Hive Case Study

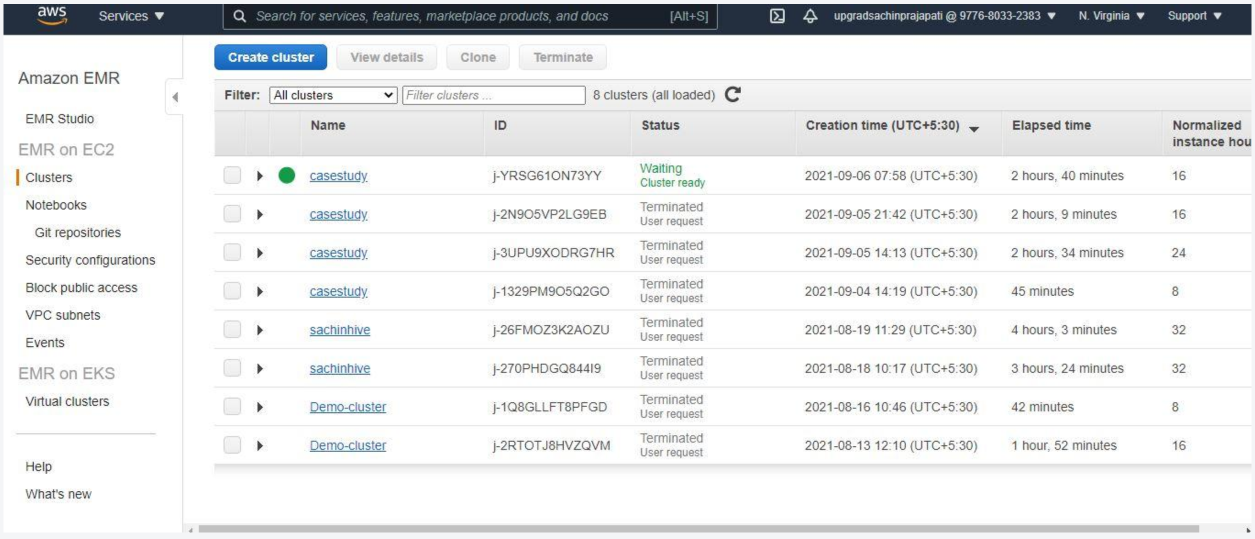
Problem Statement

With online sales gaining popularity, tech companies are exploring ways to improve their sales by analysing customer behaviour and gaining insights about product trends. Furthermore, the websites make it easier for customers to find the products they require without much scavenging. Needless to say, the role of big data analysts is among the most sought-after job profiles of this decade. Therefore, as part of this assignment, we will be challenging you, as a big data analyst, to extract data and gather insights from a real-life data set of an e-commerce company.

The implementation phase can be divided into the following parts:

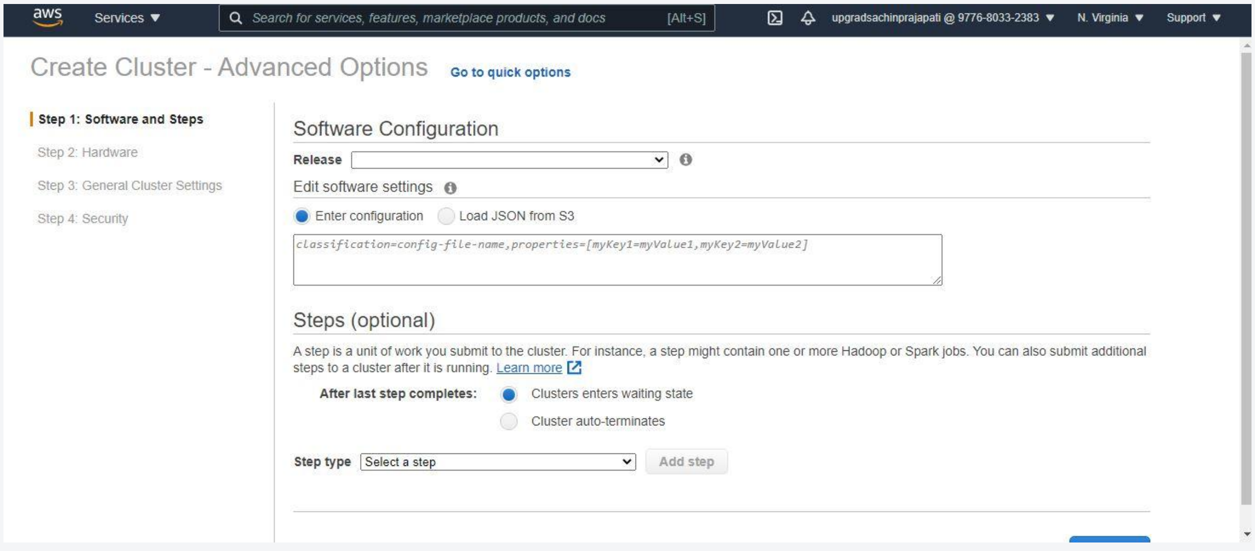
- Copying the data set into the HDFS:
    - Launch an EMR cluster that utilizes the Hive services :

Create EMR Cluster with version as 5.29.0 and having core, master as m4.large.



8 clusters (all loaded)

Name	ID	Status	Creation time (UTC+5:30)	Elapsed time	Normalized instance hou
casestudy	j-YRSG61ON73YY	Waiting Cluster ready	2021-09-06 07:58 (UTC+5:30)	2 hours, 40 minutes	16
casestudy	j-2N9O5VP2LG9EB	Terminated User request	2021-09-05 21:42 (UTC+5:30)	2 hours, 9 minutes	16
casestudy	j-3UPU9XODRG7HR	Terminated User request	2021-09-05 14:13 (UTC+5:30)	2 hours, 34 minutes	24
casestudy	j-1329PM9O5Q2GO	Terminated User request	2021-09-04 14:19 (UTC+5:30)	45 minutes	8
sachinhive	j-26FMOZ3K2AOZU	Terminated User request	2021-08-19 11:29 (UTC+5:30)	4 hours, 3 minutes	32
sachinhive	j-270PHDGQ844I9	Terminated User request	2021-08-18 10:17 (UTC+5:30)	3 hours, 24 minutes	32
Demo-cluster	j-1Q8GLLFT8PFGD	Terminated User request	2021-08-16 10:46 (UTC+5:30)	42 minutes	8
Demo-cluster	j-2RTOTJ8HVZQVM	Terminated User request	2021-08-13 12:10 (UTC+5:30)	1 hour, 52 minutes	16



aws
Services ▼
Search for services, features, marketplace products, and docs [Alt+S]
upgradsachinprajapati @ 9776-8033-2383 ▼
N. Virginia ▼
Support ▼

Create Cluster - Advanced Options **Go to quick options**

Step 1: Software and Steps
Step 2: Hardware
Step 3: General Cluster Settings
Step 4: Security

Software Configuration

Release [ ]

Edit software settings

● Enter configuration ○ Load JSON from S3

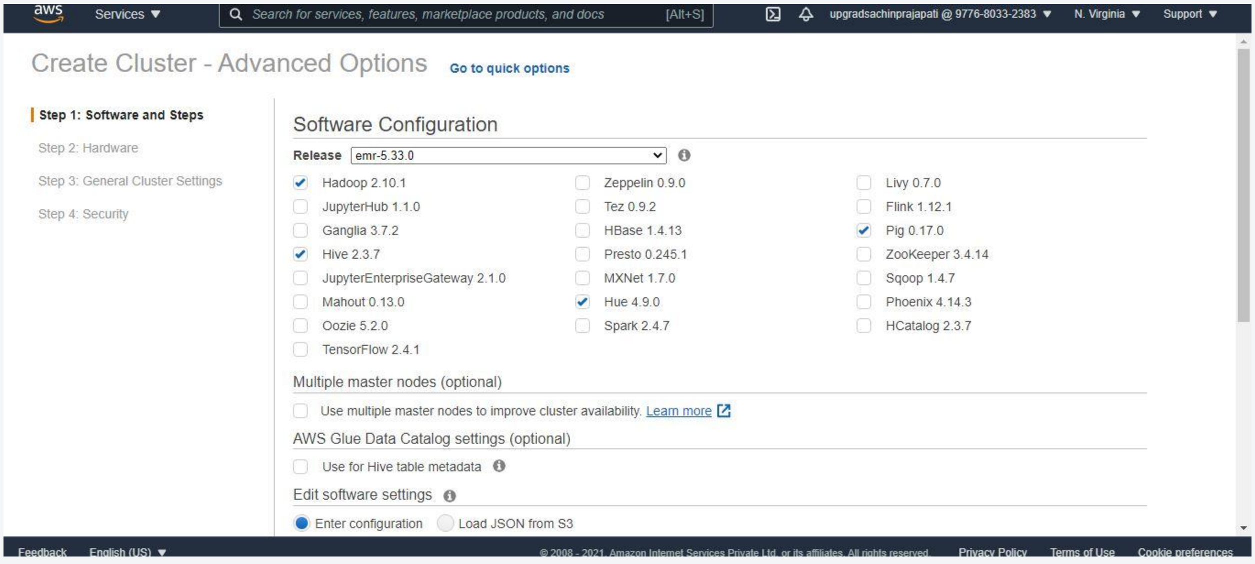
```
classification=config-file-name,properties=[myKey1=myValue1,myKey2=myValue2]
```

