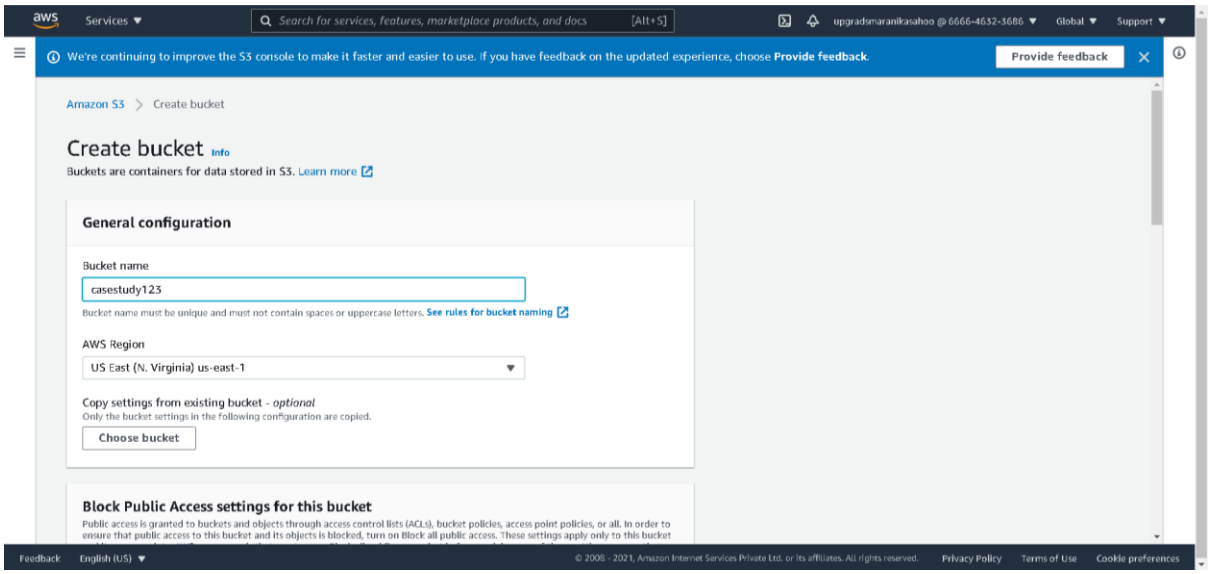
Hive case study- DA track

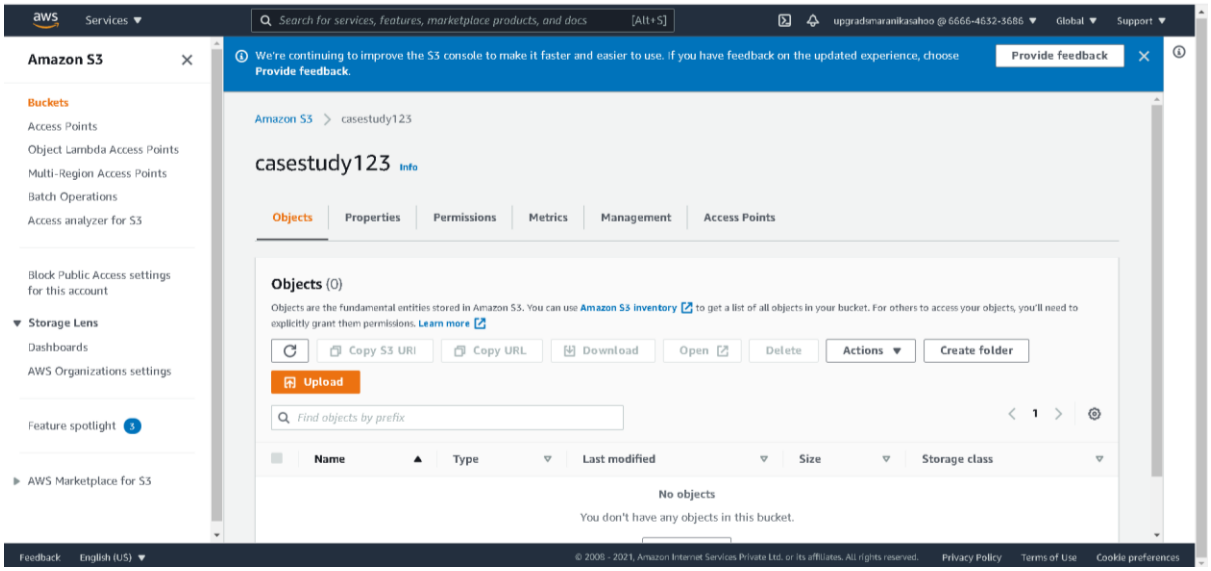
Submission By : Nitish Rathore & Smaranika Sahoo

S3 BUCKET:

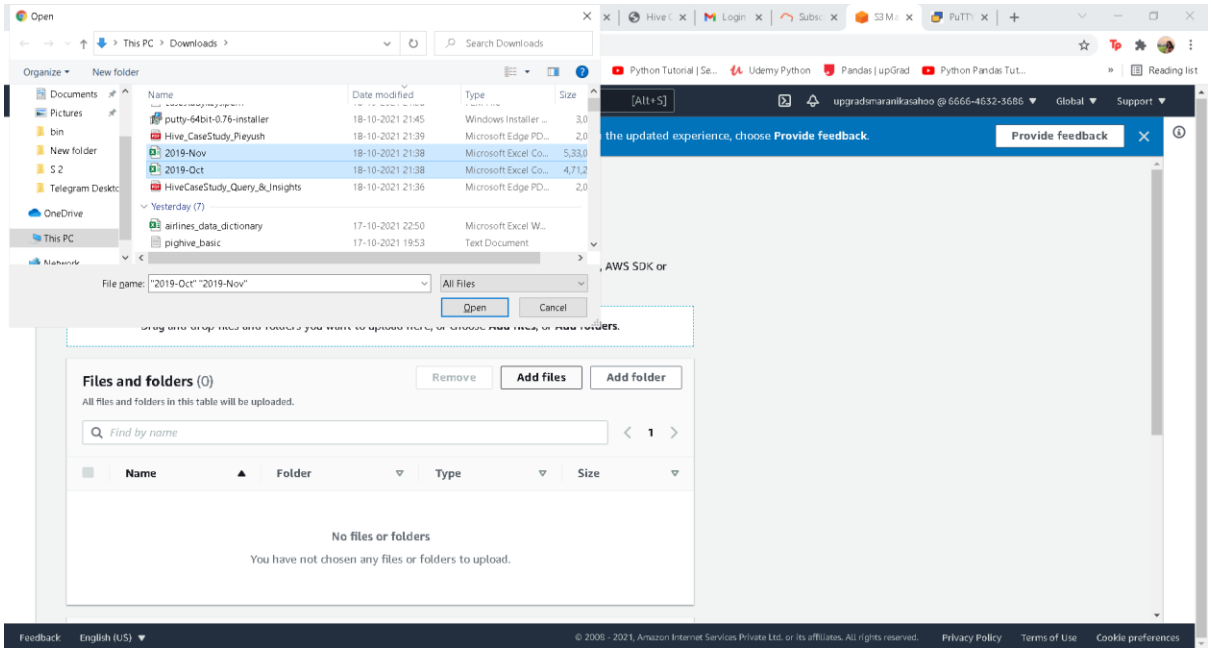
- To Store the data – Click on “Create Bucket”.

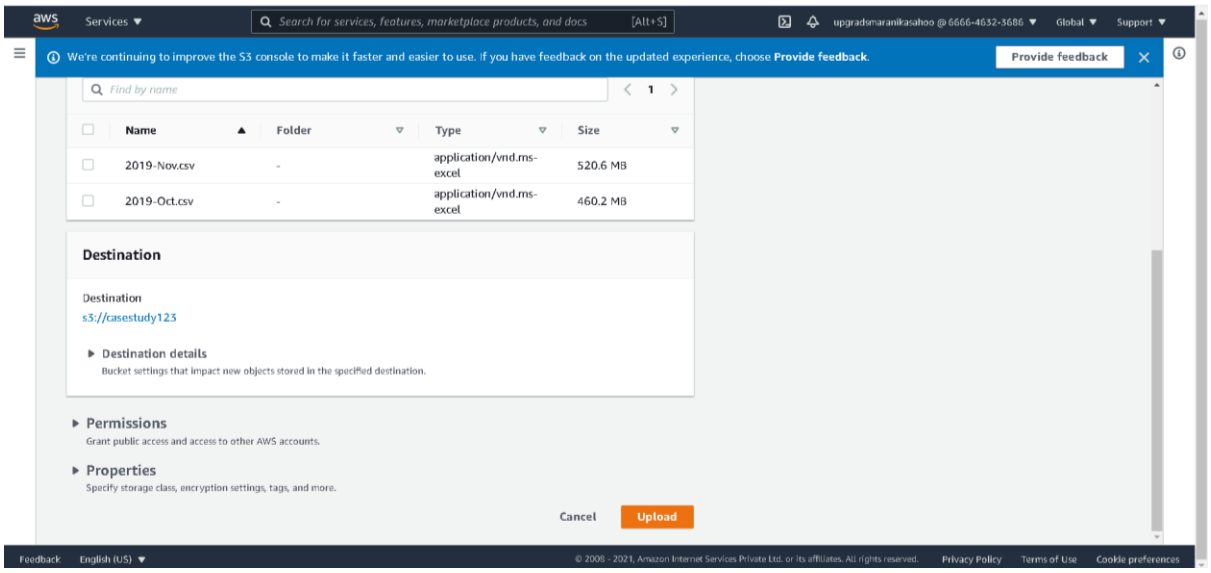


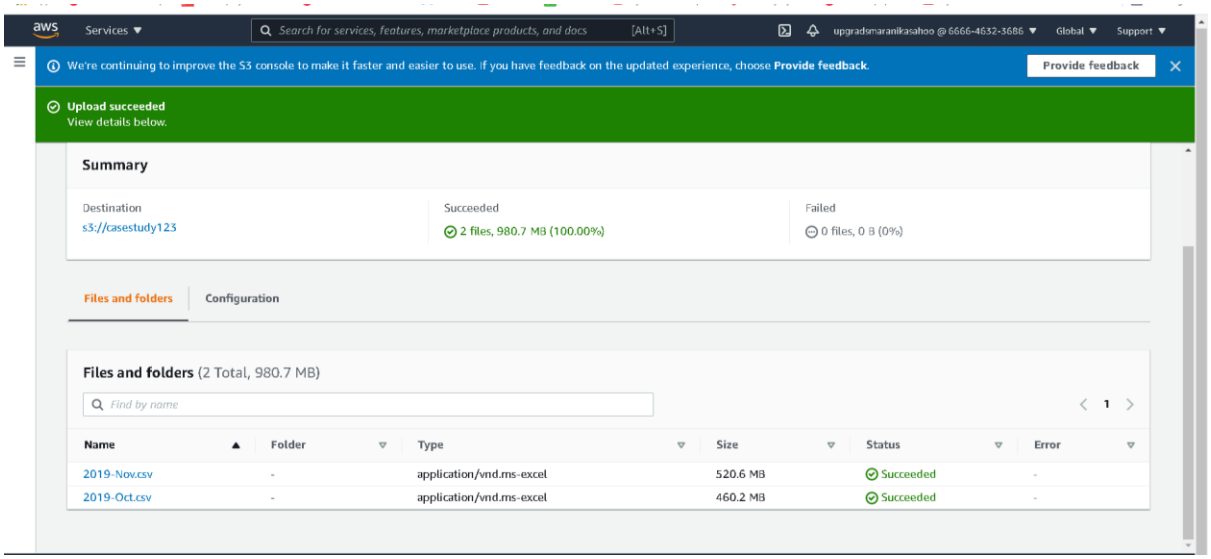
- Creating “casestudy123” with all default options.



- Bucket Successfully got created.


