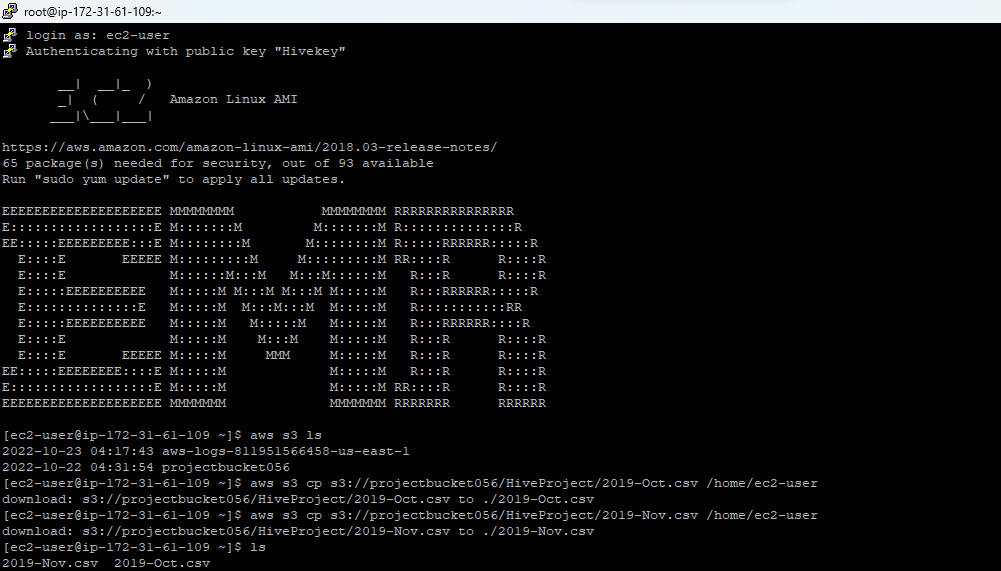
Connecting to putty

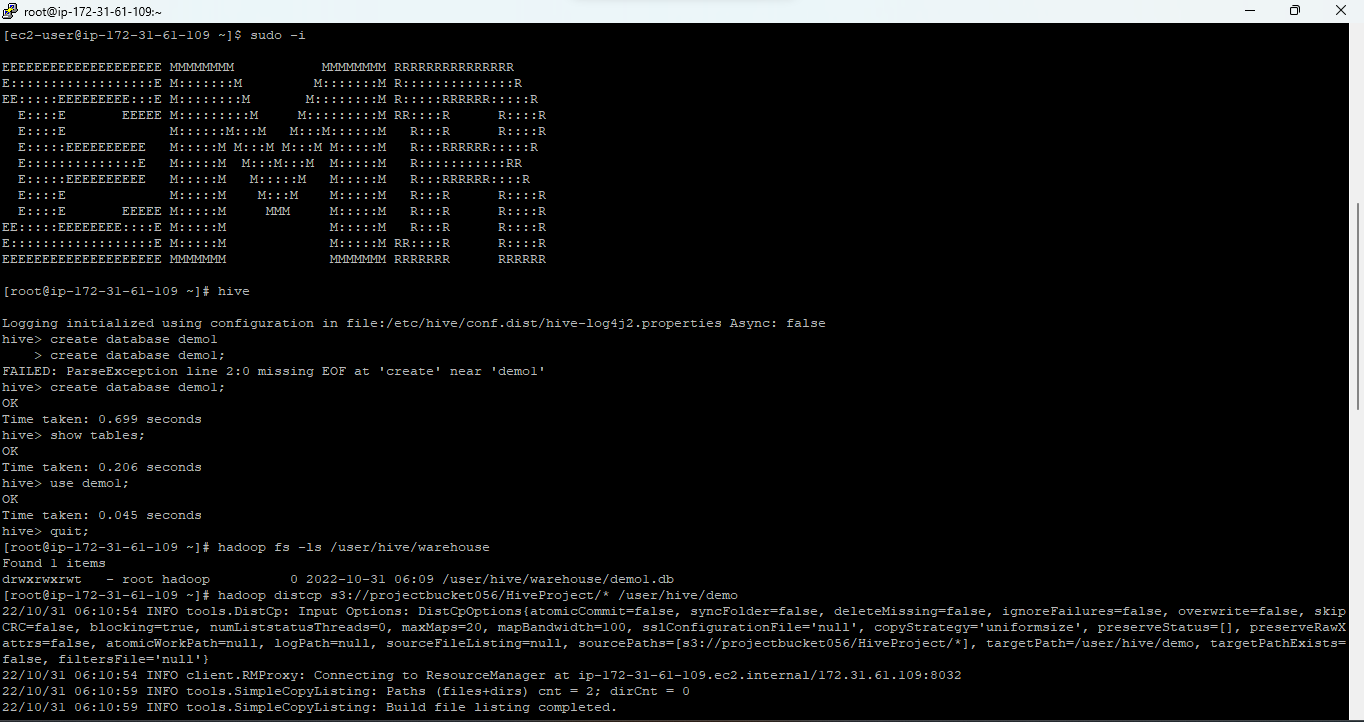


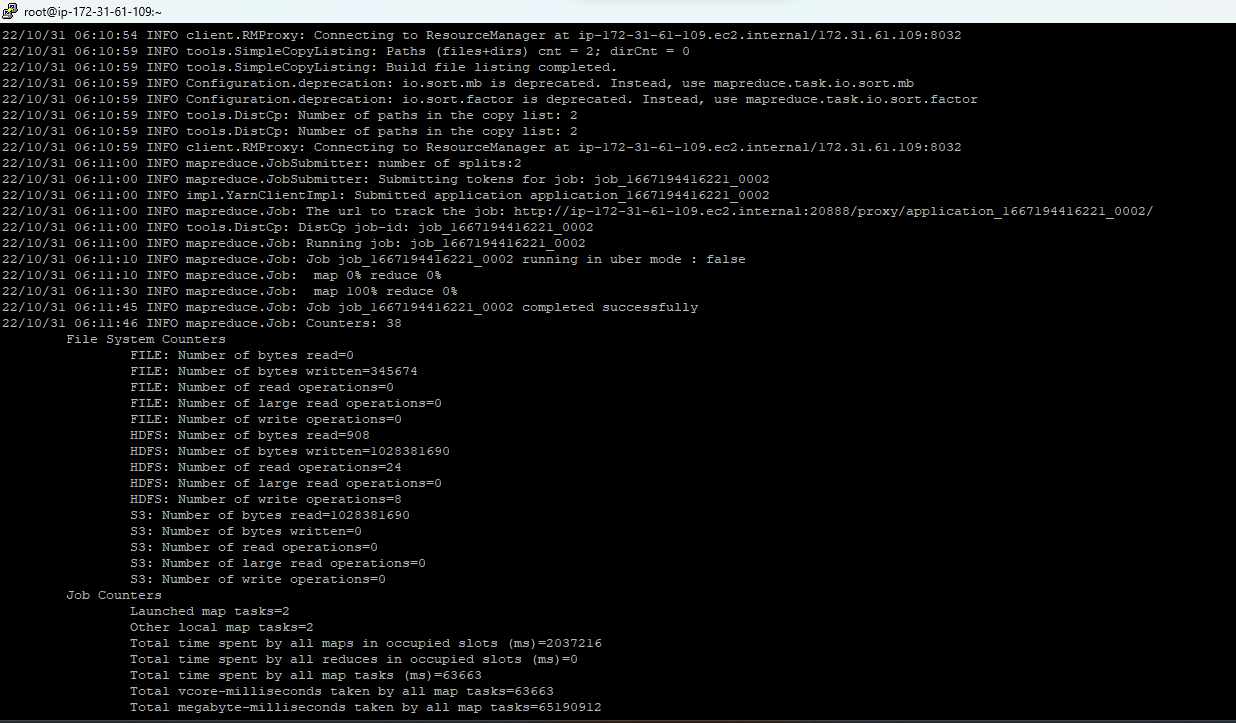


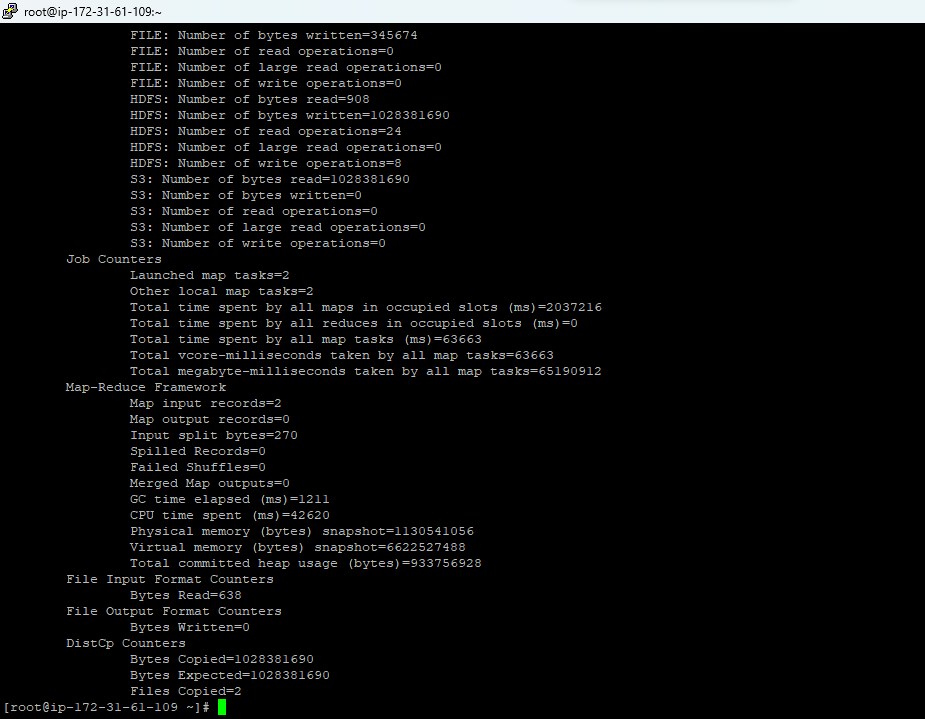
