HIVE CASE STUDY

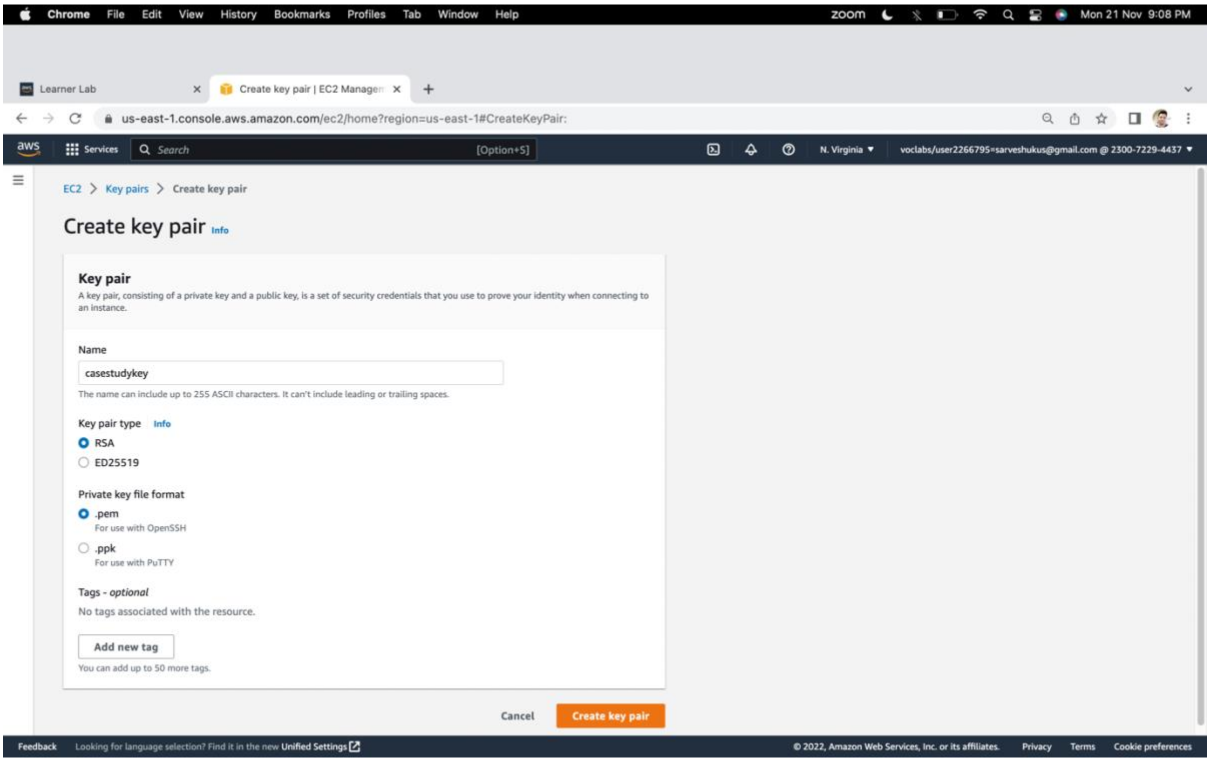
CASE STUDY

Submitted by: Aditya Punjabi, Sarvesh Sharma & Keshav Gupta

UPGRAD|DATASCIENCE-C42|COURSE-3|HIVE CASE STUDY-DA TRACK

STEPS FOR CASE STUDY

1. Creating and generating a KEY PAIR:-


Chrome File Edit View History Bookmarks Profiles Tab Window Help
zoom
Mon 21 Nov 9:08 PM
Learner Lab
Create key pair | EC2 Managem
us-east-1.console.aws.amazon.com/ec2/home?region=us-east-1#CreateKeyPair:
aws
Services
Search
[Option+S]
N. Virginia
voclabs/user2266795=sarveshukus@gmail.com @ 2300-7229-4437
EC2 > Key pairs > Create key pair

Create key pair Info

Key pair

A key pair, consisting of a private key and a public key, is a set of security credentials that you use to prove your identity when connecting to an instance.

Name

casestudykey

The name can include up to 255 ASCII characters. It can't include leading or trailing spaces.

Key pair type Info

● RSA

○ ED25519

Private key file format

● .pem
For use with OpenSSH

○ .ppk
For use with PuTTY

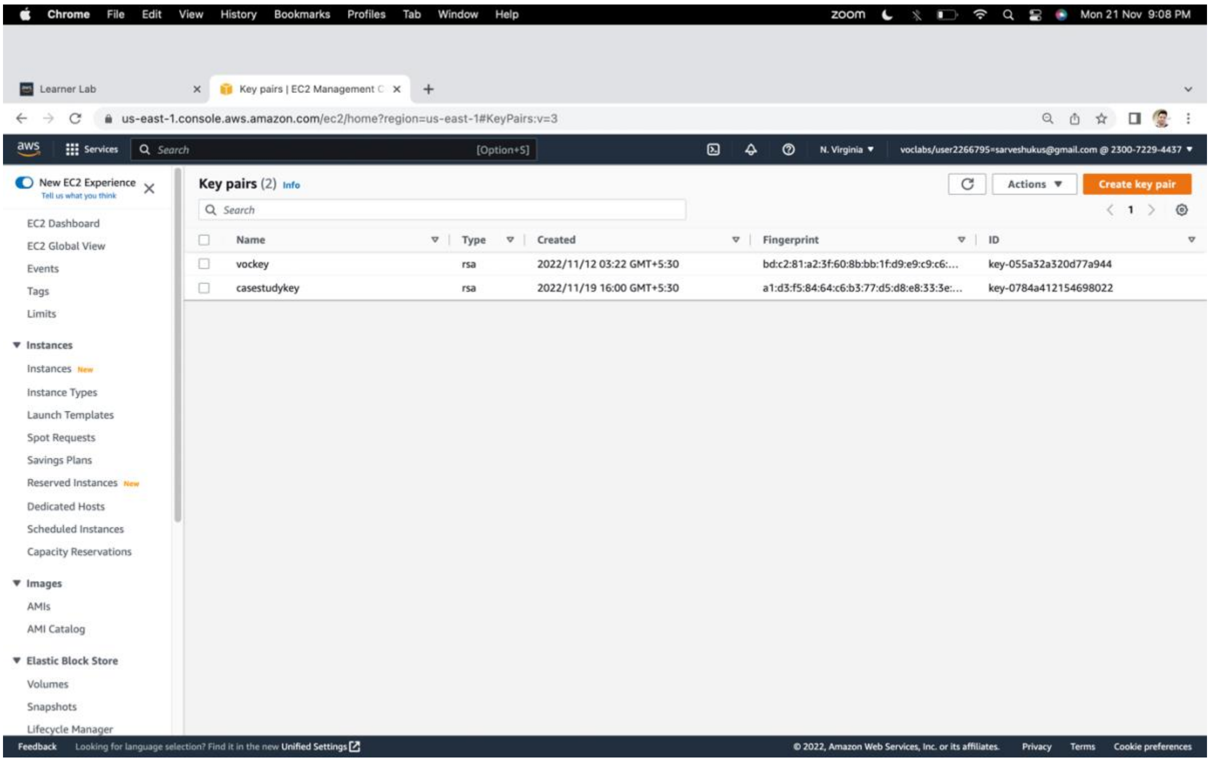
Tags - *optional*

No tags associated with the resource.

Add new tag

You can add up to 50 more tags.

Cancel
Create key pair
Feedback
Looking for language selection? Find it in the new Unified Settings
© 2022, Amazon Web Services, Inc. or its affiliates.
Privacy
Terms
Cookie preferences

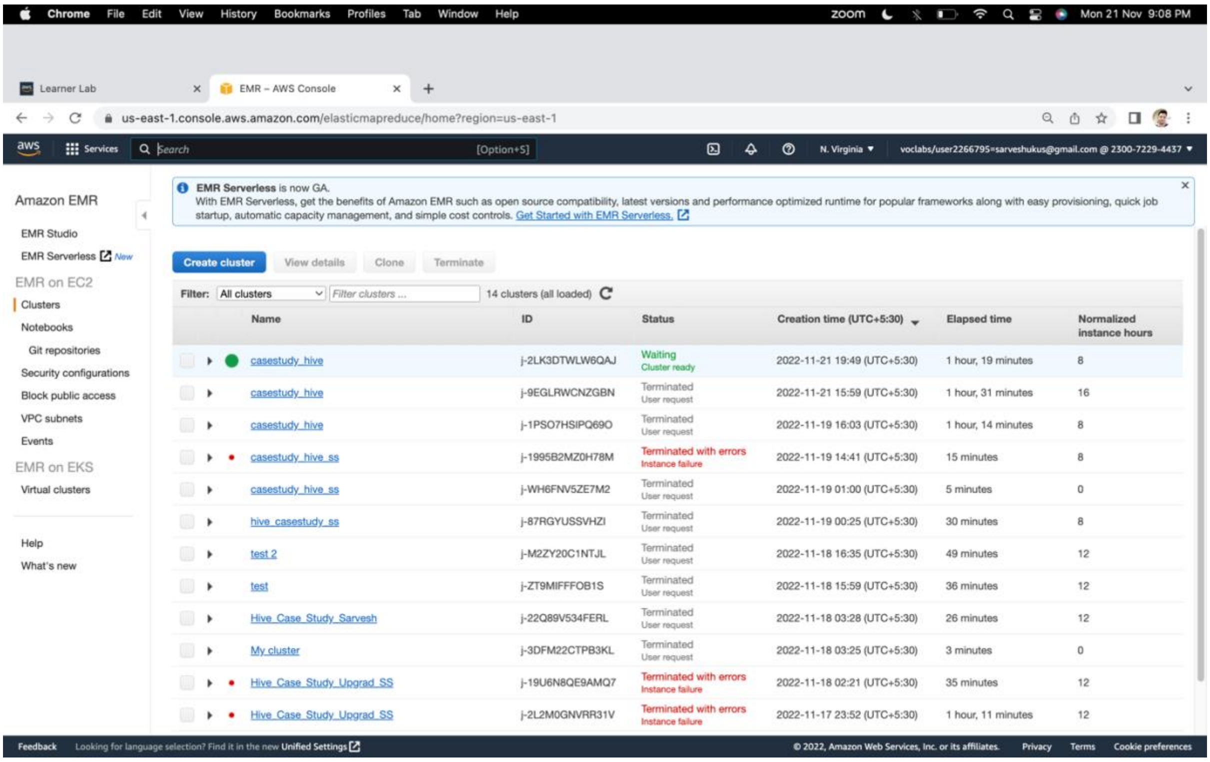


Key pairs (2) Info

	Name	Type	Created	Fingerprint	ID
□	vockey	rsa	2022/11/12 03:22 GMT+5:30	bd:c2:81:a2:3f:60:8b:bb:1f:d9:e9:c9:c6:...	key-055a32a320d77a944
□	casestudykey	rsa	2022/11/19 16:00 GMT+5:30	a1:d3:f5:84:64:c6:b3:77:d5:d8:e8:33:3e:...	key-0784a412154698022

Feedback
Looking for language selection? Find it in the new Unified Settings
© 2022, Amazon Web Services, Inc. or its affiliates.
Privacy
Terms
Cookie preferences

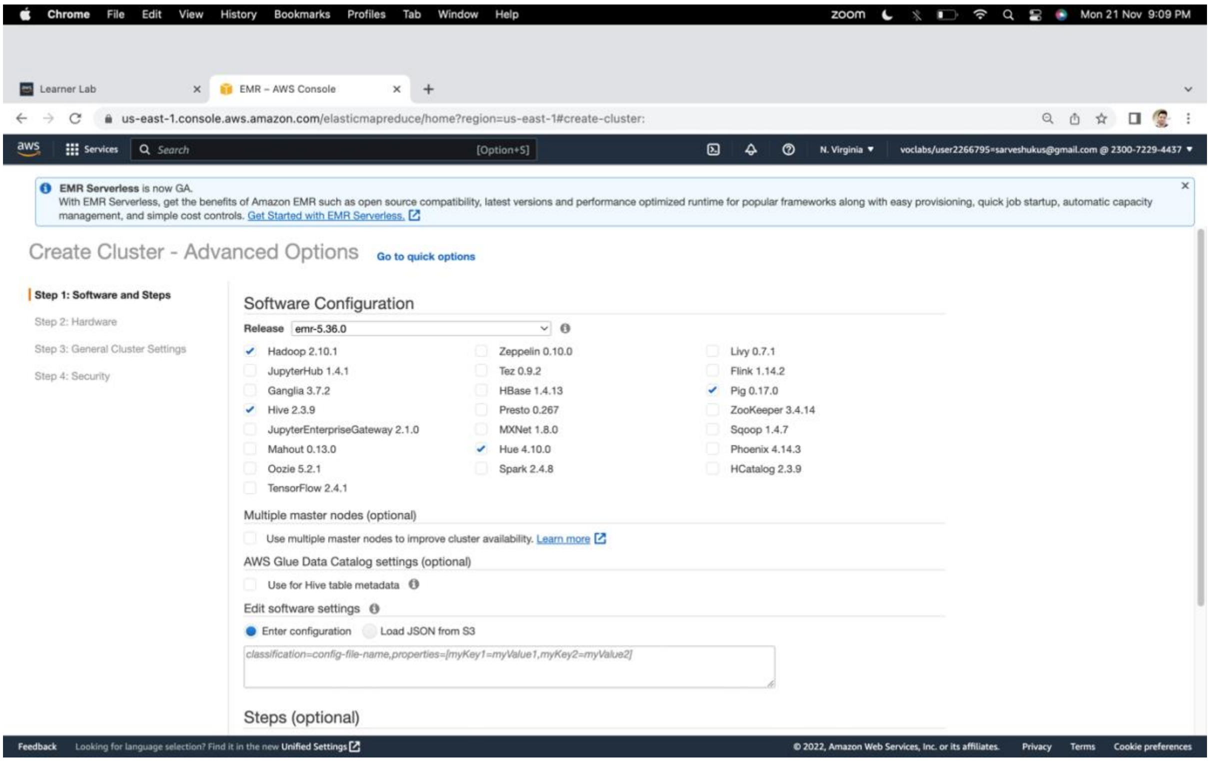
2. Creating EMR Cluster



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.

Filter: All clusters | Filter clusters ... | 14 clusters (all loaded)

Name	ID	Status	Creation time (UTC+5:30)	Elapsed time	Normalized instance hours
● casestudy_hive	j-2LK3DTWLW6QAJ	Waiting Cluster ready	2022-11-21 19:49 (UTC+5:30)	1 hour, 19 minutes	8
casestudy_hive	j-9EGLRWCNZGBN	Terminated User request	2022-11-21 15:59 (UTC+5:30)	1 hour, 31 minutes	16
casestudy_hive	j-1PSO7HSIPQ69O	Terminated User request	2022-11-19 16:03 (UTC+5:30)	1 hour, 14 minutes	8
● casestudy_hive_ss	j-1995B2MZ0H78M	Terminated with errors Instance failure	2022-11-19 14:41 (UTC+5:30)	15 minutes	8
casestudy_hive_ss	j-WH6FNV5ZE7M2	Terminated User request	2022-11-19 01:00 (UTC+5:30)	5 minutes	0
hive_casestudy_ss	j-87RGYUSSVHZI	Terminated User request	2022-11-19 00:25 (UTC+5:30)	30 minutes	8
test 2	j-M2ZY20C1NTJL	Terminated User request	2022-11-18 16:35 (UTC+5:30)	49 minutes	12
test	j-ZT9MIFFFOB1S	Terminated User request	2022-11-18 15:59 (UTC+5:30)	36 minutes	12
Hive_Case_Study_Sarvesh	j-22Q89V534FERL	Terminated User request	2022-11-18 03:28 (UTC+5:30)	26 minutes	12
My cluster	j-3DFM22CTPB3KL	Terminated User request	2022-11-18 03:25 (UTC+5:30)	3 minutes	0
● Hive_Case_Study_Upgrad_SS	j-19U6N8QE9AMQ7	Terminated with errors Instance failure	2022-11-18 02:21 (UTC+5:30)	35 minutes	12
● Hive_Case_Study_Upgrad_SS	j-2L2M0GNVRR31V	Terminated with errors Instance failure	2022-11-17 23:52 (UTC+5:30)	1 hour, 11 minutes	12


Chrome File Edit View History Bookmarks Profiles Tab Window Help
zoom
Mon 21 Nov 9:09 PM
Learner Lab
EMR – AWS Console
us-east-1.console.aws.amazon.com/elasticmapreduce/home?region=us-east-1#create-cluster:
Services
Search
[Option+S]
N. Virginia
voclabs/user2266795=sarveshukus@gmail.com @ 2300-7229-4437
Go to quick options
Step 1: Software and Steps
Step 2: Hardware
Step 3: General Cluster Settings
Step 4: Security
Feedback
Looking for language selection? Find it in the new Unified Settings
© 2022, Amazon Web Services, Inc. or its affiliates.
Privacy
Terms
Cookie preferences

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.

Create Cluster - Advanced Options

Software Configuration

Release: emr-5.36.0

☑ Hadoop 2.10.1
☐ JupyterHub 1.4.1
☐ Ganglia 3.7.2
☑ Hive 2.3.9
☐ JupyterEnterpriseGateway 2.1.0
☐ Mahout 0.13.0
☐ Oozie 5.2.1
☐ TensorFlow 2.4.1
☐ Zeppelin 0.10.0
☐ Tez 0.9.2
☐ HBase 1.4.13
☐ Presto 0.267
☐ MXNet 1.8.0
☑ Hue 4.10.0
☐ Spark 2.4.8
☐ Livy 0.7.1
☐ Flink 1.14.2
☑ Pig 0.17.0
☐ ZooKeeper 3.4.14
☐ Sqoop 1.4.7
☐ Phoenix 4.14.3
☐ HCatalog 2.3.9

Multiple master nodes (optional)

☐ Use multiple master nodes to improve cluster availability. Learn more

AWS Glue Data Catalog settings (optional)

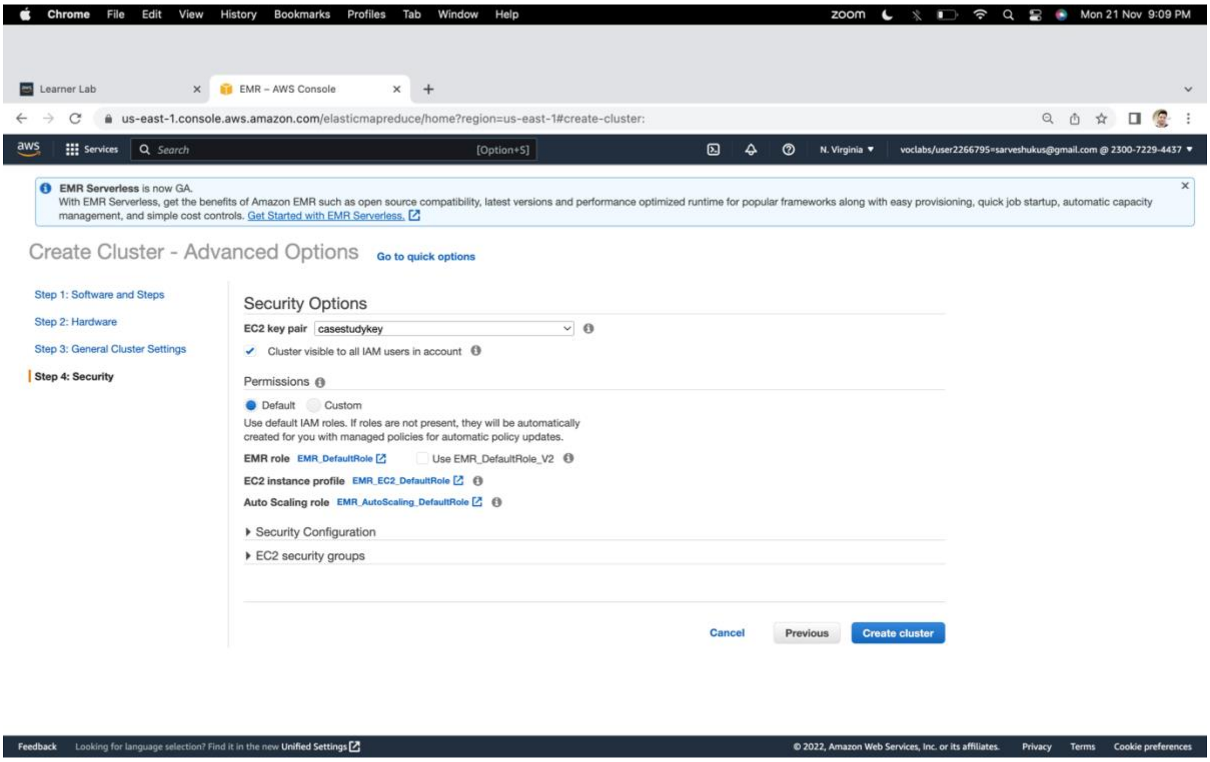
☐ Use for Hive table metadata

Edit software settings

● Enter configuration ○ Load JSON from S3

classification=config-file-name,properties=[myKey1=myValue1,myKey2=myValue2]

Steps (optional)


Chrome File Edit View History Bookmarks Profiles Tab Window Help
zoom
Mon 21 Nov 9:09 PM
Learner Lab
EMR – AWS Console
Services
Search
[Option+S]
N. Virginia
voclabs/user2266795=sarveshukus@gmail.com @ 2300-7229-4437
Step 1: Software and Steps
Step 2: Hardware
Step 3: General Cluster Settings
Step 4: Security
Cancel
Previous
Create cluster
Feedback
Looking for language selection? Find it in the new Unified Settings
© 2022, Amazon Web Services, Inc. or its affiliates.
Privacy
Terms
Cookie preferences

us-east-1.console.aws.amazon.com/elasticmapreduce/home?region=us-east-1#create-cluster:

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.