Open
This PC > Downloads >
Search Downloads
Organize
New folder
Documents
Pictures
bin
New folder
S 2
Telegram Deskto
OneDrive
This PC
Name
Date modified
Type
Size
putty-64bit-0.76-installer
18-10-2021 21:45
Windows Installer ...
Hive_CaseStudy_Piyush
18-10-2021 21:39
Microsoft Edge PD...
2019-Nov
18-10-2021 21:38
Microsoft Excel Co...
2019-Oct
18-10-2021 21:38
Microsoft Excel Co...
HiveCaseStudy_Query_&_Insights
18-10-2021 21:36
Microsoft Edge PD...
Yesterday (7)
airlines_data_dictionary
17-10-2021 22:50
Microsoft Excel W...
pighive_basic
17-10-2021 19:53
Text Document
File name: "2019-Oct" "2019-Nov"
All Files
Open
Cancel
Python Tutorial | Se...
Udemy Python
Pandas | upGrad
Python Pandas Tut...
Reading list
[Alt+S]
upgradsmaranikasahoo @ 6666-4632-3686
Global
Support
the updated experience, choose Provide feedback.
Provide feedback
, AWS SDK or
Files and folders (0)
All files and folders in this table will be uploaded.
Remove
Add files
Add folder
Find by name
1
Name
Folder
Type
Size
No files or folders
You have not chosen any files or folders to upload.
Feedback
English (US)
© 2008 - 2021, Amazon Internet Services Private Ltd. or its affiliates. All rights reserved.
Privacy Policy
Terms of Use
Cookie preferences

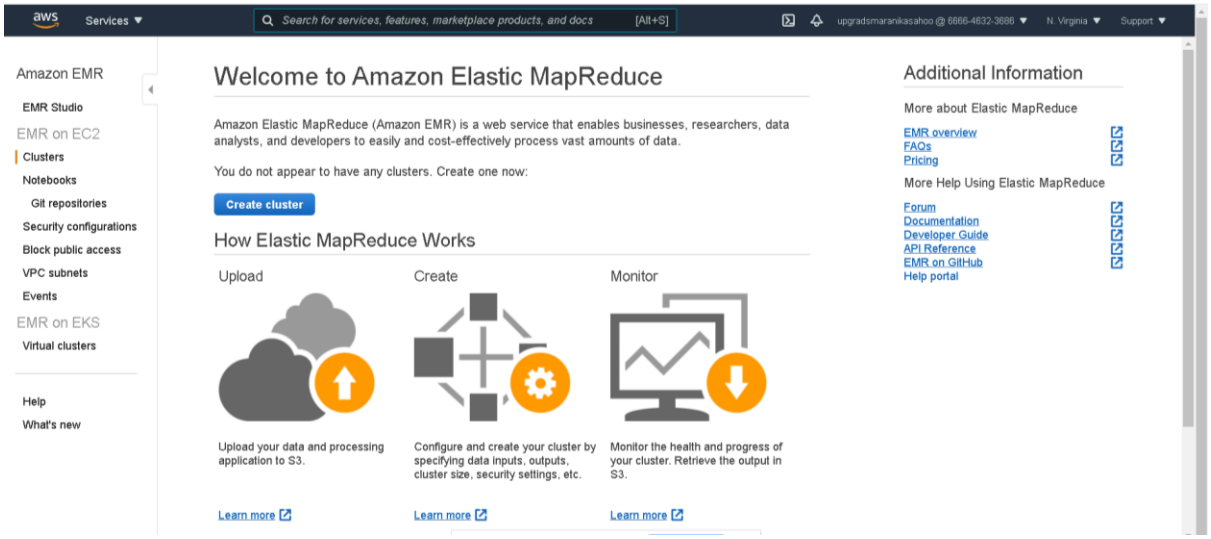

aws
Services
Search for services, features, marketplace products, and docs
[Alt+S]
upgradsmaranikasahoo @ 6666-4632-3686
Global
Support
We're continuing to improve the S3 console to make it faster and easier to use. If you have feedback on the updated experience, choose Provide feedback.
Provide feedback
Find by name
1
Name
Folder
Type
Size
2019-Nov.csv
-
application/vnd.ms-excel
520.6 MB
2019-Oct.csv
-
application/vnd.ms-excel
460.2 MB
Destination
Destination
s3://casestudy123
Destination details
Bucket settings that impact new objects stored in the specified destination.
Permissions
Grant public access and access to other AWS accounts.
Properties
Specify storage class, encryption settings, tags, and more.
Cancel
Upload
Feedback
English (US)
© 2008 - 2021, Amazon Internet Services Private Ltd. or its affiliates. All rights reserved.
Privacy Policy
Terms of Use
Cookie preferences



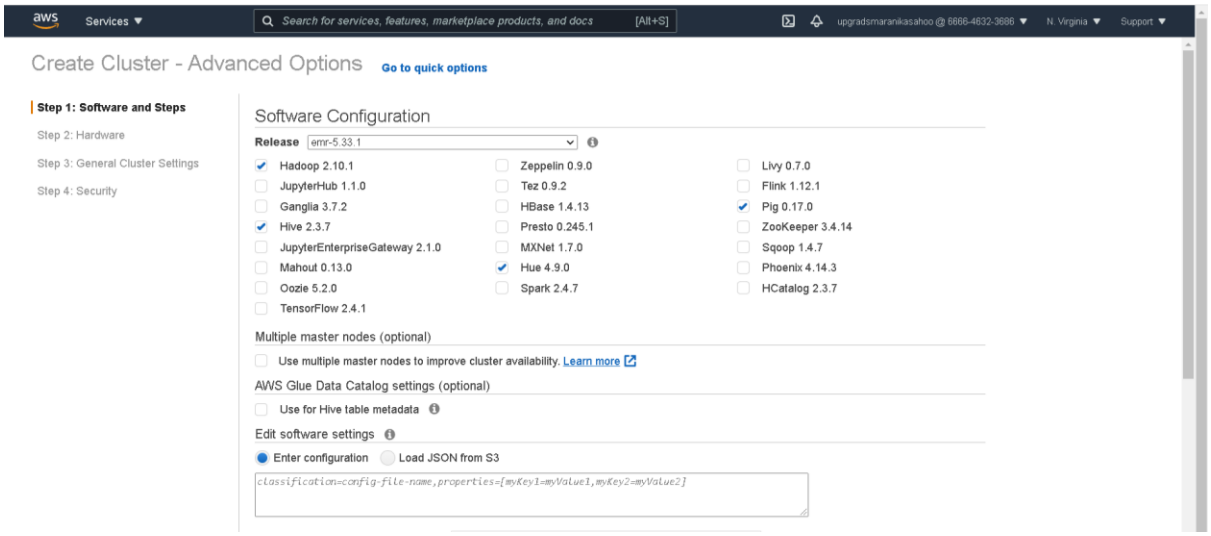
- Successfully uploaded the 2019 October and 2019 November csv file to S3 bucket.

EMR CLUSTER CREATION:

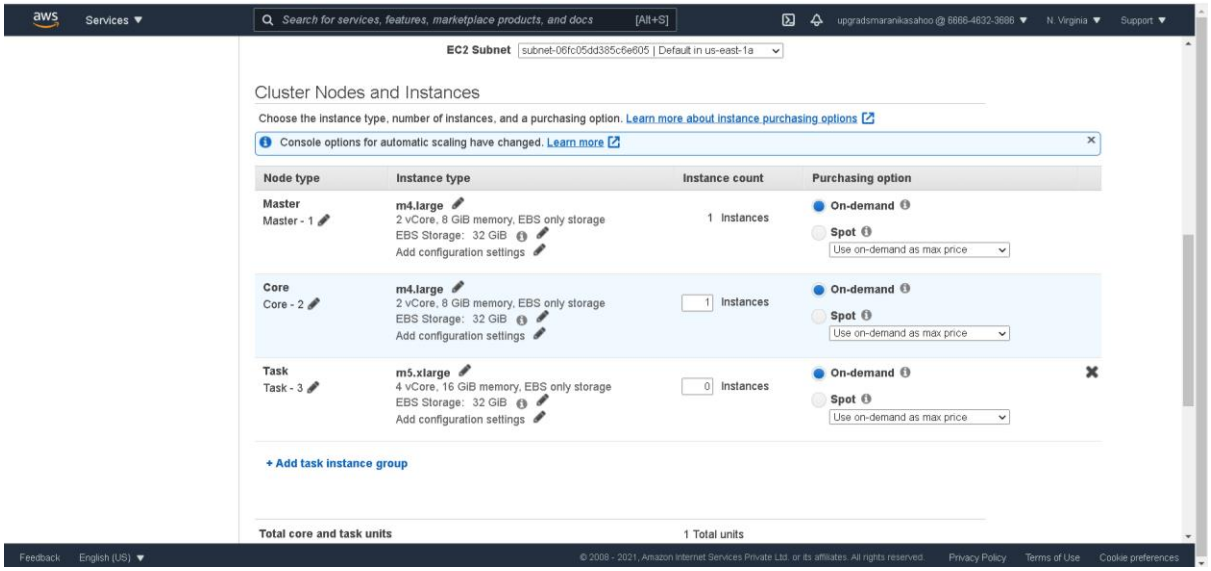
Click on “Create cluster” button to create the EMR cluster.

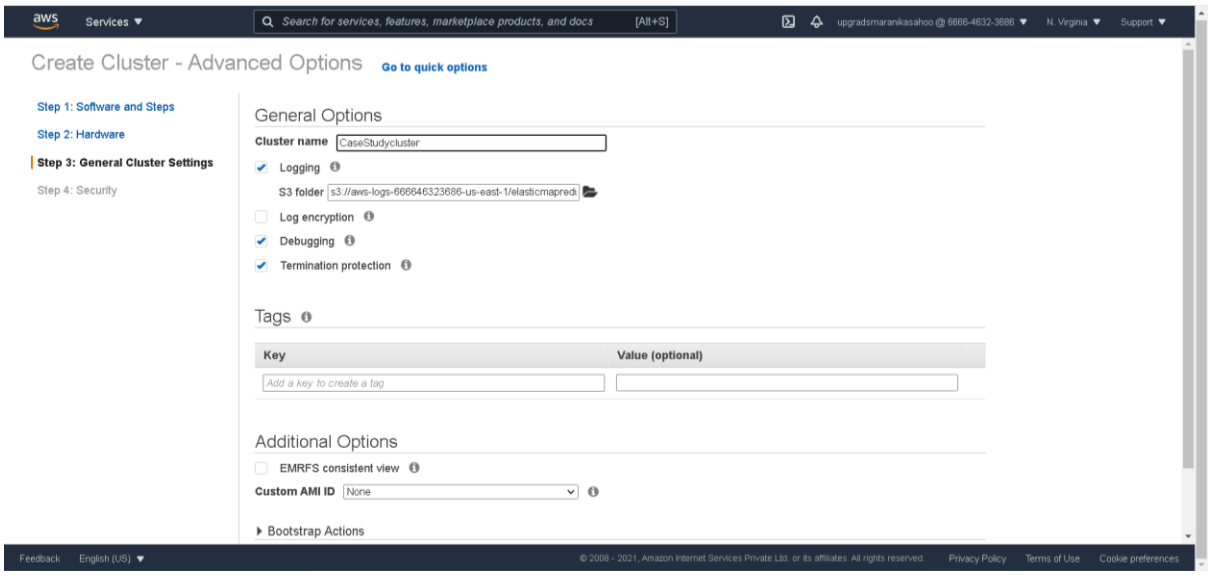


- Creating cluster with advanced options.

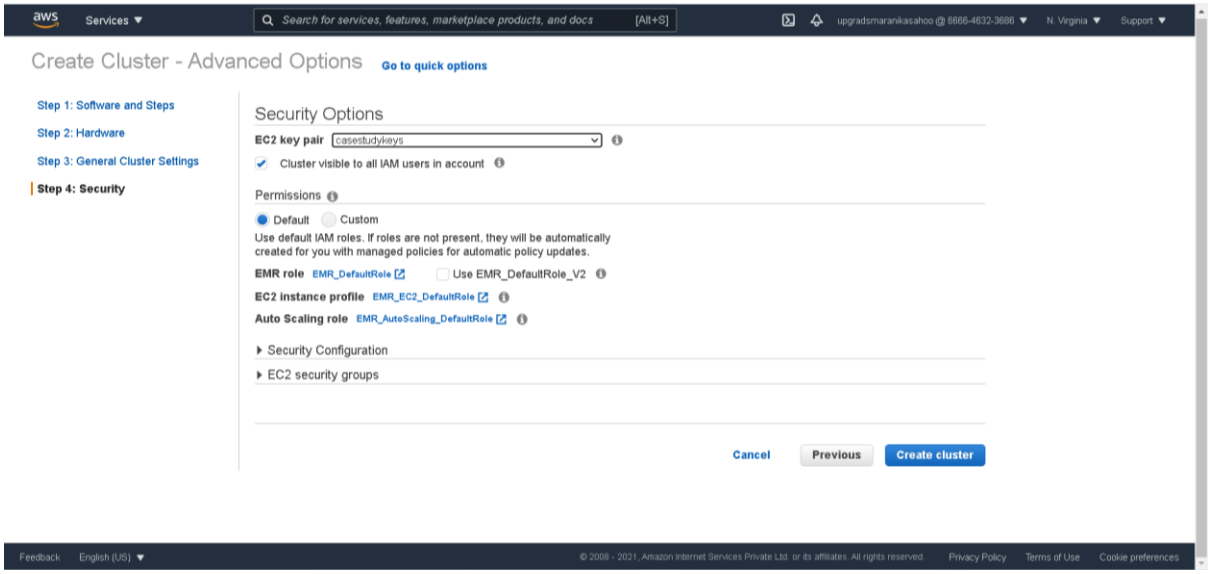


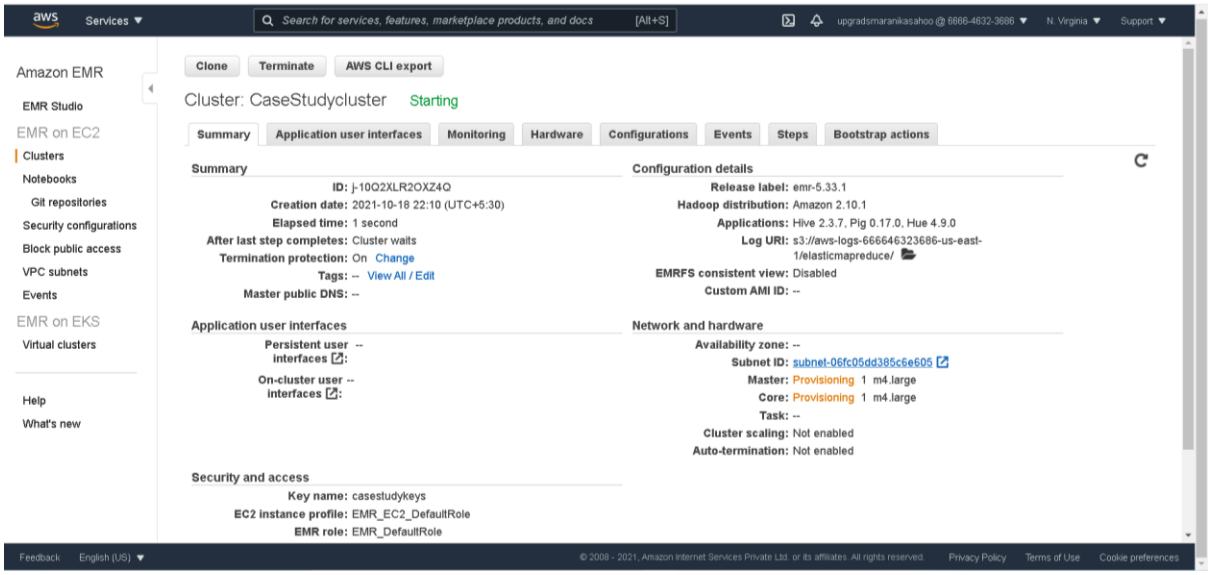
- Hardware page: Changing the Master and Core nodes from m5.xlarge to “m4.large” with 1 instance each to make 2-node EMR cluster.