Steps (optional)

A step is a unit of work you submit to the cluster. For instance, a step might contain one or more Hadoop or Spark jobs. You can also submit additional steps to a cluster after it is running. Learn more

After last step completes: ● Clusters enters waiting state ○ Cluster auto-terminates

Step type Select a step   Add step



aws
Services ▼
Search for services, features, marketplace products, and docs [Alt+S]
upgradsachinprajapati @ 9776-8033-2383 ▼
N. Virginia ▼
Support ▼

Create Cluster - Advanced Options **Go to quick options**

Step 1: Software and Steps
Step 2: Hardware
Step 3: General Cluster Settings
Step 4: Security

Software Configuration

Release emr-5.33.0

☑ Hadoop 2.10.1
□ JupyterHub 1.1.0
□ Ganglia 3.7.2
☑ Hive 2.3.7
□ JupyterEnterpriseGateway 2.1.0
□ Mahout 0.13.0
□ Oozie 5.2.0
□ TensorFlow 2.4.1
□ Zeppelin 0.9.0
□ Tez 0.9.2
□ HBase 1.4.13
□ Presto 0.245.1
□ MXNet 1.7.0
☑ Hue 4.9.0
□ Spark 2.4.7
□ Livy 0.7.0
□ Flink 1.12.1
☑ Pig 0.17.0
□ ZooKeeper 3.4.14
□ Sqoop 1.4.7
□ Phoenix 4.14.3
□ HCatalog 2.3.7

Multiple master nodes (optional)

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

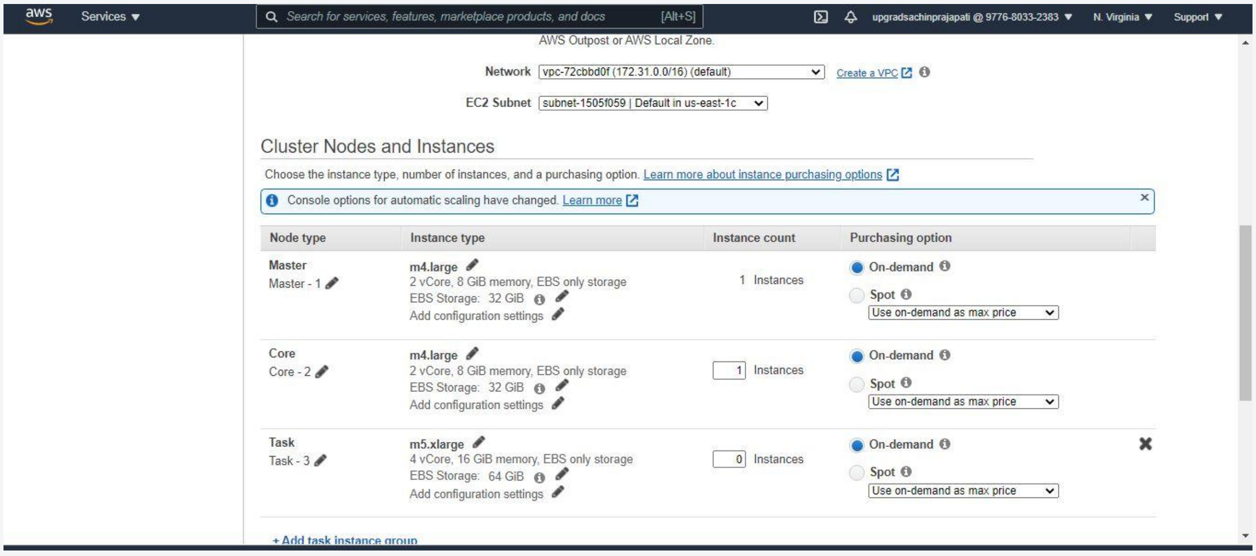
AWS Glue Data Catalog settings (optional)

□ Use for Hive table metadata

Edit software settings

● Enter configuration ○ Load JSON from S3

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]
upgradsachinprajapati @ 9776-8033-2383 ▼
N. Virginia ▼
Support ▼

AWS Outpost or AWS Local Zone.

Network: vpc-72cbbd0f (172.31.0.0/16) (default) — Create a VPC

EC2 Subnet: subnet-1505f059 | Default in us-east-1c

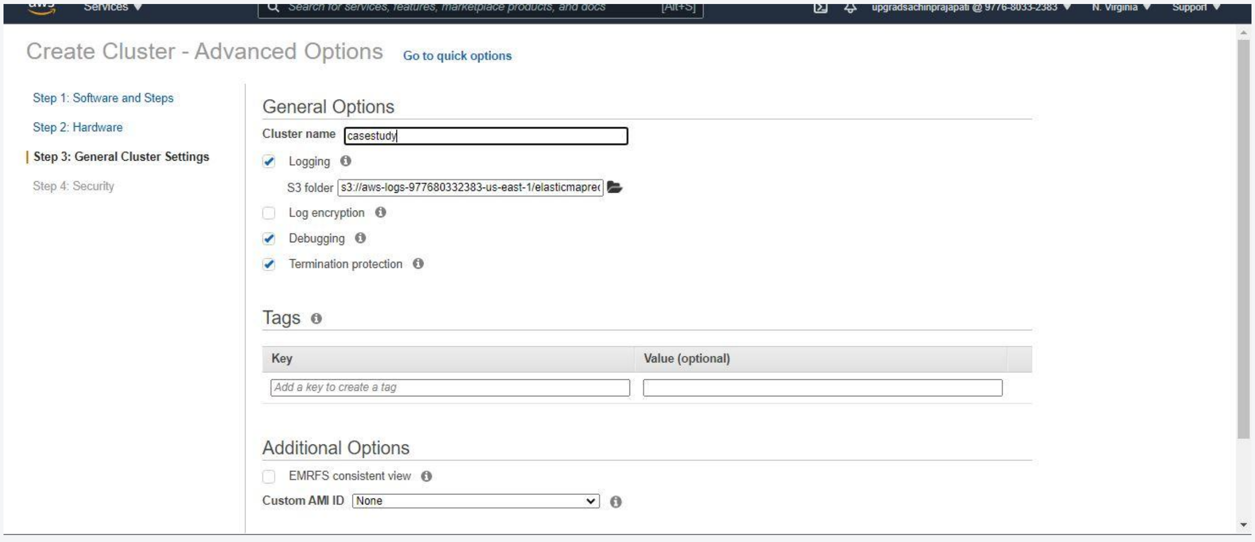
Cluster Nodes and Instances

Choose the instance type, number of instances, and a purchasing option. Learn more about instance purchasing options

Console options for automatic scaling have changed. Learn more

Node type	Instance type	Instance count	Purchasing option
Master Master - 1	m4.large 2 vCore, 8 GiB memory, EBS only storage EBS Storage: 32 GiB Add configuration settings	1 Instances	● On-demand ○ Spot Use on-demand as max price
Core Core - 2	m4.large 2 vCore, 8 GiB memory, EBS only storage EBS Storage: 32 GiB Add configuration settings	1 Instances	● On-demand ○ Spot Use on-demand as max price
Task Task - 3	m5.xlarge 4 vCore, 16 GiB memory, EBS only storage EBS Storage: 64 GiB Add configuration settings	0 Instances	● On-demand ○ Spot Use on-demand as max price

+ Add task instance group



aws
Services ▼
Search for services, features, marketplace products, and docs [Alt+S]
upgradsachinprajapati @ 9776-8033-2383 ▼
N. Virginia ▼
Support ▼

Create Cluster - Advanced Options   Go to quick options

Step 1: Software and Steps
Step 2: Hardware
Step 3: General Cluster Settings
Step 4: Security