Create Cluster - Advanced Options   Go to quick options

Security Options

EC2 key pair: casestudykey

☑ Cluster visible to all IAM users in account

Permissions

● Default ○ Custom

Use default IAM roles. If roles are not present, they will be automatically created for you with managed policies for automatic policy updates.

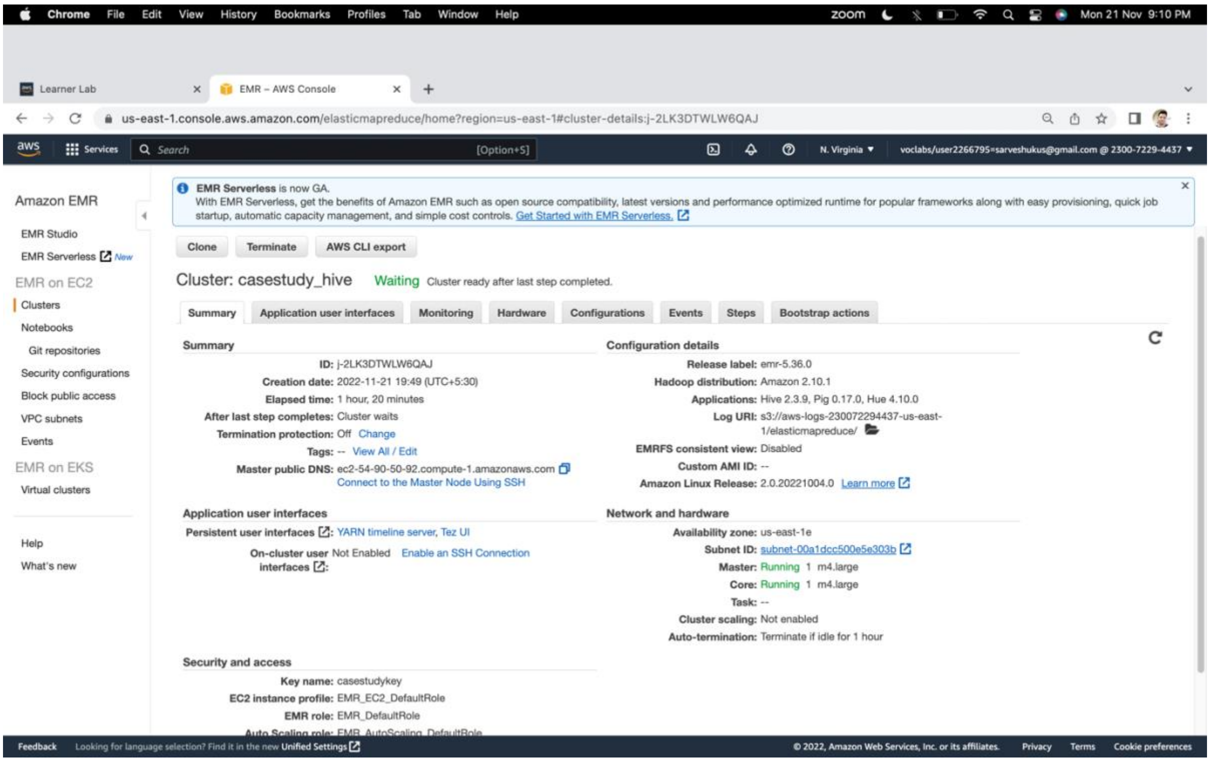
EMR role: EMR_DefaultRole   □ Use EMR_DefaultRole_V2

EC2 instance profile: EMR_EC2_DefaultRole

Auto Scaling role: EMR_AutoScaling_DefaultRole

▸ Security Configuration

▸ EC2 security groups



us-east-1.console.aws.amazon.com/elasticmapreduce/home?region=us-east-1#cluster-details:j-2LK3DTWLW6QAJ

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.

Cluster: casestudy_hive   Waiting   Cluster ready after last step completed.

Summary

ID: j-2LK3DTWLW6QAJ
Creation date: 2022-11-21 19:49 (UTC+5:30)
Elapsed time: 1 hour, 20 minutes
After last step completes: Cluster waits
Termination protection: Off Change
Tags: -- View All / Edit
Master public DNS: ec2-54-90-50-92.compute-1.amazonaws.com
Connect to the Master Node Using SSH

Configuration details

Release label: emr-5.36.0
Hadoop distribution: Amazon 2.10.1
Applications: Hive 2.3.9, Pig 0.17.0, Hue 4.10.0
Log URI: s3://aws-logs-230072294437-us-east-1/elasticmapreduce/
EMRFS consistent view: Disabled
Custom AMI ID: --
Amazon Linux Release: 2.0.20221004.0 Learn more

Application user interfaces

Persistent user interfaces: YARN timeline server, Tez UI
On-cluster user interfaces: Not Enabled Enable an SSH Connection

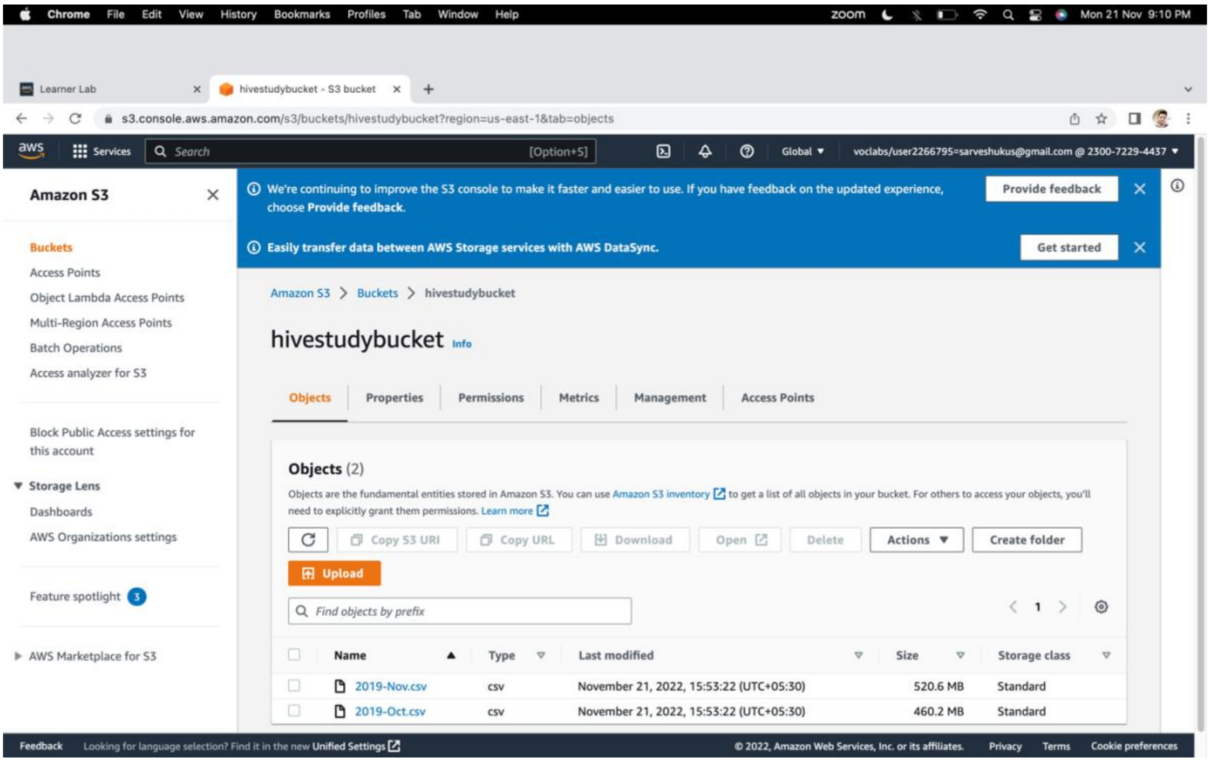
Network and hardware

Availability zone: us-east-1e
Subnet ID: subnet-00a1dcc500e5e303b
Master: Running 1 m4.large
Core: Running 1 m4.large
Task: --
Cluster scaling: Not enabled
Auto-termination: Terminate if idle for 1 hour

Security and access

Key name: casestudykey
EC2 instance profile: EMR_EC2_DefaultRole
EMR role: EMR_DefaultRole
Auto Scaling role: EMR_AutoScaling_DefaultRole

3. S3 Bucket to store data files


Chrome File Edit View History Bookmarks Profiles Tab Window Help
zoom
Mon 21 Nov 9:10 PM
Learner Lab
hivestudybucket - S3 bucket
s3.console.aws.amazon.com/s3/buckets/hivestudybucket?region=us-east-1&tab=objects
aws
Services
Search
[Option+S]
Global
voclabs/user2266795=sarveshukus@gmail.com @ 2300-7229-4437
Amazon S3
Buckets
Access Points
Object Lambda Access Points
Multi-Region Access Points
Batch Operations
Access analyzer for S3
Block Public Access settings for this account
Storage Lens
Dashboards
AWS Organizations settings
Feature spotlight 3
AWS Marketplace for S3
Provide feedback
Get started

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**.

Easily transfer data between AWS Storage services with AWS DataSync.

Amazon S3 > Buckets > hivestudybucket

hivestudybucket Info

Objects
Properties
Permissions
Metrics
Management
Access Points

Objects (2)

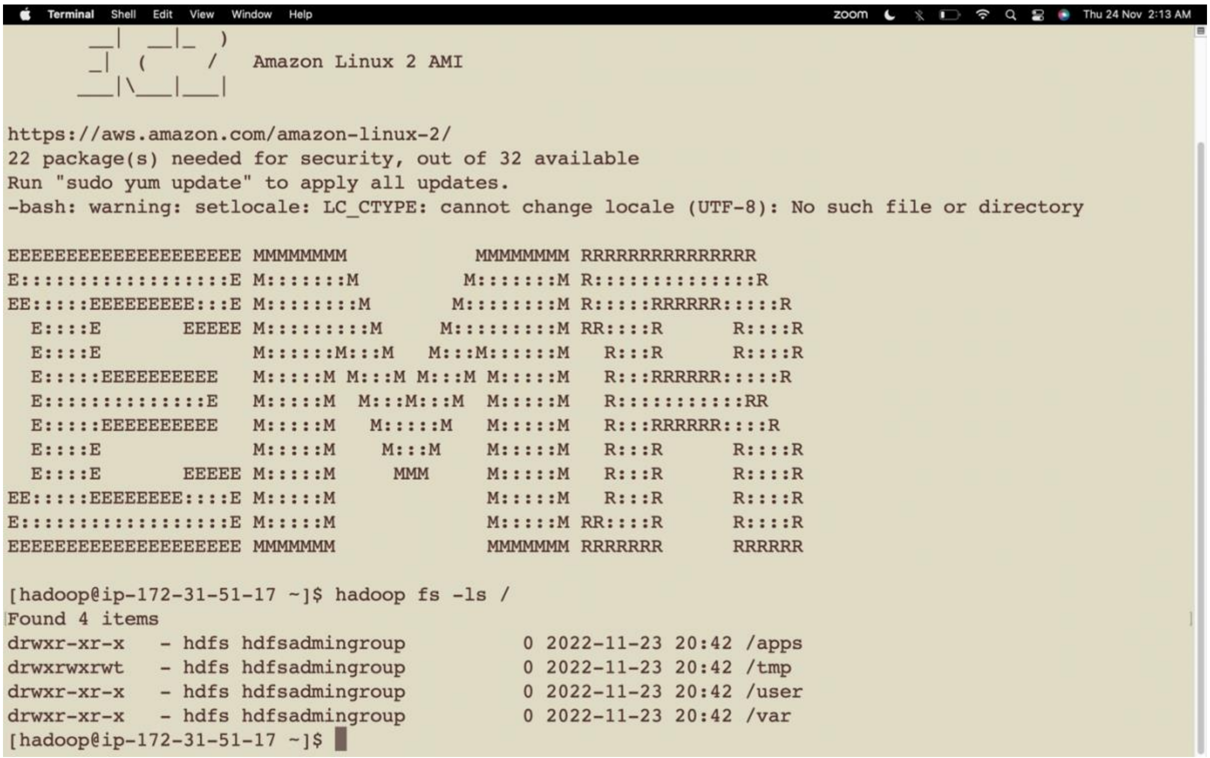
Objects are the fundamental entities stored in Amazon S3. You can use Amazon S3 inventory to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. Learn more

□	Name	Type	Last modified	Size	Storage class
□	2019-Nov.csv	csv	November 21, 2022, 15:53:22 (UTC+05:30)	520.6 MB	Standard
□	2019-Oct.csv	csv	November 21, 2022, 15:53:22 (UTC+05:30)	460.2 MB	Standard

Feedback
Looking for language selection? Find it in the new Unified Settings
© 2022, Amazon Web Services, Inc. or its affiliates.
Privacy
Terms
Cookie preferences

4. Command to check for already present directories in HDFS

- hadoop fs -ls /



```
       __|  __|_  )
       _|  (     /   Amazon Linux 2 AMI
      ___|\___|___|

https://aws.amazon.com/amazon-linux-2/
22 package(s) needed for security, out of 32 available
Run "sudo yum update" to apply all updates.
-bash: warning: setlocale: LC_CTYPE: cannot change locale (UTF-8): No such file or directory

EEEEEEEEEEEEEEEEEEEE MMMMMMMM           MMMMMMMM RRRRRRRRRRRRRRR
E::::::::::::::::::E M:::::::M         M:::::::M R::::::::::::::R
EE:::::EEEEEEEEE:::E M::::::::M       M::::::::M R:::::RRRRRR:::::R
  E::::E       EEEEE M:::::::::M     M:::::::::M RR::::R      R::::R
  E::::E             M::::::M:::M   M:::M::::::M   R:::R      R::::R
  E:::::EEEEEEEEEE   M:::::M M:::M M:::M M:::::M   R:::RRRRRR:::::R
  E::::::::::::::E   M:::::M  M:::M:::M  M:::::M   R:::::::::::RR
  E:::::EEEEEEEEEE   M:::::M   M:::::M   M:::::M   R:::RRRRRR::::R
  E::::E             M:::::M    M:::M    M:::::M   R:::R      R::::R
  E::::E       EEEEE M:::::M     MMM     M:::::M   R:::R      R::::R
EE:::::EEEEEEEE::::E M:::::M             M:::::M   R:::R      R::::R
E::::::::::::::::::E M:::::M             M:::::M RR::::R      R::::R
EEEEEEEEEEEEEEEEEEEE MMMMMMM             MMMMMMM RRRRRRR      RRRRRR

[hadoop@ip-172-31-51-17 ~]$ hadoop fs -ls /
Found 4 items
drwxr-xr-x   - hdfs hdfsadmingroup          0 2022-11-23 20:42 /apps
drwxrwxrwt   - hdfs hdfsadmingroup          0 2022-11-23 20:42 /tmp
drwxr-xr-x   - hdfs hdfsadmingroup          0 2022-11-23 20:42 /user
drwxr-xr-x   - hdfs hdfsadmingroup          0 2022-11-23 20:42 /var
[hadoop@ip-172-31-51-17 ~]$ █
```

Acumen:

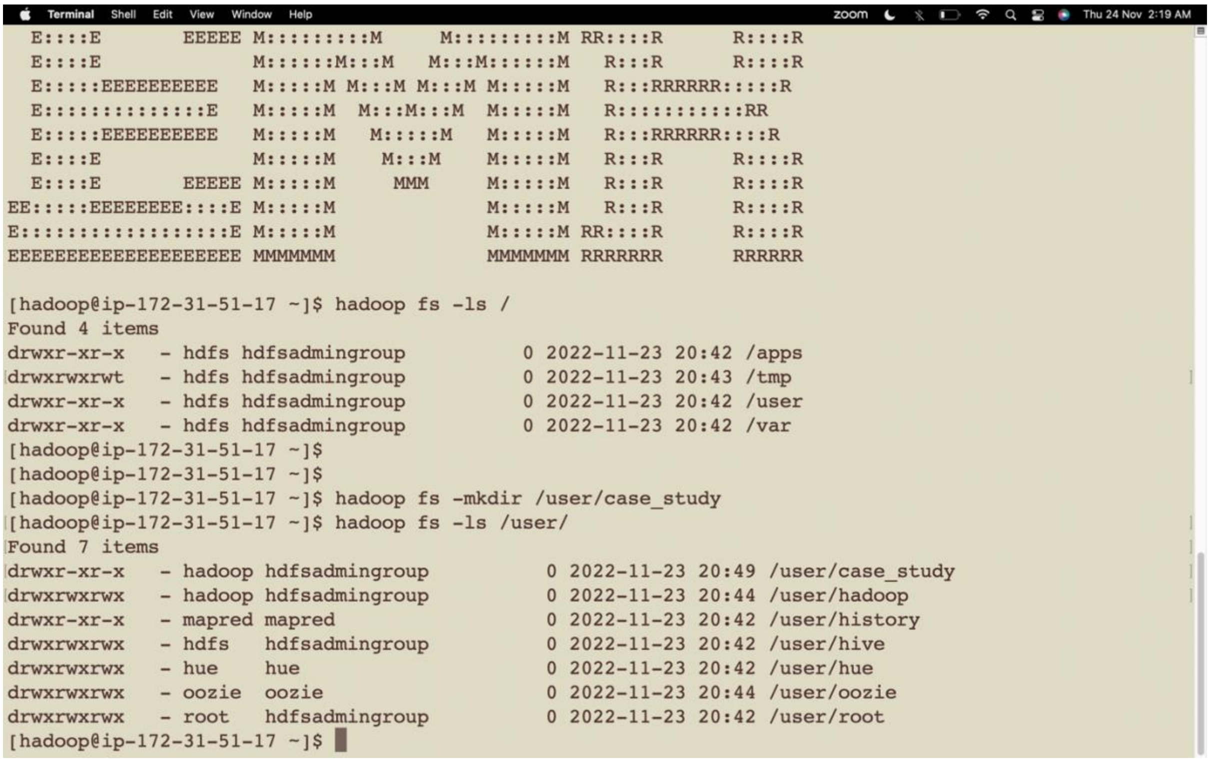
- All the above directories are in-built in HDFS.
- Either these directories can be used to create our temporary directory to store data files or create a separate temporary directory.

