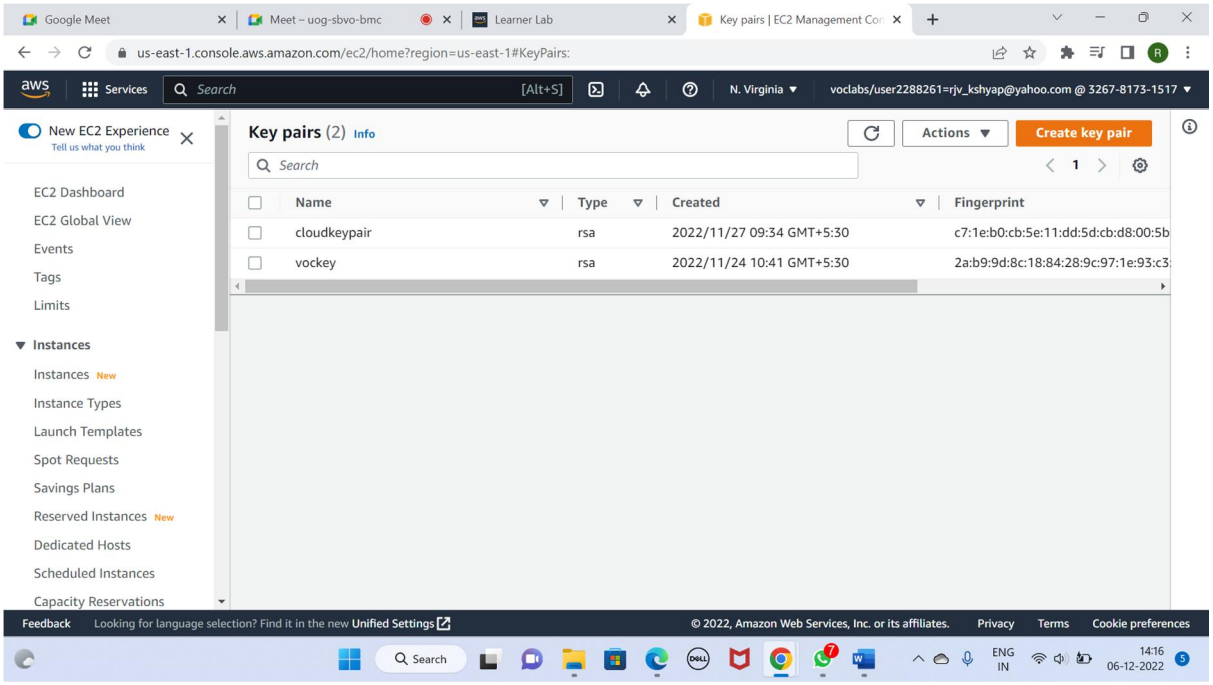
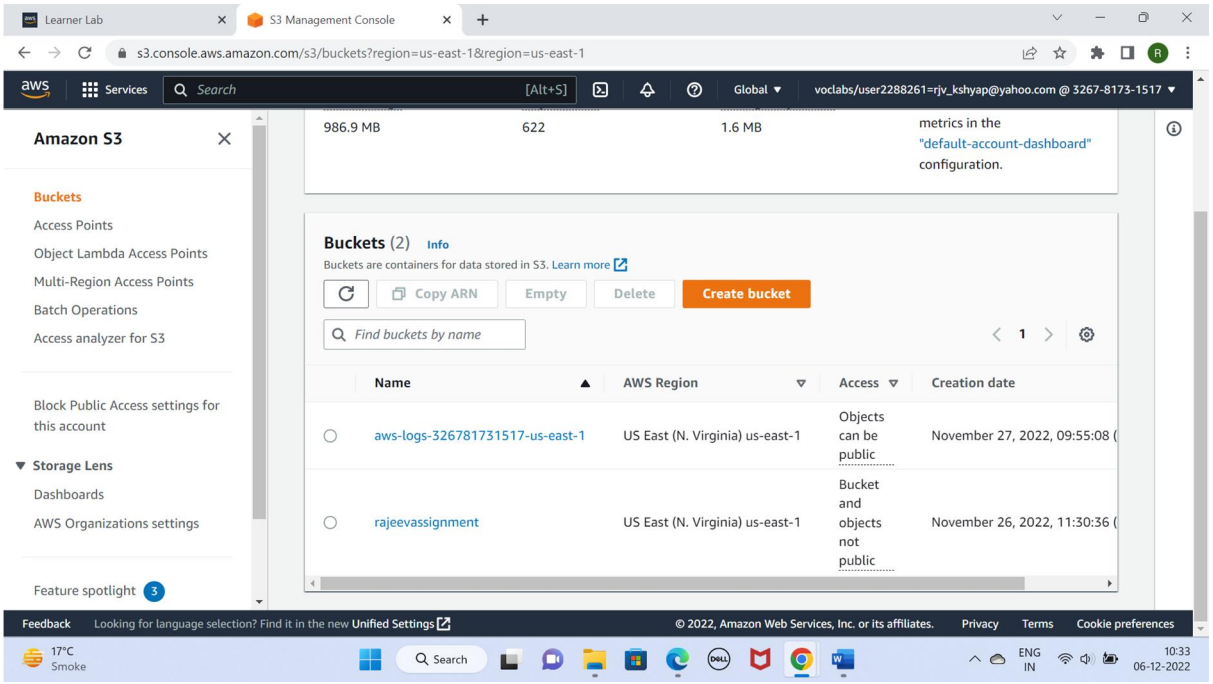
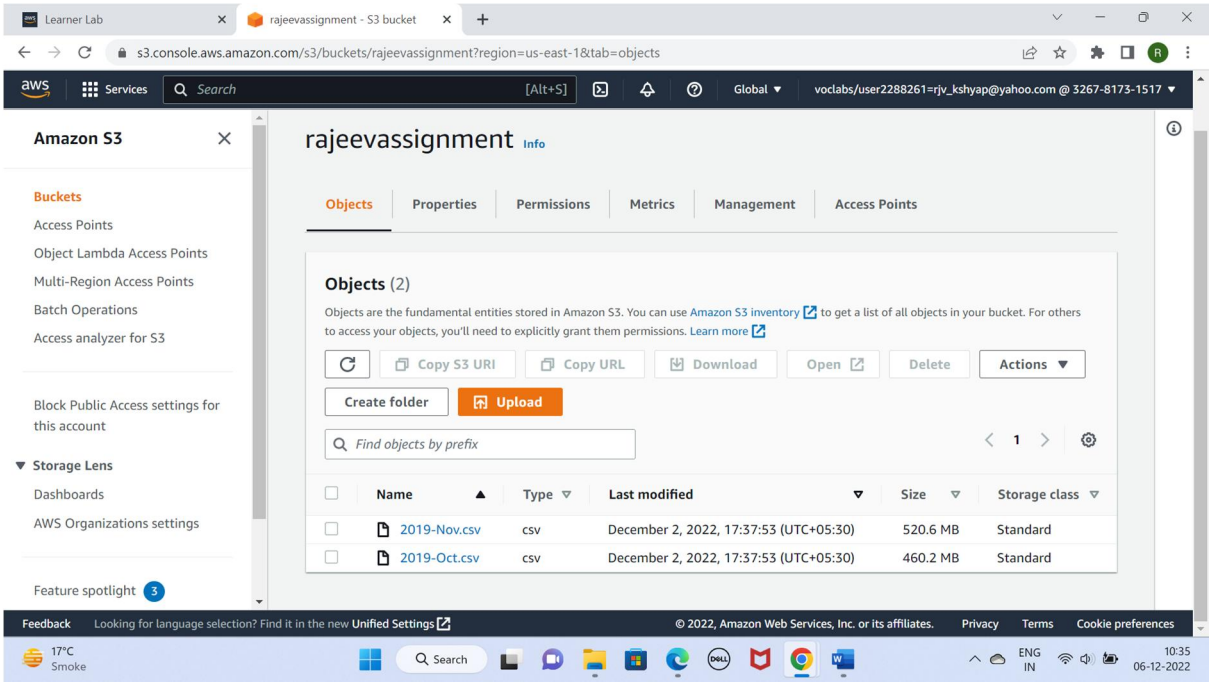
We have created keypair with a name “cloudkeypair”.



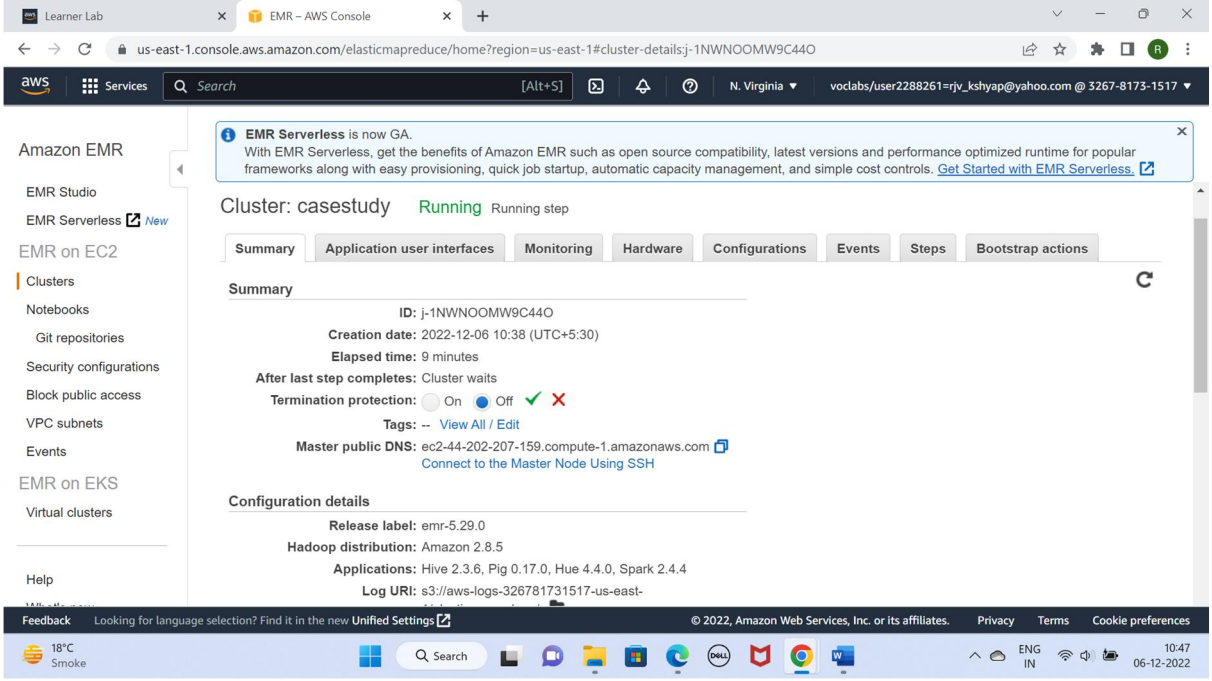
We have created bucket with name “rajeevassignment”.



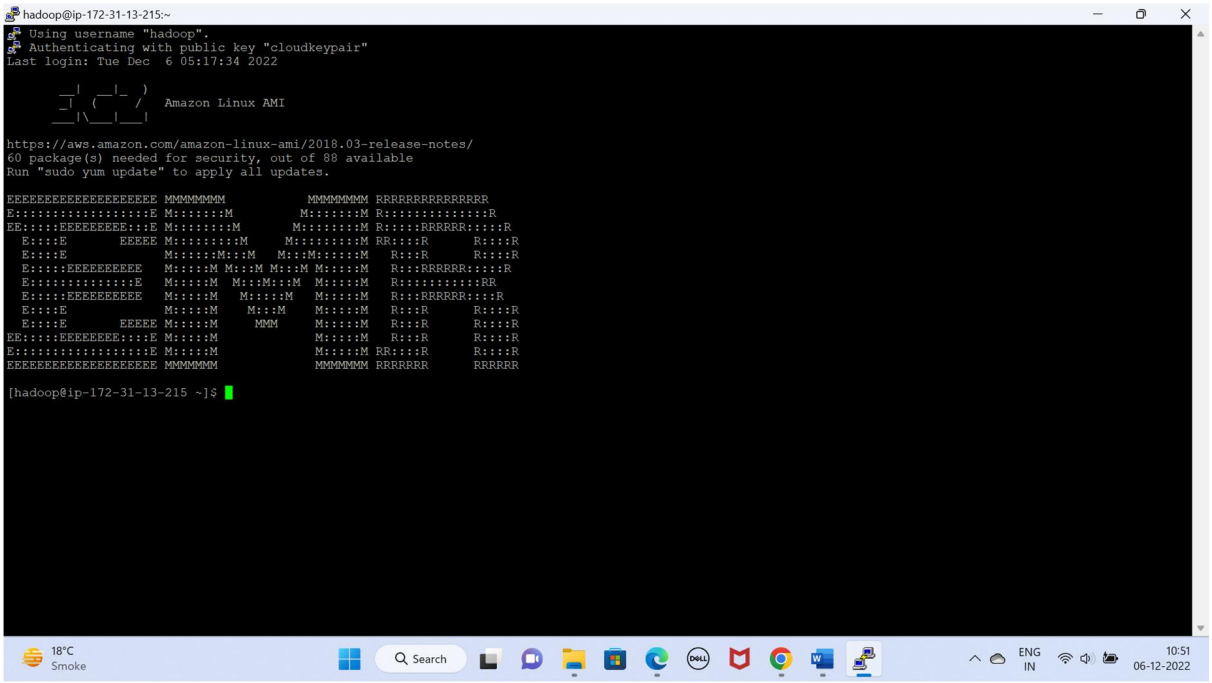
We have uploaded October and November dataset as given in case study.