Launching an EMR Cluster that utilizes the Hive services.



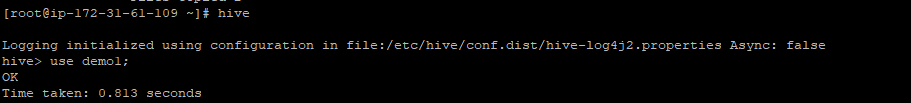
Creating the database and launching Hive queries on your EMR cluster 



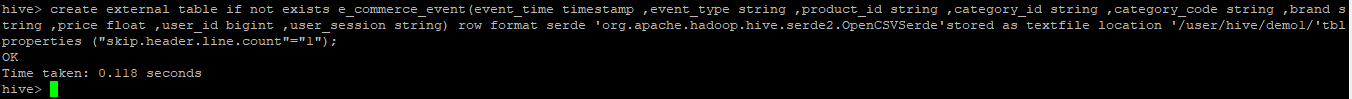


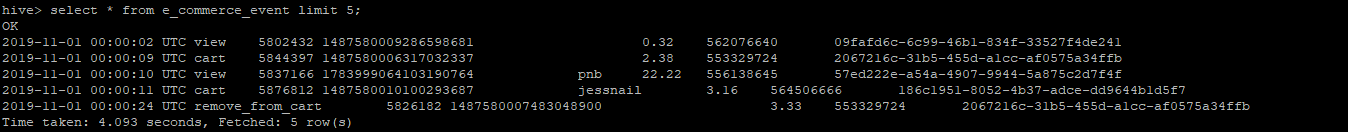


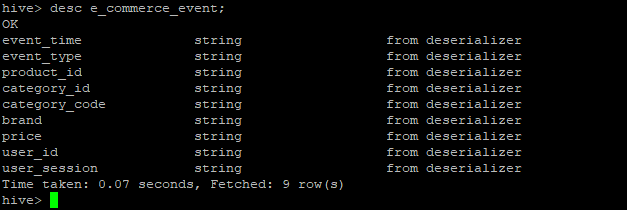
Creating the structure of database



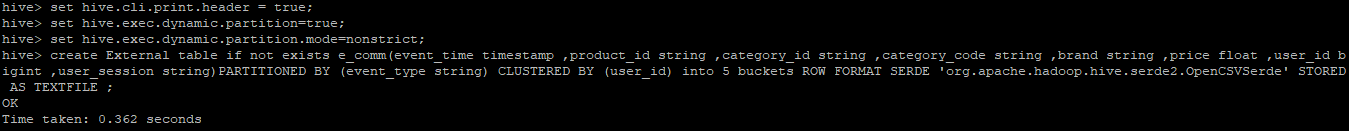
Creating an External Table e\_commerce\_event