5. Creating new temporary directory i.e., 'case_study' to store data file in the already present directory (Permanent) i.e., 'user'

- hadoop fs -mkdir /user/case_study

6. Command to check creation of new temporary Directory in ‘user’ directory

- hadoop fs -ls /user/



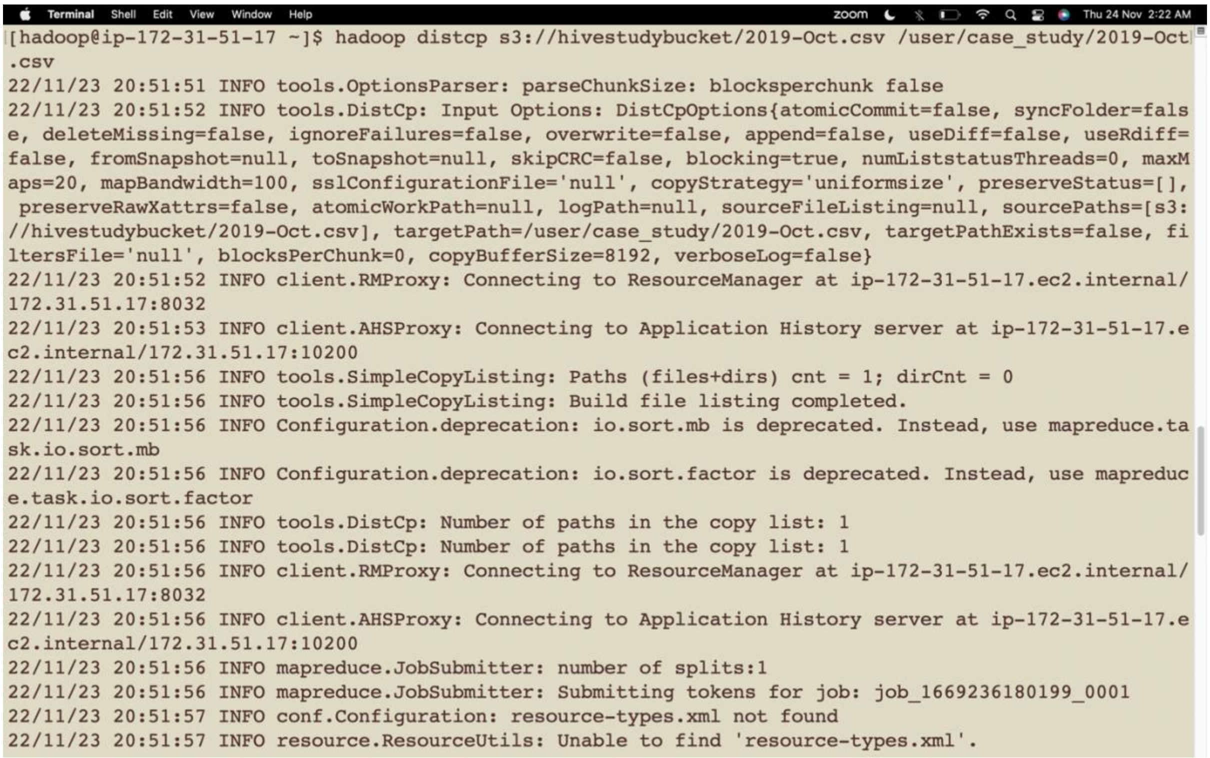
```
  E::::E       EEEEE M:::::::::M     M:::::::::M RR::::R      R::::R
  E::::E             M::::::M:::M   M:::M::::::M   R:::R      R::::R
  E:::::EEEEEEEEEE   M:::::M M:::M M:::M M:::::M   R:::RRRRRR:::::R
  E::::::::::::::E   M:::::M  M:::M:::M  M:::::M   R:::::::::::RR
  E:::::EEEEEEEEEE   M:::::M   M:::::M   M:::::M   R:::RRRRRR::::R
  E::::E             M:::::M    M:::M    M:::::M   R:::R      R::::R
  E::::E       EEEEE M:::::M     MMM     M:::::M   R:::R      R::::R
EE:::::EEEEEEEE::::E M:::::M             M:::::M   R:::R      R::::R
E::::::::::::::::::E M:::::M             M:::::M RR::::R      R::::R
EEEEEEEEEEEEEEEEEEEE MMMMMMM             MMMMMMM RRRRRRR      RRRRRR

[hadoop@ip-172-31-51-17 ~]$ hadoop fs -ls /
Found 4 items
drwxr-xr-x   - hdfs hdfsadmingroup          0 2022-11-23 20:42 /apps
drwxrwxrwt   - hdfs hdfsadmingroup          0 2022-11-23 20:43 /tmp
drwxr-xr-x   - hdfs hdfsadmingroup          0 2022-11-23 20:42 /user
drwxr-xr-x   - hdfs hdfsadmingroup          0 2022-11-23 20:42 /var
[hadoop@ip-172-31-51-17 ~]$
[hadoop@ip-172-31-51-17 ~]$
[hadoop@ip-172-31-51-17 ~]$ hadoop fs -mkdir /user/case_study
[hadoop@ip-172-31-51-17 ~]$ hadoop fs -ls /user/
Found 7 items
drwxr-xr-x   - hadoop hdfsadmingroup          0 2022-11-23 20:49 /user/case_study
drwxrwxrwx   - hadoop hdfsadmingroup          0 2022-11-23 20:44 /user/hadoop
drwxr-xr-x   - mapred mapred                  0 2022-11-23 20:42 /user/history
drwxrwxrwx   - hdfs   hdfsadmingroup          0 2022-11-23 20:42 /user/hive
drwxrwxrwx   - hue    hue                     0 2022-11-23 20:42 /user/hue
drwxrwxrwx   - oozie  oozie                   0 2022-11-23 20:44 /user/oozie
drwxrwxrwx   - root   hdfsadmingroup          0 2022-11-23 20:42 /user/root
[hadoop@ip-172-31-51-17 ~]$ 
```

Acumen:

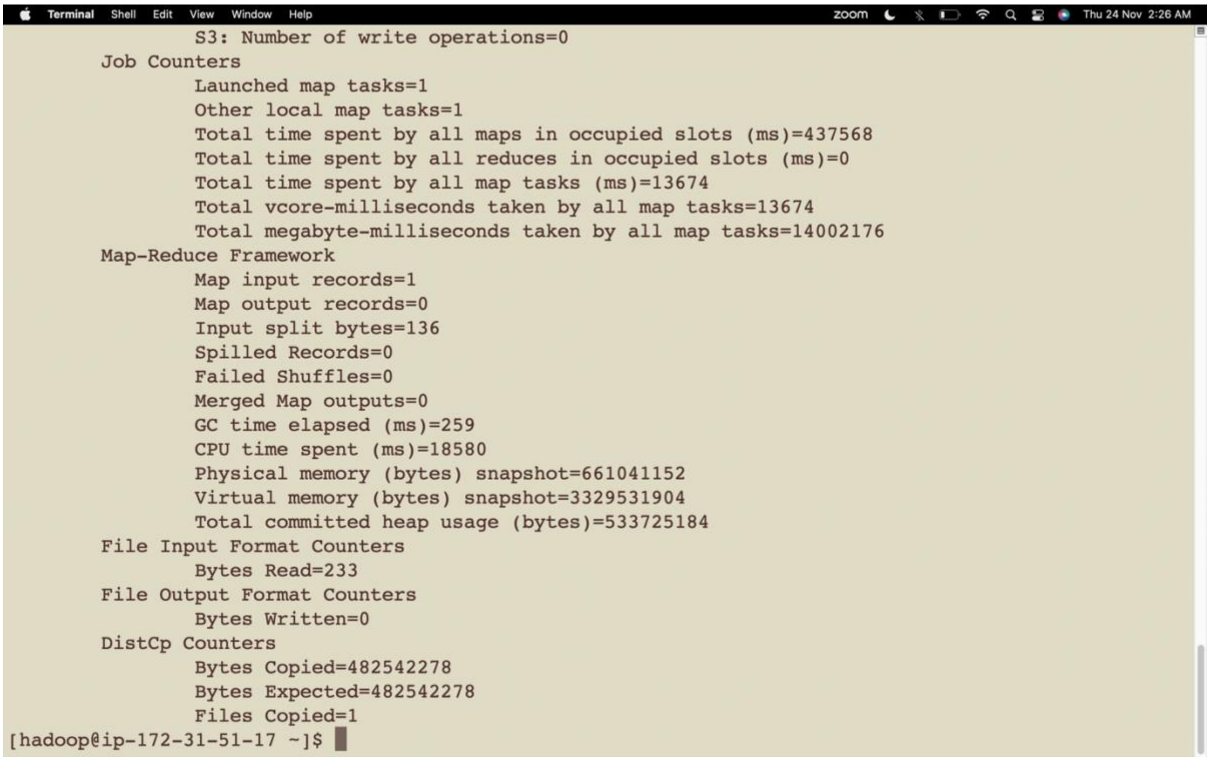
- There will always be some files within the permanent directories of the HDFS.

7. Command to load data files '2019-Oct.csv' from S3 storage into HDFS storage as '2019-Oct.csv'

- hadoop distcp s3://hivestudybucket/2019-Oct.csv /user/case_study/2019-Oct.csv



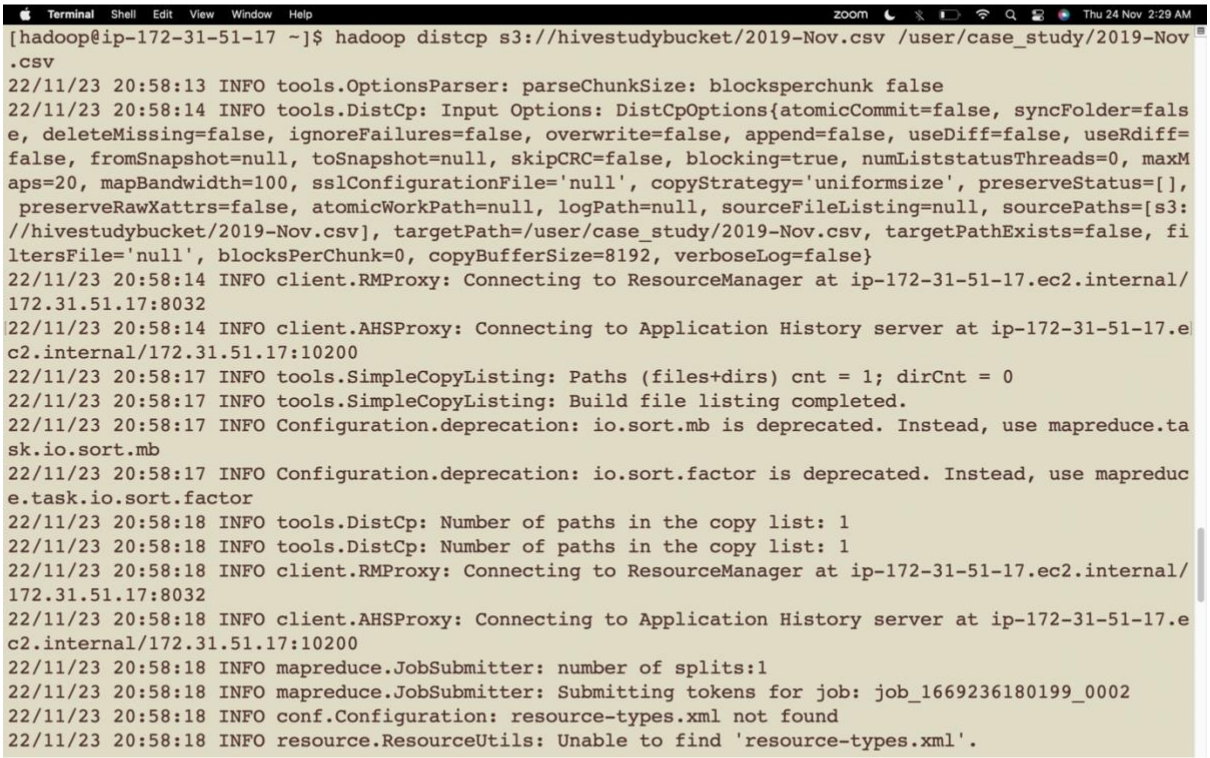
```
[hadoop@ip-172-31-51-17 ~]$ hadoop distcp s3://hivestudybucket/2019-Oct.csv /user/case_study/2019-Oct
.csv
22/11/23 20:51:51 INFO tools.OptionsParser: parseChunkSize: blocksperchunk false
22/11/23 20:51:52 INFO tools.DistCp: Input Options: DistCpOptions{atomicCommit=false, syncFolder=fals
e, deleteMissing=false, ignoreFailures=false, overwrite=false, append=false, useDiff=false, useRdiff=
false, fromSnapshot=null, toSnapshot=null, skipCRC=false, blocking=true, numListstatusThreads=0, maxM
aps=20, mapBandwidth=100, sslConfigurationFile='null', copyStrategy='uniformsize', preserveStatus=[],
 preserveRawXattrs=false, atomicWorkPath=null, logPath=null, sourceFileListing=null, sourcePaths=[s3:
//hivestudybucket/2019-Oct.csv], targetPath=/user/case_study/2019-Oct.csv, targetPathExists=false, fi
ltersFile='null', blocksPerChunk=0, copyBufferSize=8192, verboseLog=false}
22/11/23 20:51:52 INFO client.RMProxy: Connecting to ResourceManager at ip-172-31-51-17.ec2.internal/
172.31.51.17:8032
22/11/23 20:51:53 INFO client.AHSProxy: Connecting to Application History server at ip-172-31-51-17.e
c2.internal/172.31.51.17:10200
22/11/23 20:51:56 INFO tools.SimpleCopyListing: Paths (files+dirs) cnt = 1; dirCnt = 0
22/11/23 20:51:56 INFO tools.SimpleCopyListing: Build file listing completed.
22/11/23 20:51:56 INFO Configuration.deprecation: io.sort.mb is deprecated. Instead, use mapreduce.ta
sk.io.sort.mb
22/11/23 20:51:56 INFO Configuration.deprecation: io.sort.factor is deprecated. Instead, use mapreduc
e.task.io.sort.factor
22/11/23 20:51:56 INFO tools.DistCp: Number of paths in the copy list: 1
22/11/23 20:51:56 INFO tools.DistCp: Number of paths in the copy list: 1
22/11/23 20:51:56 INFO client.RMProxy: Connecting to ResourceManager at ip-172-31-51-17.ec2.internal/
172.31.51.17:8032
22/11/23 20:51:56 INFO client.AHSProxy: Connecting to Application History server at ip-172-31-51-17.e
c2.internal/172.31.51.17:10200
22/11/23 20:51:56 INFO mapreduce.JobSubmitter: number of splits:1
22/11/23 20:51:56 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1669236180199_0001
22/11/23 20:51:57 INFO conf.Configuration: resource-types.xml not found
22/11/23 20:51:57 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
```



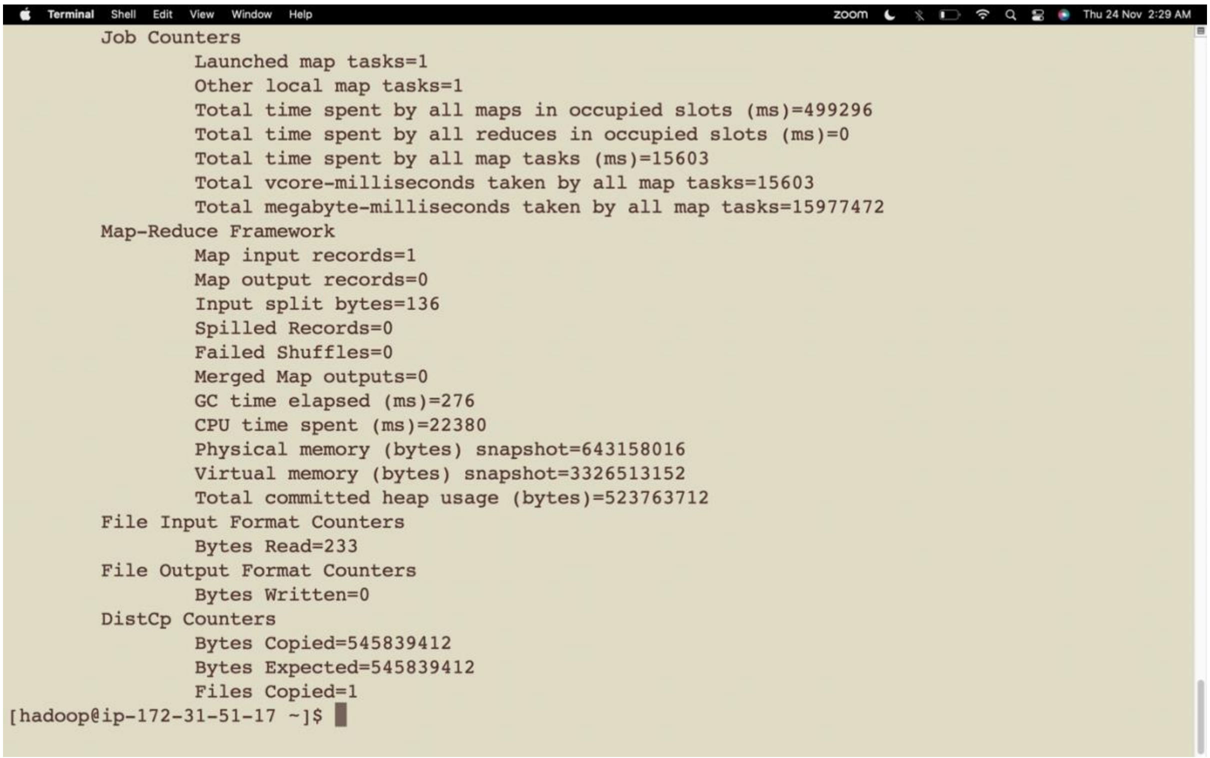
```
                S3: Number of write operations=0
        Job Counters
                Launched map tasks=1
                Other local map tasks=1
                Total time spent by all maps in occupied slots (ms)=437568
                Total time spent by all reduces in occupied slots (ms)=0
                Total time spent by all map tasks (ms)=13674
                Total vcore-milliseconds taken by all map tasks=13674
                Total megabyte-milliseconds taken by all map tasks=14002176
        Map-Reduce Framework
                Map input records=1
                Map output records=0
                Input split bytes=136
                Spilled Records=0
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=259
                CPU time spent (ms)=18580
                Physical memory (bytes) snapshot=661041152
                Virtual memory (bytes) snapshot=3329531904
                Total committed heap usage (bytes)=533725184
        File Input Format Counters
                Bytes Read=233
        File Output Format Counters
                Bytes Written=0
        DistCp Counters
                Bytes Copied=482542278
                Bytes Expected=482542278
                Files Copied=1
[hadoop@ip-172-31-51-17 ~]$ █
```

8. Command to load data files ‘2019-Nov.csv’ from S3 storage into HDFS storage as '2019-Nov.csv'

    - hadoop distcp s3://hivestudybucket/2019-Nov.csv /user/case_study/2019-Nov.csv



```
[hadoop@ip-172-31-51-17 ~]$ hadoop distcp s3://hivestudybucket/2019-Nov.csv /user/case_study/2019-Nov
.csv
22/11/23 20:58:13 INFO tools.OptionsParser: parseChunkSize: blocksperchunk false
22/11/23 20:58:14 INFO tools.DistCp: Input Options: DistCpOptions{atomicCommit=false, syncFolder=fals
e, deleteMissing=false, ignoreFailures=false, overwrite=false, append=false, useDiff=false, useRdiff=
false, fromSnapshot=null, toSnapshot=null, skipCRC=false, blocking=true, numListstatusThreads=0, maxM
aps=20, mapBandwidth=100, sslConfigurationFile='null', copyStrategy='uniformsize', preserveStatus=[],
 preserveRawXattrs=false, atomicWorkPath=null, logPath=null, sourceFileListing=null, sourcePaths=[s3:
//hivestudybucket/2019-Nov.csv], targetPath=/user/case_study/2019-Nov.csv, targetPathExists=false, fi
ltersFile='null', blocksPerChunk=0, copyBufferSize=8192, verboseLog=false}
22/11/23 20:58:14 INFO client.RMProxy: Connecting to ResourceManager at ip-172-31-51-17.ec2.internal/
172.31.51.17:8032
22/11/23 20:58:14 INFO client.AHSProxy: Connecting to Application History server at ip-172-31-51-17.e
c2.internal/172.31.51.17:10200
22/11/23 20:58:17 INFO tools.SimpleCopyListing: Paths (files+dirs) cnt = 1; dirCnt = 0
22/11/23 20:58:17 INFO tools.SimpleCopyListing: Build file listing completed.
22/11/23 20:58:17 INFO Configuration.deprecation: io.sort.mb is deprecated. Instead, use mapreduce.ta
sk.io.sort.mb
22/11/23 20:58:17 INFO Configuration.deprecation: io.sort.factor is deprecated. Instead, use mapreduc
e.task.io.sort.factor
22/11/23 20:58:18 INFO tools.DistCp: Number of paths in the copy list: 1
22/11/23 20:58:18 INFO tools.DistCp: Number of paths in the copy list: 1
22/11/23 20:58:18 INFO client.RMProxy: Connecting to ResourceManager at ip-172-31-51-17.ec2.internal/
172.31.51.17:8032
22/11/23 20:58:18 INFO client.AHSProxy: Connecting to Application History server at ip-172-31-51-17.e
c2.internal/172.31.51.17:10200
22/11/23 20:58:18 INFO mapreduce.JobSubmitter: number of splits:1
22/11/23 20:58:18 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1669236180199_0002
22/11/23 20:58:18 INFO conf.Configuration: resource-types.xml not found
22/11/23 20:58:18 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
```