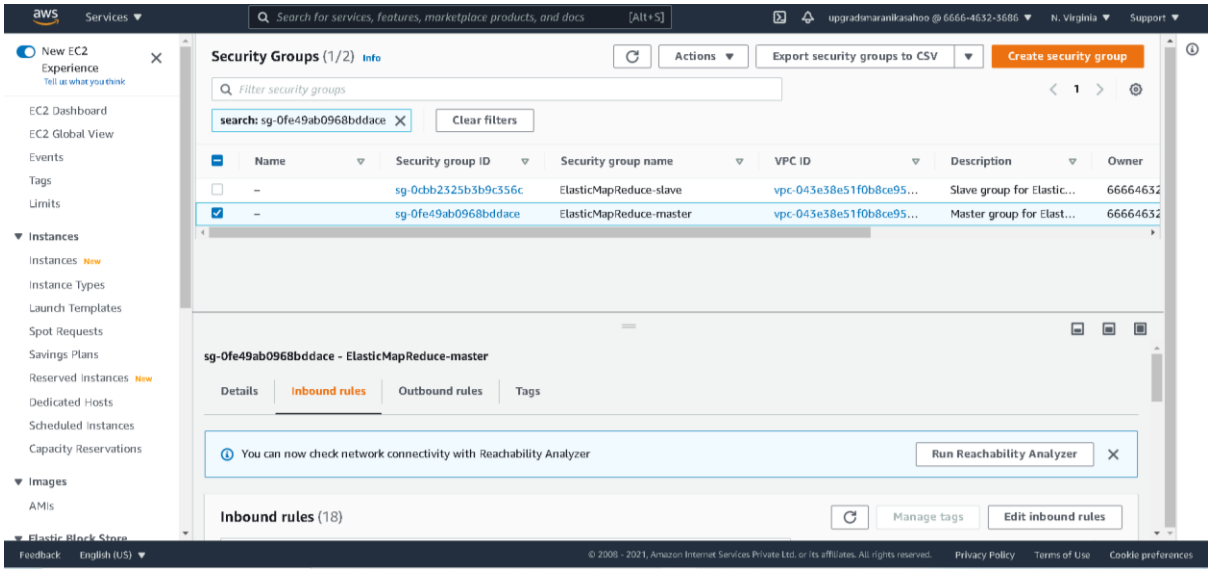


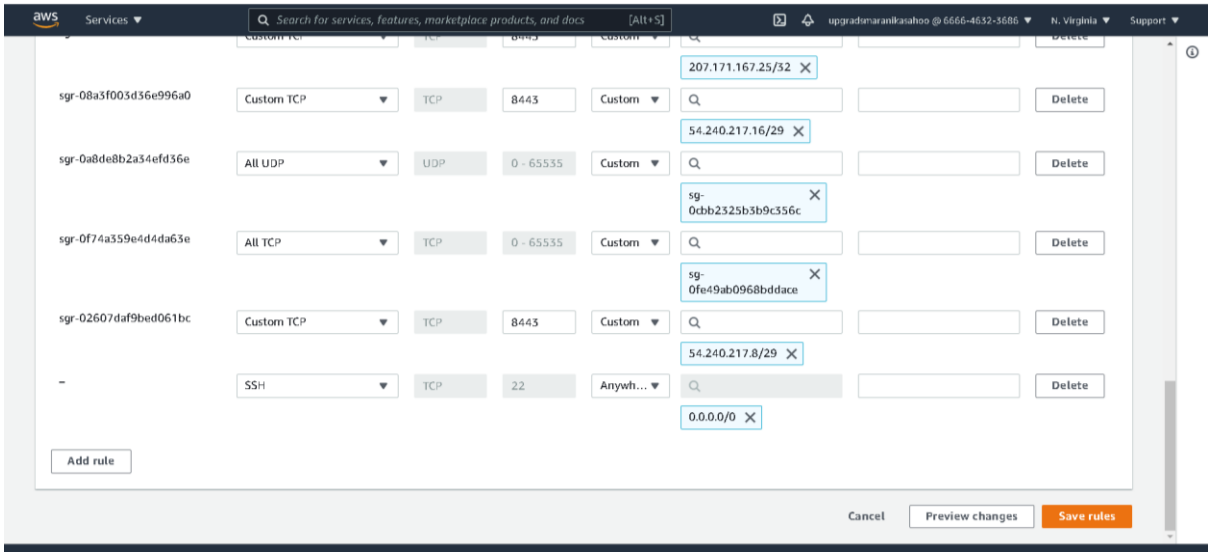


- Security page: changing the EC2 key pair option to our created key pair – “CaseStudykeys”.

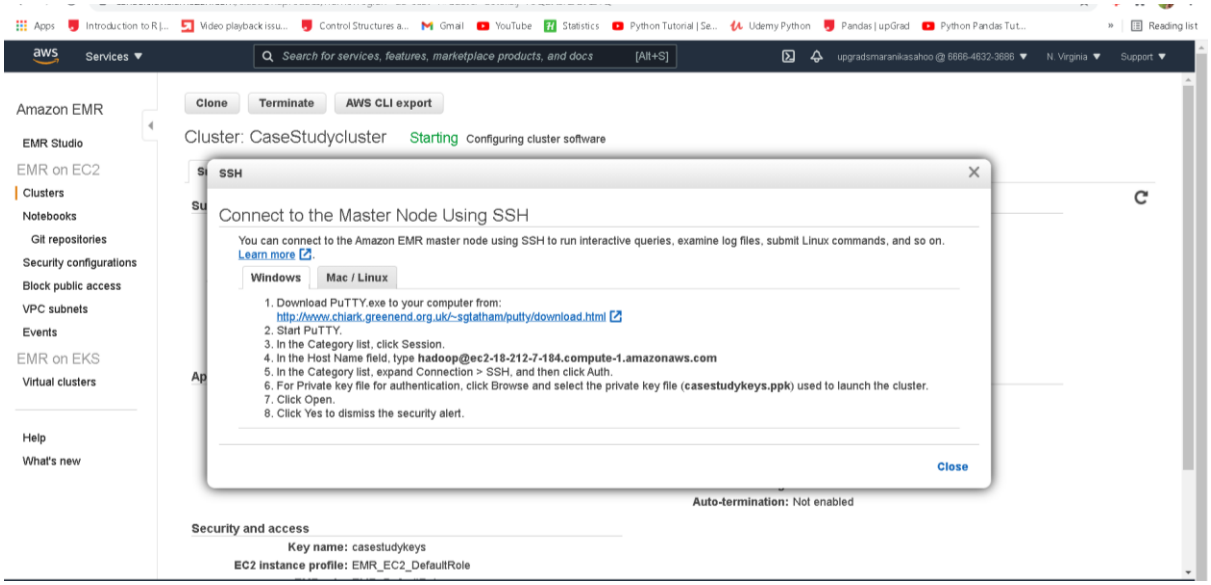


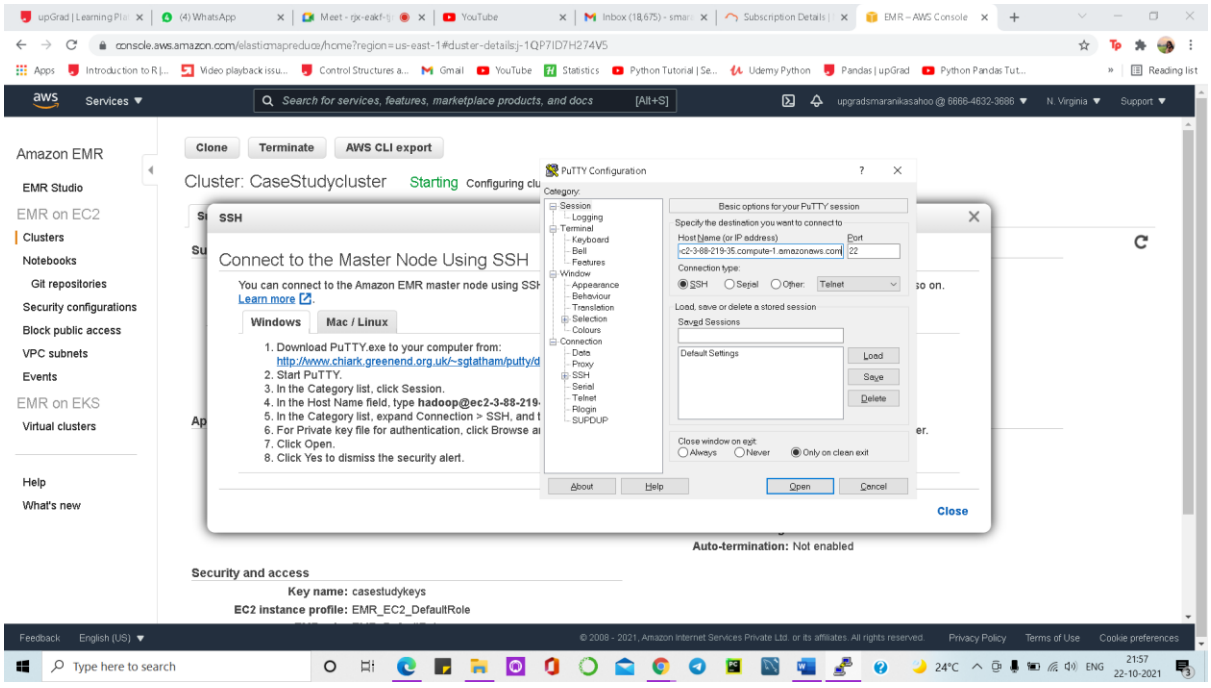

Amazon EMR
EMR Studio
EMR on EC2
Clusters
Notebooks
Git repositories
Security configurations
Block public access
VPC subnets
Events
EMR on EKS
Virtual clusters
Help
What's new
Clone
Terminate
AWS CLI export
Cluster: CaseStudycluster
Starting
Summary
Application user interfaces
Monitoring
Hardware
Configurations
Events
Steps
Bootstrap actions
Summary
ID: j-10Q2XLR2OXZ4Q
Creation date: 2021-10-18 22:10 (UTC+5:30)
Elapsed time: 1 second
After last step completes: Cluster waits
Termination protection: On Change
Tags: -- View All / Edit
Master public DNS: --
Configuration details
Release label: emr-5.33.1
Hadoop distribution: Amazon 2.10.1
Applications: Hive 2.3.7, Pig 0.17.0, Hue 4.9.0
Log URI: s3://aws-logs-666646323686-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: --
Subnet ID: subnet-06fc05dd385c6e605
Master: Provisioning 1 m4.large
Core: Provisioning 1 m4.large
Task: --
Cluster scaling: Not enabled
Auto-termination: Not enabled
Security and access
Key name: casestudykeys
EC2 instance profile: EMR_EC2_DefaultRole
EMR role: EMR_DefaultRole

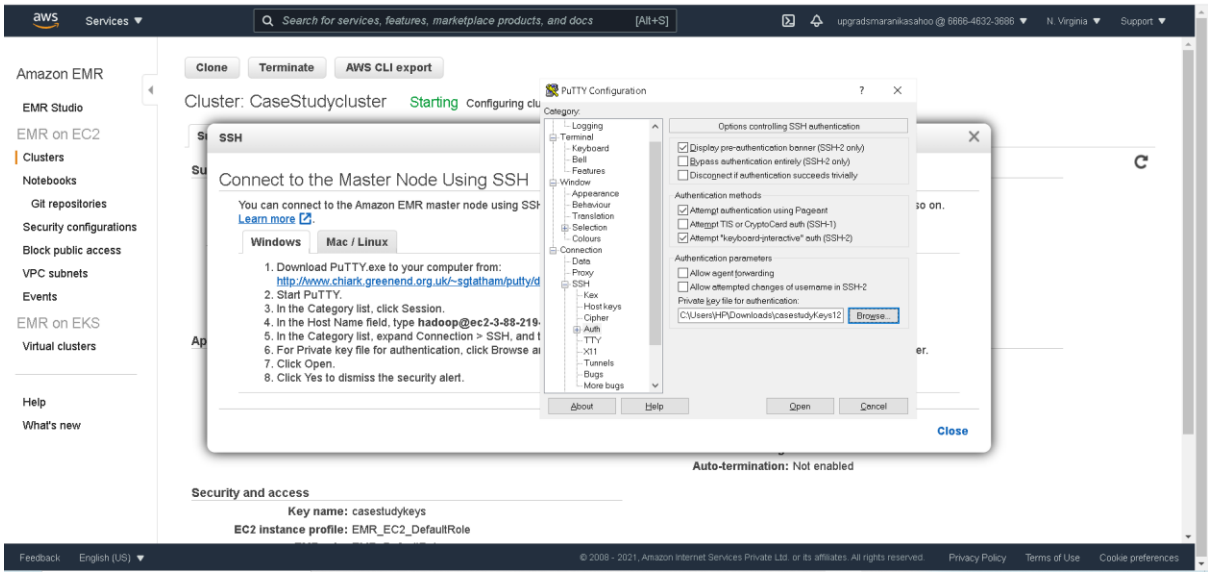

New EC2 Experience
EC2 Dashboard
EC2 Global View
Events
Tags
Limits
Instances
Instances
Instance Types
Launch Templates
Spot Requests
Savings Plans
Reserved Instances
Dedicated Hosts
Scheduled Instances
Capacity Reservations
Images
AMIs
Security Groups (1/2) Info
Actions
Export security groups to CSV
Create security group
Filter security groups
search: sg-0fe49ab0968bddace
Clear filters
Name
Security group ID
Security group name
VPC ID
Description
Owner
sg-0cbb2325b3b9c356c
ElasticMapReduce-slave
sg-0fe49ab0968bddace
ElasticMapReduce-master
sg-0fe49ab0968bddace - ElasticMapReduce-master
Details
Inbound rules
Outbound rules
Tags
You can now check network connectivity with Reachability Analyzer
Run Reachability Analyzer
Inbound rules (18)
Manage tags
Edit inbound rules