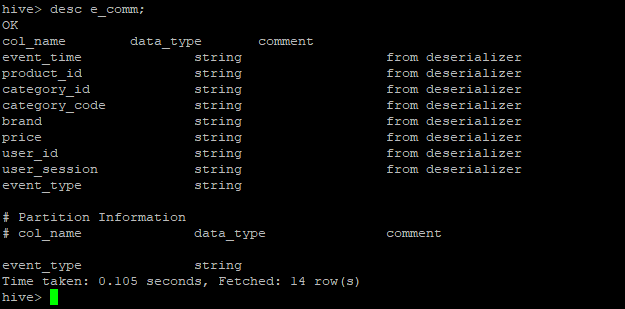


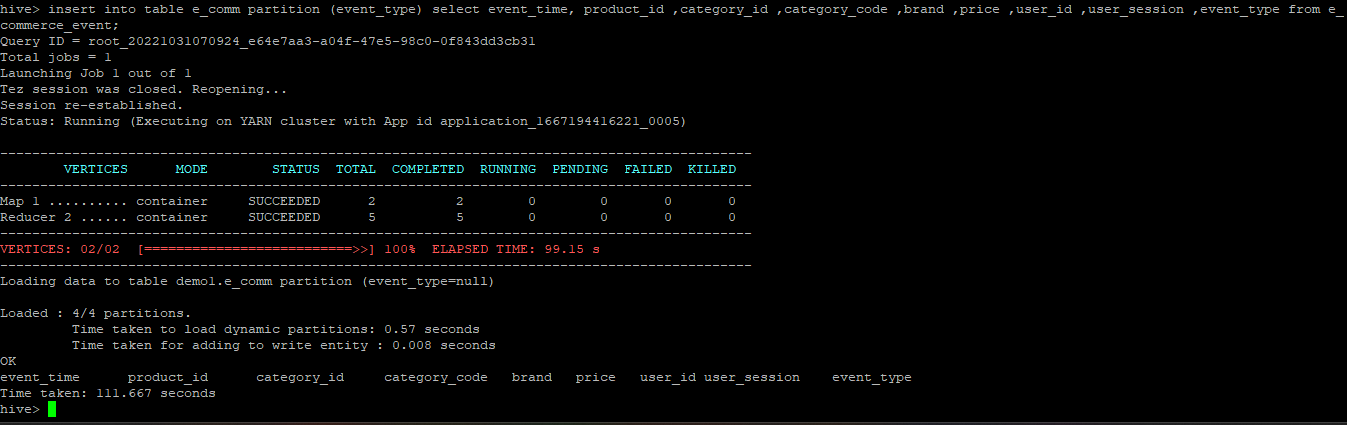




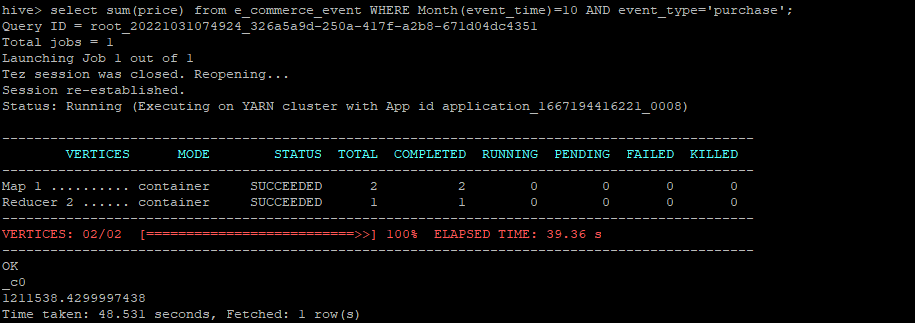
Loading the data into the new table e\_comm from the old e\_commerce\_event table