General Options

Cluster name: casestudy

☑ Logging

S3 folder: s3://aws-logs-977680332383-us-east-1/elasticmapre(

□ Log encryption

☑ Debugging

☑ Termination protection

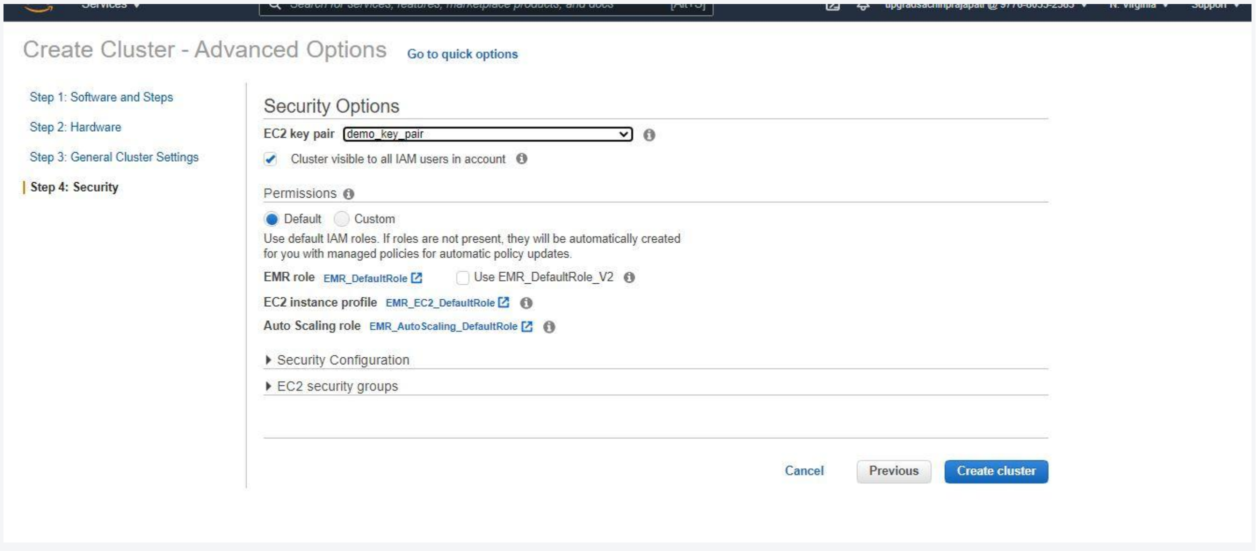
Tags

Key	Value (optional)
Add a key to create a tag	

Additional Options

□ EMRFS consistent view

Custom AMI ID: None



Create Cluster - Advanced Options Go to quick options

Security Options

EC2 key pair: demo_key_pair

☑ 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.

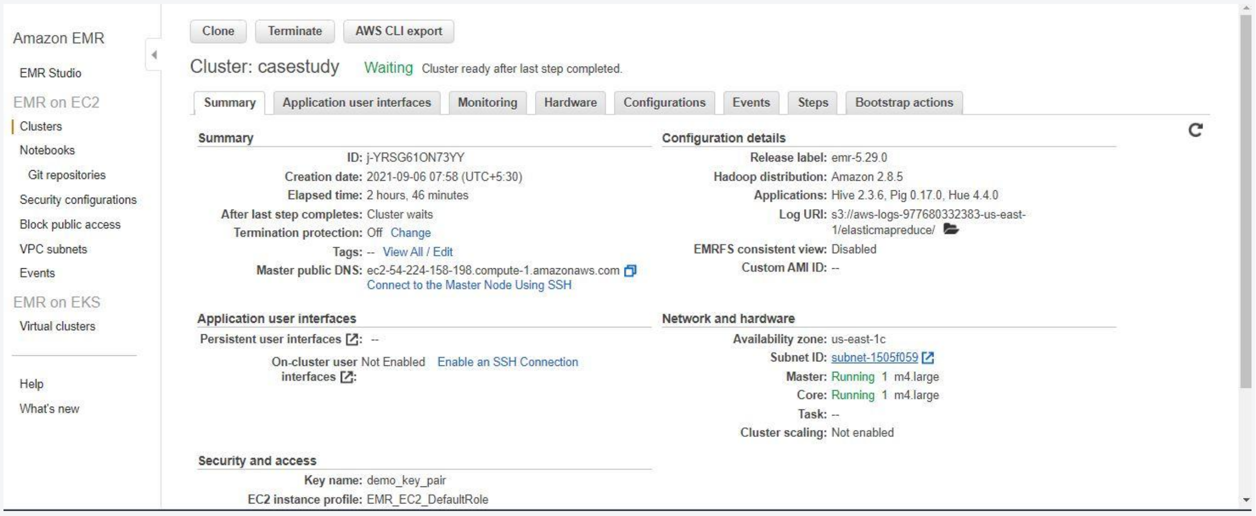
EMR role EMR_DefaultRole □ Use EMR_DefaultRole_V2

EC2 instance profile EMR_EC2_DefaultRole

Auto Scaling role EMR_AutoScaling_DefaultRole

▸ Security Configuration

▸ EC2 security groups



Cluster: casestudy Waiting Cluster ready after last step completed.

Summary

ID: j-YRSG61ON73YY
Creation date: 2021-09-06 07:58 (UTC+5:30)
Elapsed time: 2 hours, 46 minutes
After last step completes: Cluster waits
Termination protection: Off Change
Tags: -- View All / Edit
Master public DNS: ec2-54-224-158-198.compute-1.amazonaws.com
Connect to the Master Node Using SSH

Configuration details

Release label: emr-5.29.0
Hadoop distribution: Amazon 2.8.5
Applications: Hive 2.3.6, Pig 0.17.0, Hue 4.4.0
Log URI: s3://aws-logs-977680332383-us-east-1/elasticmapreduce/
EMRFS consistent view: Disabled
Custom AMI ID: --

Application user interfaces

Persistent user interfaces: --
On-cluster user interfaces: Not Enabled Enable an SSH Connection

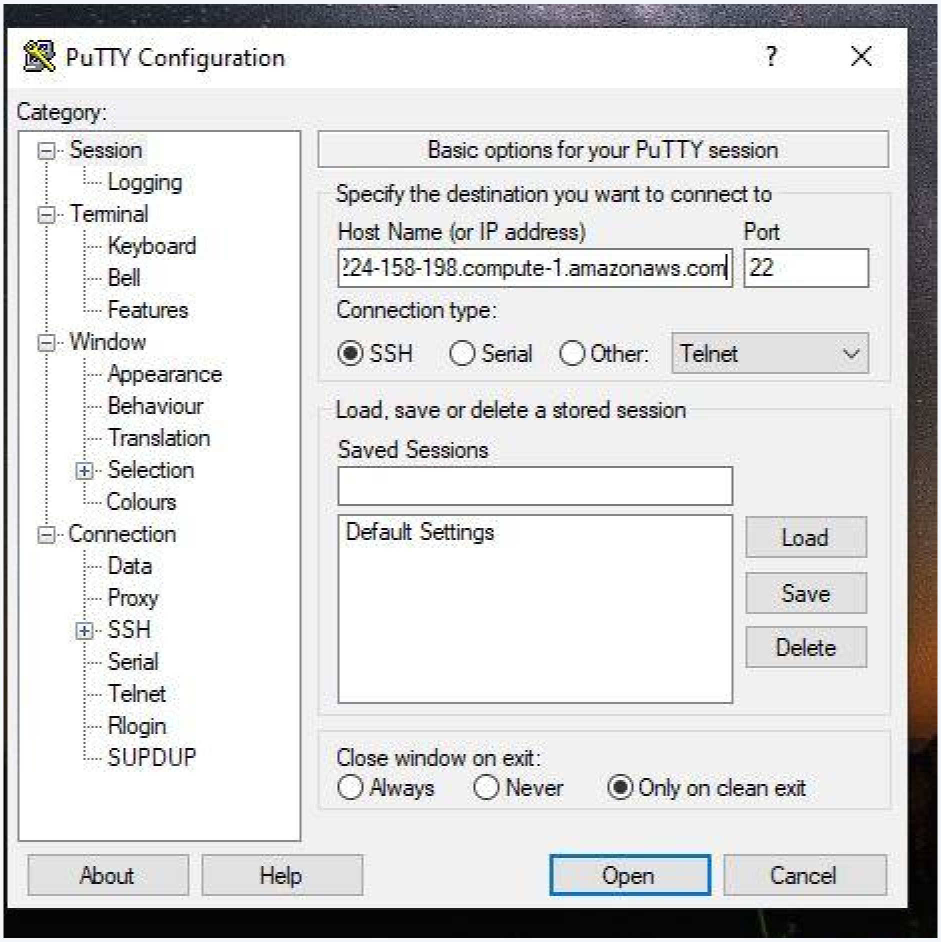
Network and hardware

Availability zone: us-east-1c
Subnet ID: subnet-1505f059
Master: Running 1 m4.large
Core: Running 1 m4.large
Task: --
Cluster scaling: Not enabled

Security and access

Key name: demo_key_pair
EC2 instance profile: EMR_EC2_DefaultRole

Connect the master node using Putty with the .ppk key value pair file .