We have created cluster named “casestudy”.It is running now as shown below:



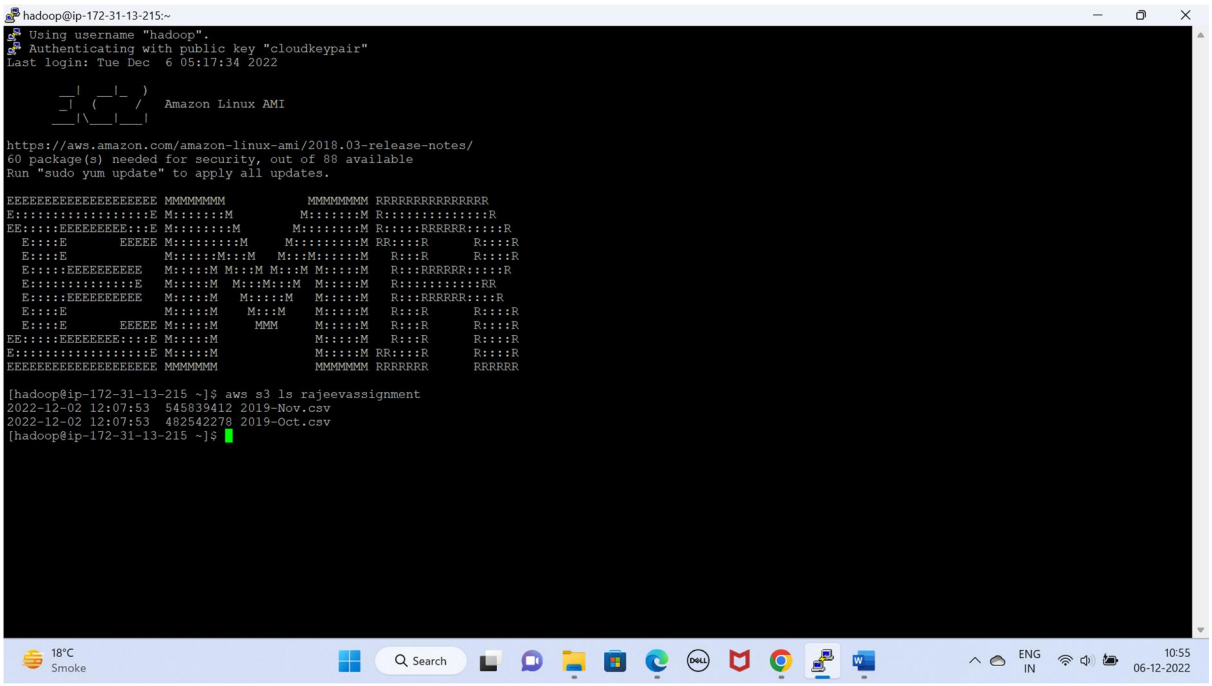
We have connected putty and it is activated now.



Data is copied into HDFS.

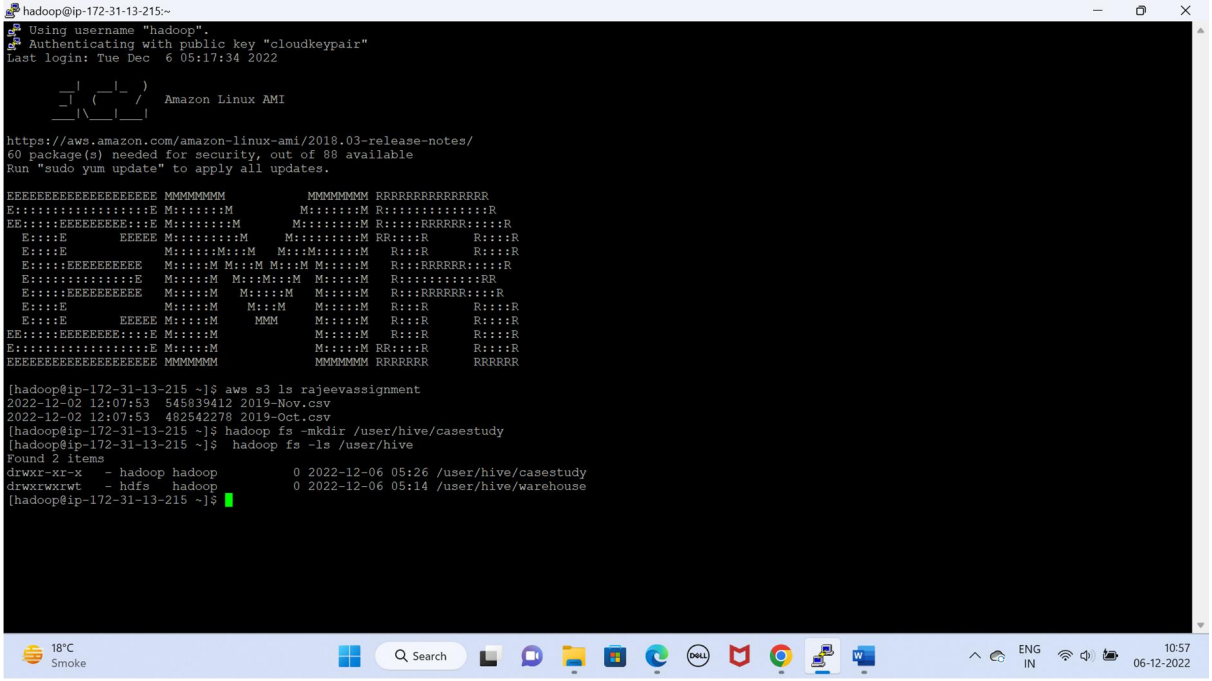
Accessing the S3 bucket.

Code: aws s3 ls rajeevassignment



Checking the availability of directory

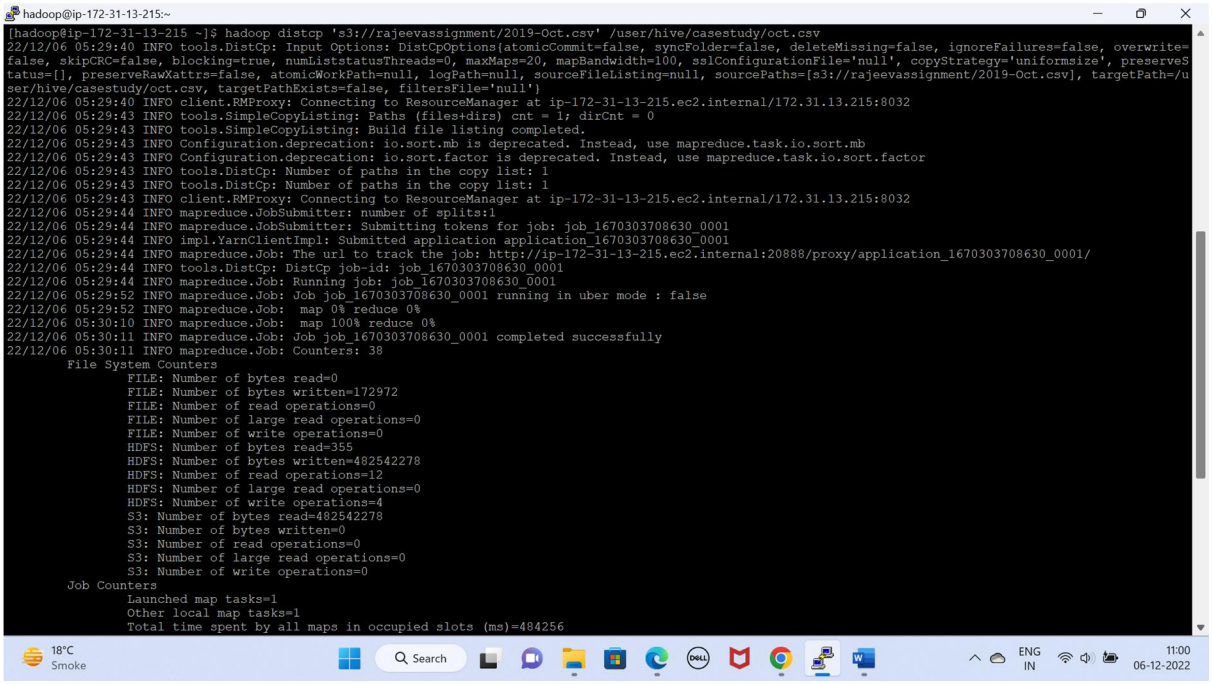
Code: Hadoop fs –ls /user/hive

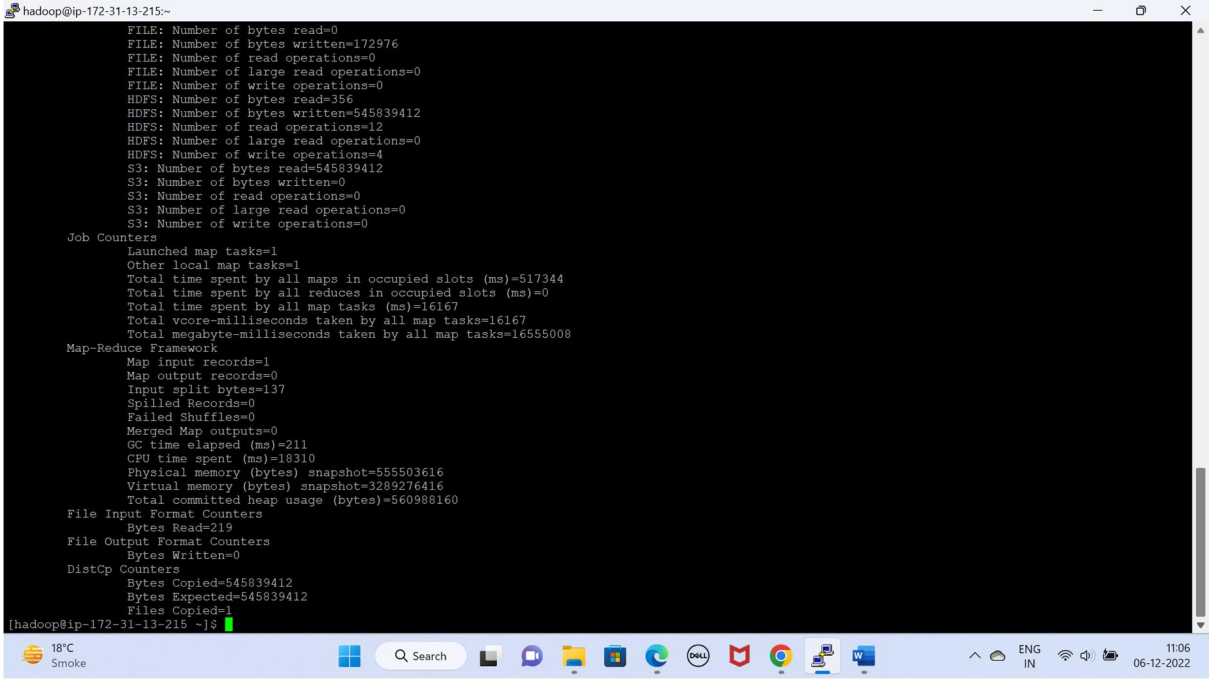


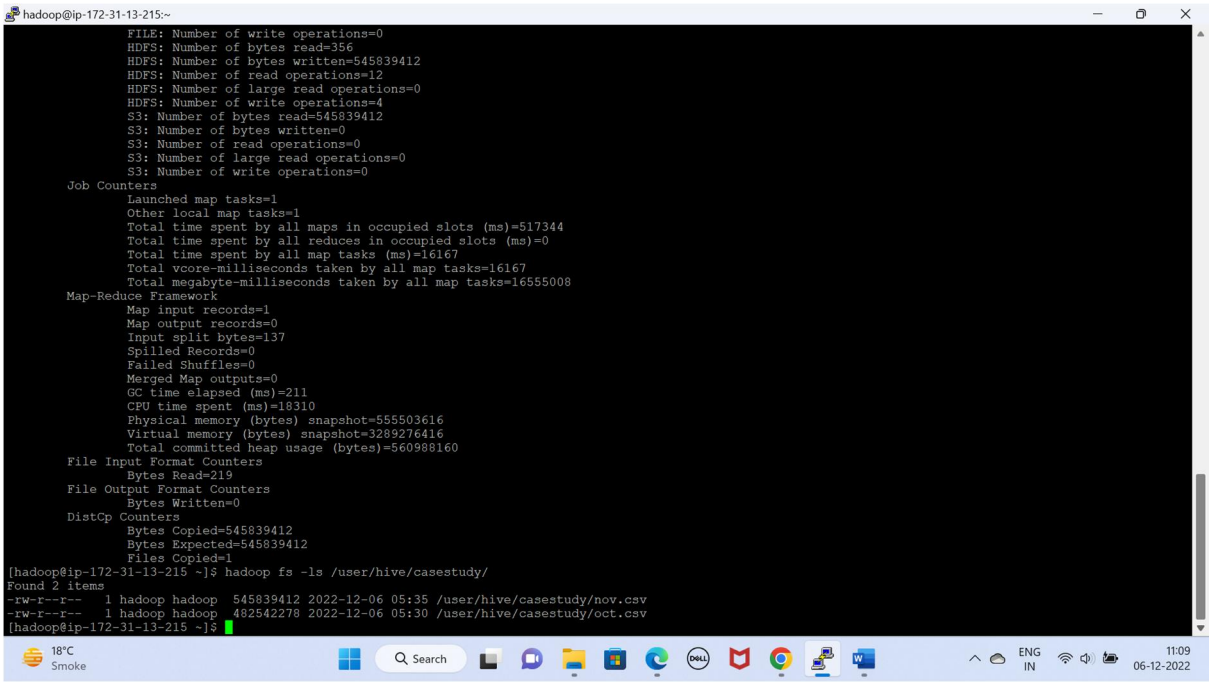
Loading the s3 public data set to created directory “casestudy” in hadoop .