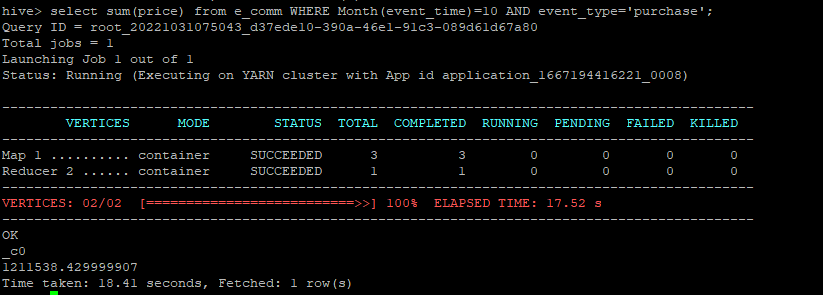


Q1. Find the total revenue generated due to purchases made in October.



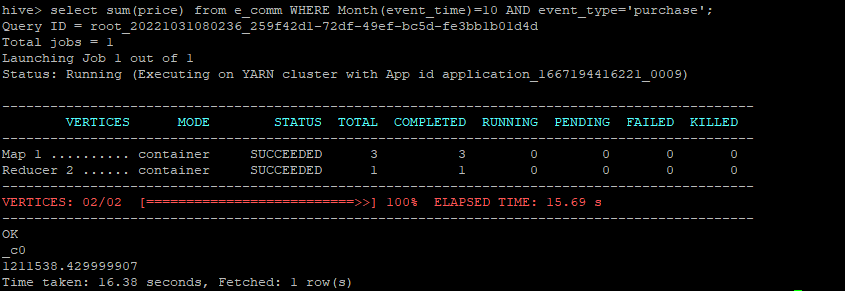
**Before partitioning** time :- **48.531 seconds**

The total sales in the month of October is **1211538.42999**.



**After partitioning** time :-**18.41 seconds**

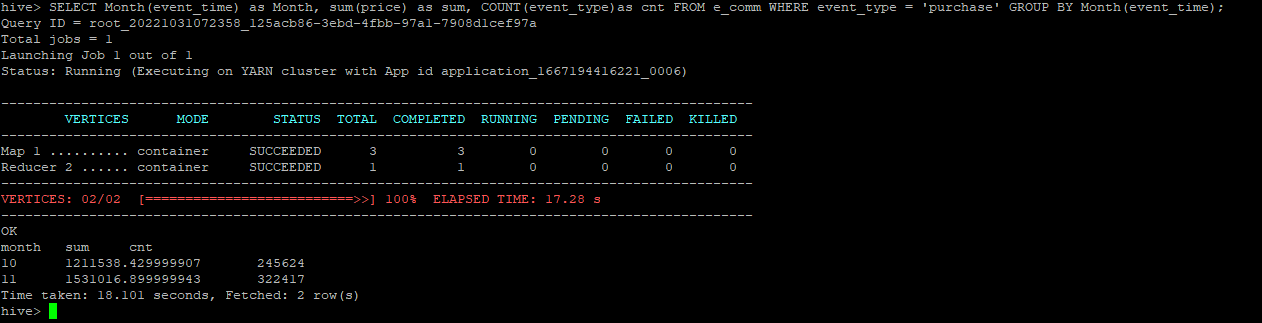
The total sales in the month of October is **1211538.42999**.



After **Dynamic partitioning** time :-**16.38 seconds.**

The total sales in the month of October is **1211538.42999**.