Specify the destination you want to connect to

Host Name (or IP address): 24-158-198.compute-1.amazonaws.com

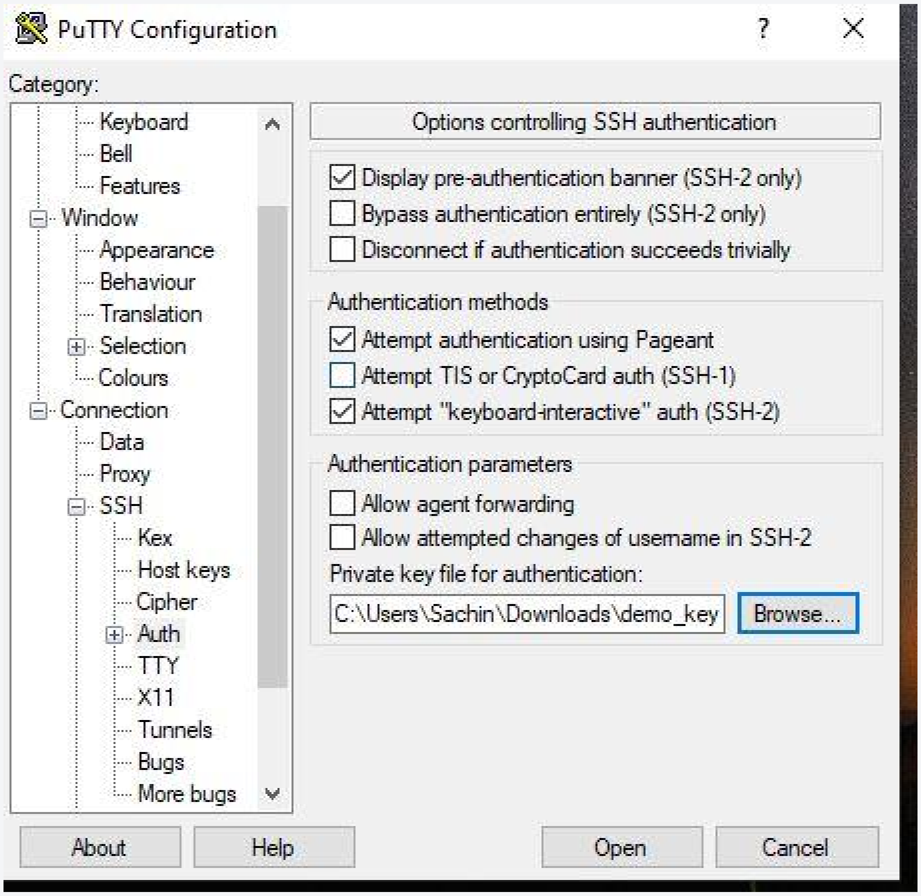
Port: 22

Connection type: ● SSH ○ Serial ○ Other: Telnet

Saved Sessions:

Default Settings

Close window on exit: ○ Always ○ Never ● Only on clean exit



Options controlling SSH authentication

☑ Display pre-authentication banner (SSH-2 only)
☐ Bypass authentication entirely (SSH-2 only)
☐ Disconnect if authentication succeeds trivially

Authentication methods

☑ Attempt authentication using Pageant
☐ Attempt TIS or CryptoCard auth (SSH-1)
☑ Attempt "keyboard-interactive" auth (SSH-2)

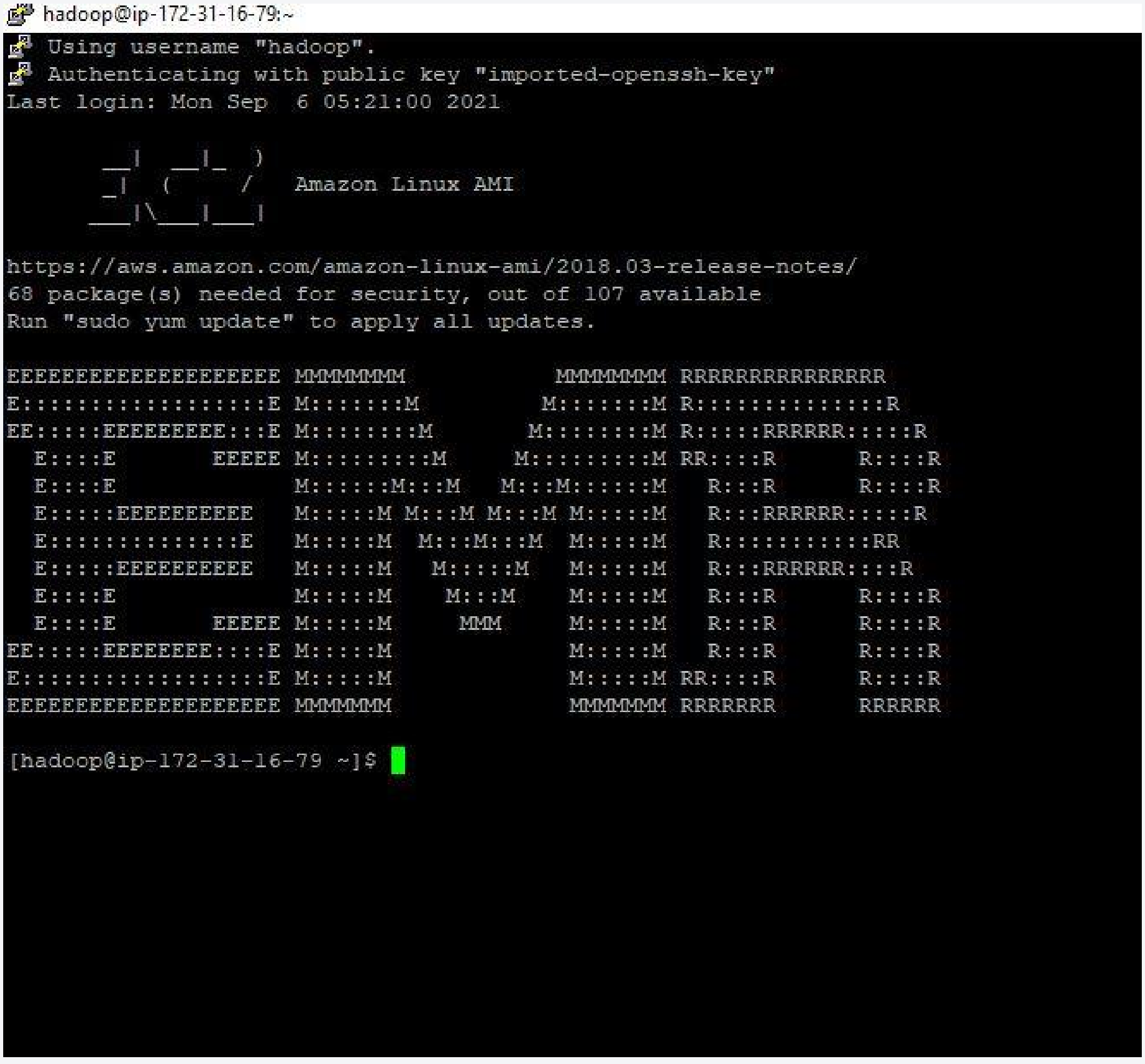
Authentication parameters

☐ Allow agent forwarding
☐ Allow attempted changes of usemame in SSH-2

Private key file for authentication:

C:\Users\Sachin\Downloads\demo_key

Browse...

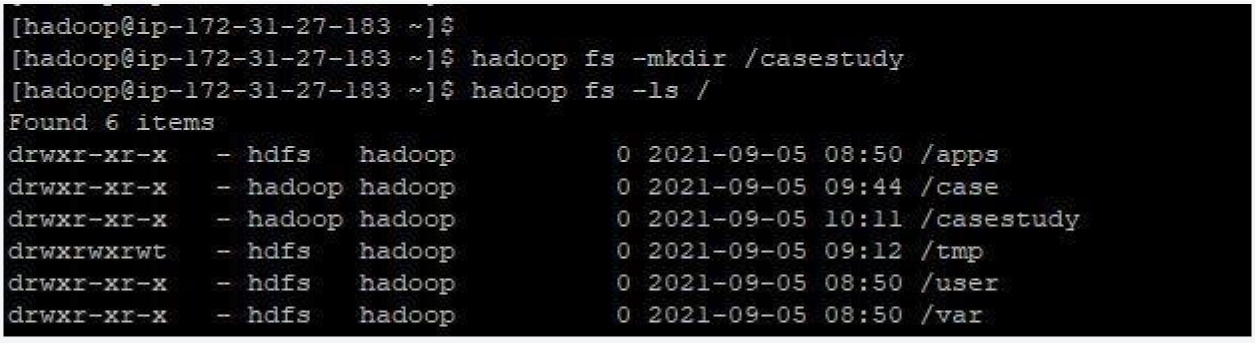


hadoop@ip-172-31-16-79:~

```
Using username "hadoop".
Authenticating with public key "imported-openssh-key"
Last login: Mon Sep  6 05:21:00 2021

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

https://aws.amazon.com/amazon-linux-ami/2018.03-release-notes/
68 package(s) needed for security, out of 107 available
Run "sudo yum update" to apply all updates.

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-172-31-16-79 ~]$
```