Code: hadoop distcp 's3://rajeevassignment/2019-Oct.csv' /user/hive/casestudy/oct.csv

hadoop distcp 's3://rajeevassignment/2019-Nov.csv' /user/hive/casestudy/nov.csv





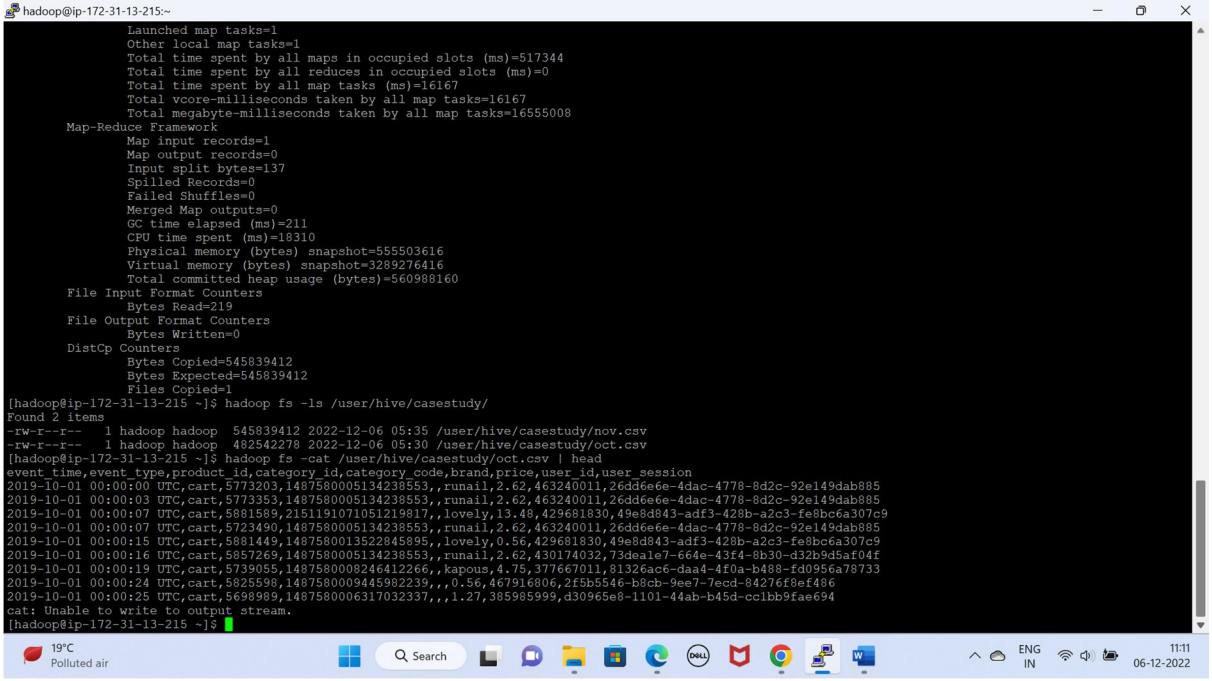


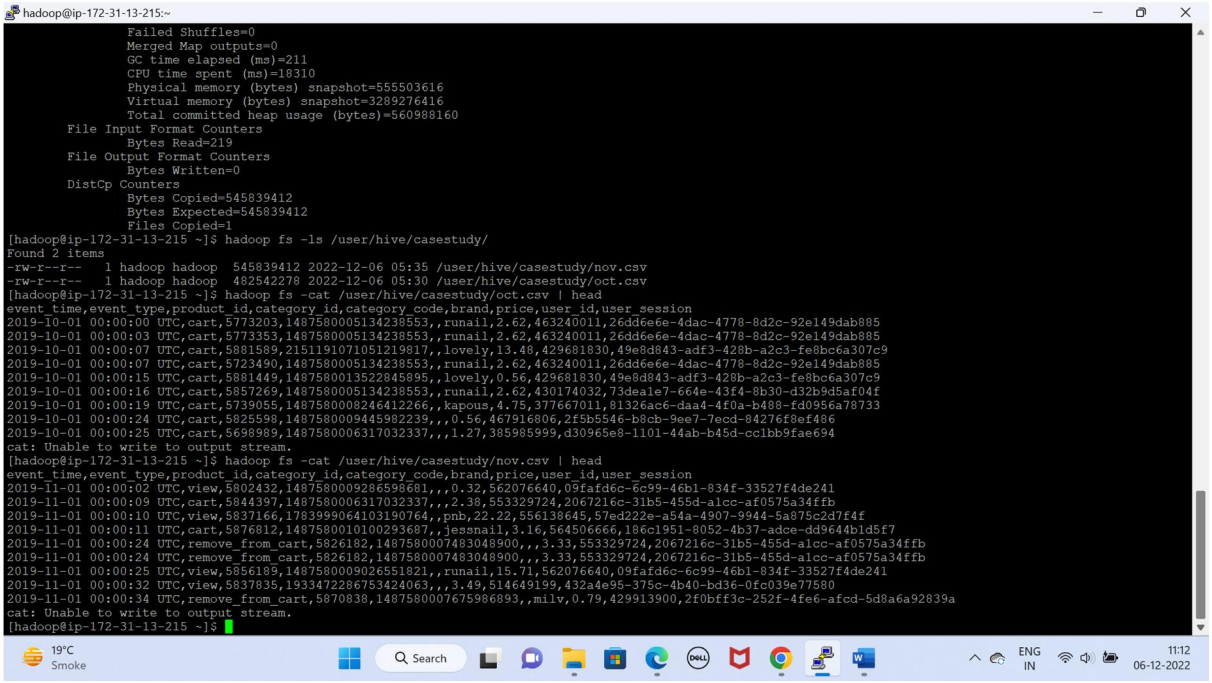
After loading the data set we check the data set file and data set in the hadoop directory.

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

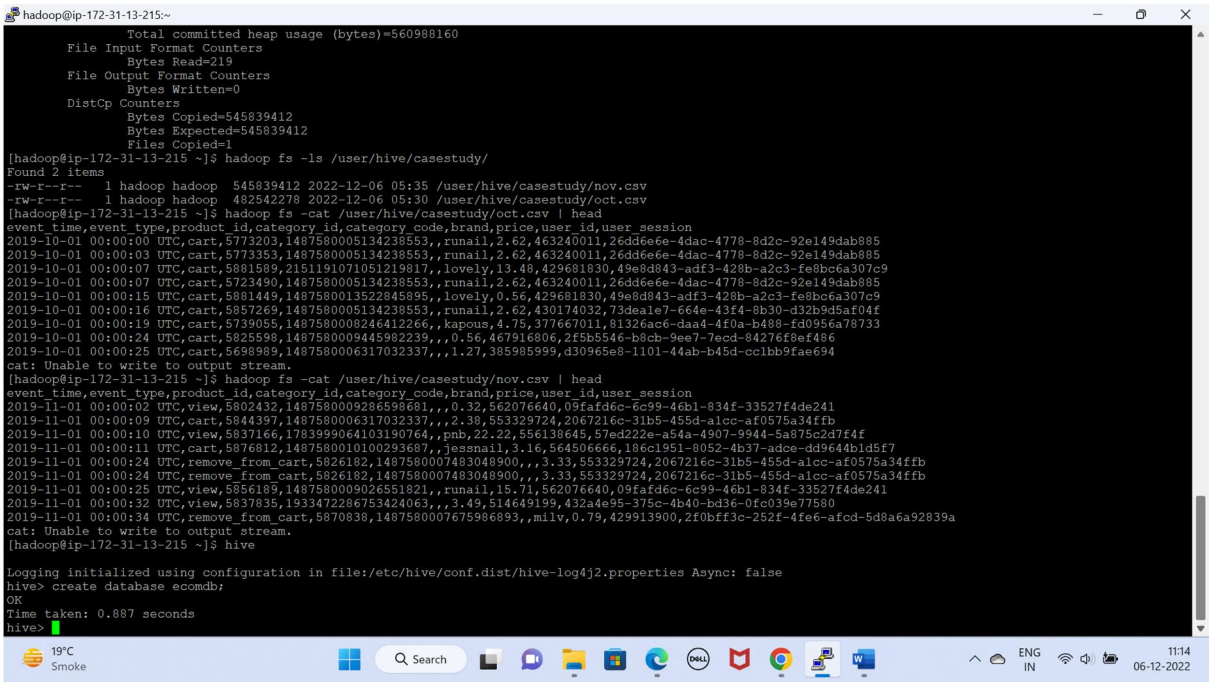
hadoop fs -cat /user/hive/casestudy/oct.csv | head

hadoop fs -cat /user/hive/casestudy/nov.csv | head


hadoop@ip-172-31-13-215:~
Launched map tasks=1
Other local map tasks=1
Total time spent by all maps in occupied slots (ms)=517344
Total time spent by all reduces in occupied slots (ms)=0
Total time spent by all map tasks (ms)=16167
Total vcore-milliseconds taken by all map tasks=16167
Total megabyte-milliseconds taken by all map tasks=16555008
Map-Reduce Framework
Map input records=1
Map output records=0
Input split bytes=137
Spilled Records=0
Failed Shuffles=0
Merged Map outputs=0
GC time elapsed (ms)=211
CPU time spent (ms)=18310
Physical memory (bytes) snapshot=555503616
Virtual memory (bytes) snapshot=3289276416
Total committed heap usage (bytes)=560988160
File Input Format Counters
Bytes Read=219
File Output Format Counters
Bytes Written=0
DistCp Counters
Bytes Copied=545839412
Bytes Expected=545839412
Files Copied=1
[hadoop@ip-172-31-13-215 ~]$ hadoop fs -ls /user/hive/casestudy/
Found 2 items
-rw-r--r-- 1 hadoop hadoop 545839412 2022-12-06 05:35 /user/hive/casestudy/nov.csv
-rw-r--r-- 1 hadoop hadoop 482542278 2022-12-06 05:30 /user/hive/casestudy/oct.csv
[hadoop@ip-172-31-13-215 ~]$ hadoop fs -cat /user/hive/casestudy/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-172-31-13-215 ~]$
19°C
Polluted air
Search
ENG
IN
11:11
06-12-2022


hadoop@ip-172-31-13-215:~
Failed Shuffles=0
Merged Map outputs=0
GC time elapsed (ms)=211
CPU time spent (ms)=18310
Physical memory (bytes) snapshot=555503616
Virtual memory (bytes) snapshot=3289276416
Total committed heap usage (bytes)=560988160
File Input Format Counters
Bytes Read=219
File Output Format Counters
Bytes Written=0
DistCp Counters
Bytes Copied=545839412
Bytes Expected=545839412
Files Copied=1
[hadoop@ip-172-31-13-215 ~]$ hadoop fs -ls /user/hive/casestudy/
Found 2 items
-rw-r--r-- 1 hadoop hadoop 545839412 2022-12-06 05:35 /user/hive/casestudy/nov.csv
-rw-r--r-- 1 hadoop hadoop 482542278 2022-12-06 05:30 /user/hive/casestudy/oct.csv
[hadoop@ip-172-31-13-215 ~]$ hadoop fs -cat /user/hive/casestudy/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-172-31-13-215 ~]$ hadoop fs -cat /user/hive/casestudy/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-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
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-5d8a6a92839a
cat: Unable to write to output stream.
[hadoop@ip-172-31-13-215 ~]$
19°C
Smoke
Search
ENG
IN
11:12
06-12-2022