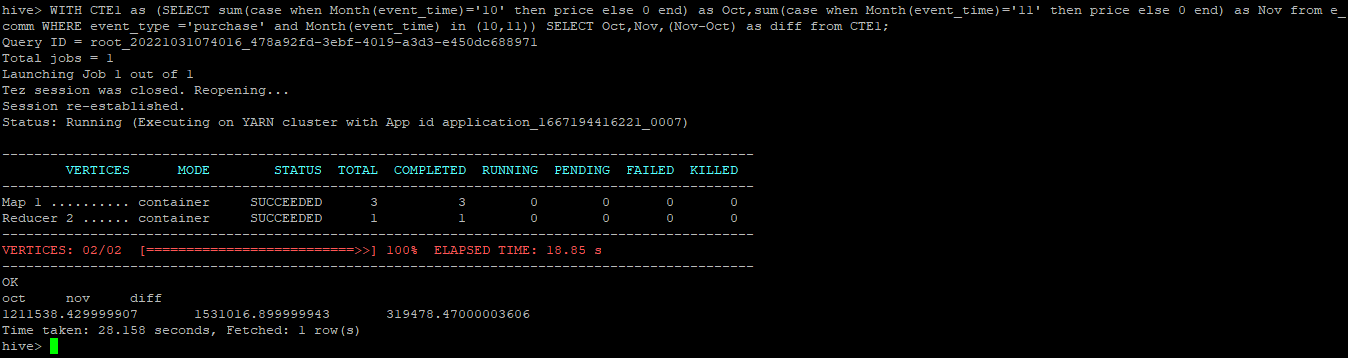
Q2. Write a query to yield the total sum of purchases per month in a single output.



In the month of **October** the total purchases are **245624** and sales is **1211538.42999**

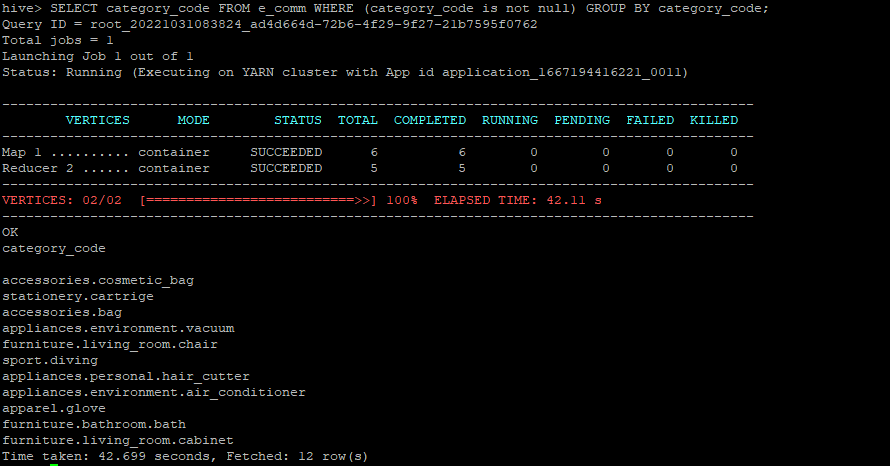
In the month of **November** the total purchases are **322417** and sales is **1531016.89999**

Q3. Write a query to find the change in revenue generated due to purchases from October to November.



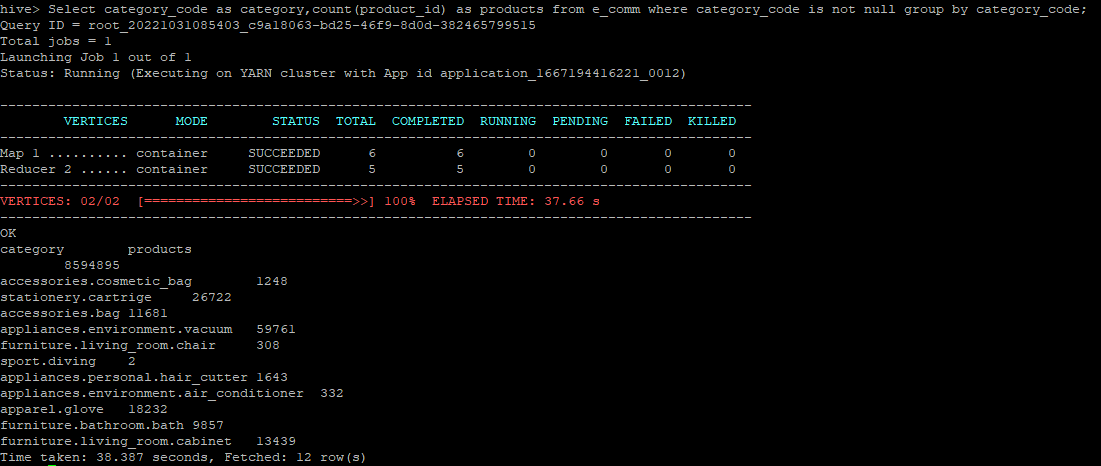
We can see that the difference in the revenue is **319478.47000.**

Q4. Find distinct categories of products. Categories with null category code can be ignored.