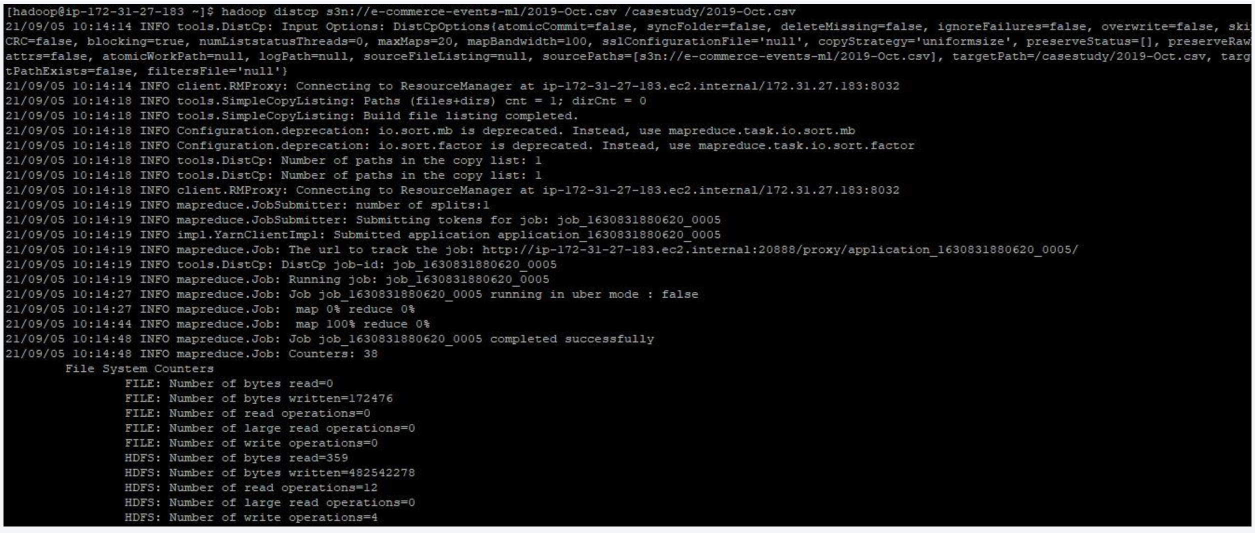
○ Move the data from the S3 bucket into the HDFS

Create directory in Hadoop HDFS using hadoop fs -mkdir /casestudy

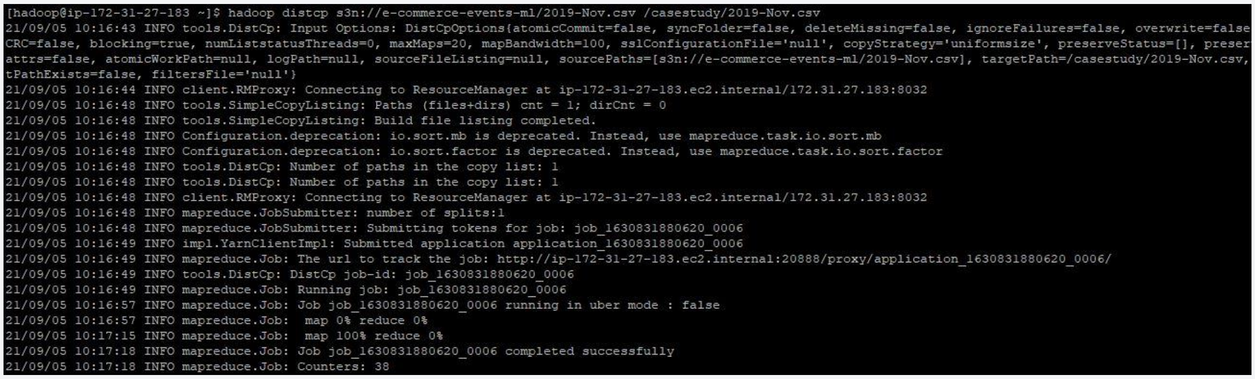


```
[hadoop@ip-172-31-27-183 ~]$
[hadoop@ip-172-31-27-183 ~]$ hadoop fs -mkdir /casestudy
[hadoop@ip-172-31-27-183 ~]$ hadoop fs -ls /
Found 6 items
drwxr-xr-x   - hdfs   hadoop          0 2021-09-05 08:50 /apps
drwxr-xr-x   - hadoop hadoop          0 2021-09-05 09:44 /case
drwxr-xr-x   - hadoop hadoop          0 2021-09-05 10:11 /casestudy
drwxrwxrwt   - hdfs   hadoop          0 2021-09-05 09:12 /tmp
drwxr-xr-x   - hdfs   hadoop          0 2021-09-05 08:50 /user
drwxr-xr-x   - hdfs   hadoop          0 2021-09-05 08:50 /var
```

Copy the data from S3 bucket to HDFS using hadoop distcp command.

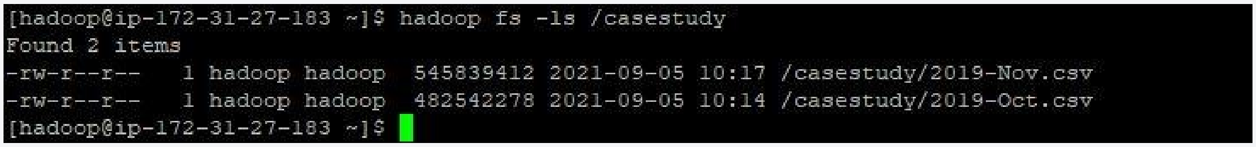


```
[hadoop@ip-172-31-27-183 ~]$ hadoop distcp s3n://e-commerce-events-ml/2019-Oct.csv /casestudy/2019-Oct.csv
21/09/05 10:14:14 INFO tools.DistCp: Input Options: DistCpOptions{atomicCommit=false, syncFolder=false, deleteMissing=false, ignoreFailures=false, overwrite=false, ski
CRC=false, blocking=true, numListstatusThreads=0, maxMaps=20, mapBandwidth=100, sslConfigurationFile='null', copyStrategy='uniformsize', preserveStatus=[], preserveRaw
attrs=false, atomicWorkPath=null, logPath=null, sourceFileListing=null, sourcePaths=[s3n://e-commerce-events-ml/2019-Oct.csv], targetPath=/casestudy/2019-Oct.csv, targ
tPathExists=false, filtersFile='null'}
21/09/05 10:14:14 INFO client.RMProxy: Connecting to ResourceManager at ip-172-31-27-183.ec2.internal/172.31.27.183:8032
21/09/05 10:14:18 INFO tools.SimpleCopyListing: Paths (files+dirs) cnt = 1; dirCnt = 0
21/09/05 10:14:18 INFO tools.SimpleCopyListing: Build file listing completed.
21/09/05 10:14:18 INFO Configuration.deprecation: io.sort.mb is deprecated. Instead, use mapreduce.task.io.sort.mb
21/09/05 10:14:18 INFO Configuration.deprecation: io.sort.factor is deprecated. Instead, use mapreduce.task.io.sort.factor
21/09/05 10:14:18 INFO tools.DistCp: Number of paths in the copy list: 1
21/09/05 10:14:18 INFO tools.DistCp: Number of paths in the copy list: 1
21/09/05 10:14:18 INFO client.RMProxy: Connecting to ResourceManager at ip-172-31-27-183.ec2.internal/172.31.27.183:8032
21/09/05 10:14:19 INFO mapreduce.JobSubmitter: number of splits:1
21/09/05 10:14:19 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1630831880620_0005
21/09/05 10:14:19 INFO impl.YarnClientImpl: Submitted application application_1630831880620_0005
21/09/05 10:14:19 INFO mapreduce.Job: The url to track the job: http://ip-172-31-27-183.ec2.internal:20888/proxy/application_1630831880620_0005/
21/09/05 10:14:19 INFO tools.DistCp: DistCp job-id: job_1630831880620_0005
21/09/05 10:14:19 INFO mapreduce.Job: Running job: job_1630831880620_0005
21/09/05 10:14:27 INFO mapreduce.Job: Job job_1630831880620_0005 running in uber mode : false
21/09/05 10:14:27 INFO mapreduce.Job:  map 0% reduce 0%
21/09/05 10:14:44 INFO mapreduce.Job:  map 100% reduce 0%
21/09/05 10:14:48 INFO mapreduce.Job: Job job_1630831880620_0005 completed successfully
21/09/05 10:14:48 INFO mapreduce.Job: Counters: 38
        File System Counters
                FILE: Number of bytes read=0
                FILE: Number of bytes written=172476
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=359
                HDFS: Number of bytes written=482542278
                HDFS: Number of read operations=12
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=4
```