Creating the database



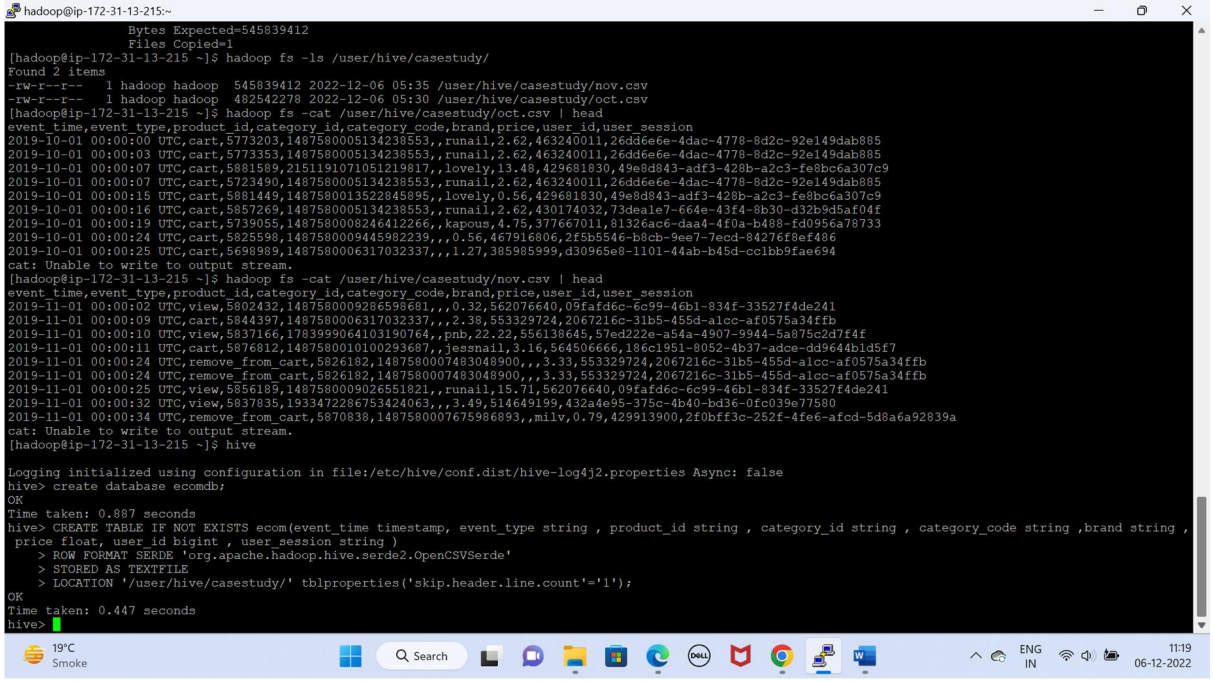
After moving the data to the directory, we create the base table(ecom) and check for the data in the table.

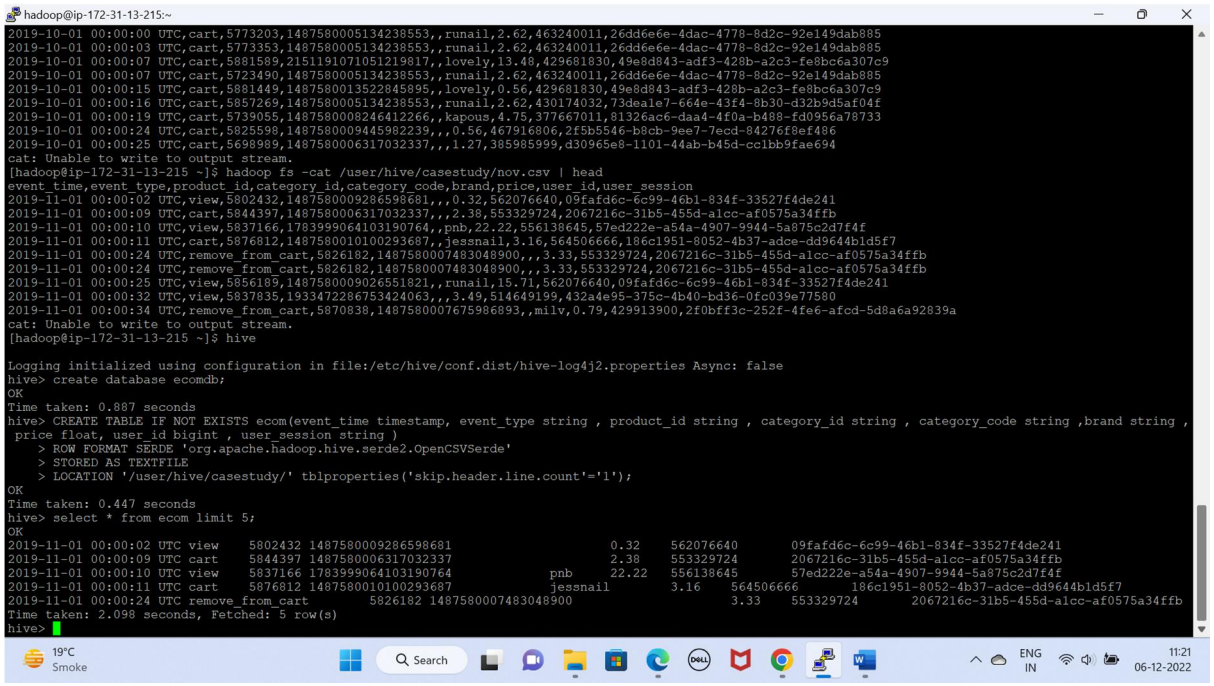
Code: CREATE TABLE IF NOT EXISTS ecom(event_time timestamp, event_type string , product_id string , category_id string , category_code string ,brand string ,price float, user_id bigint , user_session string)

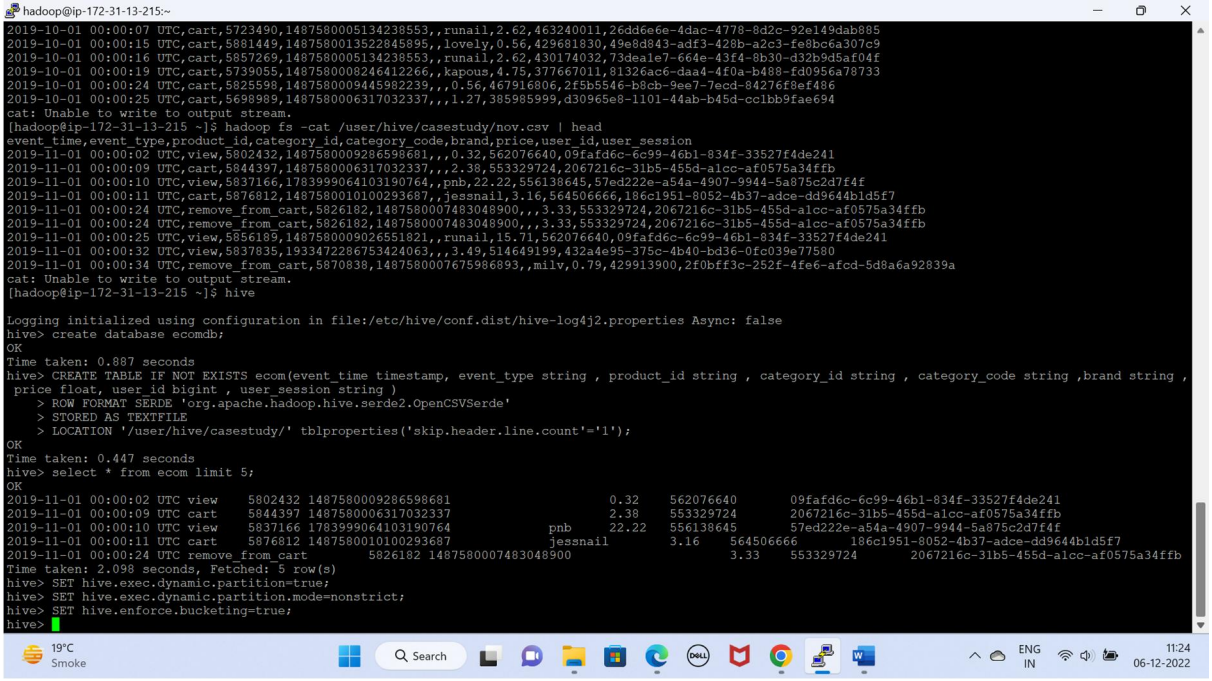
ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.OpenCSVSerde'

STORED AS TEXTFILE

LOCATION '/user/hive/casestudy/' tblproperties('skip.header.line.count'='1');


hadoop@ip-172-31-13-215:~
Bytes Expected=545839412
Files Copied=1
[hadoop@ip-172-31-13-215 ~]$ hadoop fs -ls /user/hive/casestudy/
Found 2 items
-rw-r--r-- 1 hadoop hadoop 545839412 2022-12-06 05:35 /user/hive/casestudy/nov.csv
-rw-r--r-- 1 hadoop hadoop 482542278 2022-12-06 05:30 /user/hive/casestudy/oct.csv
[hadoop@ip-172-31-13-215 ~]$ hadoop fs -cat /user/hive/casestudy/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-172-31-13-215 ~]$ hadoop fs -cat /user/hive/casestudy/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-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
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-5d8a6a92839a
cat: Unable to write to output stream.
[hadoop@ip-172-31-13-215 ~]$ hive
Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j2.properties Async: false
hive> create database ecomdb;
OK
Time taken: 0.887 seconds
hive> CREATE TABLE IF NOT EXISTS ecom(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: 0.447 seconds
hive>
19°C
Smoke
Search
ENG
IN
11:19
06-12-2022


hadoop@ip-172-31-13-215:~
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-172-31-13-215 ~]$ hadoop fs -cat /user/hive/casestudy/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-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
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-5d8a6a92839a
cat: Unable to write to output stream.
[hadoop@ip-172-31-13-215 ~]$ hive
Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j2.properties Async: false
hive> create database ecomdb;
OK
Time taken: 0.887 seconds
hive> CREATE TABLE IF NOT EXISTS ecom(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: 0.447 seconds
hive> select * from ecom 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: 2.098 seconds, Fetched: 5 row(s)
hive>
19°C
Smoke
Search
ENG
IN
11:21
06-12-2022



Once the base table is created, we need to optimize the table for quick query result through partitioning and bucketing. Our optimized table name is ecom_part.

Code:

CREATE TABLE IF NOT EXISTS ecom_part (event_time timestamp, product_id string , category_id string , category_code string ,brand string , price float, user_id bigint , user_session string)

partitioned by (event_type string)

clustered by (category_code) into 12 buckets

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

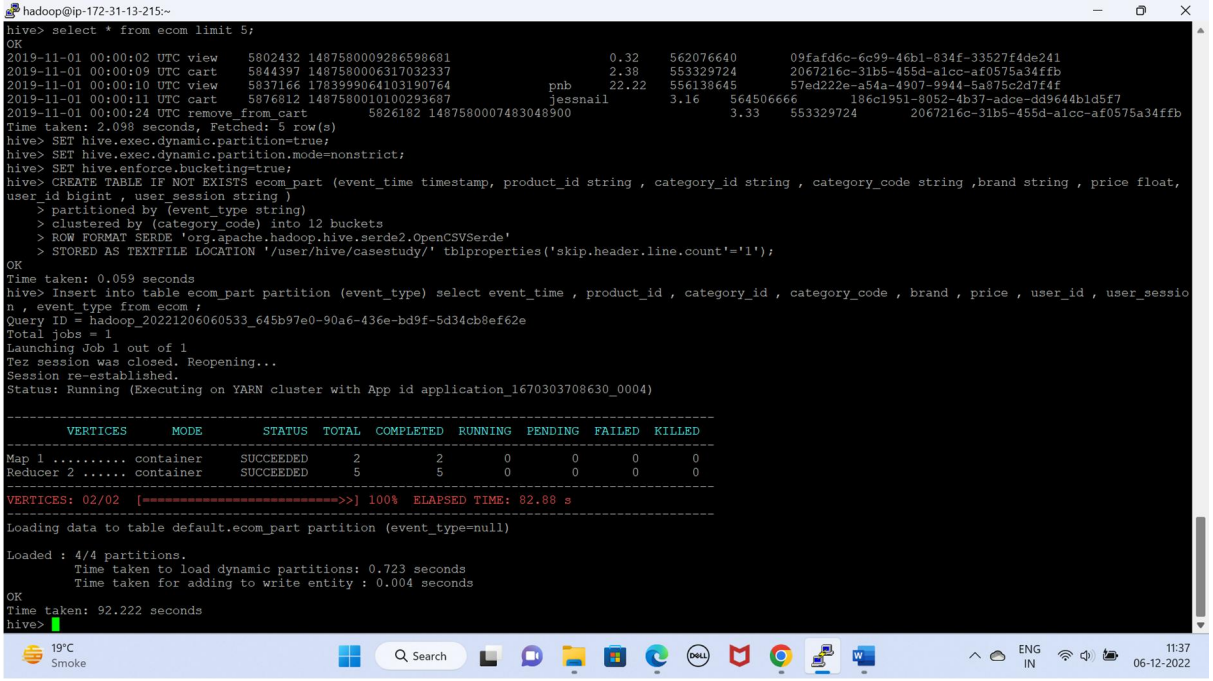
STORED AS TEXTFILE LOCATION '/user/hive/casestudy/' tblproperties('skip.header.line.count'='1');