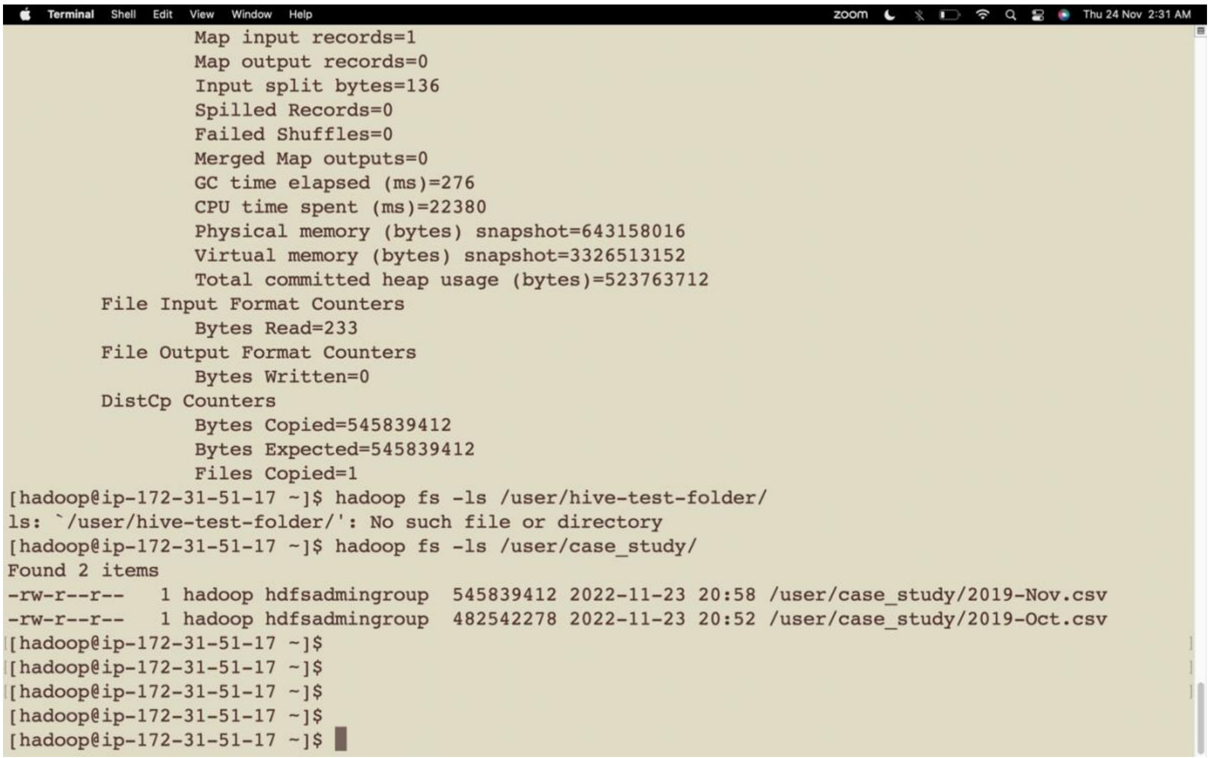
```
        Job Counters
                Launched map tasks=1
                Other local map tasks=1
                Total time spent by all maps in occupied slots (ms)=499296
                Total time spent by all reduces in occupied slots (ms)=0
                Total time spent by all map tasks (ms)=15603
                Total vcore-milliseconds taken by all map tasks=15603
                Total megabyte-milliseconds taken by all map tasks=15977472
        Map-Reduce Framework
                Map input records=1
                Map output records=0
                Input split bytes=136
                Spilled Records=0
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=276
                CPU time spent (ms)=22380
                Physical memory (bytes) snapshot=643158016
                Virtual memory (bytes) snapshot=3326513152
                Total committed heap usage (bytes)=523763712
        File Input Format Counters
                Bytes Read=233
        File Output Format Counters
                Bytes Written=0
        DistCp Counters
                Bytes Copied=545839412
                Bytes Expected=545839412
                Files Copied=1
[hadoop@ip-172-31-51-17 ~]$ █
```

9. Command to check successful loading of data files into the already created new temporary directory of HDFS i.e., 'case_study'

- hadoop fs -ls /user/case_study

Output: Found 2 items

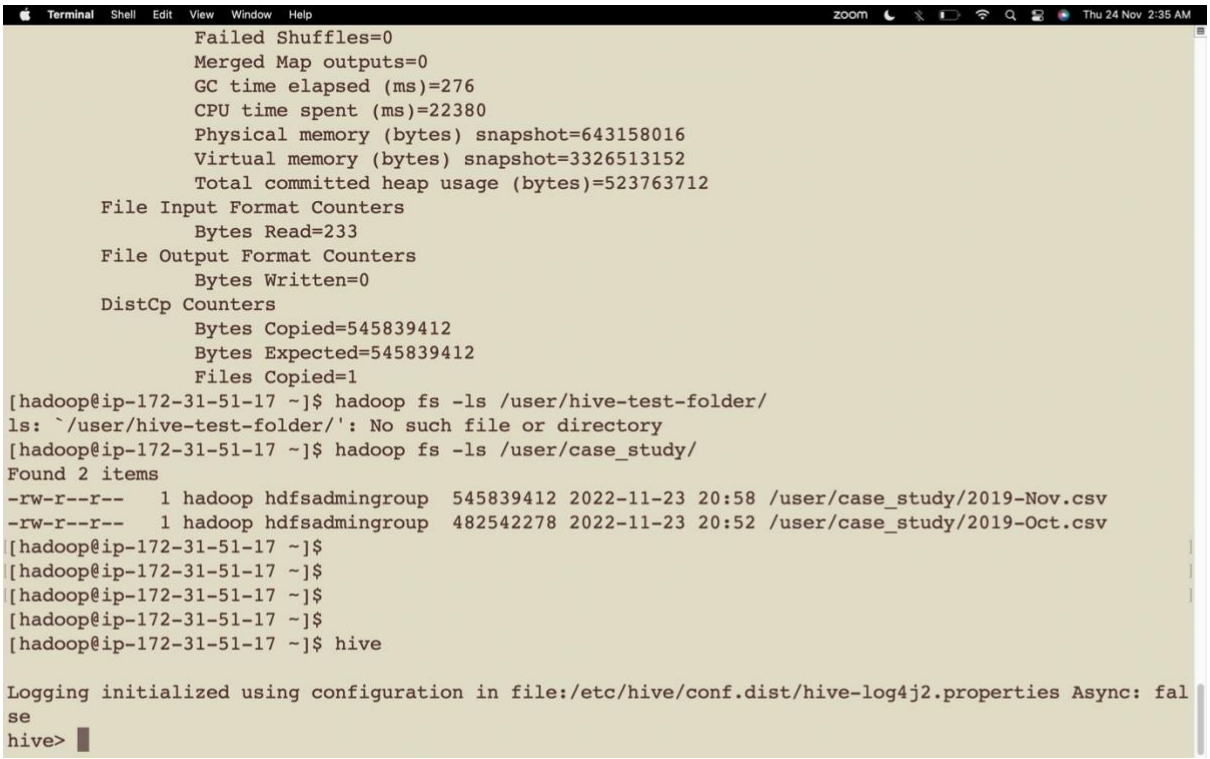
-rw-r--r-- 1 hadoop hadoop 545839412 2022-11-23 20:58 /user/case_study/2019-Nov.csv
-rw-r--r-- 1 hadoop hadoop 482542278 2022-11-23 20:52 /user/case_study/2019-Oct.csv



```
                Map input records=1
                Map output records=0
                Input split bytes=136
                Spilled Records=0
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=276
                CPU time spent (ms)=22380
                Physical memory (bytes) snapshot=643158016
                Virtual memory (bytes) snapshot=3326513152
                Total committed heap usage (bytes)=523763712
        File Input Format Counters
                Bytes Read=233
        File Output Format Counters
                Bytes Written=0
        DistCp Counters
                Bytes Copied=545839412
                Bytes Expected=545839412
                Files Copied=1
[hadoop@ip-172-31-51-17 ~]$ hadoop fs -ls /user/hive-test-folder/
ls: `/user/hive-test-folder/': No such file or directory
[hadoop@ip-172-31-51-17 ~]$ hadoop fs -ls /user/case_study/
Found 2 items
-rw-r--r--   1 hadoop hdfsadmingroup  545839412 2022-11-23 20:58 /user/case_study/2019-Nov.csv
-rw-r--r--   1 hadoop hdfsadmingroup  482542278 2022-11-23 20:52 /user/case_study/2019-Oct.csv
[hadoop@ip-172-31-51-17 ~]$
[hadoop@ip-172-31-51-17 ~]$
[hadoop@ip-172-31-51-17 ~]$
[hadoop@ip-172-31-51-17 ~]$
[hadoop@ip-172-31-51-17 ~]$
```

10. Command to start Hive system

- hive



```
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=276
                CPU time spent (ms)=22380
                Physical memory (bytes) snapshot=643158016
                Virtual memory (bytes) snapshot=3326513152
                Total committed heap usage (bytes)=523763712
        File Input Format Counters
                Bytes Read=233
        File Output Format Counters
                Bytes Written=0
        DistCp Counters
                Bytes Copied=545839412
                Bytes Expected=545839412
                Files Copied=1
[hadoop@ip-172-31-51-17 ~]$ hadoop fs -ls /user/hive-test-folder/
ls: `/user/hive-test-folder/': No such file or directory
[hadoop@ip-172-31-51-17 ~]$ hadoop fs -ls /user/case_study/
Found 2 items
-rw-r--r--   1 hadoop hdfsadmingroup  545839412 2022-11-23 20:58 /user/case_study/2019-Nov.csv
-rw-r--r--   1 hadoop hdfsadmingroup  482542278 2022-11-23 20:52 /user/case_study/2019-Oct.csv
[hadoop@ip-172-31-51-17 ~]$
[hadoop@ip-172-31-51-17 ~]$
[hadoop@ip-172-31-51-17 ~]$
[hadoop@ip-172-31-51-17 ~]$
[hadoop@ip-172-31-51-17 ~]$ hive

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

11. Creating a table i.e., 'ecommerce' which will hold the data for both the data files stored in temporary directory of HDFS.

Query:

```
create table if not exists ecommerce (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 delimited fields terminated by ','
   > lines terminated by '\n' stored as textfile
   > location '/user/case_study'
   > tblproperties ("skip.header.line.count"="1");
```

12. Command to enable heading in the output

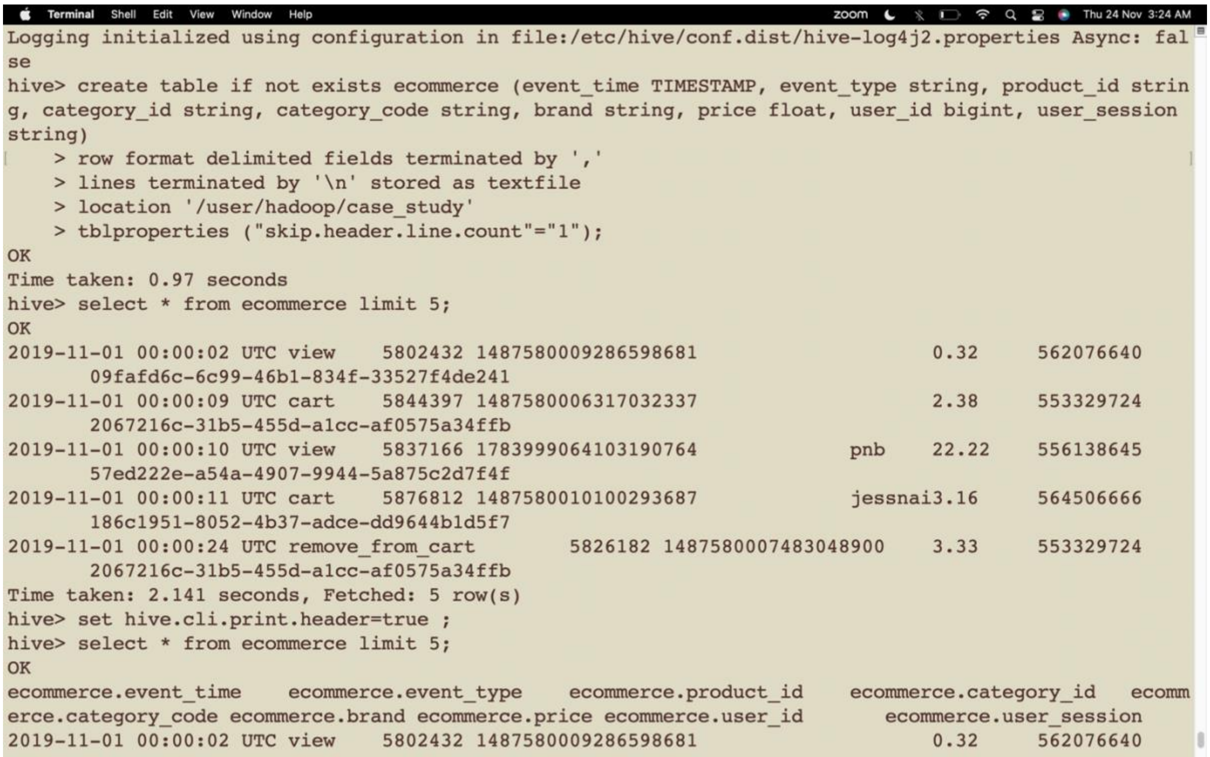
- set hive.cli.print.header=true ;

13. Simple HiveQL command to check successful creation of table and storage of data from both data files into table

    Query:

    SELECT * FROM ecommerce limit 5;

    Output:



```
Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j2.properties Async: fal
se
hive> create table if not exists ecommerce (event_time TIMESTAMP, event_type string, product_id strin
g, category_id string, category_code string, brand string, price float, user_id bigint, user_session
string)
    > row format delimited fields terminated by ','
    > lines terminated by '\n' stored as textfile
    > location '/user/hadoop/case_study'
    > tblproperties ("skip.header.line.count"="1");
OK
Time taken: 0.97 seconds
hive> select * from ecommerce 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             jessnai3.16      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: 2.141 seconds, Fetched: 5 row(s)
hive> set hive.cli.print.header=true ;
hive> select * from ecommerce limit 5;
OK
ecommerce.event_time    ecommerce.event_type    ecommerce.product_id    ecommerce.category_id   ecomm
erce.category_code ecommerce.brand ecommerce.price ecommerce.user_id       ecommerce.user_session
2019-11-01 00:00:02 UTC view    5802432 1487580009286598681                     0.32     562076640
```

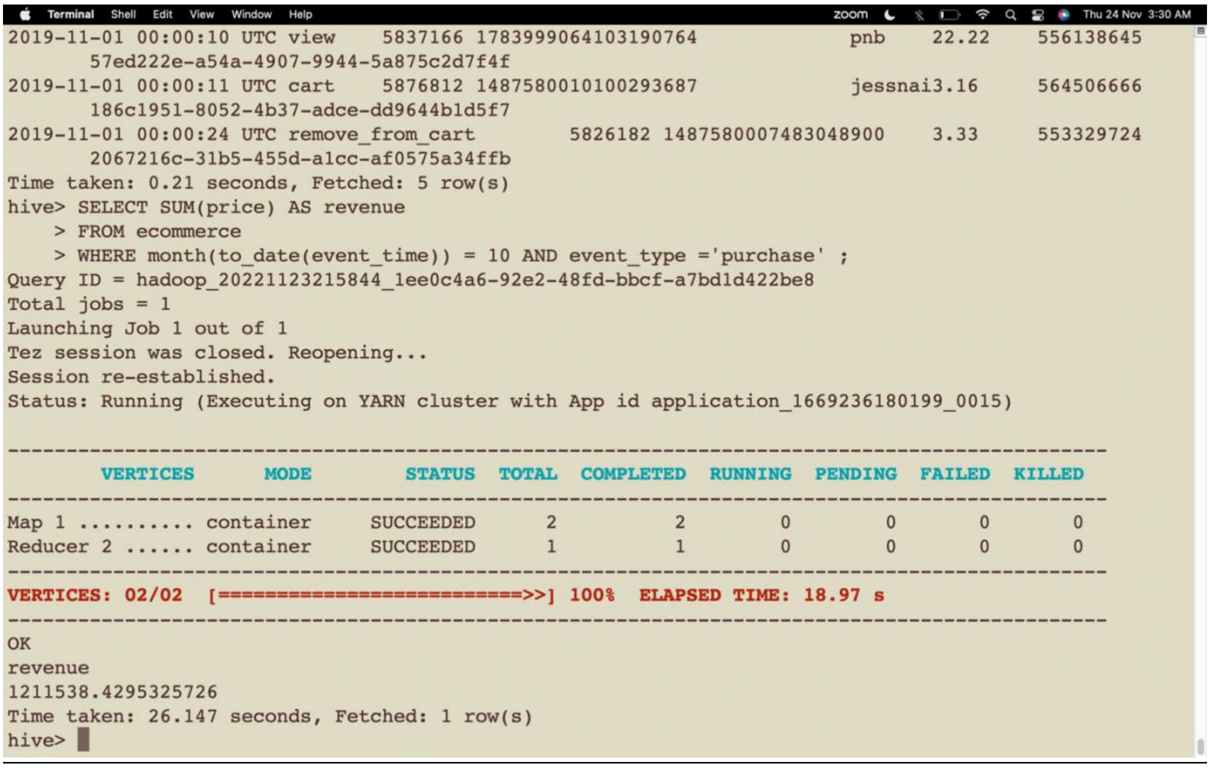
Questions and Answers using Hive Query

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

Query:

SELECT SUM(price) AS revenue
FROM ecommerce
WHERE month(to_date(event_time)) = 10 AND event_type ='purchase' ;

Output:



```
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             jessnai3.16     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: 0.21 seconds, Fetched: 5 row(s)
hive> SELECT SUM(price) AS revenue
    > FROM ecommerce
    > WHERE month(to_date(event_time)) = 10 AND event_type ='purchase' ;
Query ID = hadoop_20221123215844_1ee0c4a6-92e2-48fd-bbcf-a7bd1d422be8
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_1669236180199_0015)

----------------------------------------------------------------------------------------------
        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: 18.97 s
----------------------------------------------------------------------------------------------
OK
revenue
1211538.4295325726
Time taken: 26.147 seconds, Fetched: 1 row(s)
hive> █
```

Insights:

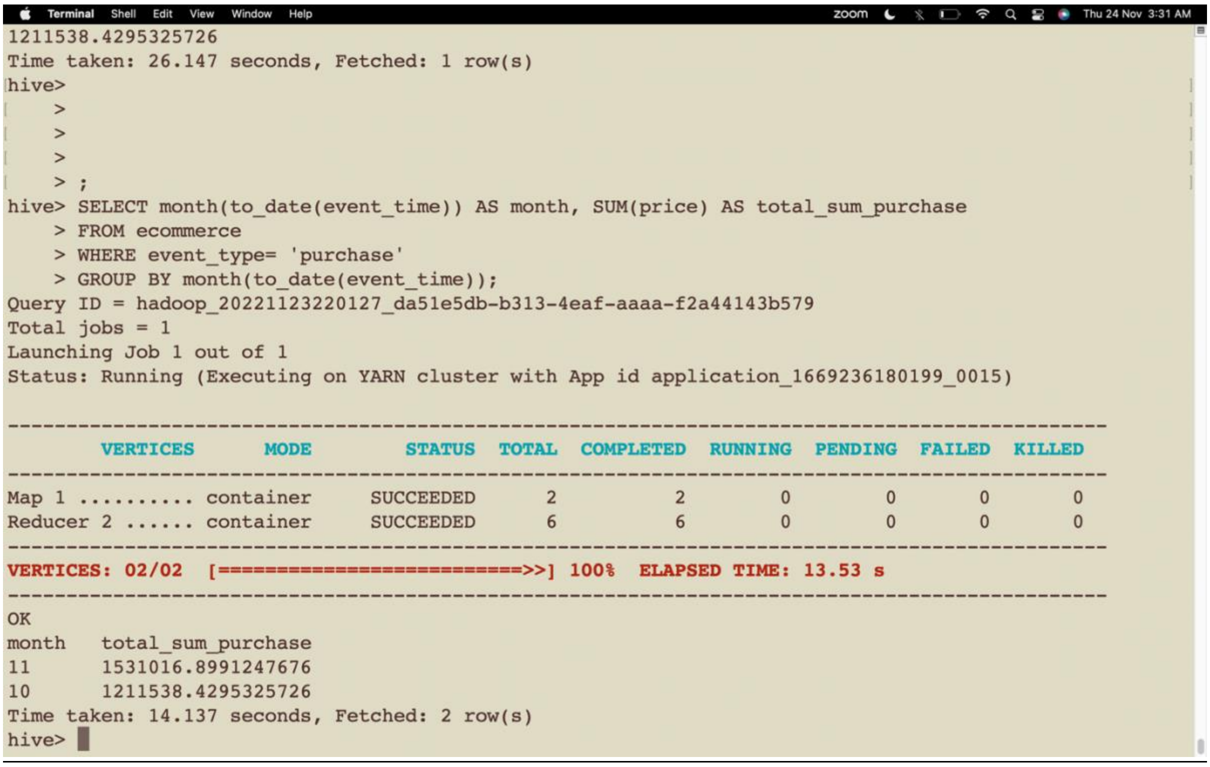
- The total revenue generated based on Purchase in the month of October of 2019 was 1,211,538.43 /-

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

Query:

SELECT month(to_date(event_time)) AS month, SUM(price) AS total_sum_purchase
FROM ecommerce
WHERE event_type= 'purchase'
GROUP BY month(to_date(event_time));

Output:



```
1211538.4295325726
Time taken: 26.147 seconds, Fetched: 1 row(s)
hive>
    >
    >
    >
    > ;
hive> SELECT month(to_date(event_time)) AS month, SUM(price) AS total_sum_purchase
    > FROM ecommerce
    > WHERE event_type= 'purchase'
    > GROUP BY month(to_date(event_time));
Query ID = hadoop_20221123220127_da51e5db-b313-4eaf-aaaa-f2a44143b579
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1669236180199_0015)

----------------------------------------------------------------------------------------------
        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: 13.53 s
----------------------------------------------------------------------------------------------
OK
month   total_sum_purchase
11      1531016.8991247676
10      1211538.4295325726
Time taken: 14.137 seconds, Fetched: 2 row(s)
hive> 
```

Insights:

- It seems to be that there was more purchase made in the month of November (11) i.e., 1,531,016 than in the month of October (10) i.e., 1,211,538.
- Looking at these figures we could assume that the month of November must be more profitable than the month of October. But we can verify our assumption by conducting further investigations.

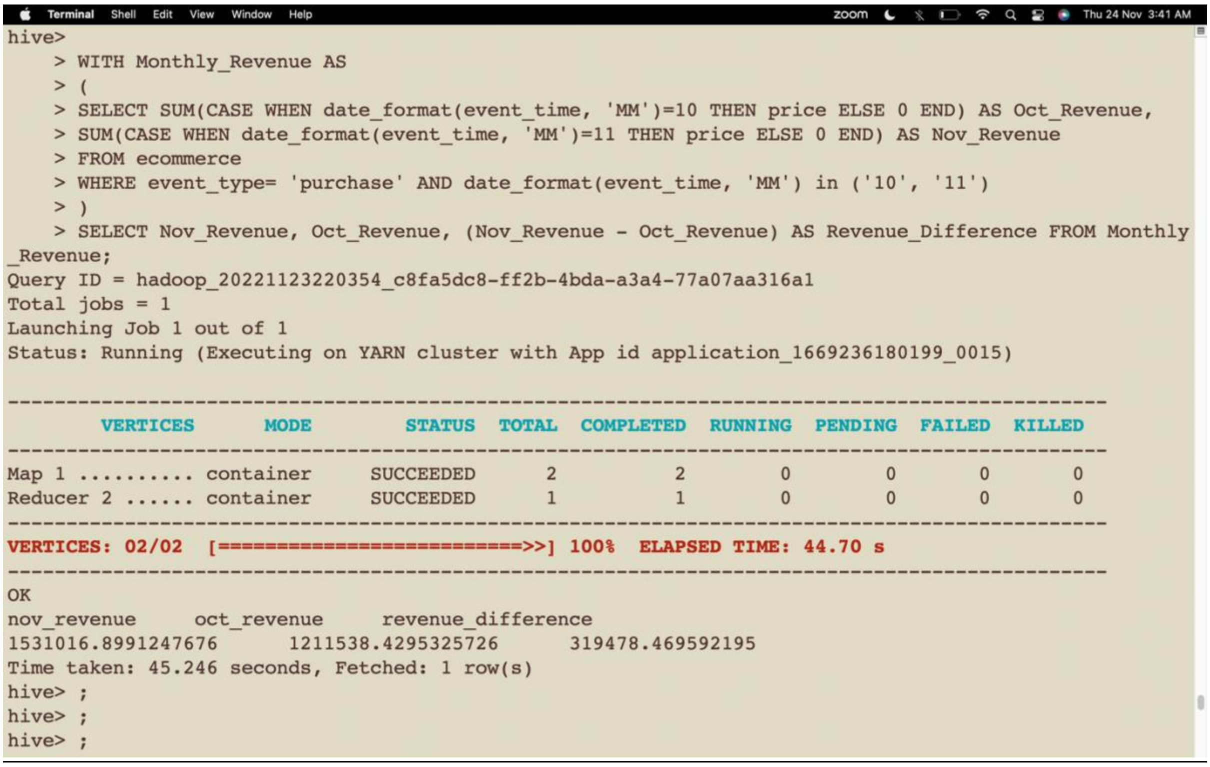
Q3: Write a query to find the change in revenue generated due to purchases from October to November.

Query:

```
WITH Monthly_Revenue AS
(
SELECT 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 ecommerce
WHERE event_type= 'purchase' AND date_format(event_time, 'MM') in ('10', '11')
)
```

SELECT Nov_Revenue, Oct_Revenue, (Nov_Revenue - Oct_Revenue) AS Revenue_Difference FROM Monthly_Revenue;

Output:



```
hive>
    > WITH Monthly_Revenue AS
    > (
    > SELECT 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 ecommerce
    > WHERE event_type= 'purchase' AND date_format(event_time, 'MM') in ('10', '11')
    > )
    > SELECT Nov_Revenue, Oct_Revenue, (Nov_Revenue - Oct_Revenue) AS Revenue_Difference FROM Monthly
_Revenue;
Query ID = hadoop_20221123220354_c8fa5dc8-ff2b-4bda-a3a4-77a07aa316a1
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1669236180199_0015)

----------------------------------------------------------------------------------------------
        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: 44.70 s
----------------------------------------------------------------------------------------------
OK
nov_revenue     oct_revenue     revenue_difference
1531016.8991247676      1211538.4295325726      319478.469592195
Time taken: 45.246 seconds, Fetched: 1 row(s)
hive> ;
hive> ;
hive> ;
```

Insights:

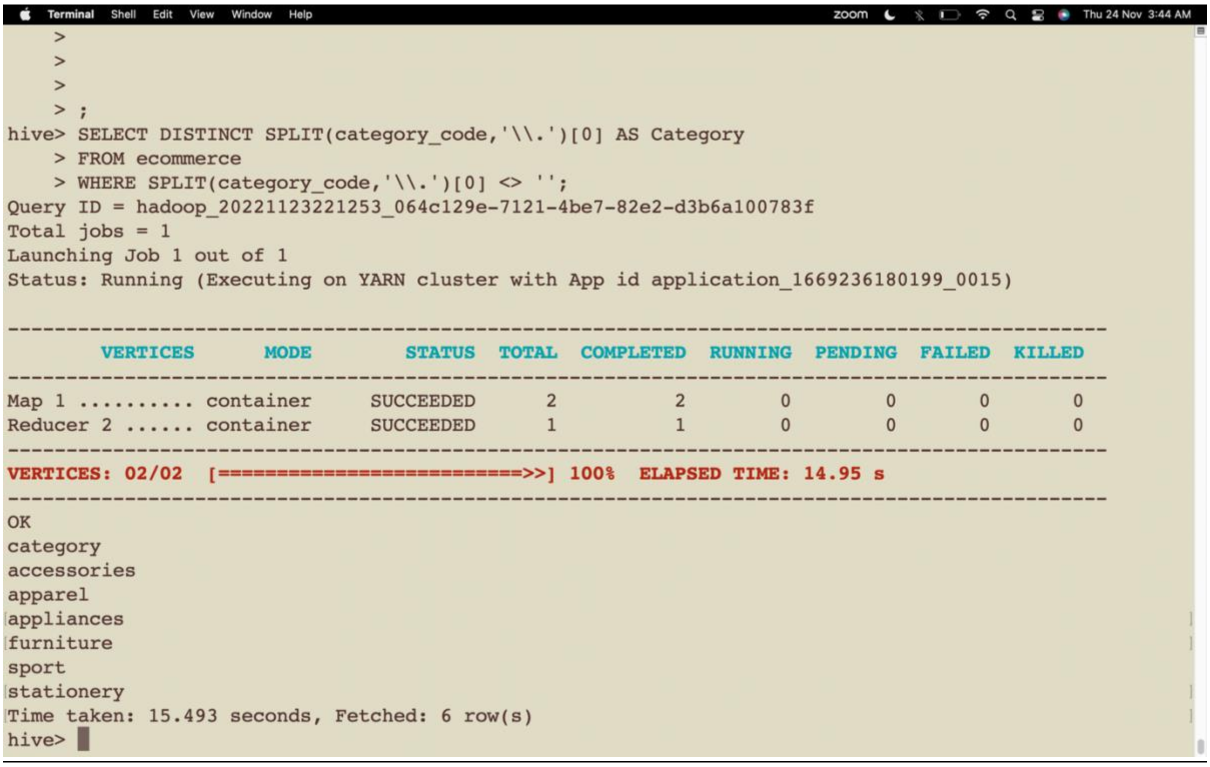
- On the basis of the results considering purchase as event, we could conclude that the revenue generated in November of 2019 was more than the revenue generated in the month of October. In other words, November was more profitable for the company than October.
- Company had a better sale in November, 2019.

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

Query:

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

Output:



```
    >
    >
    >
    > ;
hive> SELECT DISTINCT SPLIT(category_code,'\\.')[0] AS Category
    > FROM ecommerce
    > WHERE SPLIT(category_code,'\\.')[0] <> '';
Query ID = hadoop_20221123221253_064c129e-7121-4be7-82e2-d3b6a100783f
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1669236180199_0015)

----------------------------------------------------------------------------------------------
        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: 14.95 s
----------------------------------------------------------------------------------------------
OK
category
accessories
apparel
appliances
furniture
sport
stationery
Time taken: 15.493 seconds, Fetched: 6 row(s)
hive>
```

Insights:

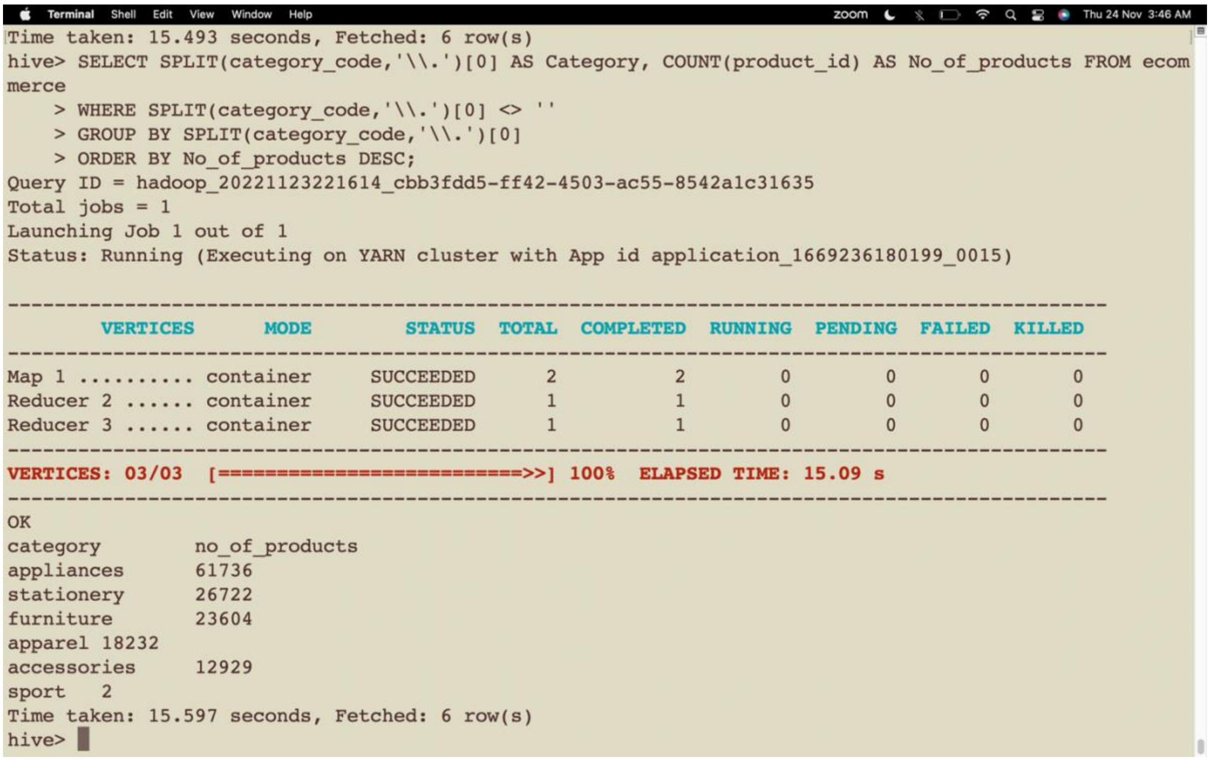
- There is total 6 different categories under which company sells their different products.
- category
  - accessories
  - apparel
  - appliances
  - furniture
  - sport
  - stationery

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

Query:

SELECT SPLIT(category_code,'\\.')[0] AS Category, COUNT(product_id) AS No_of_products
FROM ecommerce
WHERE SPLIT(category_code,'\\.')[0] <> ''
GROUP BY SPLIT(category_code,'\\.')[0]
ORDER BY No_of_products DESC;

Output:



```
Time taken: 15.493 seconds, Fetched: 6 row(s)
hive> SELECT SPLIT(category_code,'\\.')[0] AS Category, COUNT(product_id) AS No_of_products FROM ecom
merce
    > WHERE SPLIT(category_code,'\\.')[0] <> ''
    > GROUP BY SPLIT(category_code,'\\.')[0]
    > ORDER BY No_of_products DESC;
Query ID = hadoop_20221123221614_cbb3fdd5-ff42-4503-ac55-8542a1c31635
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1669236180199_0015)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      2          2        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 15.09 s
----------------------------------------------------------------------------------------------
OK
category        no_of_products
appliances      61736
stationery      26722
furniture       23604
apparel 18232
accessories     12929
sport   2
Time taken: 15.597 seconds, Fetched: 6 row(s)
hive> 
```

Insights:

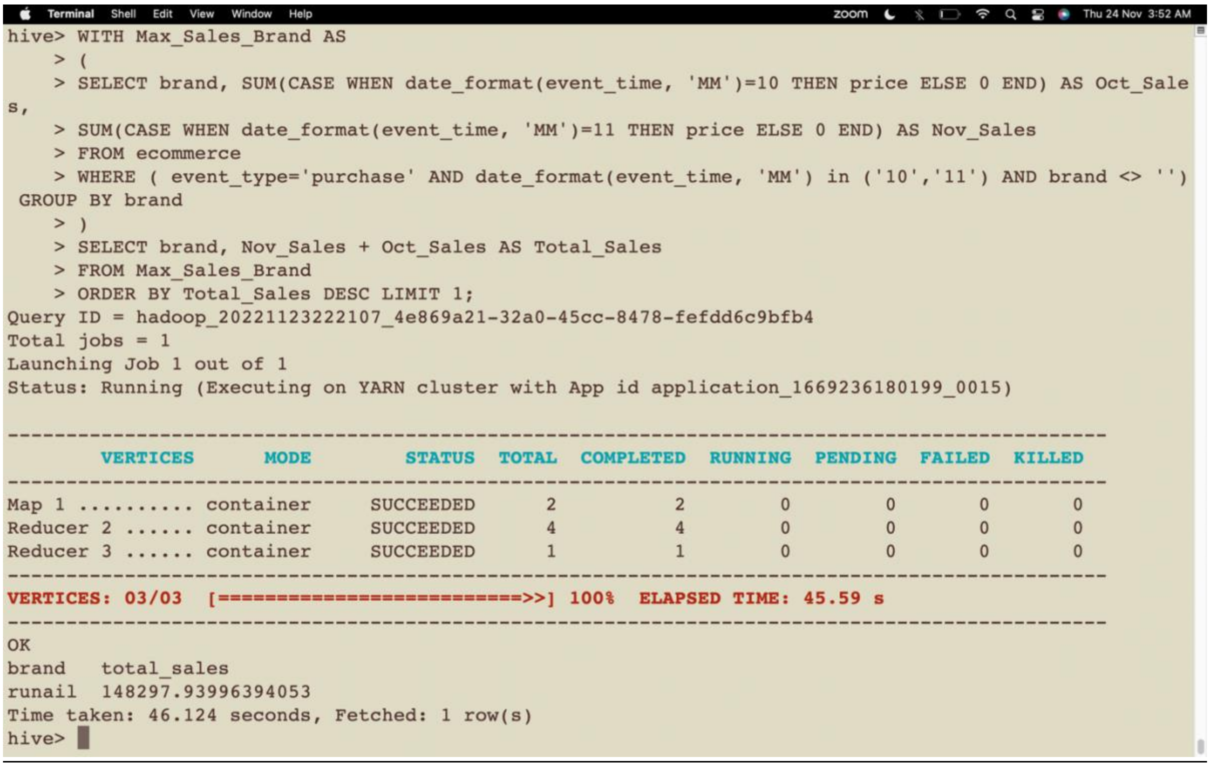
- Company has more products registered under Appliances category i.e., 61,736 products than any other categories.
- Then it is followed by stationery as second with 26,722 products, furniture as third with 23,604 products, apparel as fourth with 18232 products registered, accessories as fifth with 12929 products.
- Sports category has only 2 products registered. This must be due to low cosmetic products in the sports market.

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

Query:

WITH Max_Sales_Brand AS
(
SELECT brand, SUM(CASE WHEN date_format(event_time, 'MM')=10 THEN price ELSE 0 END) AS Oct_Sales,
SUM(CASE WHEN date_format(event_time, 'MM')=11 THEN price ELSE 0 END) AS Nov_Sales
FROM ecommerce
WHERE (event_type='purchase' AND date_format(event_time, 'MM') in ('10','11') AND brand <> '') GROUP BY brand
)
SELECT brand, Nov_Sales + Oct_Sales AS Total_Sales
FROM Max_Sales_Brand
ORDER BY Total_Sales DESC LIMIT 1;

Output:



```
hive> WITH Max_Sales_Brand AS
    > (
    > SELECT brand, SUM(CASE WHEN date_format(event_time, 'MM')=10 THEN price ELSE 0 END) AS Oct_Sale
s,
    > SUM(CASE WHEN date_format(event_time, 'MM')=11 THEN price ELSE 0 END) AS Nov_Sales
    > FROM ecommerce
    > WHERE ( event_type='purchase' AND date_format(event_time, 'MM') in ('10','11') AND brand <> '')
 GROUP BY brand
    > )
    > SELECT brand, Nov_Sales + Oct_Sales AS Total_Sales
    > FROM Max_Sales_Brand
    > ORDER BY Total_Sales DESC LIMIT 1;
Query ID = hadoop_20221123222107_4e869a21-32a0-45cc-8478-fefdd6c9bfb4
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1669236180199_0015)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      2          2        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      4          4        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 45.59 s
----------------------------------------------------------------------------------------------
OK
brand   total_sales
runail  148297.93996394053
Time taken: 46.124 seconds, Fetched: 1 row(s)
hive>
```

brand total_sales
runail 148297.93996394053

Insights:

- Runail is the brand that has highest / maximum sales in the month of October and November of 2019 combined.
- It seems that Runail brand has high popularity among cosmetic lovers and bringing in more products related to Runail brand could help in increasing their profit.

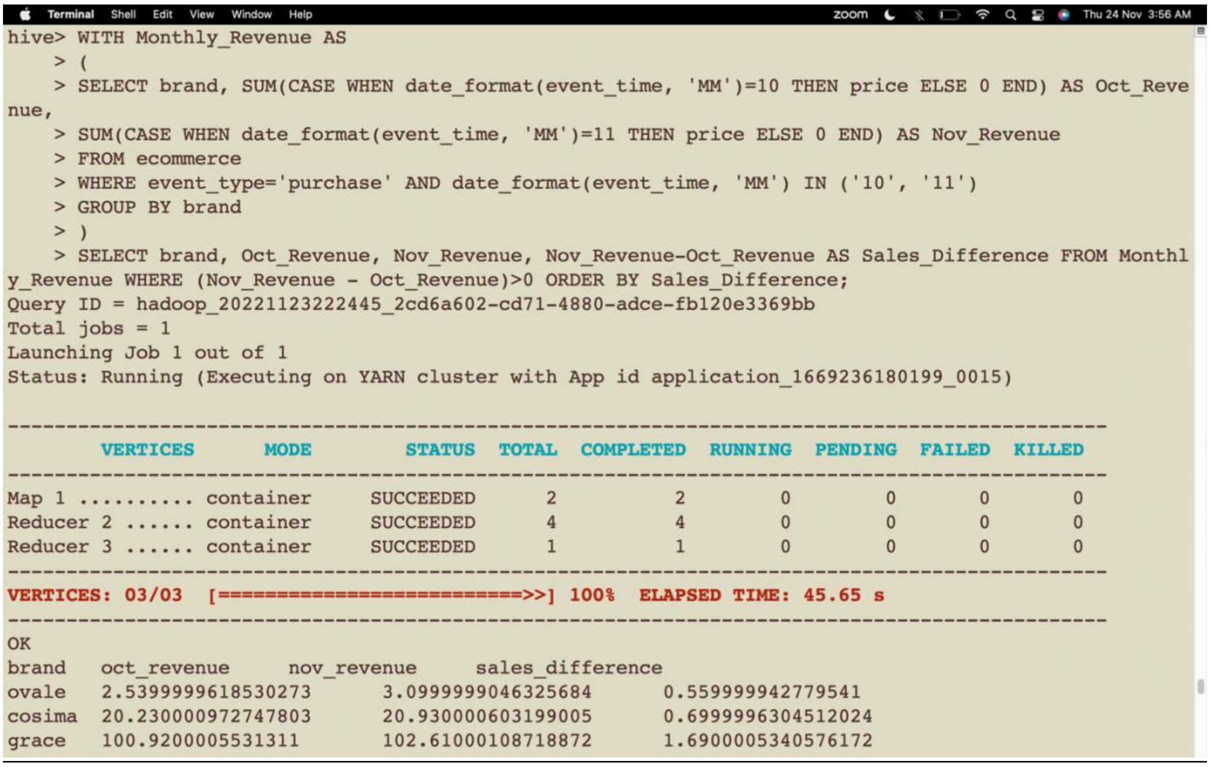
Q7: Which brands increased their sales from October to November?

Query:

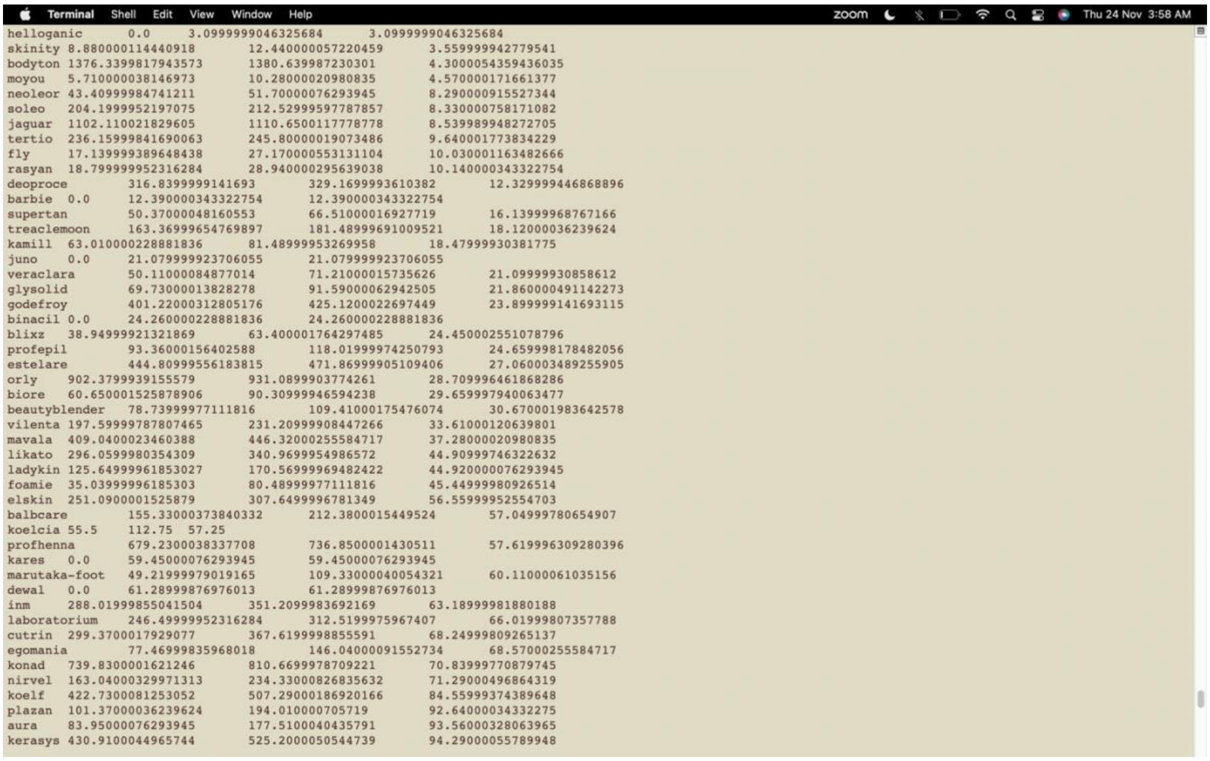
WITH Monthly_Revenue AS
(
SELECT brand, SUM(CASE WHEN date_format(event_time, 'MM')=10 THEN price ELSE 0 END) AS Oct_Revenue,
SUM(CASE WHEN date_format(event_time, 'MM')=11 THEN price ELSE 0 END) AS Nov_Revenue
FROM ecommerce
WHERE event_type='purchase' AND date_format(event_time, 'MM') IN ('10', '11')
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;

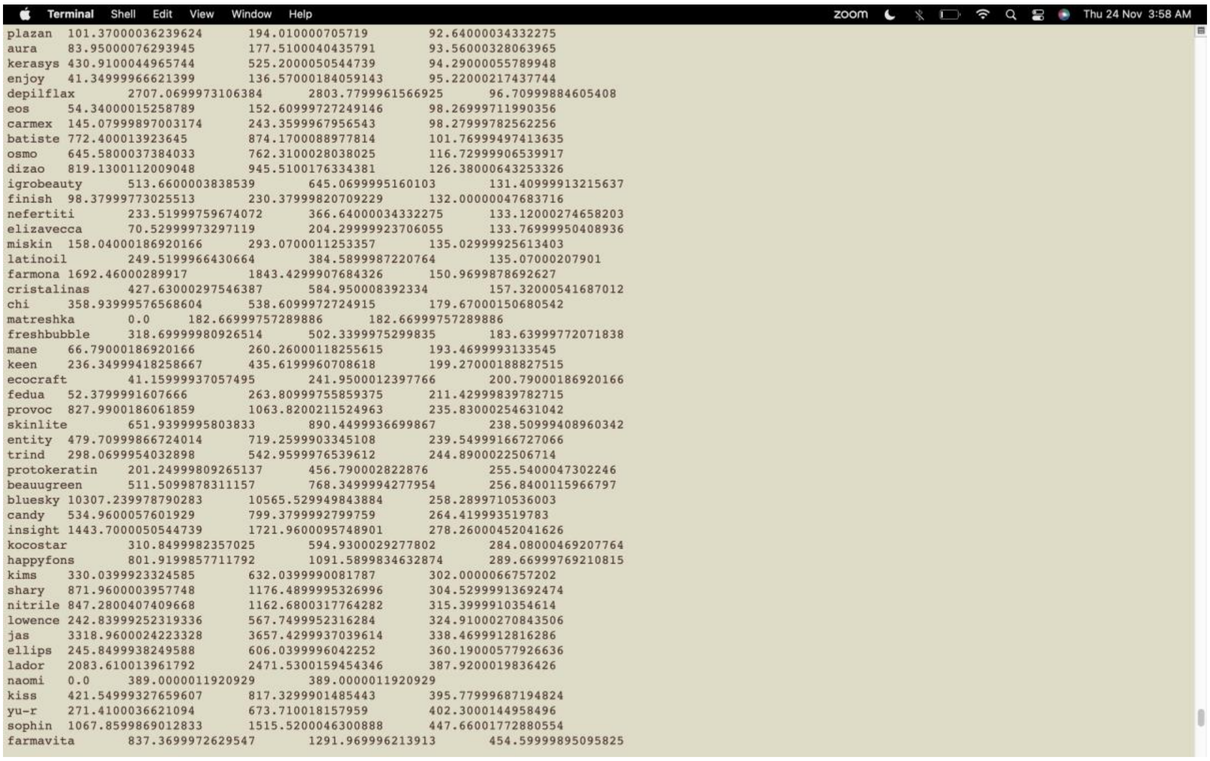
<u>Output:</u>



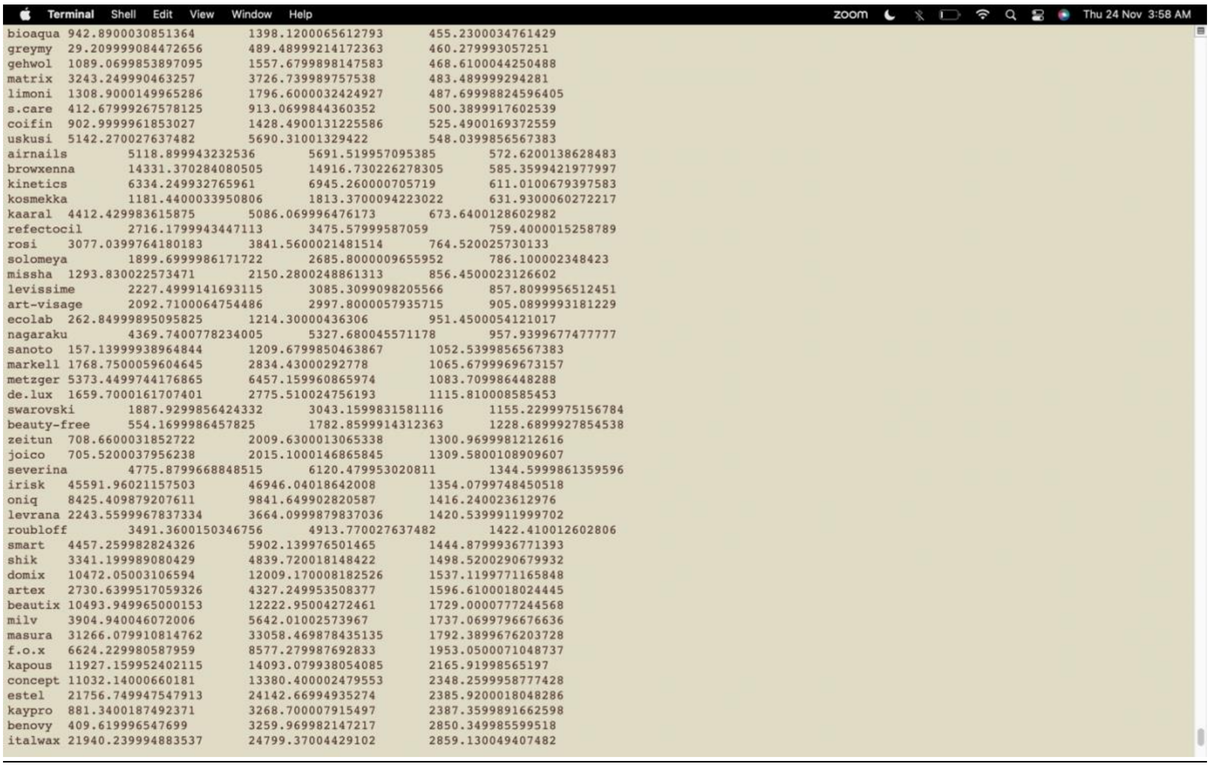
```
hive> WITH Monthly_Revenue AS
    > (
    > SELECT brand, SUM(CASE WHEN date_format(event_time, 'MM')=10 THEN price ELSE 0 END) AS Oct_Reve
nue,
    > SUM(CASE WHEN date_format(event_time, 'MM')=11 THEN price ELSE 0 END) AS Nov_Revenue
    > FROM ecommerce
    > WHERE event_type='purchase' AND date_format(event_time, 'MM') IN ('10', '11')
    > GROUP BY brand
    > )
    > SELECT brand, Oct_Revenue, Nov_Revenue, Nov_Revenue-Oct_Revenue AS Sales_Difference FROM Monthl
y_Revenue WHERE (Nov_Revenue - Oct_Revenue)>0 ORDER BY Sales_Difference;
Query ID = hadoop_20221123222445_2cd6a602-cd71-4880-adce-fb120e3369bb
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1669236180199_0015)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      2          2        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      4          4        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 45.65 s
----------------------------------------------------------------------------------------------
OK
brand   oct_revenue     nov_revenue     sales_difference
ovale   2.5399999618530273      3.0999999046325684      0.559999942779541
cosima  20.230000972747803      20.930000603199005      0.6999996304512024
grace   100.9200005531311       102.61000108718872      1.6900005340576172
```



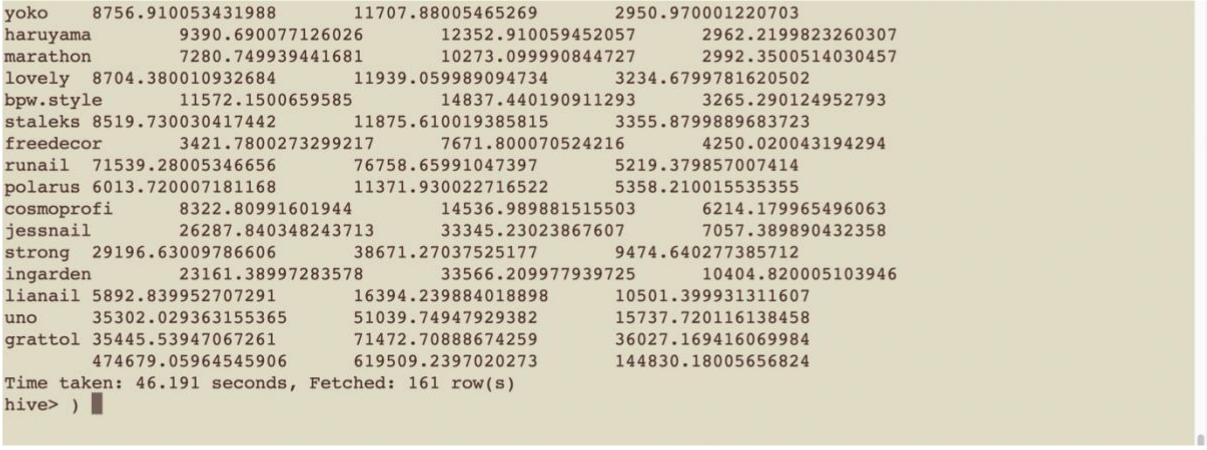
```
helloganic      0.0     3.0999999046325684      3.0999999046325684
skinity 8.880000114440918       12.440000057220459      3.559999942779541
bodyton 1376.3399817943573      1380.639987230301       4.3000054359436035
moyou   5.710000038146973       10.28000020980835       4.570000171661377
neoleor 43.40999984741211       51.70000076293945       8.290000915527344
soleo   204.1999952197075       212.52999597787857      8.330000758171082
jaguar  1102.110021829605       1110.6500117778778      8.539989948272705
tertio  236.15999841690063      245.80000019073486      9.640001773834229
fly     17.139999389648438      27.170000553131104      10.030001163482666
rasyan  18.799999952316284      28.940000295639038      10.140000343322754
deoproce        316.8399999141693       329.1699993610382       12.329999446868896
barbie  0.0     12.390000343322754      12.390000343322754
supertan        50.37000048160553       66.51000016927719       16.13999968767166
treaclemoon     163.36999654769897      181.48999691009521      18.12000036239624
kamill  63.010000228881836      81.48999953269958       18.47999930381775
juno    0.0     21.079999923706055      21.079999923706055
veraclara       50.11000084877014       71.21000015735626       21.09999930858612
glysolid        69.73000013828278       91.59000062942505       21.860000491142273
godefroy        401.22000312805176      425.1200022697449       23.899999141693115
binacil 0.0     24.260000228881836      24.260000228881836
blixz   38.94999921321869       63.400001764297485      24.450002551078796
profepil        93.36000156402588       118.01999974250793      24.659998178482056
estelare        444.80999556183815      471.86999905109406      27.060003489255905
orly    902.3799939155579       931.0899903774261       28.709996461868286
biore   60.650001525878906      90.30999946594238       29.659997940063477
beautyblender   78.73999977111816       109.41000175476074      30.670001983642578
vilenta 197.59999787807465      231.20999908447266      33.61000120639801
mavala  409.0400023460388       446.32000255584717      37.28000020980835
likato  296.0599980354309       340.9699954986572       44.90999746322632
ladykin 125.64999961853027      170.56999969482422      44.920000076293945
foamie  35.03999996185303       80.48999977111816       45.44999980926514
elskin  251.0900001525879       307.6499996781349       56.55999952554703
balbcare        155.33000373840332      212.3800015449524       57.04999780654907
koelcia 55.5    112.75  57.25
profhenna       679.2300038337708       736.8500001430511       57.619996309280396
kares   0.0     59.45000076293945       59.45000076293945
marutaka-foot   49.21999979019165       109.33000040054321      60.11000061035156
dewal   0.0     61.28999876976013       61.28999876976013
inm     288.01999855041504      351.2099983692169       63.18999981880188
laboratorium    246.49999952316284      312.5199975967407       66.01999807357788
cutrin  299.3700017929077       367.6199998855591       68.24999809265137
egomania        77.46999835968018       146.04000091552734      68.5700025584717
konad   739.8300001621246       810.6699978709221       70.83999770879745
nirvel  163.04000329971313      234.33000826835632      71.29000496864319
koelf   422.7300081253052       507.29000186920166      84.55999374389648
plazan  101.37000036239624      194.010000705719        92.64000034332275
aura    83.95000076293945       177.5100040435791       93.56000328063965
kerasys 430.9100044965744       525.2000050544739       94.29000055789948
```



```
plazan	101.37000036239624	194.010000705719	92.64000034332275
aura	83.95000076293945	177.5100040435791	93.56000328063965
kerasys	430.9100044965744	525.2000050544739	94.2900005578 9948
enjoy	41.34999966621399	136.57000184059143	95.22000217437744
depilflax	2707.0699973106384	2803.7799961566925	96.70999884605408
eos	54.34000015258789	152.60999727249146	98.26999711990356
carmex	145.07999897003174	243.3599967956543	98.27999782562256
batiste	772.400013923645	874.1700088977814	101.76999497413635
osmo	645.5800037384033	762.3100028038025	116.72999906539917
dizao	819.1300112009048	945.5100176334381	126.38000643253326
igrobeauty	513.6600003838539	645.0699995160103	131.40999913215637
finish	98.37999773025513	230.37999820709229	132.00000047683716
nefertiti	233.51999759674072	366.64000034332275	133.12000274658203
elizavecca	70.52999973297119	204.29999923706055	133.76999950408936
miskin	158.04000186920166	293.0700011253357	135.02999925613403
latinoil	249.5199966430664	384.5899987220764	135.07000207901
farmona	1692.46000289917	1843.4299907684326	150.9699878692627
cristalinas	427.63000297546387	584.950008392334	157.32000541687012
chi	358.93999576568604	538.6099972724915	179.67000150680542
matreshka	0.0	182.66999757289886	182.66999757289886
freshbubble	318.69999980926514	502.3399975299835	183.63999772071838
mane	66.79000186920166	260.26000118255615	193.4699993133545
keen	236.34999418258667	435.6199960708618	199.27000188827515
ecocraft	41.15999937057495	241.9500012397766	200.79000186920166
fedua	52.3799991607666	263.80999755859375	211.42999839782715
provoc	827.9900186061859	1063.8200211524963	235.83000254631042
skinlite	651.9399995803833	890.4499936699867	238.50999408960342
entity	479.70999866724014	719.2599903345108	239.54999166727066
trind	298.0699954032898	542.9599976539612	244.8900022506714
protokeratin	201.24999809265137	456.790002822876	255.5400047302246
beauugreen	511.5099878311157	768.3499994277954	256.8400115966797
bluesky	10307.239978790283	10565.529949843884	258.2899710536003
candy	534.9600057601929	799.3799992799759	264.419993519783
insight	1443.7000050544739	1721.9600095748901	278.26000452041626
kocostar	310.8499982357025	594.9300029277802	284.08000469207764
happyfons	801.9199857711792	1091.5899834632874	289.66999769210815
kims	330.0399923324585	632.0399990081787	302.0000066757202
shary	871.9600003957748	1176.4899995326996	304.52999913692474
nitrile	847.2800407409668	1162.6800317764282	315.3999910354614
lowence	242.83999252319336	567.7499952316284	324.91000270843506
jas	3318.9600024223328	3657.4299937039614	338.4699912816286
ellips	245.8499938249588	606.0399996042252	360.19000577926636
lador	2083.610013961792	2471.5300159454346	387.9200019836426
naomi	0.0	389.0000011920929	389.0000011920929
kiss	421.54999327659607	817.3299901485443	395.77999687194824
yu-r	271.4100036621094	673.710018157959	402.3000144958496
sophin	1067.8599869012833	1515.5200046300888	447.66001772880554
farmavita	837.3699972629547	1291.969996213913	454.59999895095825
```