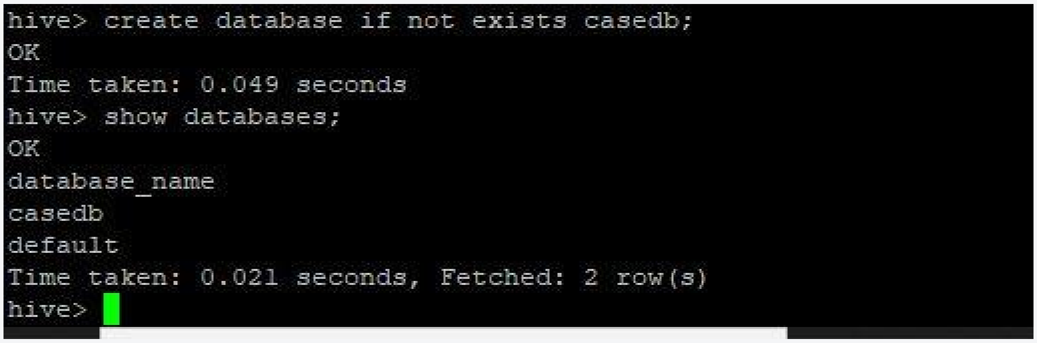
```
[hadoop@ip-172-31-27-183 ~]$ hadoop distcp s3n://e-commerce-events-ml/2019-Nov.csv /casestudy/2019-Nov.csv
21/09/05 10:16:43 INFO tools.DistCp: Input Options: DistCpOptions{atomicCommit=false, syncFolder=false, deleteMissing=false, ignoreFailures=false, overwrite=false
CRC=false, blocking=true, numListstatusThreads=0, maxMaps=20, mapBandwidth=100, sslConfigurationFile='null', copyStrategy='uniformsize', preserveStatus=[], preser
attrs=false, atomicWorkPath=null, logPath=null, sourceFileListing=null, sourcePaths=[s3n://e-commerce-events-ml/2019-Nov.csv], targetPath=/casestudy/2019-Nov.csv,
tPathExists=false, filtersFile='null'}
21/09/05 10:16:44 INFO client.RMProxy: Connecting to ResourceManager at ip-172-31-27-183.ec2.internal/172.31.27.183:8032
21/09/05 10:16:48 INFO tools.SimpleCopyListing: Paths (files+dirs) cnt = 1; dirCnt = 0
21/09/05 10:16:48 INFO tools.SimpleCopyListing: Build file listing completed.
21/09/05 10:16:48 INFO Configuration.deprecation: io.sort.mb is deprecated. Instead, use mapreduce.task.io.sort.mb
21/09/05 10:16:48 INFO Configuration.deprecation: io.sort.factor is deprecated. Instead, use mapreduce.task.io.sort.factor
21/09/05 10:16:48 INFO tools.DistCp: Number of paths in the copy list: 1
21/09/05 10:16:48 INFO tools.DistCp: Number of paths in the copy list: 1
21/09/05 10:16:48 INFO client.RMProxy: Connecting to ResourceManager at ip-172-31-27-183.ec2.internal/172.31.27.183:8032
21/09/05 10:16:48 INFO mapreduce.JobSubmitter: number of splits:1
21/09/05 10:16:48 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1630831880620_0006
21/09/05 10:16:49 INFO impl.YarnClientImpl: Submitted application application_1630831880620_0006
21/09/05 10:16:49 INFO mapreduce.Job: The url to track the job: http://ip-172-31-27-183.ec2.internal:20888/proxy/application_1630831880620_0006/
21/09/05 10:16:49 INFO tools.DistCp: DistCp job-id: job_1630831880620_0006
21/09/05 10:16:49 INFO mapreduce.Job: Running job: job_1630831880620_0006
21/09/05 10:16:57 INFO mapreduce.Job: Job job_1630831880620_0006 running in uber mode : false
21/09/05 10:16:57 INFO mapreduce.Job:  map 0% reduce 0%
21/09/05 10:17:15 INFO mapreduce.Job:  map 100% reduce 0%
21/09/05 10:17:18 INFO mapreduce.Job: Job job_1630831880620_0006 completed successfully
21/09/05 10:17:18 INFO mapreduce.Job: Counters: 38
```

To verify whether data is in HDFS or not, check using hadoop fs -ls /casestudy.



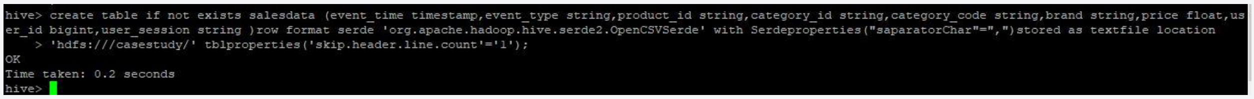
```
[hadoop@ip-172-31-27-183 ~]$ hadoop fs -ls /casestudy
Found 2 items
-rw-r--r--   1 hadoop hadoop  545839412 2021-09-05 10:17 /casestudy/2019-Nov.csv
-rw-r--r--   1 hadoop hadoop  482542278 2021-09-05 10:14 /casestudy/2019-Oct.csv
[hadoop@ip-172-31-27-183 ~]$
```

- **Creating the database and launching Hive queries on your EMR cluster:**
  - ○ Create the structure of your database



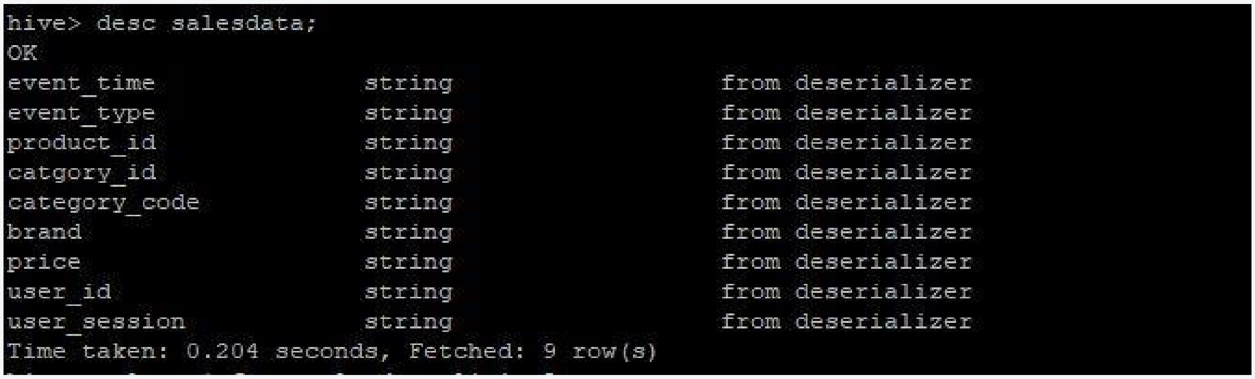
```
hive> create database if not exists casedb;
OK
Time taken: 0.049 seconds
hive> show databases;
OK
database_name
casedb
default
Time taken: 0.021 seconds, Fetched: 2 row(s)
hive>
```

Create Table for ClickStream data, here the file is of csv type so we use csvserde and as header is the first row in the data file so to skip header we use tblproperties as skip.header.line.count =1.

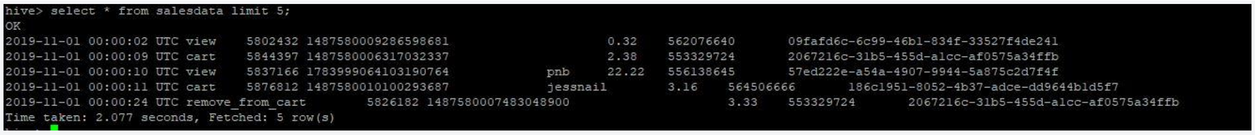


```
hive> create table if not exists salesdata (event_time timestamp,event_type string,product_id string,category_id string,category_code string,brand string,price float,us
er_id bigint,user_session string )row format serde 'org.apache.hadoop.hive.serde2.OpenCSVSerde' with Serdeproperties("saparatorChar"=",")stored as textfile location
    > 'hdfs:///casestudy/' tblproperties('skip.header.line.count'='1');
OK
Time taken: 0.2 seconds
hive>
```

The schema of external table is as follows :



```
hive> desc salesdata;
OK
event_time              string                  from deserializer
event_type              string                  from deserializer
product_id              string                  from deserializer
catgory_id              string                  from deserializer
category_code           string                  from deserializer
brand                   string                  from deserializer
price                   string                  from deserializer
user_id                 string                  from deserializer
user_session            string                  from deserializer
Time taken: 0.204 seconds, Fetched: 9 row(s)
```

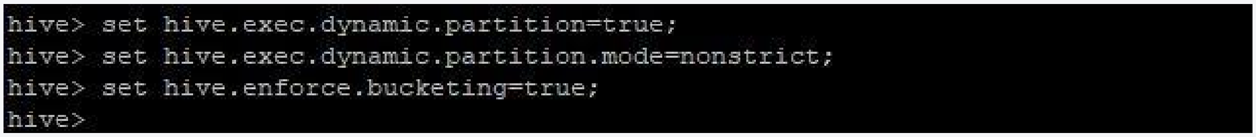


```
hive> select * from salesdata limit 5;
OK
2019-11-01 00:00:02 UTC view    5802432 1487580009286598681                     0.32    562076640       09fafd6c-6c99-46b1-834f-33527f4de241
2019-11-01 00:00:09 UTC cart    5844397 1487580006317032337                     2.38    553329724       2067216c-31b5-455d-a1cc-af0575a34ffb
2019-11-01 00:00:10 UTC view    5837166 1783999064103190764             pnb     22.22   556138645       57ed222e-a54a-4907-9944-5a875c2d7f4f
2019-11-01 00:00:11 UTC cart    5876812 1487580010100293687             jessnail        3.16    564506666       186c1951-8052-4b37-adce-dd9644b1d5f7
2019-11-01 00:00:24 UTC remove_from_cart        5826182 1487580007483048900                     3.33    553329724       2067216c-31b5-455d-a1cc-af0575a34ffb
Time taken: 2.077 seconds, Fetched: 5 row(s)
```

- ○ Use optimized techniques to run your queries as efficiently as possible

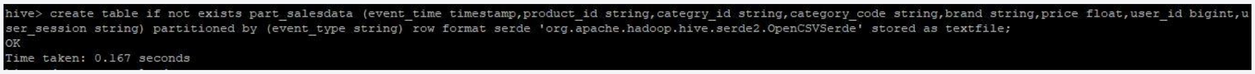
To improve query time we use partitioning and bucketing .