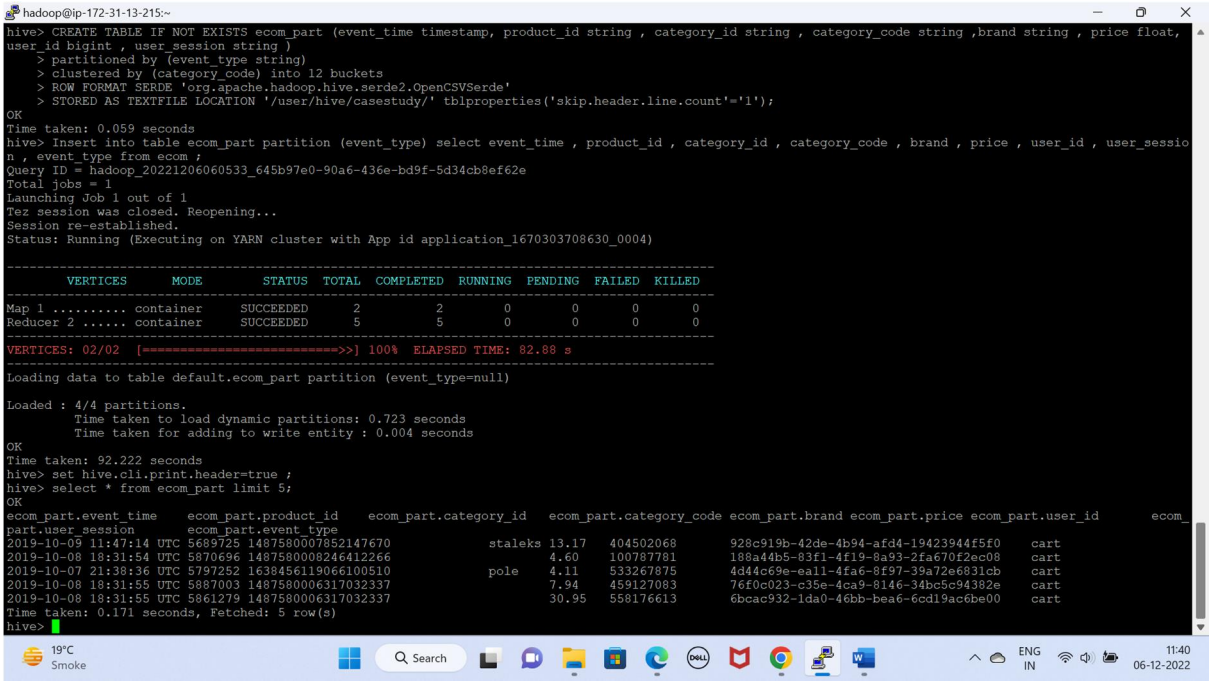
AND

Loading the data into optimize table from base table.

Code:

Insert into table ecom_part partition (event_type) select event_time , product_id , category_id , category_code , brand , price , user_id , user_session , event_type from ecom ;


hadoop@ip-172-31-13-215:~
hive> select * from ecom 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: 2.098 seconds, Fetched: 5 row(s)
hive> SET hive.exec.dynamic.partition=true;
hive> SET hive.exec.dynamic.partition.mode=nonstrict;
hive> SET hive.enforce.bucketing=true;
hive> CREATE TABLE IF NOT EXISTS ecom_part (event_time timestamp, product_id string , category_id string , category_code string ,brand string , price float,
user_id bigint , user_session string)
> partitioned by (event_type string)
> clustered by (category_code) into 12 buckets
> 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: 0.059 seconds
hive> Insert into table ecom_part partition (event_type) select event_time , product_id , category_id , category_code , brand , price , user_id , user_sessio
n , event_type from ecom ;
Query ID = hadoop_20221206060533_645b97e0-90a6-436e-bd9f-5d34cb8ef62e
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_1670303708630_0004)
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: 82.88 s
Loading data to table default.ecom_part partition (event_type=null)
Loaded : 4/4 partitions.
Time taken to load dynamic partitions: 0.723 seconds
Time taken for adding to write entity : 0.004 seconds
OK
Time taken: 92.222 seconds
hive>
19°C
Smoke
Search
ENG
IN
11:37
06-12-2022

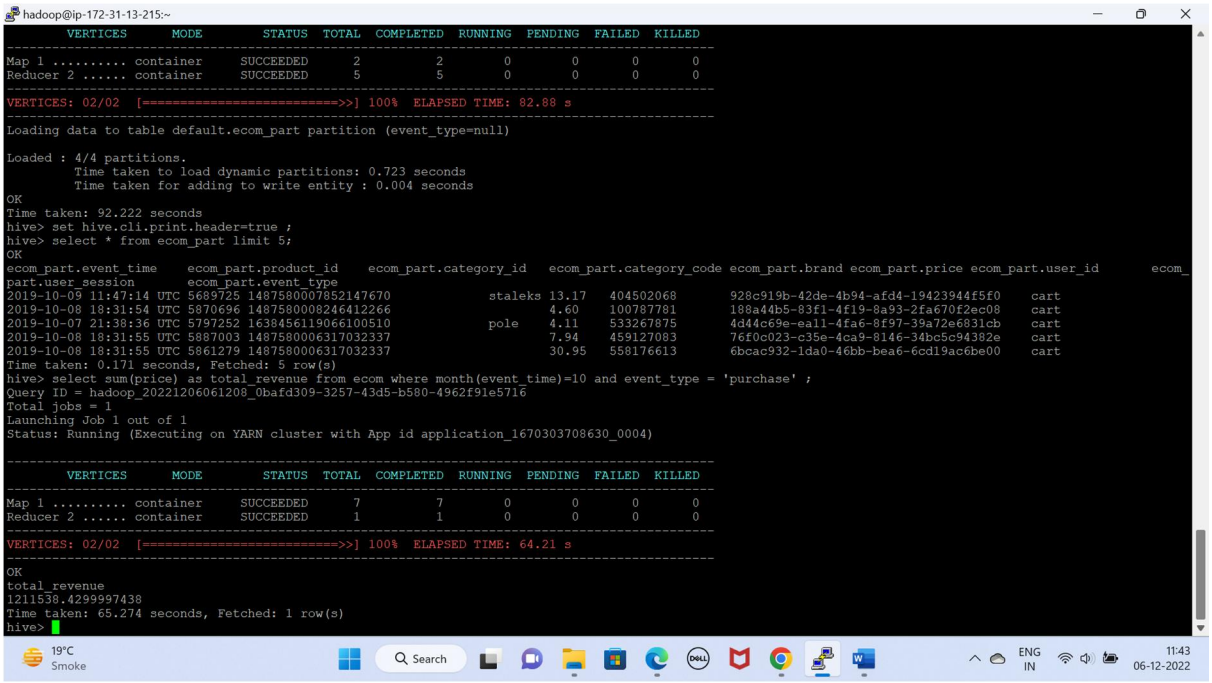

hadoop@ip-172-31-13-215:~
hive> CREATE TABLE IF NOT EXISTS ecom_part (event_time timestamp, product_id string , category_id string , category_code string ,brand string , price float,
user_id bigint , user_session string)
> partitioned by (event_type string)
> clustered by (category_code) into 12 buckets
> 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: 0.059 seconds
hive> Insert into table ecom_part partition (event_type) select event_time , product_id , category_id , category_code , brand , price , user_id , user_sessio
n , event_type from ecom ;
Query ID = hadoop_20221206060533_645b97e0-90a6-436e-bd9f-5d34cb8ef62e
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_1670303708630_0004)
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: 82.88 s
Loading data to table default.ecom_part partition (event_type=null)
Loaded : 4/4 partitions.
Time taken to load dynamic partitions: 0.723 seconds
Time taken for adding to write entity : 0.004 seconds
OK
Time taken: 92.222 seconds
hive> set hive.cli.print.header=true ;
hive> select * from ecom_part limit 5;
OK
ecom_part.event_time ecom_part.product_id ecom_part.category_id ecom_part.category_code ecom_part.brand ecom_part.price ecom_part.user_id ecom_
part.user_session ecom_part.event_type
2019-10-09 11:47:14 UTC 5689725 1487580007852147670 staleks 13.17 404502068 928c919b-42de-4b94-afd4-19423944f5f0 cart
2019-10-08 18:31:54 UTC 5870696 1487580008246412266 4.60 100787781 188a44b5-83f1-4f19-8a93-2fa670f2ec08 cart
2019-10-07 21:38:36 UTC 5797252 1638456119066100510 pole 4.11 533267875 4d44c69e-ea11-4fa6-8f97-39a72e6831cb cart
2019-10-08 18:31:55 UTC 5887003 1487580006317032337 7.94 459127083 76f0c023-c35e-4ca9-8146-34bc5c94382e cart
2019-10-08 18:31:55 UTC 5861279 1487580006317032337 30.95 558176613 6bcac932-1da0-46bb-bea6-6cd19ac6be00 cart
Time taken: 0.171 seconds, Fetched: 5 row(s)
hive>
19°C
Smoke
Search
ENG
IN
11:40
06-12-2022

Analysis of Queries

QUESTION 1

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

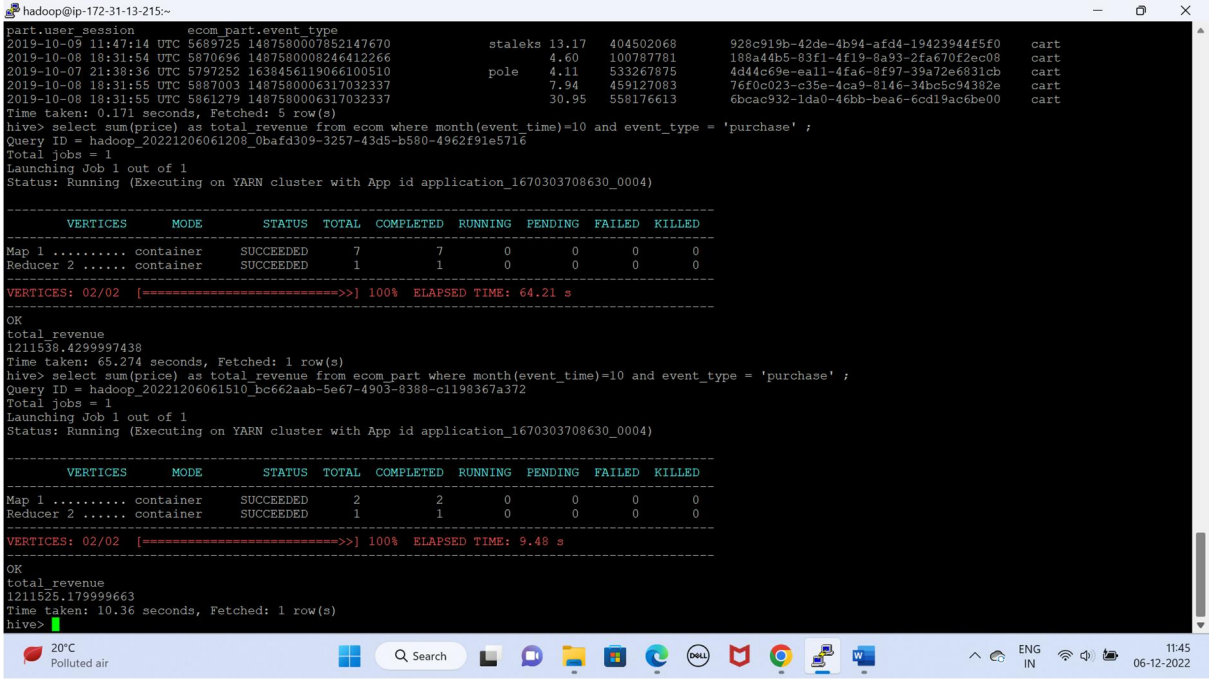
Code: select sum(price) as total_revenue from ecom where month(event_time)=10 and event_type = 'purchase' ;



Note: The below screenshot of the same query from both the base table and the bucketed table. When compared the bucketed table takes less time to query the result than the base table.

Code:

select sum(price) as total_revenue from ecom_part where month(event_time)=10 and event_type = 'purchase' ;



Insights:

Total revenue generated on the basis of Purchase in the month of October 2019 was 1,211,525/- We tried getting the result using both base table[ecom] and optimized table[ecom_part].We figured out that Bucketed table reduces the query time when compared to the base table approximately by 55 seconds.