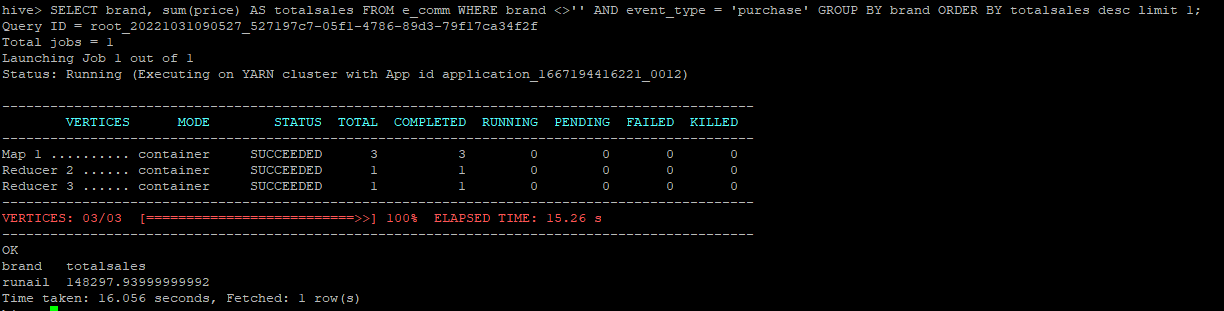


The distinct categories are Furniture, Appliances, Accessories, Apparel, sport, and stationery.

Q5. Find the total number of products available under each category.