To set the dynamic partitioning & bucketing we have to use below syntax .



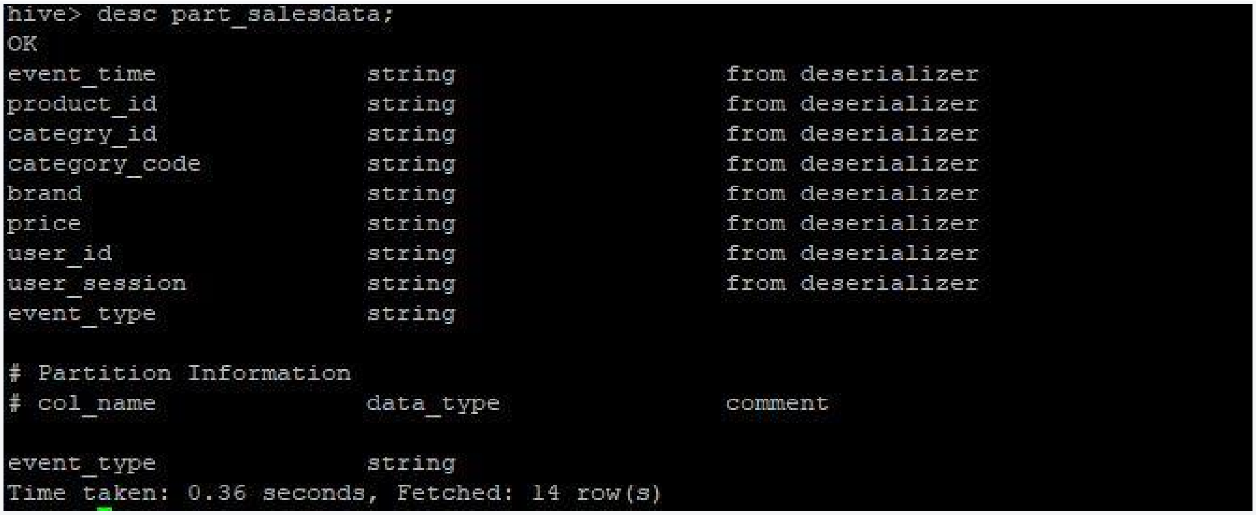
```
hive> set hive.exec.dynamic.partition=true;
hive> set hive.exec.dynamic.partition.mode=nonstrict;
hive> set hive.enforce.bucketing=true;
hive>
```

Partitioning table created using event_type ‘column’.



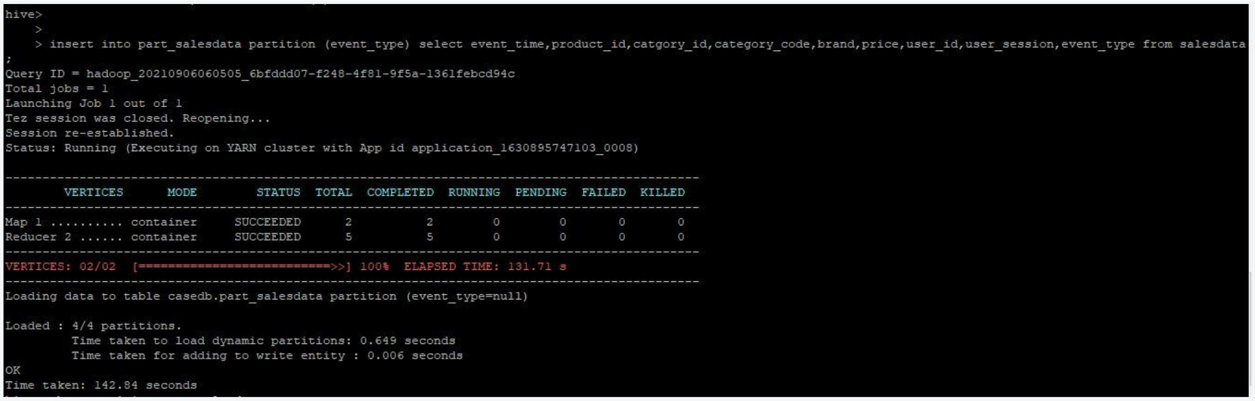
```
hive> create table if not exists part_salesdata (event_time timestamp,product_id string,categry_id string,category_code string,brand string,price float,user_id bigint,u
ser_session string) partitioned by (event_type string) row format serde 'org.apache.hadoop.hive.serde2.OpenCSVSerde' stored as textfile;
OK
Time taken: 0.167 seconds
```

The schema of partition table is as follows :



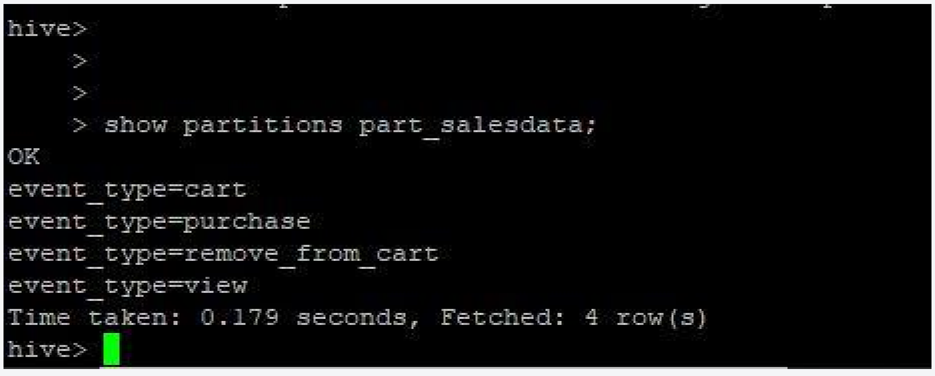
```
hive> desc part_salesdata;
OK
event_time              string                  from deserializer
product_id              string                  from deserializer
categry_id              string                  from deserializer
category_code           string                  from deserializer
brand                   string                  from deserializer
price                   string                  from deserializer
user_id                 string                  from deserializer
user_session            string                  from deserializer
event_type              string

# Partition Information
# col_name              data_type               comment

event_type              string
Time taken: 0.36 seconds, Fetched: 14 row(s)
```

Insert values in the partition table using the external salesdata table.



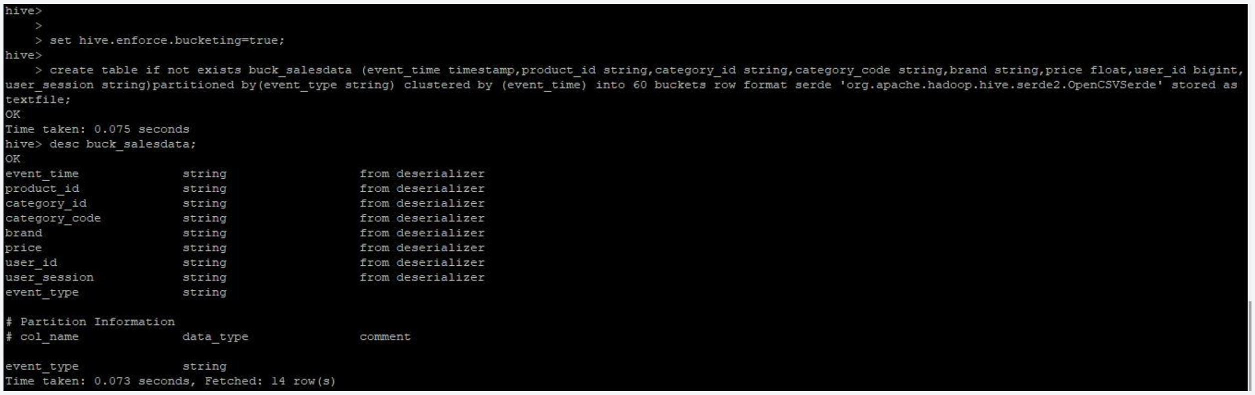
```
hive>
    >
    > insert into part_salesdata partition (event_type) select event_time,product_id,catgory_id,category_code,brand,price,user_id,user_session,event_type from salesdata
;
Query ID = hadoop_20210906060505_6bfddd07-f248-4f81-9f5a-1361febcd94c
Total jobs = 1
Launching Job 1 out of 1
Tez session was closed. Reopening...
Session re-established.
Status: Running (Executing on YARN cluster with App id application_1630895747103_0008)

----------------------------------------------------------------------------------------------
        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: 131.71 s
----------------------------------------------------------------------------------------------
Loading data to table casedb.part_salesdata partition (event_type=null)

Loaded : 4/4 partitions.
         Time taken to load dynamic partitions: 0.649 seconds
         Time taken for adding to write entity : 0.006 seconds
OK
Time taken: 142.84 seconds
```

To check whether the partition table has a partition or not. We can check in the below command .