207.171.167.25/32
sgr-08a3f003d36e996a0
Custom TCP
TCP
8443
Custom
54.240.217.16/29
Delete
sgr-0a8de8b2a34efd36e
All UDP
UDP
0 - 65535
sg-0cbb2325b3b9c356c
sgr-0f74a359e4d4da63e
All TCP
TCP
0 - 65535
sg-0fe49ab0968bddace
sgr-02607daf9bed061bc
54.240.217.8/29
-
SSH
TCP
22
Anywh...
0.0.0.0/0
Add rule
Cancel
Preview changes
Save rules

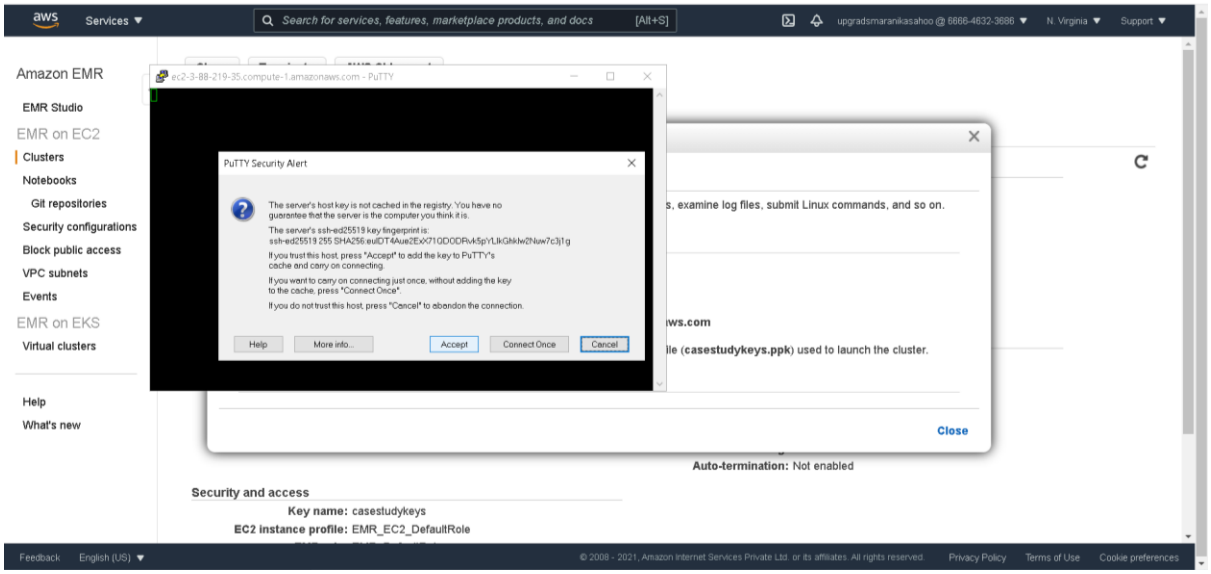
- Then save the SSH rule to the inbound rules CONNECT TO MASTER NODE: Open the putty and enter the Host Name as hadoop@ec2-18-212-7-184.compute-1.amazonaws.com and navigate to Connection > SSH > Auth then browse and select the private key, which we creating initially.

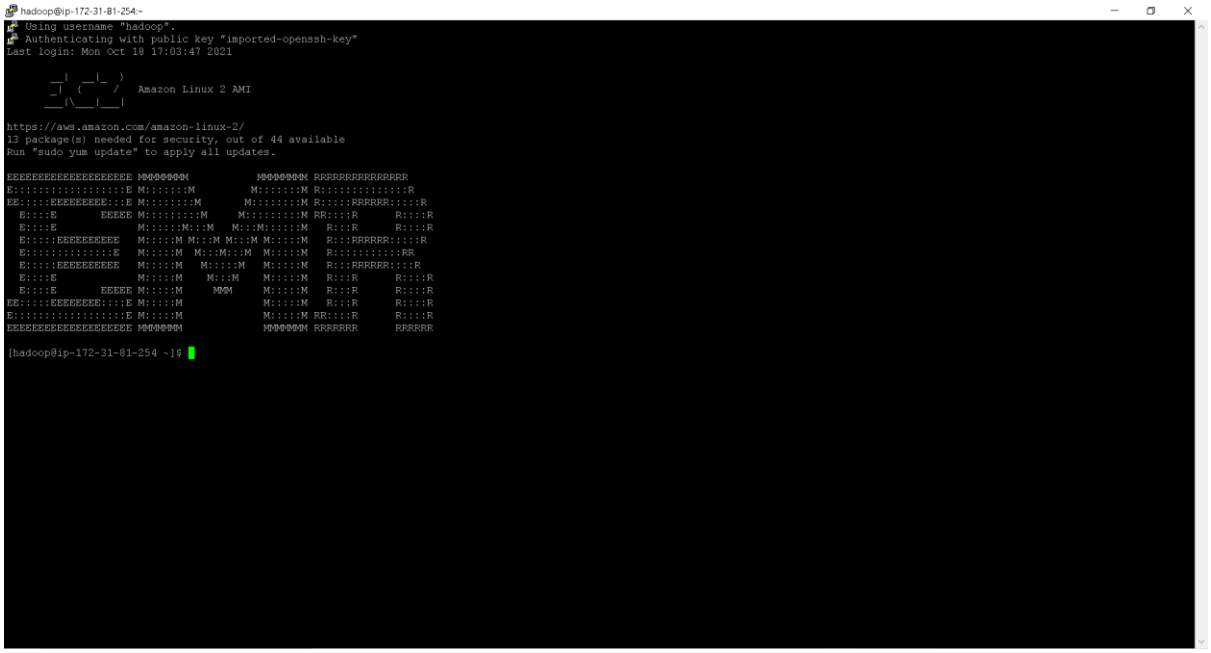




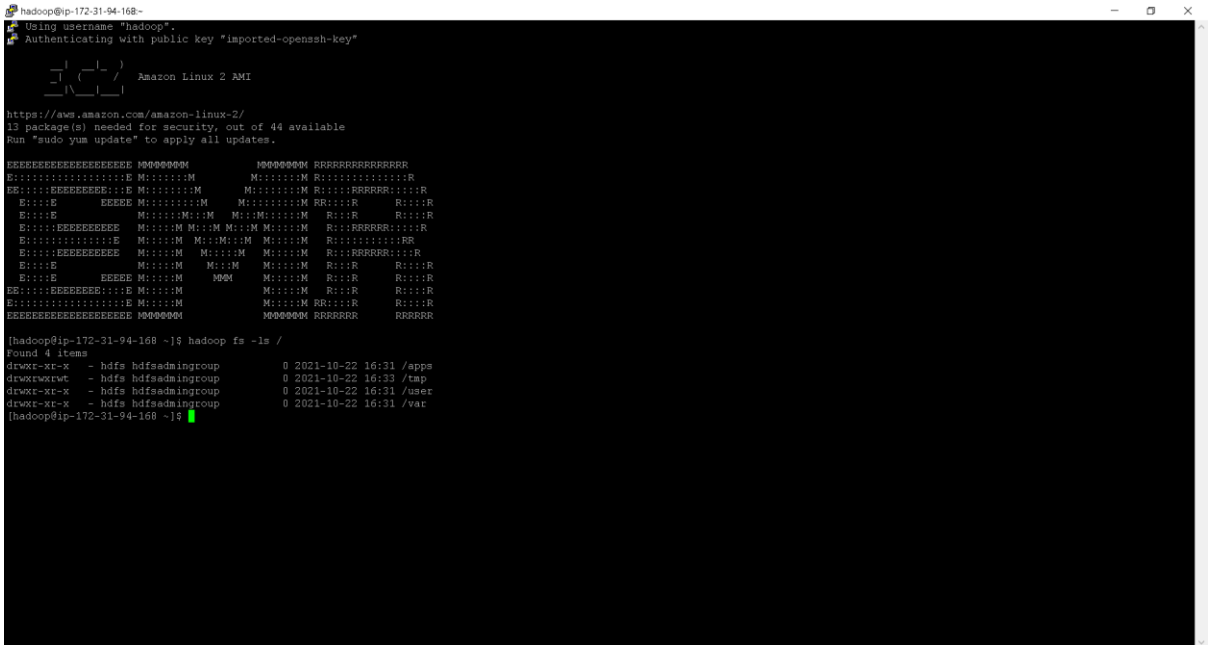


- Click on “open” and then Accept the connection.





- Checking the Hadoop file system with command “hadoop fs -ls /”



CREATING A NEW DIRECTORY FOR HIVE CASE STUDY:

- Creating a new directory under user for Hive case study to store the data files and directory name creating is “HiveCaseStudy” and verifying whether the new directory is listed in Hadoop file hadoop fs -mkdir /user/hive/HiveCaseStudy.

```
[hadoop@ip-172-31-81-254 ~]$ hadoop fs -mkdir /user/HiveCaseStudy/
[hadoop@ip-172-31-81-254 ~]$ hadoop fs -ls /user/
Found 7 items
drwxr-xr-x   - hadoop hdfsadmingroup          0 2021-10-18 17:13 /user/HiveCaseStudy
drwxrwxrwx   - hadoop hdfsadmingroup          0 2021-10-18 16:48 /user/hadoop
drwxr-xr-x   - mapred mapred                  0 2021-10-18 16:48 /user/history
drwxrwxrwx   - hdfs   hdfsadmingroup          0 2021-10-18 16:48 /user/hive
drwxrwxrwx   - hue    hue                     0 2021-10-18 16:48 /user/hue
drwxrwxrwx   - oozie  oozie                   0 2021-10-18 16:51 /user/oozie
drwxrwxrwx   - root   hdfsadmingroup          0 2021-10-18 16:48 /user/root
[hadoop@ip-172-31-81-254 ~]$
```

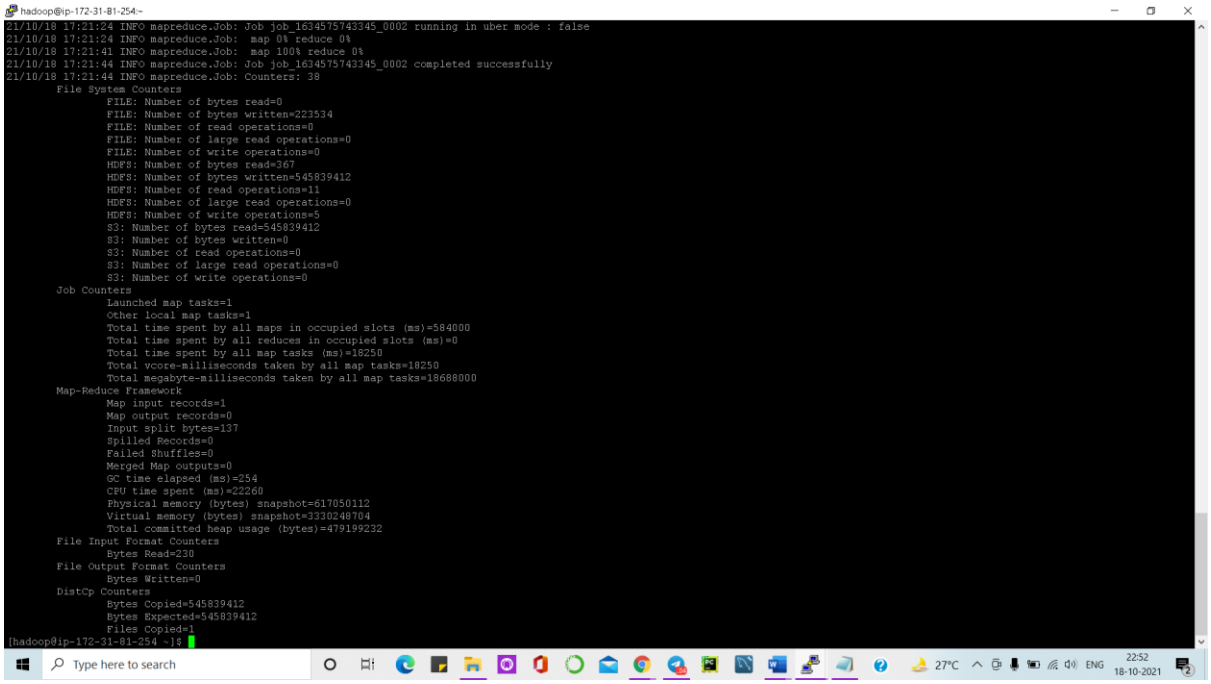
- New directory is successfully created.