QUESTION 2:

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

Code:

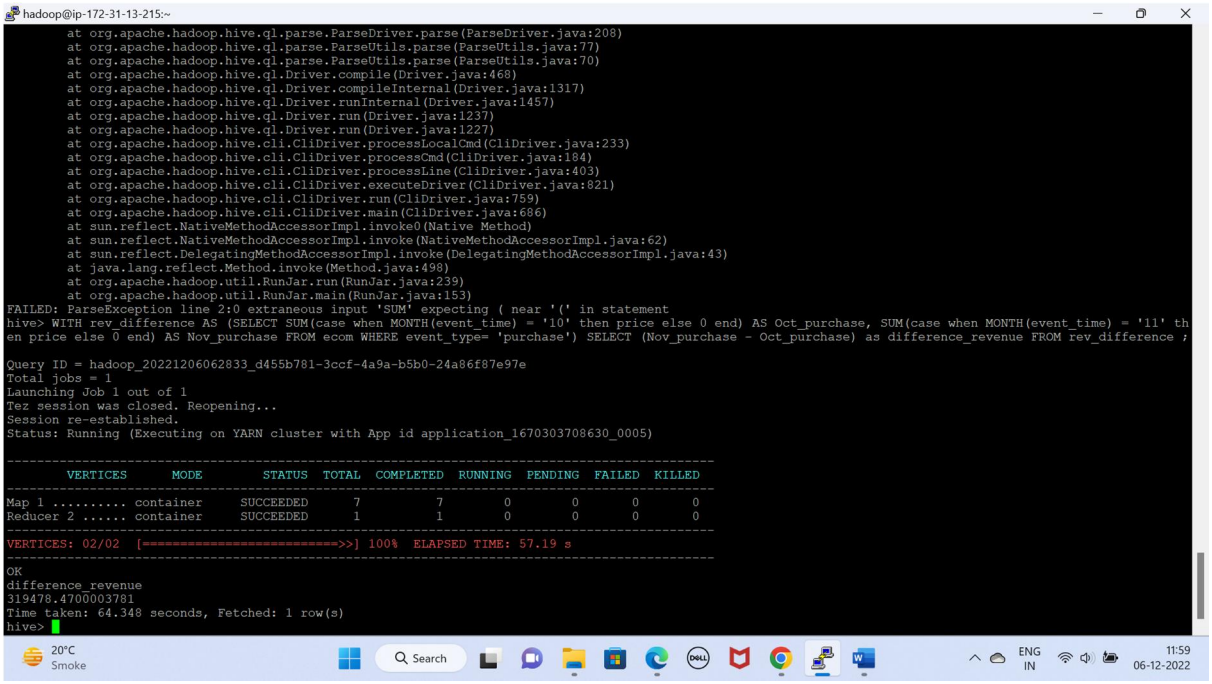
```
SELECT date_format(event_time, 'MM') AS Months, COUNT(event_type) AS Sum_of_Purchases

FROM ecom_data_part WHERE event_type='purchase'

GROUP BY date_format(event_time, 'MM');
```

QUESTION 3:

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

Code:

WITH rev_difference 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 ecom WHERE event_type= 'purchase') SELECT (Nov_purchase - Oct_purchase) as difference_revenue FROM rev_difference ;



Insights:

The difference in revenue between month of October and November

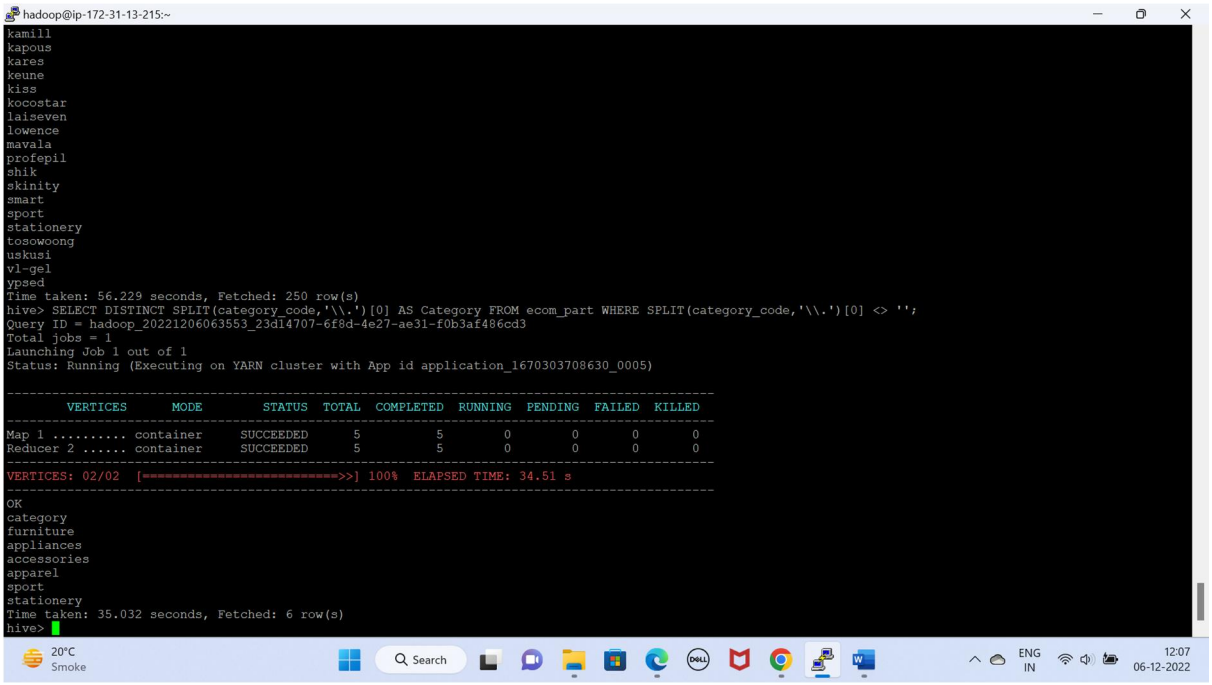
is 319478.4700. So, we conclude that the revenue generated in the month of November was more than the revenue generated in the month of October.

QUESTION 4:

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

Code:

```
SELECT DISTINCT SPLIT(category_code,'\\.')[0] AS Category
FROM ecom_part
WHERE SPLIT(category_code,'\\.')[0] <> '';
```



Insights:

There are six different categories under which company sells it’s

different products and they are furniture, appliances, accessories, apparel, sport and

stationery.

QUESTION 5:

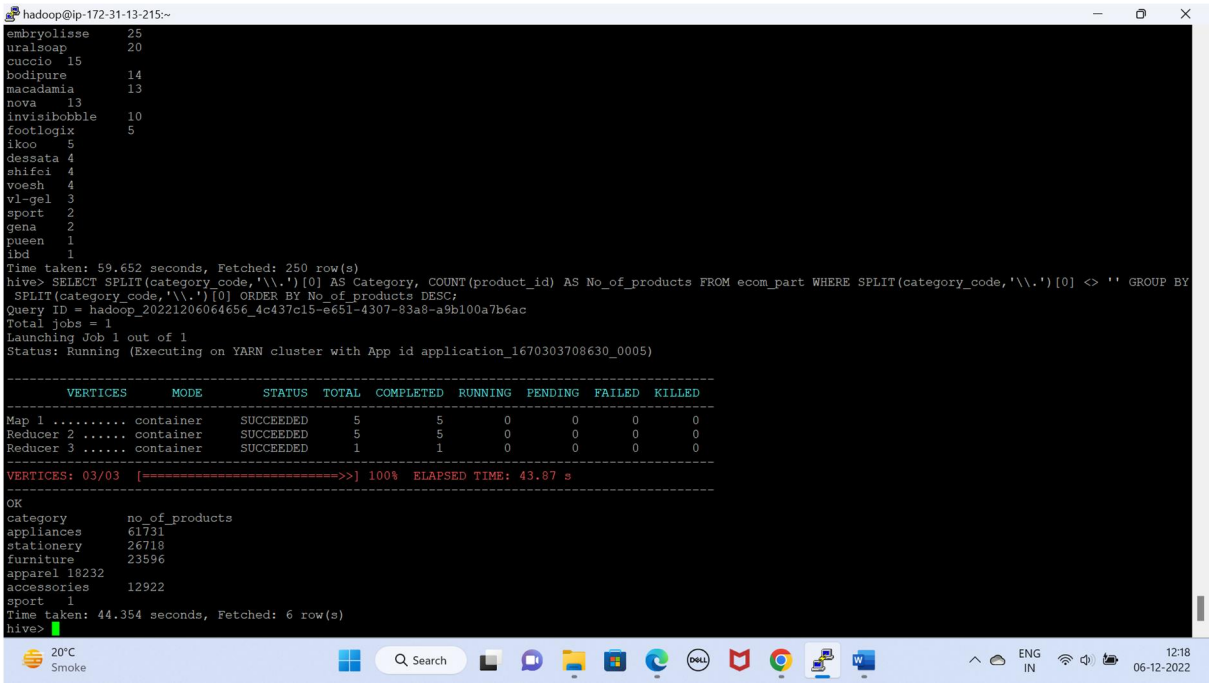
Find the total number of products available under each category.

Code:

```
SELECT SPLIT(category_code,'\\.')[0] AS Category, COUNT(product_id) AS No_of_products
FROM ecom_part
WHERE SPLIT(category_code,'\\.')[0] <> ''
```

GROUP BY SPLIT(category_code,'\\.')[0]

ORDER BY No_of_products DESC;



Insights:

The highest no. of products are available under appliances category which is 61731 and lowest products under sport category which is 1.

QUESTION 6:

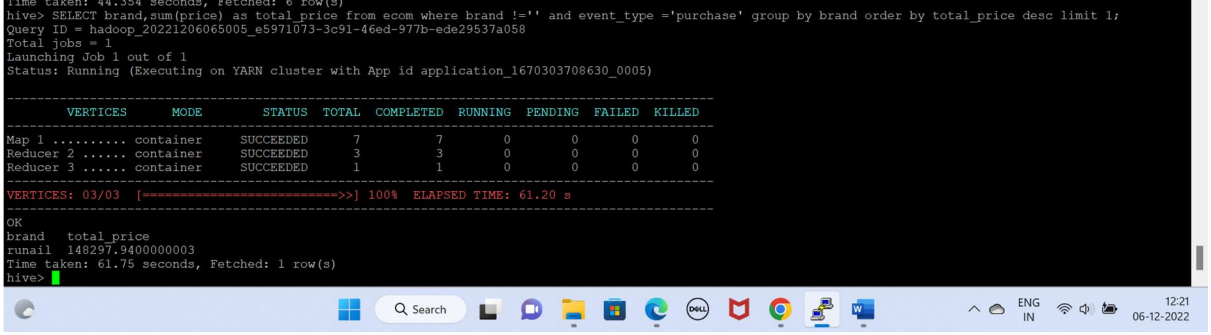
Which brand had the maximum sales in October and November combined?

Code:

SELECT brand,sum(price) as total_price from ecom

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

group by brand order by total_price desc limit 1;



Insights:

Runail is the brand which has maximum sales in the month of October and November of 2019 combined.

QUESTION 7:

Which brands increased their sales from October to November?

Code:

WITH Monthly_Revenue AS (

SELECT brand,

SUM(CASE WHEN date_format(event_time, 'MM')=10 THEN price ELSE 0 END)

AS Oct_Revenue, SUM(CASE WHEN date_format(event_time, 'MM')=11 THEN

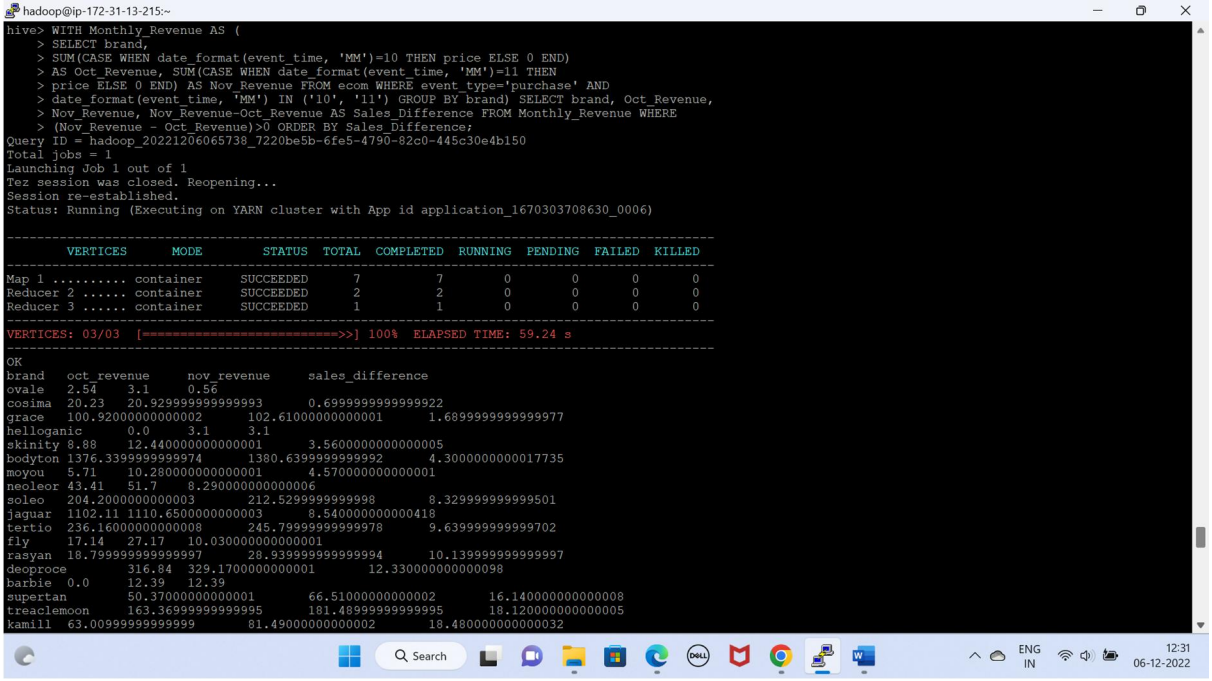
price ELSE 0 END) AS Nov_Revenue FROM ecom WHERE event_type='purchase' AND date_format(event_time, 'MM') IN ('10', '11') GROUP BY brand)