```
bioaqua	942.8900030851364	1398.1200065612793	455.2300034761429
greymy	29.209999084472656	489.48999214172363	460.279993057251
gehwol	1089.0699853897095	1557.6799898147583	468.6100044250488
matrix	3243.249990463257	3726.739989757538	483.489999294281
limoni	1308.9000149965286	1796.6000032424927	487.69998824596405
s.care	412.67999267578125	913.0699844360352	500.3899917602539
coifin	902.9999961853027	1428.4900131225586	525.4900169372559
uskusi	5142.270027637482	5690.31001329422	548.0399856567383
airnails	5118.899943232536	5691.519957095385	572.6200138628483
browxenna	14331.370284080505	14916.730226278305	585.3599421977997
kinetics	6334.249932765961	6945.260000705719	611.0100679397583
kosmekka	1181.4400033950806	1813.3700094223022	631.9300060272217
kaaral	4412.429983615875	5086.069996476173	673.6400128602982
refectocil	2716.1799943447113	3475.57999587059	759.4000015258789
rosi	3077.0399764180183	3841.5600021481514	764.520025730133
solomeya	1899.6999986171722	2685.8000009655952	786.100002348423
missha	1293.830022573471	2150.2800248861313	856.4500023126602
levissime	2227.4999141693115	3085.3099098205566	857.8099956512451
art-visage	2092.7100064754486	2997.8000057935715	905.0899993181229
ecolab	262.84999895095825	1214.30000436306	951.4500054121017
nagaraku	4369.7400778234005	5327.680045571178	957.9399677477777
sanoto	157.13999938964844	1209.6799850463867	1052.5399856567383
markell	1768.7500059604645	2834.43000292778	1065.6799969673157
metzger	5373.4499744176865	6457.159960865974	1083.709986448288
de.lux	1659.7000161707401	2775.510024756193	1115.810008585453
swarovski	1887.9299856424332	3043.1599831581116	1155.2299975156784
beauty-free	554.1699986457825	1782.8599914312363	1228.6899927854538
zeitun	708.6600031852722	2009.6300013065338	1300.9699981212616
joico	705.5200037956238	2015.1000146865845	1309.5800108909607
severina	4775.8799668848515	6120.479953020811	1344.5999861359596
irisk	45591.96021157503	46946.04018642008	1354.0799748450518
oniq	8425.409879207611	9841.649902820587	1416.240023612976
levrana	2243.5599967837334	3664.0999879837036	1420.5399911999702
roubloff	3491.3600150346756	4913.770027637482	1422.410012602806
smart	4457.259982824326	5902.139976501465	1444.8799936771393
shik	3341.199989080429	4839.720018148422	1498.5200290679932
domix	10472.05003106594	12009.170008182526	1537.1199771165848
artex	2730.6399517059326	4327.249953508377	1596.6100018024445
beautix	10493.949965000153	12222.95004272461	1729.0000777244568
milv	3904.940046072006	5642.01002573967	1737.0699796676636
masura	31266.079910814762	33058.469878435135	1792.3899676203728
f.o.x	6624.229980587959	8577.279987692833	1953.0500071048737
kapous	11927.159952402115	14093.079938054085	2165.91998565197
concept	11032.14000660181	13380.400002479553	2348.2599958777428
estel	21756.749947547913	24142.66994935274	2385.9200018048286
kaypro	881.3400187492371	3268.700007915497	2387.3599891662598
benovy	409.619996547699	3259.969982147217	2850.349985599518
italwax	21940.239994883537	24799.37004429102	2859.130049407482
```



```
yoko    8756.910053431988       11707.88005465269       2950.970001220703
haruyama        9390.690077126026       12352.910059452057      2962.2199823260307
marathon        7280.749939441681       10273.099990844727      2992.3500514030457
lovely  8704.380010932684       11939.059989094734      3234.6799781620502
bpw.style       11572.1500659585        14837.440190911293      3265.290124952793
staleks 8519.730030417442       11875.610019385815      3355.8799889683723
freedecor       3421.7800273299217      7671.800070524216       4250.020043194294
runail  71539.28005346656       76758.65991047397       5219.379857007414
polarus 6013.720007181168       11371.930022716522      5358.210015535355
cosmoprofi      8322.80991601944        14536.989881515503      6214.179965496063
jessnail        26287.840348243713      33345.23023867607       7057.389890432358
strong  29196.63009786606       38671.27037525177       9474.640277385712
ingarden        23161.38997283578       33566.209977939725      10404.820005103946
lianail 5892.839952707291       16394.239884018898      10501.399931311607
uno     35302.029363155365      51039.74947929382       15737.720116138458
grattol 35445.53947067261       71472.70888674259       36027.169416069984
        474679.05964545906      619509.2397020273       144830.18005656824
Time taken: 46.191 seconds, Fetched: 161 row(s)
hive> )
```

Insights:

- Here are some 161 brands with increment in the selling from October to November.
- ‘Grattol’ brand has the highest total increment i.e., 36,027 /- and ‘Ovale’ seems to have least increment of 0.56 /- from October to November.
- Among all these brands list, ‘Runail’ which was the best brand in terms of selling in October and November combined is also in the top 10 brands with high increment for October (71539.28 /-) to November (76758.61 /-) i.e., increment of total 5219.38 /-.
- This implies that ‘Runail’ is the best and popular brand among all other brands within people.

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

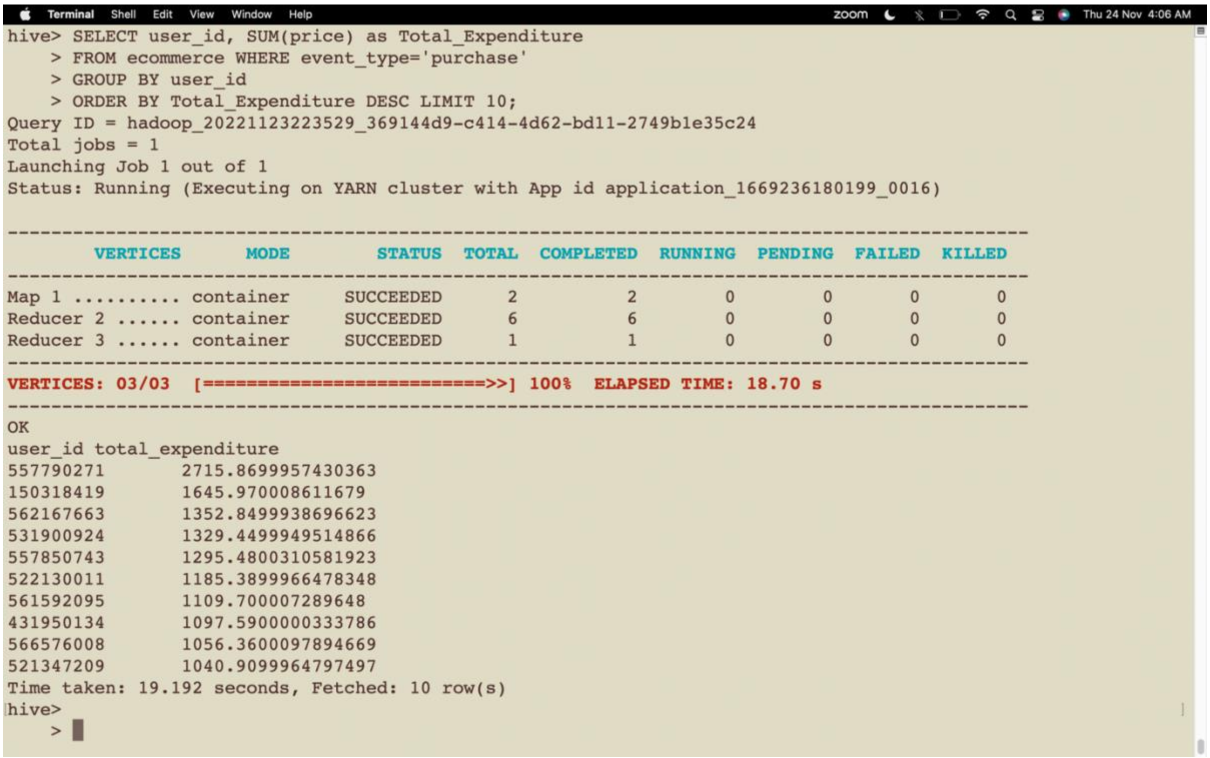
Query:

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

Output:

user_id	total_expenditure
557790271	2715.8699957430363
150318419	1645.970008611679
562167663	1352.8499938696623
531900924	1329.4499949514866
557850743	1295.4800310581923
522130011	1185.3899966478348
561592095	1109.700007289648
431950134	1097.5900000333786
566576008	1056.3600097894669
521347209	1040.9099964797497

Time taken: 19.192 seconds, Fetched: 10 row(s)



```
hive> SELECT user_id, SUM(price) as Total_Expenditure
    > FROM ecommerce WHERE event_type='purchase'
    > GROUP BY user_id
    > ORDER BY Total_Expenditure DESC LIMIT 10;
Query ID = hadoop_20221123223529_369144d9-c414-4d62-bd11-2749b1e35c24
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1669236180199_0016)

----------------------------------------------------------------------------------------------
        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: 18.70 s
----------------------------------------------------------------------------------------------
OK
user_id total_expenditure
557790271       2715.8699957430363
150318419       1645.970008611679
562167663       1352.8499938696623
531900924       1329.4499949514866
557850743       1295.4800310581923
522130011       1185.3899966478348
561592095       1109.700007289648
431950134       1097.5900000333786
566576008       1056.3600097894669
521347209       1040.9099964797497
Time taken: 19.192 seconds, Fetched: 10 row(s)
hive>
    > 
```

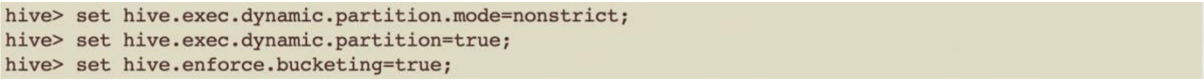
Insights:

- Here is the list of the top 10 users or buyers who have spent the most and could be rewarded with a Golden Customer plan to attract more people in the coming future.
- We are selecting this query to be executed using Optimized table to check that does optimized table reduces execution time with proper partitioning and bucketing.
- Time taken to execute this query on Base table (non-optimized table) is 19.192 seconds.

OPTIMIZED QUERIES

1. **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;



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

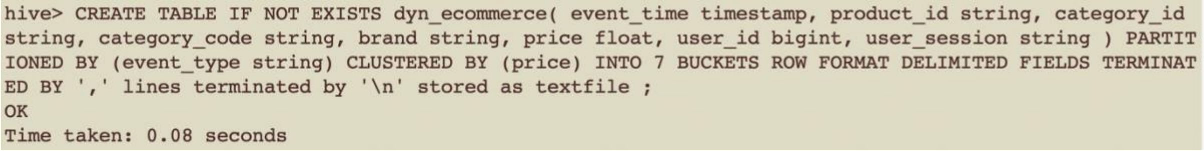
Table Optimization Steps:

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

Query:

CREATE TABLE IF NOT EXISTS dyn_ecommerce(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 DELIMITED FIELDS TERMINATED BY ',' lines terminated by '\n' stored as textfile
;

Output:



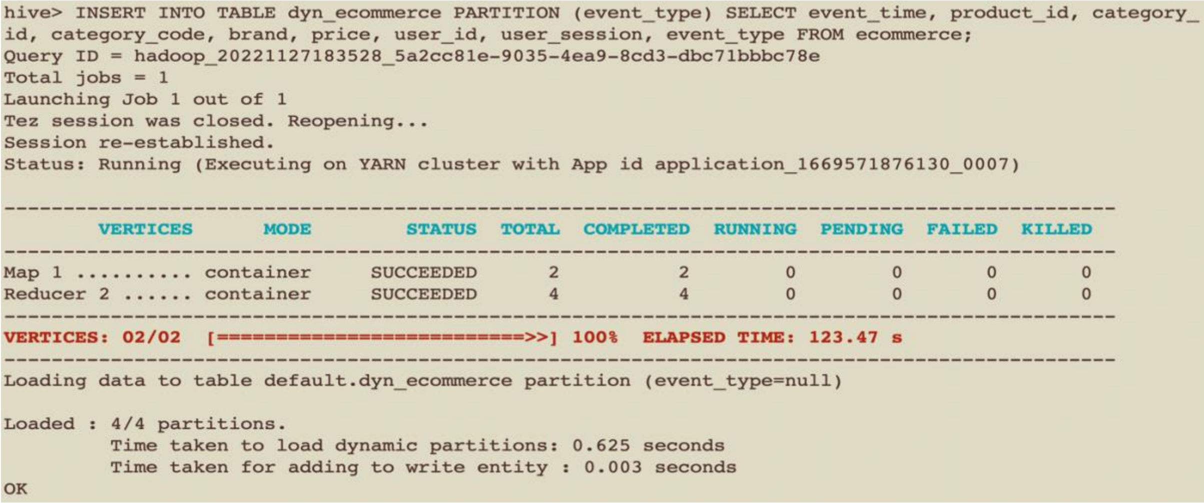
```
hive> CREATE TABLE IF NOT EXISTS dyn_ecommerce( event_time timestamp, product_id string, category_id
string, category_code string, brand string, price float, user_id bigint, user_session string ) PARTIT
IONED BY (event_type string) CLUSTERED BY (price) INTO 7 BUCKETS ROW FORMAT DELIMITED FIELDS TERMINAT
ED BY ',' lines terminated by '\n' stored as textfile ;
OK
Time taken: 0.08 seconds
```

3. To add data into partitioned and bucketed table we need to get it from already created table i.e., 'ecommerce'

Query:

INSERT INTO TABLE dyn_ecommerce PARTITION (event_type) SELECT event_time, product_id, category_id, category_code, brand, price, user_id, user_session, event_type FROM ecommerce;

Output:



```
hive> INSERT INTO TABLE dyn_ecommerce PARTITION (event_type) SELECT event_time, product_id, category_
id, category_code, brand, price, user_id, user_session, event_type FROM ecommerce;
Query ID = hadoop_20221127183528_5a2cc81e-9035-4ea9-8cd3-dbc71bbbc78e
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_1669571876130_0007)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      2          2        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      4          4        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 123.47 s
----------------------------------------------------------------------------------------------
Loading data to table default.dyn_ecommerce partition (event_type=null)

Loaded : 4/4 partitions.
         Time taken to load dynamic partitions: 0.625 seconds
         Time taken for adding to write entity : 0.003 seconds
OK
```

4. 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 commands.

a. **Command to exit Hive environment**

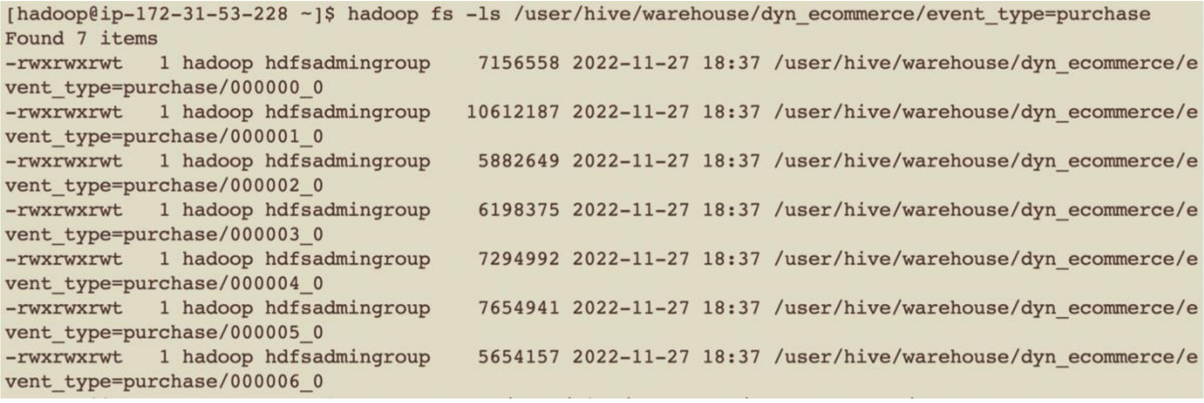
- EXIT;

b. **Command to check existence of partitions (event_type = purchase) in the table**

<u>Query:</u>

hadoop fs -ls /user/hive/warehouse/dyn_ecommerce/event_type=purchase

<u>Output:</u>



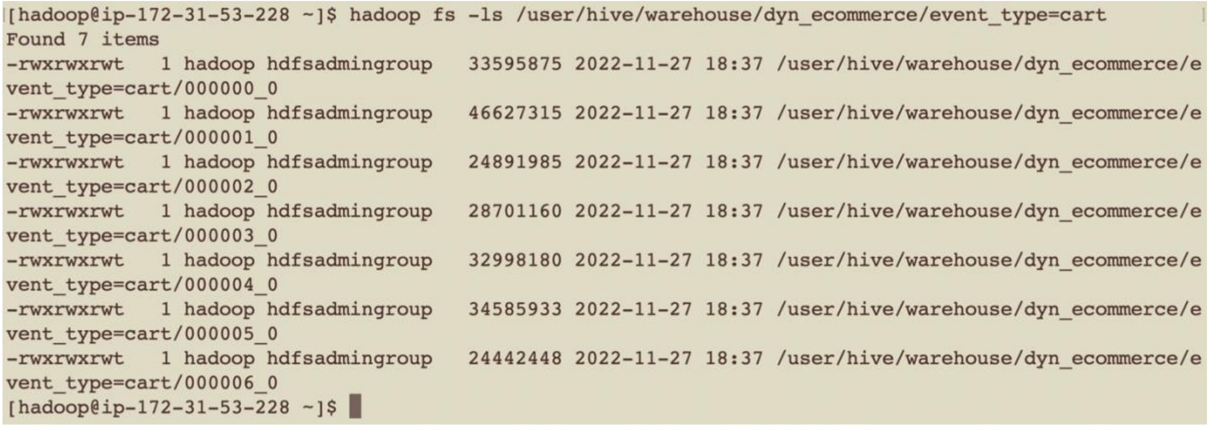
```
[hadoop@ip-172-31-53-228 ~]$ hadoop fs -ls /user/hive/warehouse/dyn_ecommerce/event_type=purchase
Found 7 items
-rwxrwxrwt   1 hadoop hdfsadmingroup    7156558 2022-11-27 18:37 /user/hive/warehouse/dyn_ecommerce/e
vent_type=purchase/000000_0
-rwxrwxrwt   1 hadoop hdfsadmingroup   10612187 2022-11-27 18:37 /user/hive/warehouse/dyn_ecommerce/e
vent_type=purchase/000001_0
-rwxrwxrwt   1 hadoop hdfsadmingroup    5882649 2022-11-27 18:37 /user/hive/warehouse/dyn_ecommerce/e
vent_type=purchase/000002_0
-rwxrwxrwt   1 hadoop hdfsadmingroup    6198375 2022-11-27 18:37 /user/hive/warehouse/dyn_ecommerce/e
vent_type=purchase/000003_0
-rwxrwxrwt   1 hadoop hdfsadmingroup    7294992 2022-11-27 18:37 /user/hive/warehouse/dyn_ecommerce/e
vent_type=purchase/000004_0
-rwxrwxrwt   1 hadoop hdfsadmingroup    7654941 2022-11-27 18:37 /user/hive/warehouse/dyn_ecommerce/e
vent_type=purchase/000005_0
-rwxrwxrwt   1 hadoop hdfsadmingroup    5654157 2022-11-27 18:37 /user/hive/warehouse/dyn_ecommerce/e
vent_type=purchase/000006_0
```

c. Command to check existence of partitions (event_type = cart) in the table

<u>Query:</u>

hadoop fs -ls /user/hive/warehouse/dyn_ecommerce/event_type=cart

<u>Output:</u>