LOADING THE DATA FROM S3 BUCKET to HDFS:

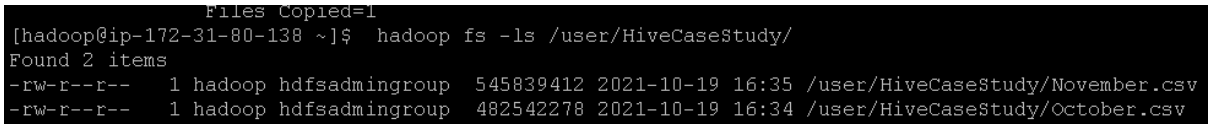
- Distributed copy command is used to copy the data from S3 to HDFS.
- For 2019 October:
  “hadoop distcp s3://casestudy123/2019-Oct.csv /user/HiveCaseStudy/October.csv”
- For 2019 November:
  “hadoop distcp s3://casestudy123/2019-Nov.csv /user/HiveCaseStudy/November.csv”
- Below are the screenshots for copying October 2019 and November 2019 data individually.

```
[hadoop@ip-172-31-81-254 ~]$ hadoop distcp s3://casestudy123/2019-Oct.csv /user/HiveCaseStudy/October.csv
21/10/18 17:18:27 INFO tools.OptionsParser: parseChunkSize: blocksperchunk false
21/10/18 17:18:28 INFO tools.DistCp: Input Options: DistCpOptions{atomicCommit=false, syncFolder=false, deleteMissing=false, ignoreFailures=false, overwrite=false, append=false, useDiff=fal
se, useRdiff=false, fromSnapshot=null, toSnapshot=null, skipCRC=false, blocking=true, numListstatusThreads=0, maxMaps=20, mapBandwidth=100, sslConfigurationFile='null', copyStrategy='unifor
msize', preserveStatus=[], preserveRawXattrs=false, atomicWorkPath=null, logPath=null, sourceFileListing=null, sourcePaths=[s3://casestudy123/2019-Oct.csv], targetPath=/user/HiveCaseStudy/O
ctober.csv, targetPathExists=false, filtersFile='null', blocksPerChunk=0, copyBufferSize=8192, verboseLog=false}
21/10/18 17:18:28 INFO client.RMProxy: Connecting to ResourceManager at ip-172-31-81-254.ec2.internal/172.31.81.254:8032
21/10/18 17:18:28 INFO client.AHSProxy: Connecting to Application History server at ip-172-31-81-254.ec2.internal/172.31.81.254:10200
21/10/18 17:18:32 INFO tools.SimpleCopyListing: Paths (files+dirs) cnt = 1; dirCnt = 0
21/10/18 17:18:32 INFO tools.SimpleCopyListing: Build file listing completed.
21/10/18 17:18:32 INFO Configuration.deprecation: io.sort.mb is deprecated. Instead, use mapreduce.task.io.sort.mb
21/10/18 17:18:32 INFO Configuration.deprecation: io.sort.factor is deprecated. Instead, use mapreduce.task.io.sort.factor
21/10/18 17:18:32 INFO tools.DistCp: Number of paths in the copy list: 1
21/10/18 17:18:33 INFO tools.DistCp: Number of paths in the copy list: 1
21/10/18 17:18:33 INFO client.RMProxy: Connecting to ResourceManager at ip-172-31-81-254.ec2.internal/172.31.81.254:8032
21/10/18 17:18:33 INFO client.AHSProxy: Connecting to Application History server at ip-172-31-81-254.ec2.internal/172.31.81.254:10200
21/10/18 17:18:33 INFO mapreduce.JobSubmitter: number of splits:1
21/10/18 17:18:33 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1634575743345_0001
21/10/18 17:18:34 INFO conf.Configuration: resource-types.xml not found
21/10/18 17:18:34 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
21/10/18 17:18:34 INFO resource.ResourceUtils: Adding resource type - name = memory-mb, units = Mi, type = COUNTABLE
21/10/18 17:18:34 INFO resource.ResourceUtils: Adding resource type - name = vcores, units = , type = COUNTABLE
21/10/18 17:18:34 INFO impl.YarnClientImpl: Submitted application application_1634575743345_0001
21/10/18 17:18:34 INFO mapreduce.Job: The url to track the job: http://ip-172-31-81-254.ec2.internal:20888/proxy/application_1634575743345_0001/
21/10/18 17:18:34 INFO tools.DistCp: DistCp job-id: job_1634575743345_0001
21/10/18 17:18:34 INFO mapreduce.Job: Running job: job_1634575743345_0001
```

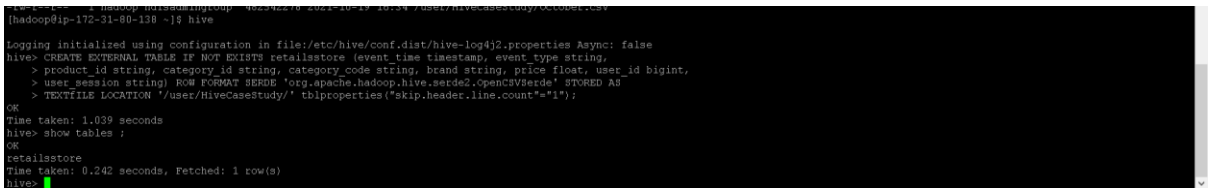
```
hadoop@ip-172-31-81-254:~
21/10/18 17:18:44 INFO mapreduce.Job: Job job_1634575743345_0001 running in uber mode : false
21/10/18 17:18:44 INFO mapreduce.Job:  map 0% reduce 0%
21/10/18 17:19:02 INFO mapreduce.Job:  map 100% reduce 0%
21/10/18 17:19:06 INFO mapreduce.Job: Job job_1634575743345_0001 completed successfully
21/10/18 17:19:06 INFO mapreduce.Job: Counters: 38
        File System Counters
                FILE: Number of bytes read=0
                FILE: Number of bytes written=223523
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=364
                HDFS: Number of bytes written=482542278
                HDFS: Number of read operations=11
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=5
                S3: Number of bytes read=482542278
                S3: Number of bytes written=0
                S3: Number of read operations=0
                S3: Number of large read operations=0
                S3: Number of write operations=0
        Job Counters
                Launched map tasks=1
                Other local map tasks=1
                Total time spent by all maps in occupied slots (ms)=606680
                Total time spent by all reduces in occupied slots (ms)=0
                Total time spent by all map tasks (ms)=18959
                Total vcore-milliseconds taken by all map tasks=18959
                Total megabyte-milliseconds taken by all map tasks=19414016
        Map-Reduce Framework
                Map input records=1
                Map output records=0
                Input split bytes=134
                Spilled Records=0
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=344
                CPU time spent (ms)=20210
                Physical memory (bytes) snapshot=629489664
                Virtual memory (bytes) snapshot=3326230528
                Total committed heap usage (bytes)=503840768
        File Input Format Counters
                Bytes Read=230
        File Output Format Counters
                Bytes Written=0
        DistCp Counters
                Bytes Copied=482542278
                Bytes Expected=482542278
                Files Copied=1
[hadoop@ip-172-31-81-254 ~]$
```



- Verifying whether the data is successfully copied into HDFS from S3 buckets Command: hadoop fs -ls /user/ HiveCaseStudy



Moving to hive:



Creating RetailDB Database:

```
hive> create database if not exists RetailDB;
OK
Time taken: 0.566 seconds
hive> describe database RetailDB ;
OK
retaildb                hdfs://ip-172-31-91-69.ec2.internal:8020/user/hive/warehouse/retaildb.db        hadoop  USER
Time taken: 0.203 seconds, Fetched: 1 row(s)
```

```
hive> use RetailDB;
OK
Time taken: 0.035 seconds
```

CREATING AN EXTERNAL TABLE IN HIVE:

- CREATE EXTERNAL TABLE IF NOT EXISTS retailsstore (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/HiveCaseStudy/' tblproperties("skip.header.line.count"="1");

```
[hadoop@ip-172-31-80-138 ~]$ hive

Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j2.properties Async: false
hive> CREATE EXTERNAL TABLE IF NOT EXISTS retailsstore (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/HiveCaseStudy/' tblproperties("skip.header.line.count"="1");
OK
Time taken: 1.039 seconds
hive> show tables ;
OK
retailsstore
Time taken: 0.242 seconds, Fetched: 1 row(s)
hive> set hive.cli.print.header=True;
hive> select * from retailsstore limit 3 ;
OK
retailsstore.event_time retailsstore.event_type retailsstore.product_id retailsstore.category_id       retailsstore.category_code      retailsstore.brand      retailsstore.price      retai
lsstore.user_id retailsstore.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
Time taken: 2.502 seconds, Fetched: 3 row(s)
hive>
```

- set hive.cli.print.header=True;
- Checked the data in the table by querying the below.

```
hive> set hive.cli.print.header=True;
hive> select * from retailsstore limit 3 ;
OK
retailsstore.event_time retailsstore.event_type retailsstore.product_id retailsstore.category_id       retailsstore.category_code      retailsstore.brand      retailsstore.price      retai
lsstore.user_id retailsstore.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
Time taken: 0.207 seconds, Fetched: 3 row(s)
```

Questions & Answers

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

Query: SELECT SUM(price) AS total_revenue_October FROM retailsstore WHERE date_format(event_time,'MM')=10 AND event_type='purchase';

```
hive> SELECT SUM(price) AS total_revenue_October FROM retailsstore WHERE date_format(event_time,'MM')=10 AND event_type='purchase';
Query ID = hadoop_20211019165603_10765737-92bc-45b7-9335-79e30cd447e8
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1634660723820_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
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 126.79 s
--------------------------------------------------------------------------------
OK
1211538.4299997438
Time taken: 132.79 seconds, Fetched: 1 row(s)
```

Insight:

The total revenue generated based on Purchase in the month of October of 2019 was 1,211,538.42999/

<u>Question 2:</u> Write a query to yield the total sum of purchases per month in a single output.

Query: SELECT date_format(event_time, 'MM') AS Months, COUNT(event_type) AS Sum_of_Purchases FROM retailsstore WHERE event_type='purchase' GROUP BY date_format(event_time, 'MM');

```
hive> SELECT date_format(event_time, 'MM') AS Months, COUNT(event_type) AS Sum_of_Purchases FROM retailsstore WHERE event_type='purchase' GROUP BY date_format(event_time, 'MM');
Query ID = hadoop_20211019170526_58e26359-7b4b-4438-ab0e-2f178c76fc82
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1634660723820_0005)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      2          2        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      6          6        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 62.31 s
--------------------------------------------------------------------------------
OK
10      245624
11      322417
Time taken: 66.994 seconds, Fetched: 2 row(s)
```

Insight:

The total sum of purchases per month, October (10) i.e., 245,624, November (11) i.e., 322,417

<u>Question 3:</u> Write a query to find the change in revenue generated due to purchases from October to November.

Query: with diff_revenue as(select sum(case when MONTH(event_time) ='10' then price else 0 end) as oct_revenue,sum(case when MONTH(event_time) ='11' then price else 0 end) as nov_revenue from retailsstore where event_type='purchase') select (nov_revenue - oct_revenue) as revenue_difference from diff_revenue;

```
hive> with diff_revenue as(select sum(case when MONTH(event_time) ='10' then price else 0 end) as oct_revenue,
    > sum(case when MONTH(event_time) ='11' then price else 0 end) as nov_revenue
    > from retailsstore where event_type='purchase')
    > select (nov_revenue - oct_revenue) as revenue_difference from diff_revenue;
Query ID = hadoop_20211024130838_fa0c4c6c-4714-4576-8b09-1f5675626936
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1635075264890_0006)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      2          2        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 56.73 s
--------------------------------------------------------------------------------
OK
319478.4700003781
Time taken: 61.75 seconds, Fetched: 1 row(s)
```

Insight:

From the query we got the difference in revenue from October to November is 319478.47

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

Query: select distinct(category_code) from retailsstore where category_code!="";

```
hive> select distinct(category_code)from retailsstore where category_code != '' ;
Query ID = hadoop_20211019172124_d9695756-ab8a-46eb-9fe5-180045826ff9
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1634660723820_0007)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      2          2        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 55.86 s
--------------------------------------------------------------------------------
OK
accessories.bag
accessories.cosmetic_bag
apparel.glove
appliances.environment.air_conditioner
appliances.environment.vacuum
appliances.personal.hair_cutter
furniture.bathroom.bath
furniture.living_room.cabinet
furniture.living_room.chair
sport.diving
stationery.cartrige
Time taken: 61.594 seconds, Fetched: 11 row(s)
```

Insight:

- Total we got 6 distinct categories are – furniture, appliances, accessories, apparel, sport, stationary.