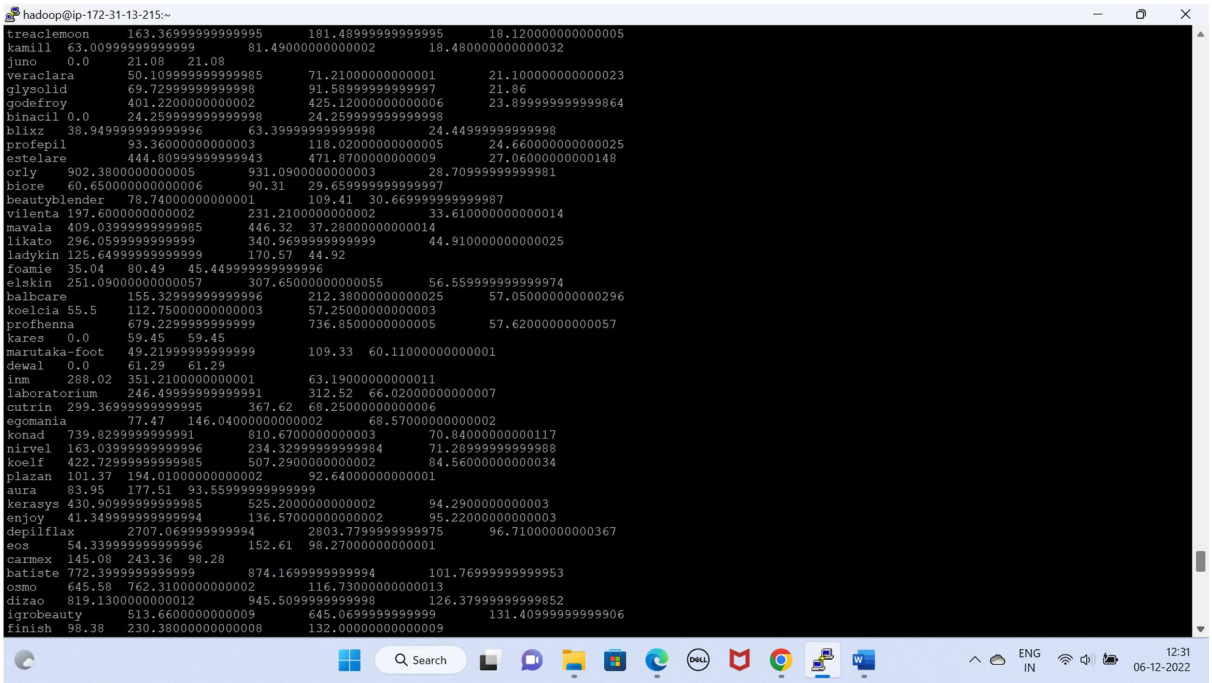
SELECT brand, Oct_Revenue, Nov_Revenue, Nov_Revenue-Oct_Revenue AS

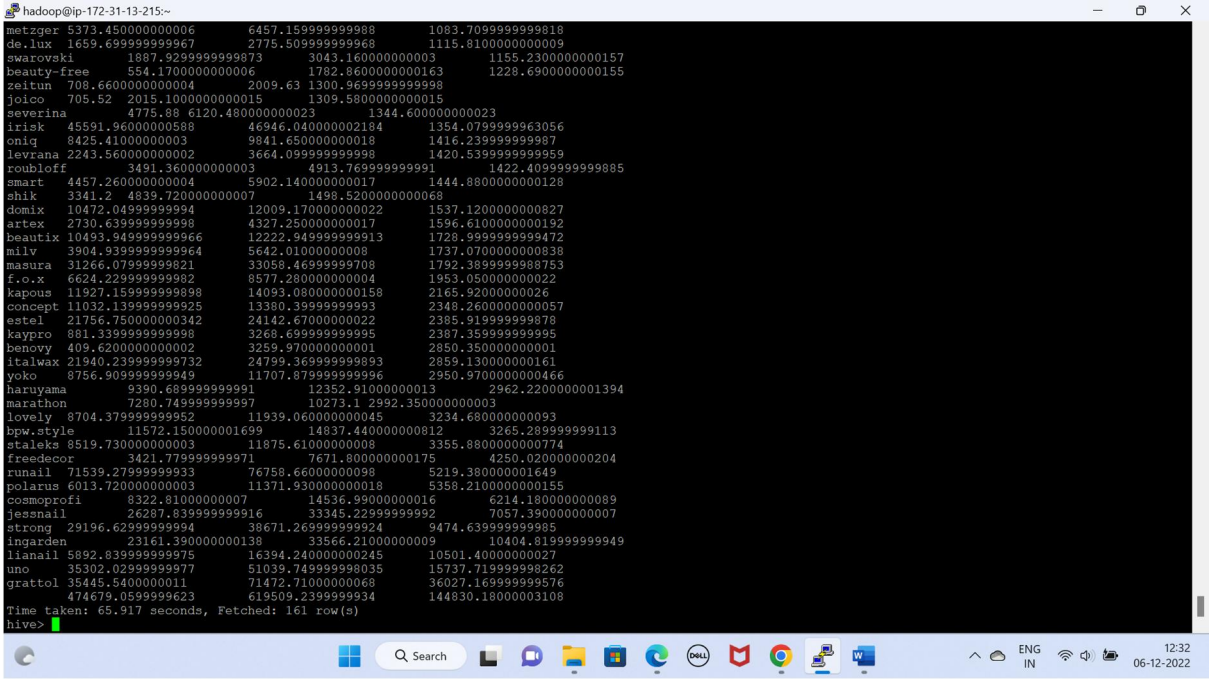
Sales_Difference FROM Monthly_Revenue

WHERE (Nov_Revenue - Oct_Revenue)>0

ORDER BY Sales_Difference;


hadoop@ip-172-31-13-215:~
hive> WITH Monthly_Revenue AS (
> SELECT brand,
> SUM(CASE WHEN date_format(event_time, 'MM')=10 THEN price ELSE 0 END)
> AS Oct_Revenue, SUM(CASE WHEN date_format(event_time, 'MM')=11 THEN
> price ELSE 0 END) AS Nov_Revenue FROM ecom WHERE event_type='purchase' AND
> date_format(event_time, 'MM') IN ('10', '11') GROUP BY brand) SELECT brand, Oct_Revenue,
> Nov_Revenue, Nov_Revenue-Oct_Revenue AS Sales_Difference FROM Monthly_Revenue WHERE
> (Nov_Revenue - Oct_Revenue)>0 ORDER BY Sales_Difference;
Query ID = hadoop_20221206065738_7220be5b-6fe5-4790-82c0-445c30e4b150
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_1670303708630_0006)
VERTICES MODE STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED
Map 1 container SUCCEEDED 7 7 0 0 0 0
Reducer 2 container SUCCEEDED 2 2 0 0 0 0
Reducer 3 container SUCCEEDED 1 1 0 0 0 0
VERTICES: 03/03 [==>>] 100% ELAPSED TIME: 59.24 s
OK
brand oct_revenue nov_revenue sales_difference
ovale 2.54 3.1 0.56
cosima 20.23 20.929999999999993 0.6999999999999922
grace 100.92000000000002 102.61000000000001 1.6899999999999977
helloganic 0.0 3.1 3.1
skinity 8.88 12.440000000000001 3.5600000000000005
bodyton 1376.3399999999974 1380.6399999999992 4.3000000000017735
moyou 5.71 10.280000000000001 4.570000000000001
neoleor 43.41 51.7 8.290000000000006
soleo 204.2000000000003 212.5299999999998 8.329999999999501
jaguar 1102.11 1110.6500000000003 8.540000000000418
tertio 236.16000000000008 245.79999999999978 9.639999999999702
fly 17.14 27.17 10.030000000000001
rasyan 18.799999999999997 28.939999999999994 10.139999999999997
deoproce 316.84 329.1700000000001 12.330000000000098
barbie 0.0 12.39 12.39
supertan 50.37000000000001 66.51000000000002 16.140000000000008
treaclemoon 163.36999999999995 181.48999999999995 18.120000000000005
kamill 63.00999999999999 81.49000000000002 18.480000000000032
ENG IN
12:31
06-12-2022


hadoop@ip-172-31-13-215:~
treaclemoon 163.36999999999995 181.48999999999995 18.120000000000005
kamill 63.00999999999999 81.49000000000002 18.480000000000032
juno 0.0 21.08 21.08
veraclara 50.109999999999985 71.21000000000001 21.100000000000023
glysolid 69.72999999999998 91.58999999999997 21.86
godefroy 401.2200000000002 425.12000000000006 23.899999999999864
binacil 0.0 24.259999999999998 24.259999999999998
blixz 38.949999999999996 63.39999999999998 24.44999999999998
profepil 93.36000000000003 118.02000000000005 24.660000000000025
estelare 444.80999999999943 471.8700000000009 27.06000000000148
orly 902.3800000000005 931.0900000000003 28.70999999999981
biore 60.650000000000006 90.31 29.659999999999997
beautyblender 78.74000000000001 109.41 30.669999999999987
vilenta 197.6000000000002 231.2100000000002 33.610000000000014
mavala 409.0399999999985 446.32 37.28000000000014
likato 296.0599999999999 340.9699999999999 44.910000000000025
ladykin 125.64999999999999 170.57 44.92
foamie 35.04 80.49 45.449999999999996
elskin 251.09000000000057 307.65000000000055 56.559999999999974
balbcare 155.32999999999996 212.38000000000025 57.050000000000296
koelcia 55.5 112.75000000000003 57.25000000000003
profhenna 679.2299999999999 736.8500000000005 57.62000000000057
kares 0.0 59.45 59.45
marutaka-foot 49.21999999999999 109.33 60.11000000000001
dewal 0.0 61.29 61.29
inm 288.02 351.2100000000001 63.19000000000011
laboratorium 246.49999999999991 312.52 66.02000000000007
cutrin 299.36999999999995 367.62 68.25000000000006
egomania 77.47 146.04000000000002 68.57000000000002
konad 739.8299999999991 810.6700000000003 70.84000000000117
nirvel 163.03999999999996 234.32999999999984 71.28999999999988
koelf 422.72999999999985 507.2900000000002 84.56000000000034
plazan 101.37 194.01000000000002 92.64000000000001
aura 83.95 177.51 93.55999999999999
kerasys 430.90999999999985 525.2000000000002 94.29000000000003
enjoy 41.349999999999994 136.57000000000002 95.22000000000003
depilflax 2707.069999999994 2803.7799999999975 96.71000000000367
eos 54.339999999999996 152.61 98.27000000000001
carmex 145.08 243.36 98.28
batiste 772.3999999999999 874.1699999999994 101.76999999999953
osmo 645.58 762.3100000000002 116.73000000000013
dizao 819.1300000000012 945.5099999999998 126.37999999999852
igrobeauty 513.6600000000009 645.0699999999999 131.40999999999906
finish 98.38 230.38000000000008 132.00000000000009
ENG IN
12:31
06-12-2022



Insights:

'Grattol' brand has maximum increment of 36,027.16 /- and 'Ovale' has least increment of 0.56 /- from October to November.