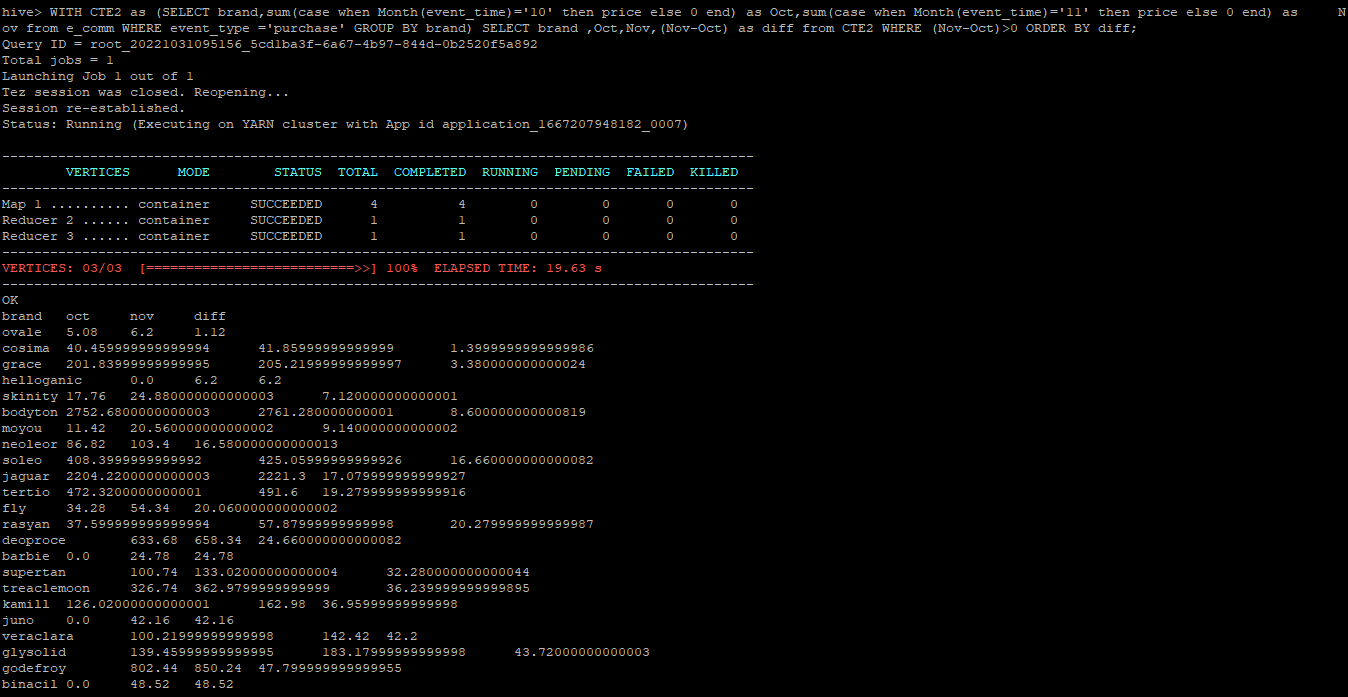


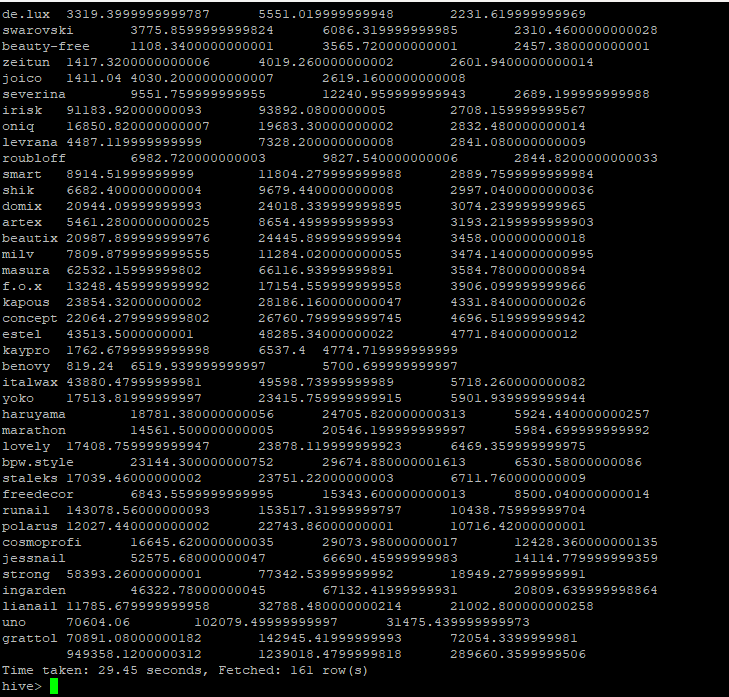
Q6. Which brand had the maximum sales in October and November combined?



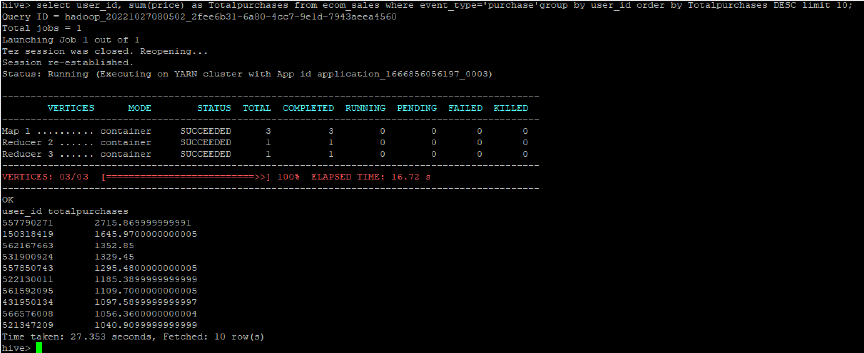
**Runail** is the brand with the maximum sales for October and November combined with total **148297.9399**

Q7. Which brands increased their sales from October to November?



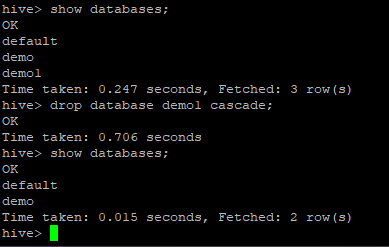


Q8. Your company wants to reward the top 10 users of its website with a Golden Customer plan. Write a query to generate a list of top 10 users who spend the most.



**Cleaning up**

Dropping database demo1



Terminating the cluster