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

Query: select count(product_id) , category_code from retailsstore where category_code IS NOT NULL group by category_code;

```
hive> select count(product_id) , category_code from retailsstore where category_code IS NOT NULL  group by category_code;
Query ID = hadoop_20211019172400_837299f4-ba1d-453e-ba84-95d85564a7e6
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1634660723820_0007)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      2          2        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 51.61 s
----------------------------------------------------------------------------------------------
OK
8594895
11681   accessories.bag
1248    accessories.cosmetic_bag
18232   apparel.glove
332     appliances.environment.air_conditioner
59761   appliances.environment.vacuum
1643    appliances.personal.hair_cutter
9857    furniture.bathroom.bath
13439   furniture.living_room.cabinet
308     furniture.living_room.chair
2       sport.diving
26722   stationery.cartrige
Time taken: 52.376 seconds, Fetched: 12 row(s)
```

Insight:

- Company has more products registered under Appliances category i.e., 61,736 products than any other categories.
- Then it is followed by stationery as second with 26,722 products, furniture as third with 23,604 products, apparel as fourth with 18232 products registered, accessories as fifth with 12929 products
- Sports category is least available with 2 products.

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

Query: select brand ,sum(price) as sales from retailsstore where brand != "" and event_type= "purchase" group by brand order by sales desc limit 1;

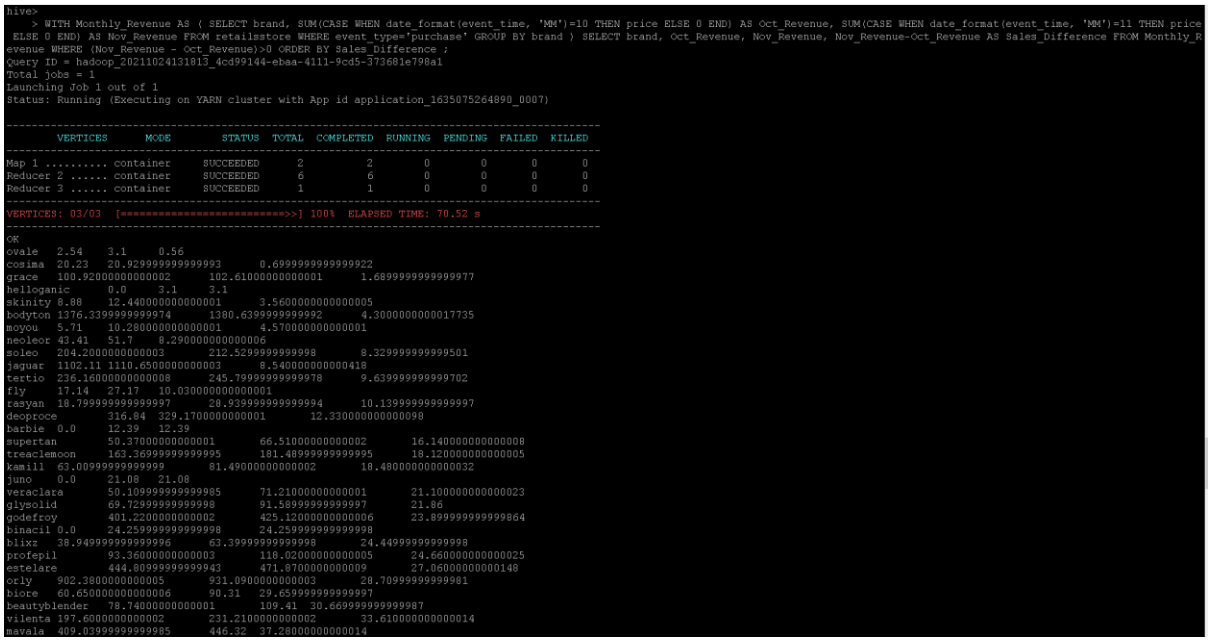
```
hive> select brand ,sum(price) as sales from retailsstore where brand != "" and event_type= "purchase" group by brand order by sales desc limit 1;
Query ID = hadoop_20211024131531_c4a5a511-39da-429d-8112-ddc5f782ed53
Total jobs = 1
Launching Job 1 out of 1
Tez session was closed. Reopening...
Session re-established.
Status: Running (Executing on YARN cluster with App id application_1635075264890_0007)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      2          2        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      6          6        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 59.49 s
----------------------------------------------------------------------------------------------
OK
runail  148297.9400000003
Time taken: 68.936 seconds, Fetched: 1 row(s)
```

Insight:

From the query we got that runail had the maximum sales in October and November combined.

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

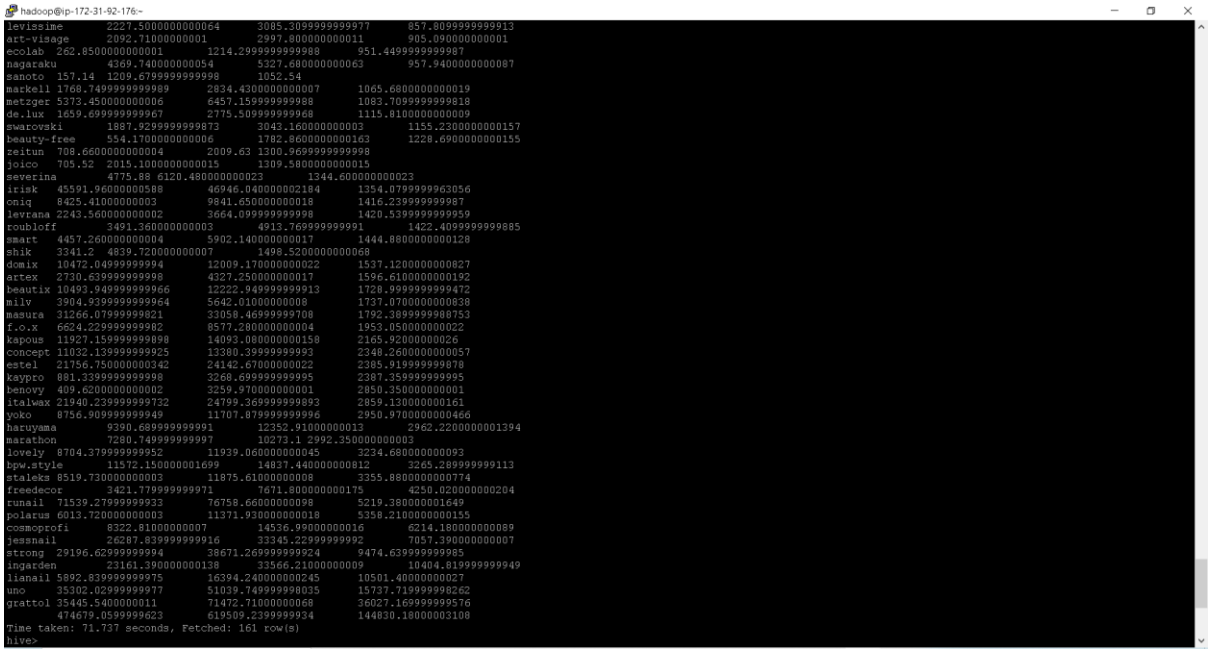
Query: WITH Monthly_Revenue AS (SELECT brand, SUM(CASE WHEN date_format(event_time, 'MM')=10 THEN price ELSE 0 END) AS Oct_Revenue, SUM(CASE WHEN date_format(event_time, 'MM')=11 THEN price ELSE 0 END) AS Nov_Revenue FROM retailsstore WHERE event_type='purchase' GROUP BY brand) SELECT brand, Oct_Revenue, Nov_Revenue, Nov_Revenue-Oct_Revenue AS Sales_Difference FROM Monthly_Revenue WHERE (Nov_Revenue - Oct_Revenue)>0 ORDER BY Sales_Difference ;



```
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 retailsstore WHERE event_type='purchase' GROUP BY brand ) SELECT brand, Oct_Revenue, Nov_Revenue, Nov_Revenue-Oct_Revenue AS Sales_Difference FROM Monthly_R
evenue WHERE (Nov_Revenue - Oct_Revenue)>0 ORDER BY Sales_Difference ;
Query ID = hadoop_20211024131813_4cd99144-ebaa-4111-9cd5-373681e798a1
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1635075264890_0007)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 ..........   container     SUCCEEDED      2          2        0        0       0       0
Reducer 2 ......   container     SUCCEEDED      6          6        0        0       0       0
Reducer 3 ......   container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 70.52 s
----------------------------------------------------------------------------------------------
OK
ovale   2.54    3.1     0.56
cosima  20.23   20.929999999999993      0.6999999999999922
grace   100.92000000000002      102.61000000000001      1.6899999999999977
helloganiс      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
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       27.060000000000148
oily    902.3800000000005       931.0900000000003       28.709999999999981
biore   60.650000000000006      90.31   29.659999999999997
beautyblender   78.74000000000001       109.41  30.669999999999987
vilenta 197.6000000000002       231.2100000000002       33.610000000000014
mavala  409.03999999999985      446.32  37.28000000000014
```

```
mavala  409.03999999999985      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       94.2900000000003
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
nefertiti       233.52000000000007      366.64  133.11999999999992
elizavecca      70.53   204.3   133.77
miskin  158.04  293.07000000000005      135.03000000000006
latinoil        249.52  384.59  135.06999999999996
farmona 1692.4599999999996      1843.4300000000007      150.97000000000116
cristalinas     427.6299999999999       584.9499999999999       157.31999999999914
chi     358.9400000000002       538.6100000000002       179.67000000000002
matreshka       0.0     182.67000000000002      182.67000000000002
freshbubble     318.70000000000001      502.34000000000015      183.64000000000004
mane    66.78999999999999       260.26  193.47
keen    236.35000000000005      435.62  199.26999999999995
ecocraft        41.160000000000004      241.95  200.79
fedua   52.38   263.81000000000006      211.43000000000006
provoc  827.9900000000009       1063.8200000000006      235.8299999999997
skinlite        651.9400000000002       890.4499999999979       238.50999999999772
entity  479.7100000000015       719.2599999999993       239.5499999999978
trind   298.07000000000005      542.9600000000002       244.8900000000001
protokeratin    201.25000000000003      456.79000000000013      255.5400000000001
beauugreen      511.5099999999999       768.35  256.84000000000015
bluesky 10307.239999999858      10565.529999999713      258.28999999985535
```

```
bluesky 10307.239999999858      10565.529999999713      258.28999999985535
candy   534.9599999999999       799.3799999999993       264.4199999999994
insight 1443.7000000000012      1721.9600000000003      278.2599999999991
kocostar        310.8500000000001       594.9300000000003       284.0800000000002
happyfons       801.9200000000006       1091.5900000000001      289.6699999999995
kims    330.03999999999996      632.0400000000001       302.0000000000001
shary   871.9599999999994       1176.4899999999989      304.5299999999995
nitrile 847.279999999999        1162.679999999999       315.4
lowence 242.84  567.7499999999997       324.9099999999996
jas     3318.959999999995       3657.4300000000026      338.47000000000753
ellips  245.8499999999999       606.0399999999996       360.1899999999997
lador   2083.610000000004       2471.530000000007       387.9200000000028
naomi   0.0     389.0   389.0
kiss    421.54999999999944      817.3299999999994       395.7799999999999
yu-r    271.41  673.7099999999998       402.2999999999998
sophin  1067.8600000000001      1515.5200000000011      447.660000000001
farmavita       837.3699999999984       1291.9700000000003      454.60000000000184
bioaqua 942.8899999999996       1398.1199999999997      455.23
greymy  29.21   489.49  460.28000000000003
gehwol  1089.07 1557.6799999999982      468.6099999999983
matrix  3243.249999999999       3726.7400000000007      483.4900000000016
limoni  1308.9000000000003      1796.5999999999997      487.69999999999936
s.care  412.68  913.0699999999999       500.38999999999993
coifin  903.0000000000001       1428.4899999999998      525.4899999999997
uskusi  5142.270000000017       5690.310000000005       548.0399999999881
airnails        5118.899999999939       5691.519999999996       572.6200000000572
browxenna       14331.36999999995       14916.729999999976      585.3600000000026
kinetics        6334.2499999999945      6945.260000000017       611.0100000000022
kosmekka        1181.4400000000003      1813.37 631.9299999999996
kaaral  4412.4299999999985      5086.069999999992       673.6399999999994
refectocil      2716.180000000005       3475.5800000000007      759.4000000000024
rosi    3077.0399999999927      3841.5600000000013      764.5200000000204
solomeya        1899.699999999992       2685.799999999991       786.0999999999999
missha  1293.8299999999995      2150.2799999999984      856.4499999999989
levissime       2227.5000000000064      3085.3099999999977      857.8099999999913
art-visage      2092.71000000001        2997.8000000000011      905.0900000000001
ecolab  262.8500000000001       1214.2999999999988      951.4499999999987
nagaraku        4369.740000000054       5327.680000000063       957.9400000000087
sanoto  157.14  1209.6799999999998      1052.54
markell 1768.7499999999989      2834.4300000000007      1065.6800000000019
metzger 5373.450000000006       6457.159999999988       1083.7099999999818
de.lux  1659.699999999967       2775.509999999968       1115.8100000000009
swarovski       1887.9299999999873      3043.1600000000003      1155.2300000000157
beauty-free     554.1700000000006       1782.8600000000163      1228.6900000000155
zeitun  708.6600000000004       2009.63 1300.9699999999998
joico   705.52  2015.1000000000015      1309.5800000000015
severina        4775.88 6120.480000000023       1344.600000000023
```

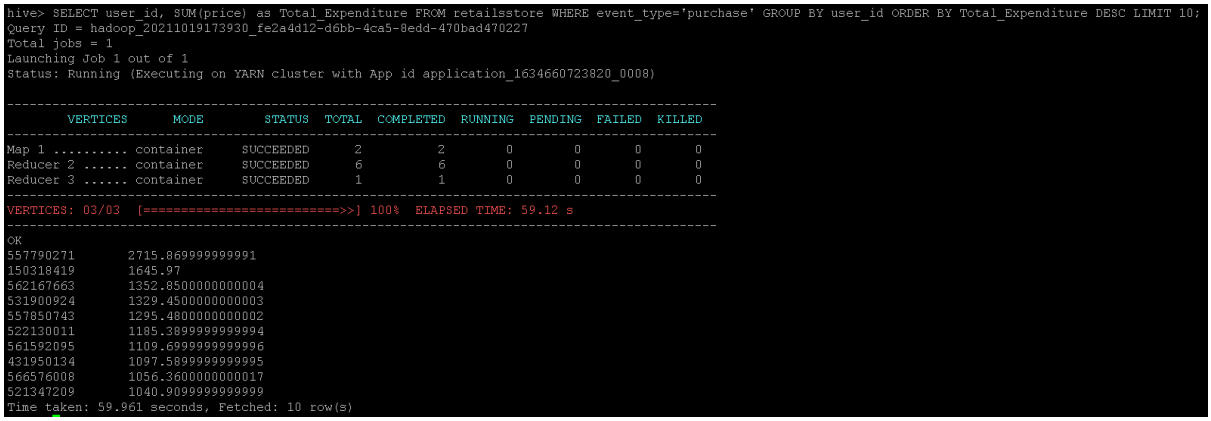


Insight:

- There are so many brands which have the increment from October to November. Among which ‘Grattol’ brand has the highest total increment and ‘Ovale’ seems to have least increment.