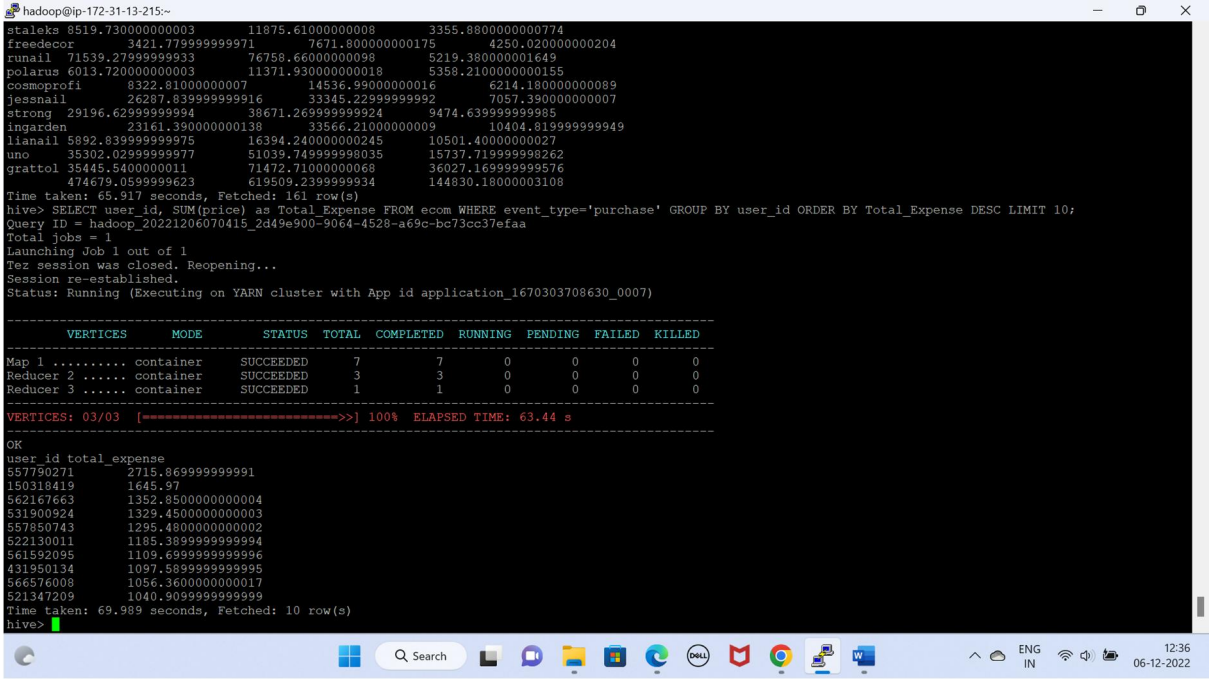
QUESTION 8:

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

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

Code:

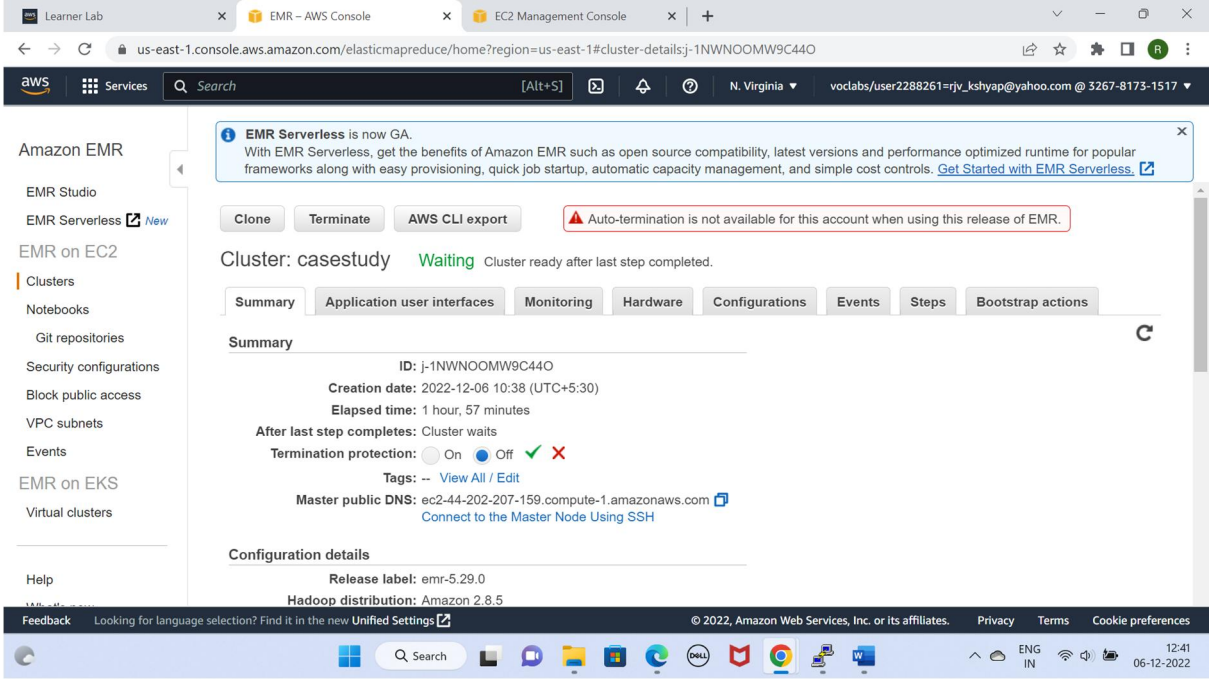
```
SELECT user_id, SUM(price) as Total_Expense

FROM ecom

WHERE event_type='purchase'

GROUP BY user_id

ORDER BY Total_Expense DESC LIMIT 10;
```

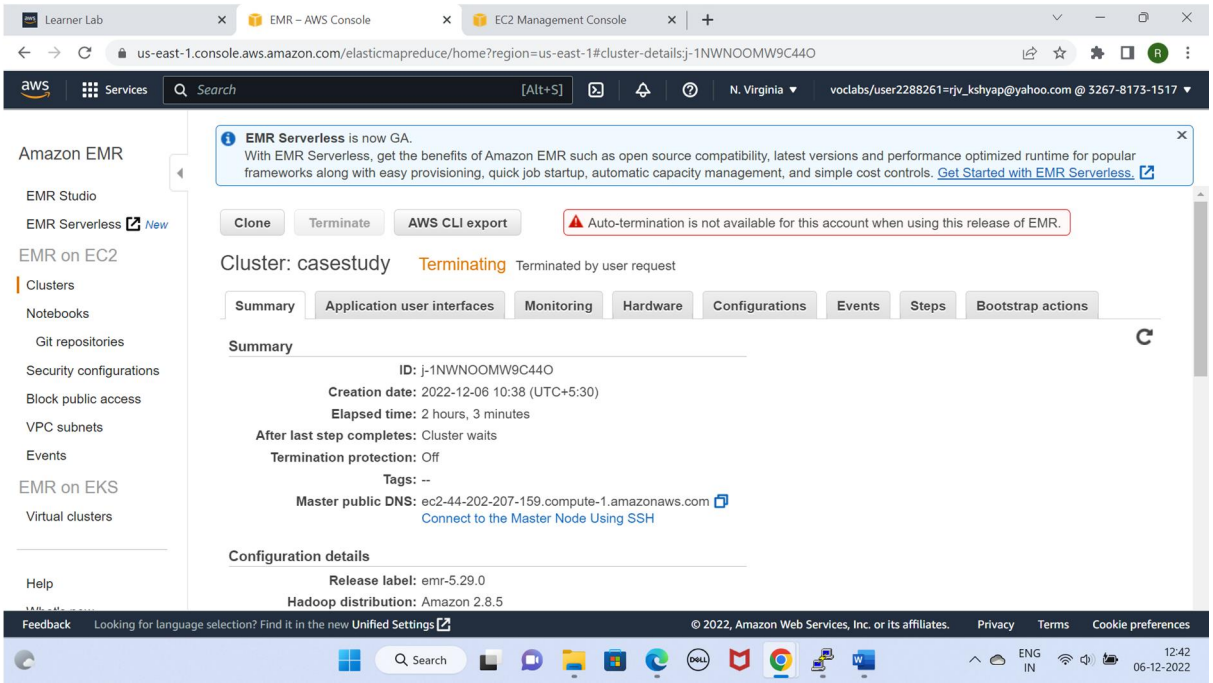


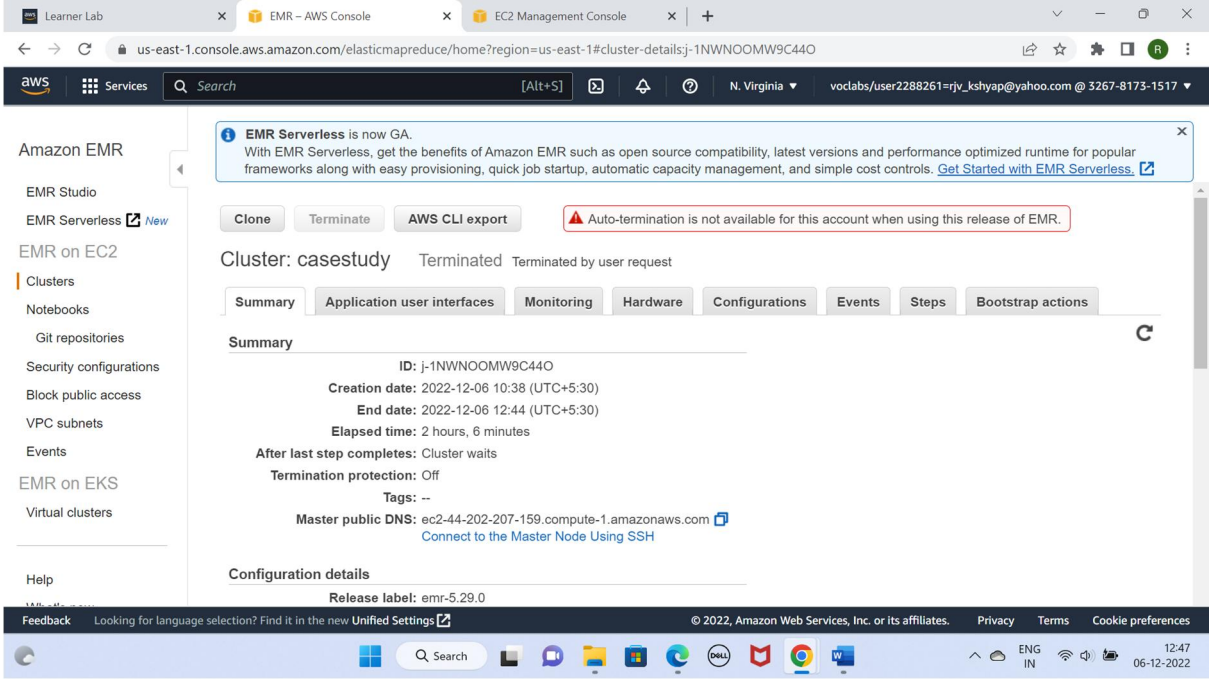
Insight:

Above is the list of top 10 users to which company can reward golden customer plan.

Once the analysis is done, we can terminate the cluster by changing the Termination protection from ON to OFF and then click on the terminate button.


Learner Lab
EMR – AWS Console
EC2 Management Console
us-east-1.console.aws.amazon.com/elasticmapreduce/home?region=us-east-1#cluster-details:j-1NWNOOMW9C44O
Services
Search
[Alt+S]
N. Virginia
voclabs/user2288261=rjv_kshyap@yahoo.com @ 3267-8173-1517
Amazon EMR
EMR Studio
EMR Serverless New
EMR on EC2
Clusters
Notebooks
Git repositories
Security configurations
Block public access
VPC subnets
Events
EMR on EKS
Virtual clusters
Help
EMR Serverless is now GA.
With EMR Serverless, get the benefits of Amazon EMR such as open source compatibility, latest versions and performance optimized runtime for popular frameworks along with easy provisioning, quick job startup, automatic capacity management, and simple cost controls. Get Started with EMR Serverless.
Clone
Terminate
AWS CLI export
Auto-termination is not available for this account when using this release of EMR.
Cluster: casestudy
Waiting
Cluster ready after last step completed.
Summary
Application user interfaces
Monitoring
Hardware
Configurations
Events
Steps
Bootstrap actions
Summary
ID: j-1NWNOOMW9C44O
Creation date: 2022-12-06 10:38 (UTC+5:30)
Elapsed time: 1 hour, 57 minutes
After last step completes: Cluster waits
Termination protection: On Off
Tags: -- View All / Edit
Master public DNS: ec2-44-202-207-159.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
Feedback
Looking for language selection? Find it in the new Unified Settings
© 2022, Amazon Web Services, Inc. or its affiliates.
Privacy
Terms
Cookie preferences
ENG
IN
12:41
06-12-2022


Learner Lab
EMR – AWS Console
EC2 Management Console
us-east-1.console.aws.amazon.com/elasticmapreduce/home?region=us-east-1#cluster-details:j-1NWNOOMW9C44O
Services
Search
[Alt+S]
N. Virginia
voclabs/user2288261=rjv_kshyap@yahoo.com @ 3267-8173-1517
Amazon EMR
EMR Studio
EMR Serverless New
EMR on EC2
Clusters
Notebooks
Git repositories
Security configurations
Block public access
VPC subnets
Events
EMR on EKS
Virtual clusters
Help
EMR Serverless is now GA.
With EMR Serverless, get the benefits of Amazon EMR such as open source compatibility, latest versions and performance optimized runtime for popular frameworks along with easy provisioning, quick job startup, automatic capacity management, and simple cost controls. Get Started with EMR Serverless.
Clone
Terminate
AWS CLI export
Auto-termination is not available for this account when using this release of EMR.
Cluster: casestudy
Terminating
Terminated by user request
Summary
Application user interfaces
Monitoring
Hardware
Configurations
Events
Steps
Bootstrap actions
Summary
ID: j-1NWNOOMW9C44O
Creation date: 2022-12-06 10:38 (UTC+5:30)
Elapsed time: 2 hours, 3 minutes
After last step completes: Cluster waits
Termination protection: Off
Tags: --
Master public DNS: ec2-44-202-207-159.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
Feedback
Looking for language selection? Find it in the new Unified Settings
© 2022, Amazon Web Services, Inc. or its affiliates.
Privacy
Terms
Cookie preferences
ENG
IN
12:42
06-12-2022



SUMMARY

For this case study we worked with public clickstream dataset of a cosmetics store.

The complete steps are given below:

1.Copying the data set into the HDFS from S3 Bucket.

2.Creating the database and launching Hive queries on EMR cluster.

[use CSVSerde with the default properties value for loading the dataset

into Hive table]

3.Cleaning up.