```
[hadoop@ip-172-31-53-228 ~]$ hadoop fs -ls /user/hive/warehouse/dyn_ecommerce/event_type=cart
Found 7 items
-rwxrwxrwt   1 hadoop hdfsadmingroup   33595875 2022-11-27 18:37 /user/hive/warehouse/dyn_ecommerce/e
vent_type=cart/000000_0
-rwxrwxrwt   1 hadoop hdfsadmingroup   46627315 2022-11-27 18:37 /user/hive/warehouse/dyn_ecommerce/e
vent_type=cart/000001_0
-rwxrwxrwt   1 hadoop hdfsadmingroup   24891985 2022-11-27 18:37 /user/hive/warehouse/dyn_ecommerce/e
vent_type=cart/000002_0
-rwxrwxrwt   1 hadoop hdfsadmingroup   28701160 2022-11-27 18:37 /user/hive/warehouse/dyn_ecommerce/e
vent_type=cart/000003_0
-rwxrwxrwt   1 hadoop hdfsadmingroup   32998180 2022-11-27 18:37 /user/hive/warehouse/dyn_ecommerce/e
vent_type=cart/000004_0
-rwxrwxrwt   1 hadoop hdfsadmingroup   34585933 2022-11-27 18:37 /user/hive/warehouse/dyn_ecommerce/e
vent_type=cart/000005_0
-rwxrwxrwt   1 hadoop hdfsadmingroup   24442448 2022-11-27 18:37 /user/hive/warehouse/dyn_ecommerce/e
vent_type=cart/000006_0
[hadoop@ip-172-31-53-228 ~]$ 
```

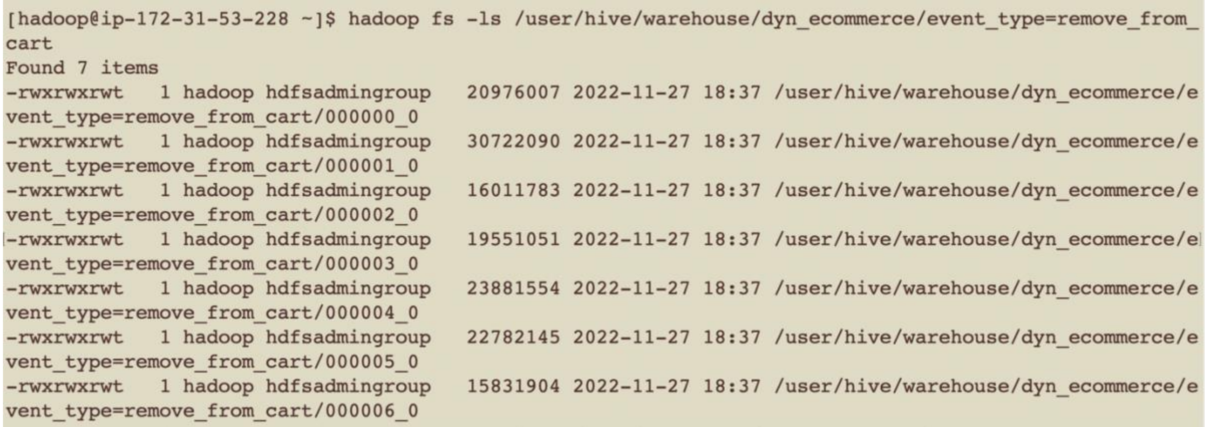
d. Command to check existence of partitions (event_type remove_from_ cart) in the table

Query:

hadoop fs -ls /user/hive/warehouse/dyn_ecommerce/event_type=remove

_from_cart

Output:



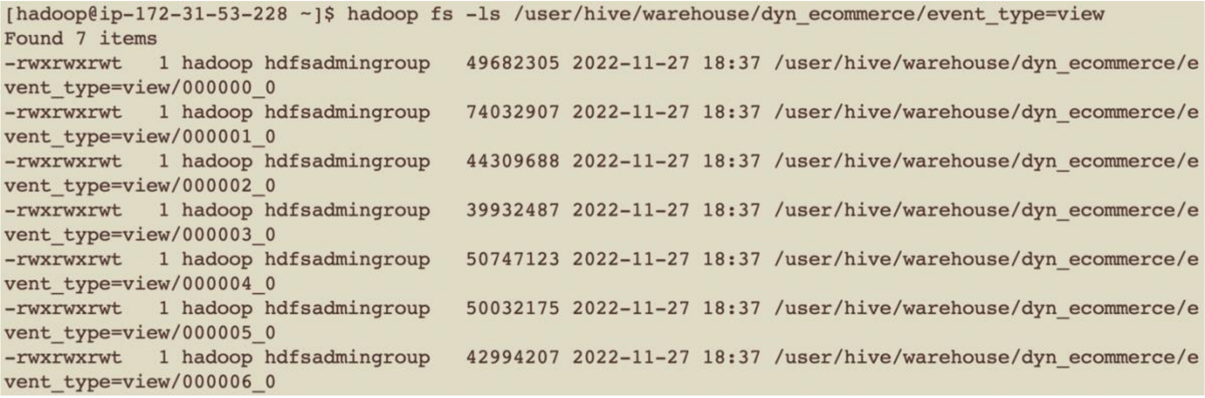
```
[hadoop@ip-172-31-53-228 ~]$ hadoop fs -ls /user/hive/warehouse/dyn_ecommerce/event_type=remove_from_
cart
Found 7 items
-rwxrwxrwt   1 hadoop hdfsadmingroup   20976007 2022-11-27 18:37 /user/hive/warehouse/dyn_ecommerce/e
vent_type=remove_from_cart/000000_0
-rwxrwxrwt   1 hadoop hdfsadmingroup   30722090 2022-11-27 18:37 /user/hive/warehouse/dyn_ecommerce/e
vent_type=remove_from_cart/000001_0
-rwxrwxrwt   1 hadoop hdfsadmingroup   16011783 2022-11-27 18:37 /user/hive/warehouse/dyn_ecommerce/e
vent_type=remove_from_cart/000002_0
-rwxrwxrwt   1 hadoop hdfsadmingroup   19551051 2022-11-27 18:37 /user/hive/warehouse/dyn_ecommerce/e
vent_type=remove_from_cart/000003_0
-rwxrwxrwt   1 hadoop hdfsadmingroup   23881554 2022-11-27 18:37 /user/hive/warehouse/dyn_ecommerce/e
vent_type=remove_from_cart/000004_0
-rwxrwxrwt   1 hadoop hdfsadmingroup   22782145 2022-11-27 18:37 /user/hive/warehouse/dyn_ecommerce/e
vent_type=remove_from_cart/000005_0
-rwxrwxrwt   1 hadoop hdfsadmingroup   15831904 2022-11-27 18:37 /user/hive/warehouse/dyn_ecommerce/e
vent_type=remove_from_cart/000006_0
```

e. Command to check existence of partitions (event_type = view) in the table

Query:

hadoop fs -ls /user/hive/warehouse/dyn_ecommerce/event_type=view

Output:



```
[hadoop@ip-172-31-53-228 ~]$ hadoop fs -ls /user/hive/warehouse/dyn_ecommerce/event_type=view
Found 7 items
-rwxrwxrwt   1 hadoop hdfsadmingroup   49682305 2022-11-27 18:37 /user/hive/warehouse/dyn_ecommerce/e
vent_type=view/000000_0
-rwxrwxrwt   1 hadoop hdfsadmingroup   74032907 2022-11-27 18:37 /user/hive/warehouse/dyn_ecommerce/e
vent_type=view/000001_0
-rwxrwxrwt   1 hadoop hdfsadmingroup   44309688 2022-11-27 18:37 /user/hive/warehouse/dyn_ecommerce/e
vent_type=view/000002_0
-rwxrwxrwt   1 hadoop hdfsadmingroup   39932487 2022-11-27 18:37 /user/hive/warehouse/dyn_ecommerce/e
vent_type=view/000003_0
-rwxrwxrwt   1 hadoop hdfsadmingroup   50747123 2022-11-27 18:37 /user/hive/warehouse/dyn_ecommerce/e
vent_type=view/000004_0
-rwxrwxrwt   1 hadoop hdfsadmingroup   50032175 2022-11-27 18:37 /user/hive/warehouse/dyn_ecommerce/e
vent_type=view/000005_0
-rwxrwxrwt   1 hadoop hdfsadmingroup   42994207 2022-11-27 18:37 /user/hive/warehouse/dyn_ecommerce/e
vent_type=view/000006_0
```

5. Running the same query for *Question 8* on Optimized as executed on Base table to understand the execution time of same query on Base table and Optimized table.

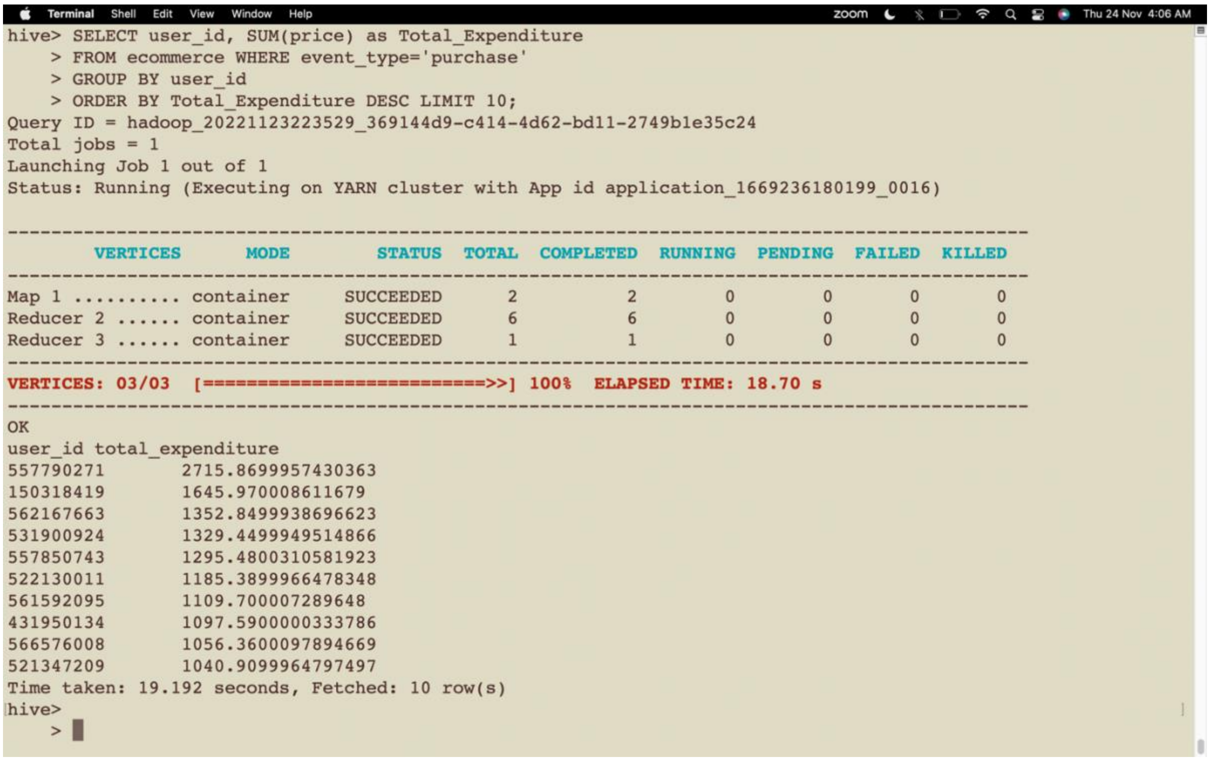
Running the Optimized query using Hive.

Q8 with (optimization): 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.

a. Normal Query using Non-Optimized table "ecommerce"

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

Output:

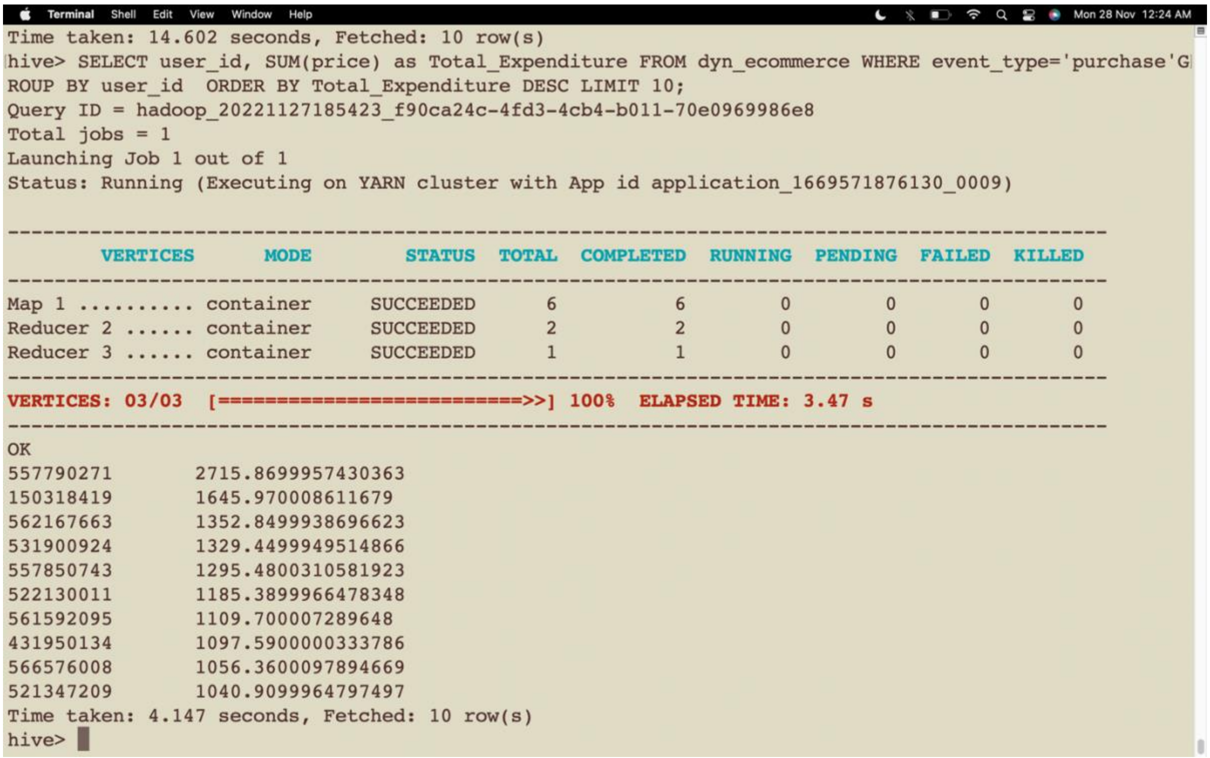


```
hive> SELECT user_id, SUM(price) as Total_Expenditure
    > FROM ecommerce WHERE event_type='purchase'
    > GROUP BY user_id
    > ORDER BY Total_Expenditure DESC LIMIT 10;
Query ID = hadoop_20221123223529_369144d9-c414-4d62-bd11-2749b1e35c24
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1669236180199_0016)

----------------------------------------------------------------------------------------------
        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: 18.70 s
----------------------------------------------------------------------------------------------
OK
user_id total_expenditure
557790271       2715.8699957430363
150318419       1645.970008611679
562167663       1352.8499938696623
531900924       1329.4499949514866
557850743       1295.4800310581923
522130011       1185.3899966478348
561592095       1109.700007289648
431950134       1097.5900000333786
566576008       1056.3600097894669
521347209       1040.9099964797497
Time taken: 19.192 seconds, Fetched: 10 row(s)
hive>
    > 
```

b. Optimised Query using dynamic table "dyn_ecommerce"

```
SELECT user_id, SUM(price) AS Total_Expenditure
FROM dyn_ecommerce
WHERE event_type='purchase'
GROUP BY user_id
ORDER BY Total_Expenditure DESC LIMIT 10;
```

Output:



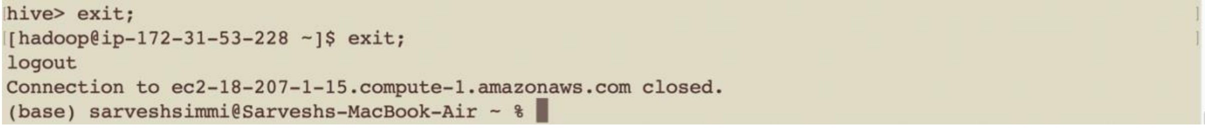
```
Time taken: 14.602 seconds, Fetched: 10 row(s)
hive> SELECT user_id, SUM(price) as Total_Expenditure FROM dyn_ecommerce WHERE event_type='purchase'G
ROUP BY user_id  ORDER BY Total_Expenditure DESC LIMIT 10;
Query ID = hadoop_20221127185423_f90ca24c-4fd3-4cb4-b011-70e0969986e8
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1669571876130_0009)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      6          6        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: 3.47 s
----------------------------------------------------------------------------------------------
OK
557790271       2715.8699957430363
150318419       1645.970008611679
562167663       1352.8499938696623
531900924       1329.4499949514866
557850743       1295.4800310581923
522130011       1185.3899966478348
561592095       1109.700007289648
431950134       1097.5900000333786
566576008       1056.3600097894669
521347209       1040.9099964797497
Time taken: 4.147 seconds, Fetched: 10 row(s)
hive> 
```

Insights:

- After creating an optimized table by Partitioning on 'event_type' attribute and Bucketing (Clustering) on 'price' we have executed same query of **Question No. 8** on this table.
- We can see the result is same as we have got when executed on Base table (Non-Optimized table).
- Secondly, most importantly we can see there is significant drop in the execution time of the same query i.e., previously the execution was measured as 19.192 seconds and now it is 4.147 seconds with the difference of 15.045 seconds.
- Hence, with proper partitioning and bucketing on table we can reduce execution time of the query.

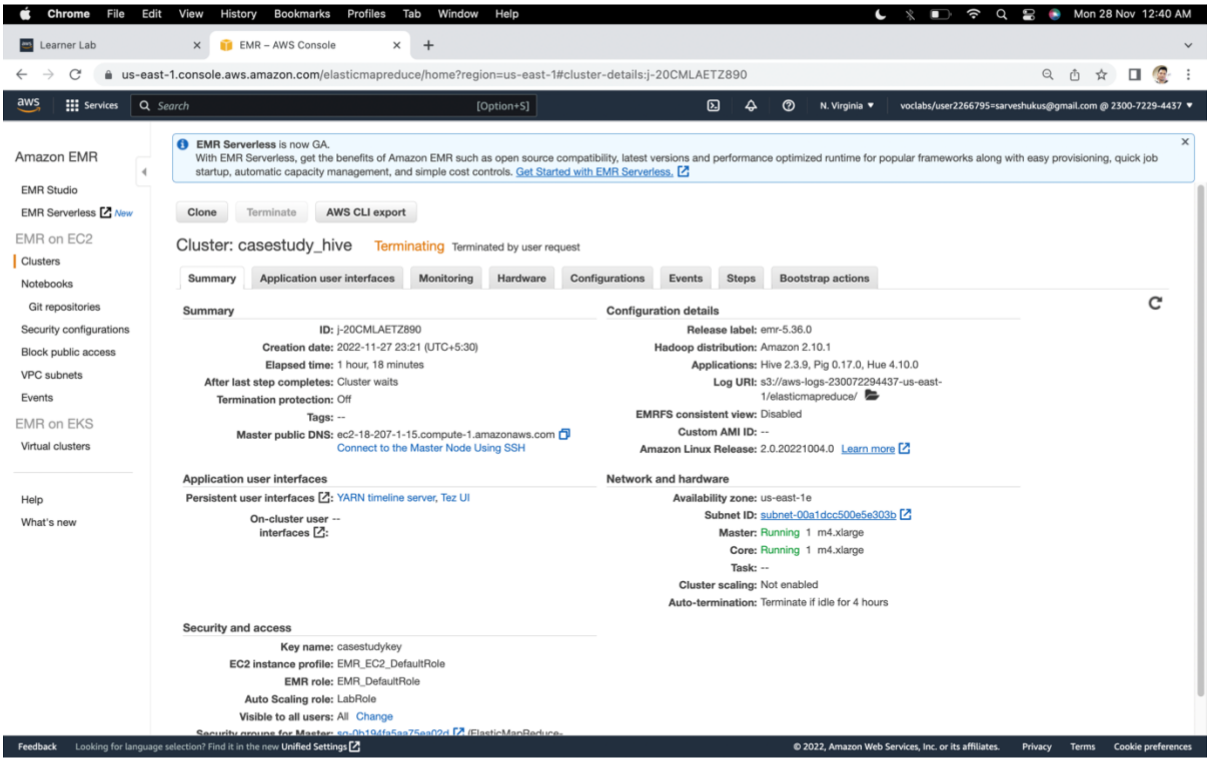
6. Exit and Terminating EMR cluster "casestudy_hive"

a. Exit from Hive and hadoop using command 'exit;'



```
hive> exit;
[hadoop@ip-172-31-53-228 ~]$ exit;
logout
Connection to ec2-18-207-1-15.compute-1.amazonaws.com closed.
(base) sarveshsimmi@Sarveshs-MacBook-Air ~ % 
```

b. Terminate EMR cluster



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.

Cluster: casestudy_hive Terminating Terminated by user request

Summary

ID: j-20CMLAETZ890
Creation date: 2022-11-27 23:21 (UTC+5:30)
Elapsed time: 1 hour, 18 minutes
After last step completes: Cluster waits
Termination protection: Off
Tags: --
Master public DNS: ec2-18-207-1-15.compute-1.amazonaws.com
Connect to the Master Node Using SSH

Configuration details

Release label: emr-5.36.0
Hadoop distribution: Amazon 2.10.1
Applications: Hive 2.3.9, Pig 0.17.0, Hue 4.10.0
Log URI: s3://aws-logs-230072294437-us-east-1/elasticmapreduce/
EMRFS consistent view: Disabled
Custom AMI ID: --
Amazon Linux Release: 2.0.20221004.0 Learn more

Application user interfaces

Persistent user interfaces: YARN timeline server, Tez UI
On-cluster user interfaces: --

Network and hardware

Availability zone: us-east-1e
Subnet ID: subnet-00a1dcc500e5e303b
Master: Running 1 m4.xlarge
Core: Running 1 m4.xlarge
Task: --
Cluster scaling: Not enabled
Auto-termination: Terminate if idle for 4 hours

Security and access

Key name: casestudykey
EC2 instance profile: EMR_EC2_DefaultRole
EMR role: EMR_DefaultRole
Auto Scaling role: LabRole
Visible to all users: All Change
Security groups for Master: sg-0b194fa5aa75ea02d (ElasticMapReduce-

End of Procedure