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.

Query: SELECT user_id, SUM(price) as Total_Expenditure FROM retailsstore WHERE event_type='purchase' GROUP BY user_id ORDER BY Total_Expenditure DESC LIMIT 10;



Insight:

- Here is the list of the top 10 users who have spent the most.

To create table with Partitioning and Bucketing below commands need to be executed one by one separately.

- set hive.exec.dynamic.partition.mode=nonstrict;
- set hive.exec.dynamic.partition=true;
- set hive.enforce.bucketing=true;

```
[hadoop@ip-172-31-84-44 ~]$ set hive.exec.dynamic.partition.mode=nonstrict;
[hadoop@ip-172-31-84-44 ~]$ set hive.exec.dynamic.partition=true;
[hadoop@ip-172-31-84-44 ~]$ set hive.enforce.bucketing=true;
```

Table optimization steps:-

- Command to create table **'Dyn_retailsstore_part'** with partition on **'event_type'** attribute and bucket(cluster) on **'price'** attribute.

Query: CREATE TABLE IF NOT EXISTS Dyn_retailsstore_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 (price) INTO 7 BUCKETS ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.OpenCSVSerde' STORED AS TEXTFILE ;

```
hive> CREATE TABLE IF NOT EXISTS Dyn_retailsstore_part ( event_time timestamp, product_id string, category_id string, category_code string, brand string, price float, user_id bigint, user_s
ession string ) PARTITIONED BY (event_type string) CLUSTERED BY (price) INTO 7 BUCKETS ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.OpenCSVSerde' STORED AS TEXTFILE ;
OK
Time taken: 0.202 seconds
hive> INSERT INTO TABLE Dyn_retailsstore_part PARTITION (event_type) SELECT event_time, product_id, category_id, category_code, brand, price, user_id, user_session, event_type FROM retailss
tore ;
FAILED: SemanticException [Error 10096]: Dynamic partition strict mode requires at least one static partition column. To turn this off set hive.exec.dynamic.partition.mode=nonstrict
hive> set hive.exec.dynamic.partition.mode=nonstrict;
hive> set hive.exec.dynamic.partition=true;
hive> set hive.enforce.bucketing=true;
hive> INSERT INTO TABLE Dyn_retailsstore_part PARTITION (event_type) SELECT event_time, product_id, category_id, category_code, brand, price, user_id, user_session, event_type FROM retailss
tore ;
Query ID = hadoop_20211020164051_03c02e7b-2f13-4468-ae42-73eb310e1904
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_1634746553704_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: 165.81 s
--------------------------------------------------------------------------------
Loading data to table default.dyn_retailsstore_part partition (event_type=null)

Loaded : 4/4 partitions.
        Time taken to load dynamic partitions: 0.817 seconds
        Time taken for adding to write entity : 0.006 seconds
OK
event_time      product_id      category_id     category_code   brand   price   user_id user_session    event_type
Time taken: 178.87 seconds
```

- To add data into partitioned and bucketed table we need to get it from already created table 'Dyn_retailsstore_part'.

Query: INSERT INTO TABLE Dyn_retailsstore_part PARTITION (event_type) SELECT event_time, product_id, category_id, category_code, brand, price, user_id, user_session, event_type FROM retailsstore ;

- Command to check the successful creation of partitioned and bucketed table first we need to exit from Hive environment by executing 'EXIT;' command and then run below mentioned command.

```
hive> exit;
[hadoop@ip-172-31-92-176 ~]$
```

```
[hadoop@ip-172-31-84-44 ~]$ hadoop fs -ls /user/hive/warehouse/dyn_retailsstore_part
Found 4 items
drwxrwxrwt   - hadoop hdfsadmingroup          0 2021-10-20 16:43 /user/hive/warehouse/dyn_retailsstore_part/event_type=cart
drwxrwxrwt   - hadoop hdfsadmingroup          0 2021-10-20 16:43 /user/hive/warehouse/dyn_retailsstore_part/event_type=purchase
drwxrwxrwt   - hadoop hdfsadmingroup          0 2021-10-20 16:43 /user/hive/warehouse/dyn_retailsstore_part/event_type=remove_from_cart
drwxrwxrwt   - hadoop hdfsadmingroup          0 2021-10-20 16:43 /user/hive/warehouse/dyn_retailsstore_part/event_type=view
```

```
[hadoop@ip-172-31-84-44 ~]$ hadoop fs -ls /user/hive/warehouse/dyn_retailsstore_part/event_type=purchase
Found 7 items
-rwxrwxrwt   1 hadoop hdfsadmingroup   13052654 2021-10-20 16:43 /user/hive/warehouse/dyn_retailsstore_part/event_type=purchase/000000_0
-rwxrwxrwt   1 hadoop hdfsadmingroup    9399111 2021-10-20 16:43 /user/hive/warehouse/dyn_retailsstore_part/event_type=purchase/000001_0
-rwxrwxrwt   1 hadoop hdfsadmingroup   12636711 2021-10-20 16:43 /user/hive/warehouse/dyn_retailsstore_part/event_type=purchase/000002_0
-rwxrwxrwt   1 hadoop hdfsadmingroup   10650131 2021-10-20 16:42 /user/hive/warehouse/dyn_retailsstore_part/event_type=purchase/000003_0
-rwxrwxrwt   1 hadoop hdfsadmingroup    7226455 2021-10-20 16:42 /user/hive/warehouse/dyn_retailsstore_part/event_type=purchase/000004_0
-rwxrwxrwt   1 hadoop hdfsadmingroup   10737803 2021-10-20 16:42 /user/hive/warehouse/dyn_retailsstore_part/event_type=purchase/000005_0
-rwxrwxrwt   1 hadoop hdfsadmingroup    7825305 2021-10-20 16:42 /user/hive/warehouse/dyn_retailsstore_part/event_type=purchase/000006_0
```

```
[hadoop@ip-172-31-84-44 ~]$ hadoop fs -ls /user/hive/warehouse/dyn_retailsstore_part/event_type=cart
Found 7 items
-rwxrwxrwt   1 hadoop hdfsadmingroup   57724286 2021-10-20 16:43 /user/hive/warehouse/dyn_retailsstore_part/event_type=cart/000000_0
-rwxrwxrwt   1 hadoop hdfsadmingroup   43094161 2021-10-20 16:43 /user/hive/warehouse/dyn_retailsstore_part/event_type=cart/000001_0
-rwxrwxrwt   1 hadoop hdfsadmingroup   56823661 2021-10-20 16:43 /user/hive/warehouse/dyn_retailsstore_part/event_type=cart/000002_0
-rwxrwxrwt   1 hadoop hdfsadmingroup   49030059 2021-10-20 16:43 /user/hive/warehouse/dyn_retailsstore_part/event_type=cart/000003_0
-rwxrwxrwt   1 hadoop hdfsadmingroup   31050141 2021-10-20 16:42 /user/hive/warehouse/dyn_retailsstore_part/event_type=cart/000004_0
-rwxrwxrwt   1 hadoop hdfsadmingroup   48253679 2021-10-20 16:42 /user/hive/warehouse/dyn_retailsstore_part/event_type=cart/000005_0
-rwxrwxrwt   1 hadoop hdfsadmingroup   34272441 2021-10-20 16:42 /user/hive/warehouse/dyn_retailsstore_part/event_type=cart/000006_0
[hadoop@ip-172-31-84-44 ~]$ hadoop fs -ls /user/hive/warehouse/dyn_retailsstore_part/event_type=remove_from_cart
Found 7 items
-rwxrwxrwt   1 hadoop hdfsadmingroup   39017824 2021-10-20 16:43 /user/hive/warehouse/dyn_retailsstore_part/event_type=remove_from_cart/000000_0
-rwxrwxrwt   1 hadoop hdfsadmingroup   29421828 2021-10-20 16:43 /user/hive/warehouse/dyn_retailsstore_part/event_type=remove_from_cart/000001_0
-rwxrwxrwt   1 hadoop hdfsadmingroup   38713899 2021-10-20 16:43 /user/hive/warehouse/dyn_retailsstore_part/event_type=remove_from_cart/000002_0
-rwxrwxrwt   1 hadoop hdfsadmingroup   31959876 2021-10-20 16:43 /user/hive/warehouse/dyn_retailsstore_part/event_type=remove_from_cart/000003_0
-rwxrwxrwt   1 hadoop hdfsadmingroup   19751571 2021-10-20 16:42 /user/hive/warehouse/dyn_retailsstore_part/event_type=remove_from_cart/000004_0
-rwxrwxrwt   1 hadoop hdfsadmingroup   31335021 2021-10-20 16:43 /user/hive/warehouse/dyn_retailsstore_part/event_type=remove_from_cart/000005_0
-rwxrwxrwt   1 hadoop hdfsadmingroup   22175799 2021-10-20 16:42 /user/hive/warehouse/dyn_retailsstore_part/event_type=remove_from_cart/000006_0
[hadoop@ip-172-31-84-44 ~]$ hadoop fs -ls /user/hive/warehouse/dyn_retailsstore_part/event_type=view
Found 7 items
-rwxrwxrwt   1 hadoop hdfsadmingroup   88831872 2021-10-20 16:43 /user/hive/warehouse/dyn_retailsstore_part/event_type=view/000000_0
-rwxrwxrwt   1 hadoop hdfsadmingroup   73953212 2021-10-20 16:42 /user/hive/warehouse/dyn_retailsstore_part/event_type=view/000001_0
-rwxrwxrwt   1 hadoop hdfsadmingroup   85620113 2021-10-20 16:43 /user/hive/warehouse/dyn_retailsstore_part/event_type=view/000002_0
-rwxrwxrwt   1 hadoop hdfsadmingroup   71874121 2021-10-20 16:43 /user/hive/warehouse/dyn_retailsstore_part/event_type=view/000003_0
-rwxrwxrwt   1 hadoop hdfsadmingroup   48335545 2021-10-20 16:43 /user/hive/warehouse/dyn_retailsstore_part/event_type=view/000004_0
-rwxrwxrwt   1 hadoop hdfsadmingroup   72515614 2021-10-20 16:43 /user/hive/warehouse/dyn_retailsstore_part/event_type=view/000005_0
-rwxrwxrwt   1 hadoop hdfsadmingroup   56694677 2021-10-20 16:43 /user/hive/warehouse/dyn_retailsstore_part/event_type=view/000006_0
```

<u>Question 3:</u> Write a query to find the change in revenue generated due to purchases from October to November.

```
hive> select sum(price) AS Total_Revenue_Oct from dyn_retailsstore_part  where month(event_time)=10 and event_type='purchase' MINUS select sum(price) AS Total_Revenue_Nov from dyn_retailsst
ore_part  where month(event_time)=11 and event_type='purchase' ;
Query ID = hadoop_20211020170255_0a9c44b8-4514-4a6b-a304-c058bc862063
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1634746553704_0005)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      7          7        0        0       0       0
Map 6 .......... container     SUCCEEDED      7          7        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      2          2        0        0       0       0
Reducer 5 ...... container     SUCCEEDED      2          2        0        0       0       0
Reducer 7 ...... container     SUCCEEDED      1          1        0        0       0       0
Reducer 8 ...... container     SUCCEEDED      2          2        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 07/07  [==========================>>] 100%  ELAPSED TIME: 33.31 s
--------------------------------------------------------------------------------
OK
1211538.4300000786
Time taken: 34.723 seconds, Fetched: 1 row(s)
```

Insight:

Earlier the query was taking 134 seconds to complete but with the partitioning and bucketing the query took only 35 seconds to complete.

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.

```
hive> SELECT user_id, SUM(price) as Total_Expenditure FROM dyn_retailsstore_part  WHERE event_type='purchase' GROUP BY user_id ORDER BY Total_Expenditure DESC LIMIT 10;
Query ID = hadoop_20211020170810_98d31001-3983-4a1d-ac09-273ee796e16d
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1634746553704_0005)

----------------------------------------------------------------------------------------------
        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: 25.00 s
----------------------------------------------------------------------------------------------
OK
557790271       2715.869999999996
150318419       1645.97
562167663       1352.85
531900924       1329.45
557850743       1295.48
522130011       1185.3900000000003
561592095       1109.7000000000003
431950134       1097.59
566576008       1056.3599999999997
521347209       1040.9099999999999
Time taken: 26.545 seconds, Fetched: 10 row(s)
```

Insight:

Earlier the query was taking 60 seconds to complete but with the partitioning and bucketing the query took only 27 seconds to complete. Hence, optimized table gives better performance in execution time.

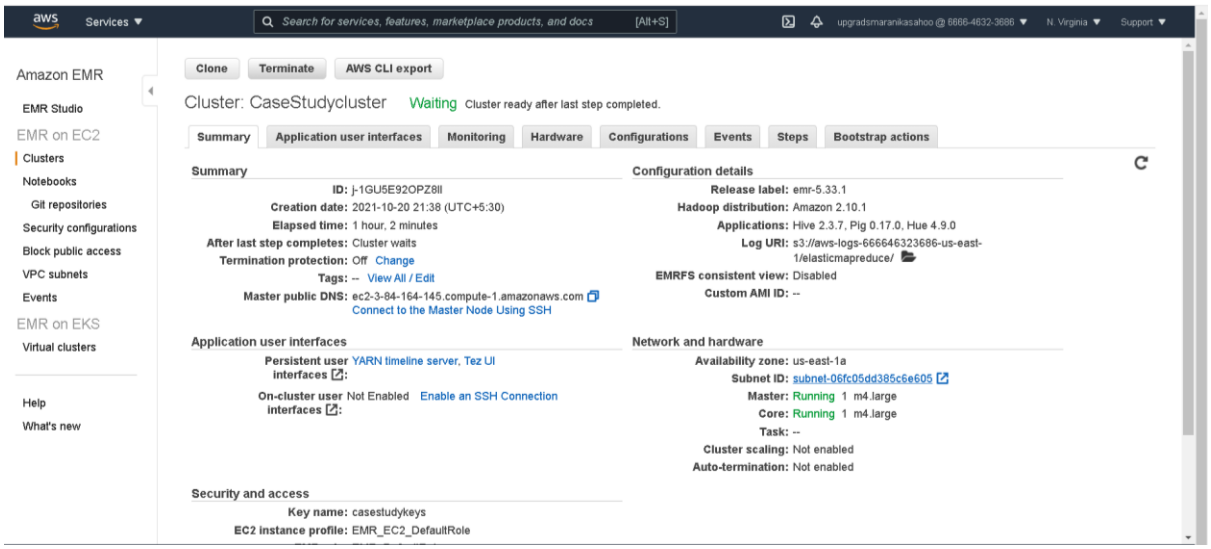
TERMINATION PROCESS:

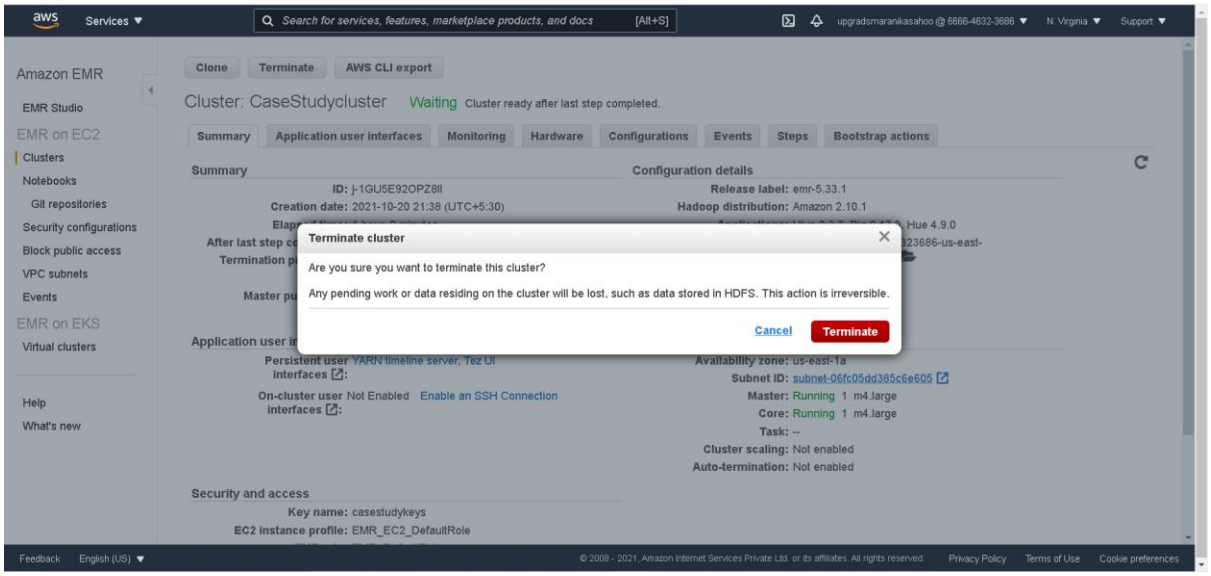
Dropping Database:

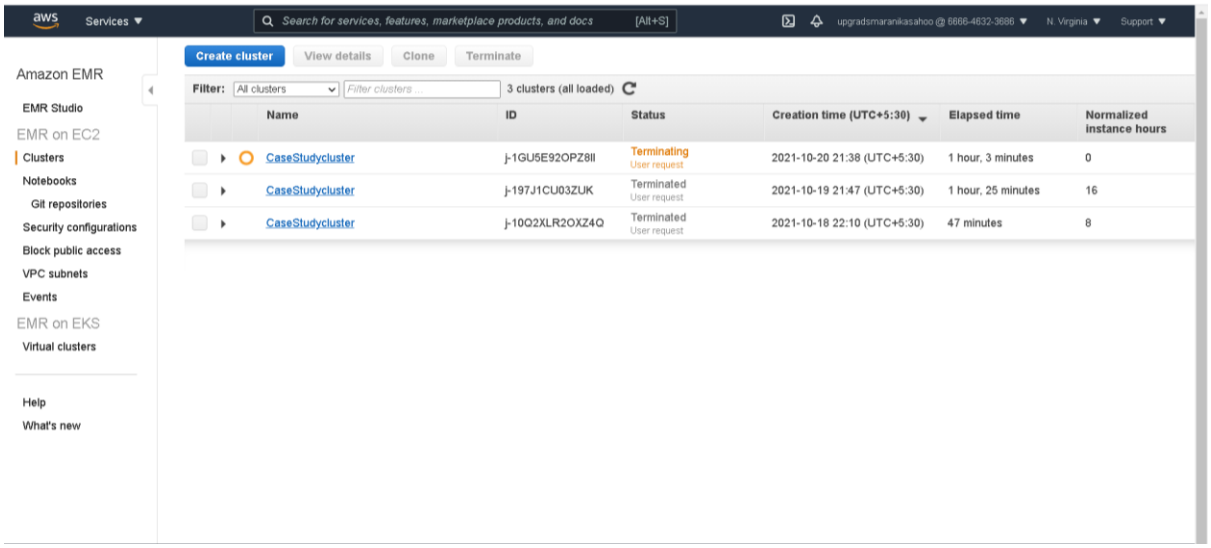
```
hive> drop database RetailDB CASCADE;
OK
Time taken: 0.706 seconds
hive> show databases ;
NoViableAltException(24@[846:1: ddlStatement : ( createDatabaseStatement | switchDatabaseStatement | dropDatabaseStatement | createTableStatement | dropTableStatement | truncateTableStatement | alterStatement | descStatement | showStatement | metastoreCheck | createViewStatement | createMaterializedViewStatement | dropViewStatement | dropMaterializedViewStatement | createFunctionStatement | createMacroStatement | createIndexStatement | dropIndexStatement | dropFunctionStatement | reloadFunctionStatement | dropMacroStatement | analyzeStatement | lockStatement | unlockStatement | lockDatabase | unlockDatabase | createRoleStatement | dropRoleStatement | ( grantPrivileges )=> grantPrivileges | ( revokePrivileges )=> revokePrivileges | showGrants | showRoleGrants | showRolePrincipals | showRoles | grantRole | revokeRole | setRole | showCurrentRole | abortTransactionStatement );])
        at org.antlr.runtime.DFA.noViableAlt(DFA.java:158)
        at org.antlr.runtime.DFA.predict(DFA.java:144)
        at org.apache.hadoop.hive.ql.parse.HiveParser.ddlStatement(HiveParser.java:3757)
        at org.apache.hadoop.hive.ql.parse.HiveParser.execStatement(HiveParser.java:2382)
        at org.apache.hadoop.hive.ql.parse.HiveParser.statement(HiveParser.java:1333)
        at org.apache.hadoop.hive.ql.parse.ParseDriver.parse(ParseDriver.java:208)
        at org.apache.hadoop.hive.ql.parse.ParseUtils.parse(ParseUtils.java:77)
        at org.apache.hadoop.hive.ql.parse.ParseUtils.parse(ParseUtils.java:70)
        at org.apache.hadoop.hive.ql.Driver.compile(Driver.java:468)
        at org.apache.hadoop.hive.ql.Driver.compileInternal(Driver.java:1317)
        at org.apache.hadoop.hive.ql.Driver.runInternal(Driver.java:1457)
        at org.apache.hadoop.hive.ql.Driver.run(Driver.java:1237)
        at org.apache.hadoop.hive.ql.Driver.run(Driver.java:1227)
        at org.apache.hadoop.hive.cli.CliDriver.processLocalCmd(CliDriver.java:233)
        at org.apache.hadoop.hive.cli.CliDriver.processCmd(CliDriver.java:184)
        at org.apache.hadoop.hive.cli.CliDriver.processLine(CliDriver.java:403)
        at org.apache.hadoop.hive.cli.CliDriver.executeDriver(CliDriver.java:821)
        at org.apache.hadoop.hive.cli.CliDriver.run(CliDriver.java:759)
        at org.apache.hadoop.hive.cli.CliDriver.main(CliDriver.java:686)
        at sun.reflect.NativeMethodAccessorImpl.invoke0(Native Method)
        at sun.reflect.NativeMethodAccessorImpl.invoke(NativeMethodAccessorImpl.java:62)
        at sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccessorImpl.java:43)
        at java.lang.reflect.Method.invoke(Method.java:498)
        at org.apache.hadoop.util.RunJar.run(RunJar.java:244)
        at org.apache.hadoop.util.RunJar.main(RunJar.java:158)
FAILED: ParseException line 1:5 cannot recognize input near 'show' 'databses' '<EOF>' in ddl statement
hive> show databases;
OK
default
Time taken: 0.025 seconds, Fetched: 1 row(s)
hive>
```

Terminating the EMR Cluster:

- After completing our analysis, we should terminate the EMR cluster.


Services
Search for services, features, marketplace products, and docs [Alt+S]
upgradsmaranikasahoo @ 6666-4632-3686
N. Virginia
Support
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
Cluster: CaseStudycluster
Waiting
Cluster ready after last step completed.
Summary
Application user interfaces
Monitoring
Hardware
Configurations
Events
Steps
Bootstrap actions
Summary
ID: j-1GU5E92OPZ8II
Creation date: 2021-10-20 21:38 (UTC+5:30)
Elapsed time: 1 hour, 2 minutes
After last step completes: Cluster waits
Termination protection: Off Change
Tags: -- View All / Edit
Master public DNS: ec2-3-84-164-145.compute-1.amazonaws.com
Connect to the Master Node Using SSH
Configuration details
Release label: emr-5.33.1
Hadoop distribution: Amazon 2.10.1
Applications: Hive 2.3.7, Pig 0.17.0, Hue 4.9.0
Log URI: s3://aws-logs-666646323686-us-east-1/elasticmapreduce/
EMRFS consistent view: Disabled
Custom AMI ID: --
Application user interfaces
Persistent user interfaces: YARN timeline server, Tez UI
On-cluster user interfaces: Not Enabled Enable an SSH Connection
Network and hardware
Availability zone: us-east-1a
Subnet ID: subnet-06fc05dd385c6e605
Master: Running 1 m4.large
Core: Running 1 m4.large
Task: --
Cluster scaling: Not enabled
Auto-termination: Not enabled
Security and access
Key name: casestudykeys
EC2 instance profile: EMR_EC2_DefaultRole


Terminate cluster
Are you sure you want to terminate this cluster?
Any pending work or data residing on the cluster will be lost, such as data stored in HDFS. This action is irreversible.
Cancel
Terminate
Feedback
English (US)
© 2008 - 2021, Amazon Internet Services Private Ltd. or its affiliates. All rights reserved.
Privacy Policy
Terms of Use
Cookie preferences


aws
Services
Search for services, features, marketplace products, and docs [Alt+S]
upgradsmaranikasahoo @ 6666-4632-3686
N. Virginia
Support
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
Create cluster
View details
Clone
Terminate
Filter:
All clusters
Filter clusters...
3 clusters (all loaded)
Name
ID
Status
Creation time (UTC+5:30)
Elapsed time
Normalized instance hours
CaseStudycluster
j-1GU5E92OPZ8II
Terminating
User request
2021-10-20 21:38 (UTC+5:30)
1 hour, 3 minutes
0
CaseStudycluster
j-197J1CU03ZUK
Terminated
User request
2021-10-19 21:47 (UTC+5:30)
1 hour, 25 minutes
16
CaseStudycluster
j-10Q2XLR2OXZ4Q
Terminated
User request
2021-10-18 22:10 (UTC+5:30)
47 minutes
8