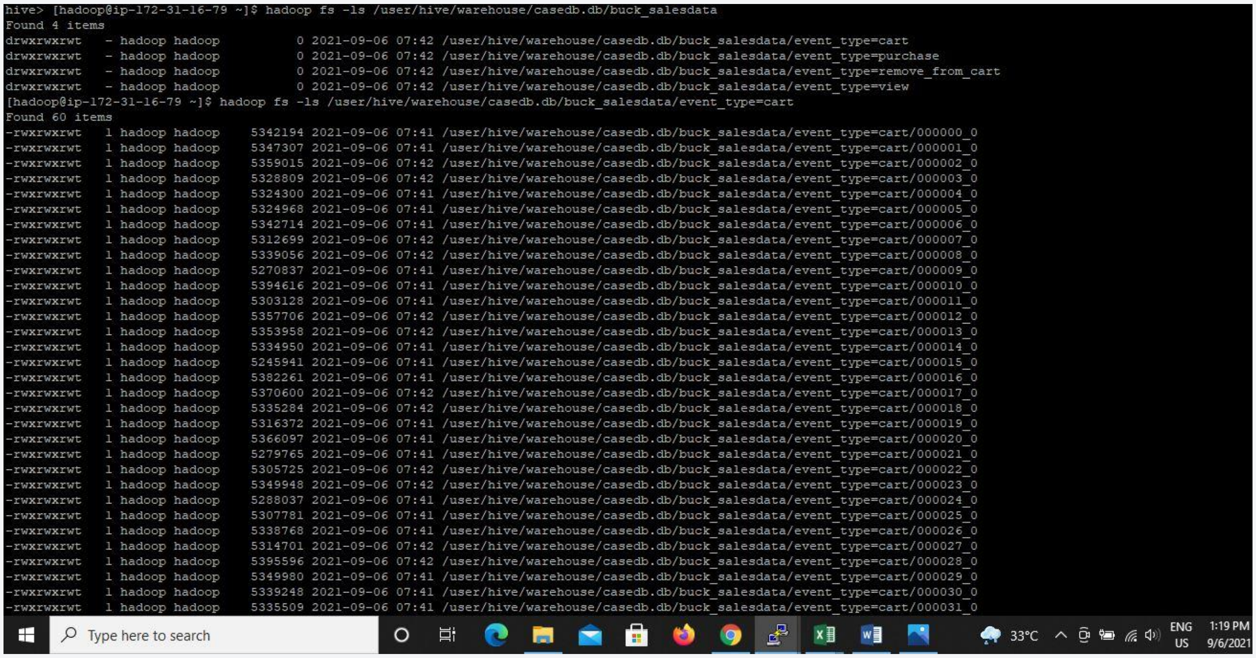
```
hive>
    >
    >
    > show partitions part_salesdata;
OK
event_type=cart
event_type=purchase
event_type=remove_from_cart
event_type=view
Time taken: 0.179 seconds, Fetched: 4 row(s)
hive>
```

Bucketing table created using clustered by keyword on column event_time.

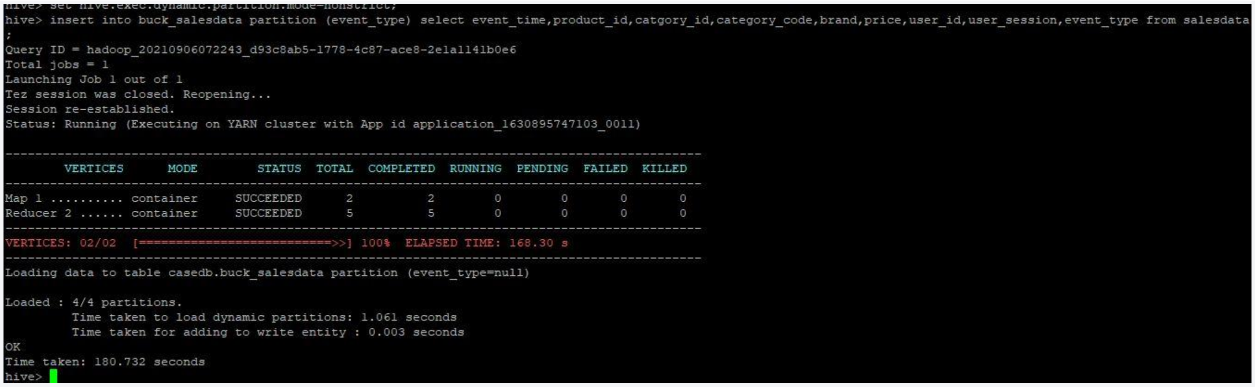


```
hive>
    >
    > set hive.enforce.bucketing=true;
hive>
    > create table if not exists buck_salesdata (event_time timestamp,product_id string,category_id string,category_code string,brand string,price float,user_id bigint,
user_session string)partitioned by(event_type string) clustered by (event_time) into 60 buckets row format serde 'org.apache.hadoop.hive.serde2.OpenCSVSerde' stored as
textfile;
OK
Time taken: 0.075 seconds
hive> desc buck_salesdata;
OK
event_time              string                  from deserializer
product_id              string                  from deserializer
category_id             string                  from deserializer
category_code           string                  from deserializer
brand                   string                  from deserializer
price                   string                  from deserializer
user_id                 string                  from deserializer
user_session            string                  from deserializer
event_type              string

# Partition Information
# col_name              data_type               comment

event_type              string
Time taken: 0.073 seconds, Fetched: 14 row(s)
```

To check whether a bucket is created or not, we can achieve this as follows.

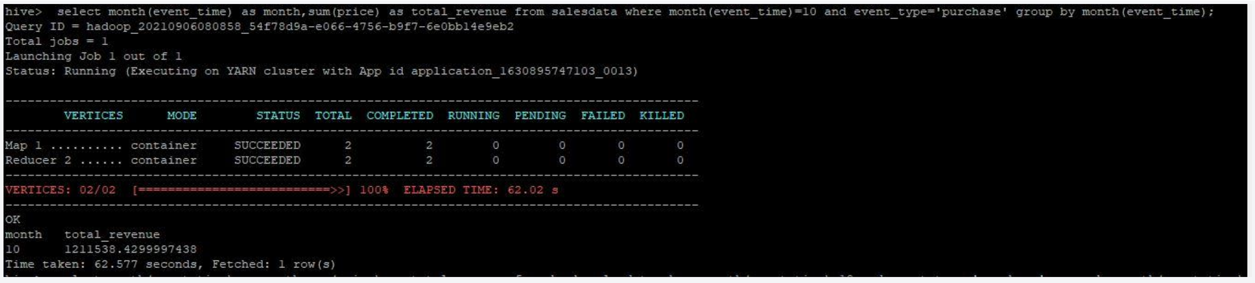


```
hive> [hadoop@ip-172-31-16-79 ~]$ hadoop fs -ls /user/hive/warehouse/casedb.db/buck_salesdata
Found 4 items
drwxrwxrwt   - hadoop hadoop          0 2021-09-06 07:42 /user/hive/warehouse/casedb.db/buck_salesdata/event_type=cart
drwxrwxrwt   - hadoop hadoop          0 2021-09-06 07:42 /user/hive/warehouse/casedb.db/buck_salesdata/event_type=purchase
drwxrwxrwt   - hadoop hadoop          0 2021-09-06 07:42 /user/hive/warehouse/casedb.db/buck_salesdata/event_type=remove_from_cart
drwxrwxrwt   - hadoop hadoop          0 2021-09-06 07:42 /user/hive/warehouse/casedb.db/buck_salesdata/event_type=view
[hadoop@ip-172-31-16-79 ~]$ hadoop fs -ls /user/hive/warehouse/casedb.db/buck_salesdata/event_type=cart
Found 60 items
-rwxrwxrwt   1 hadoop hadoop    5342194 2021-09-06 07:41 /user/hive/warehouse/casedb.db/buck_salesdata/event_type=cart/000000_0
-rwxrwxrwt   1 hadoop hadoop    5347307 2021-09-06 07:41 /user/hive/warehouse/casedb.db/buck_salesdata/event_type=cart/000001_0
-rwxrwxrwt   1 hadoop hadoop    5359015 2021-09-06 07:42 /user/hive/warehouse/casedb.db/buck_salesdata/event_type=cart/000002_0
-rwxrwxrwt   1 hadoop hadoop    5328809 2021-09-06 07:42 /user/hive/warehouse/casedb.db/buck_salesdata/event_type=cart/000003_0
-rwxrwxrwt   1 hadoop hadoop    5324300 2021-09-06 07:41 /user/hive/warehouse/casedb.db/buck_salesdata/event_type=cart/000004_0
-rwxrwxrwt   1 hadoop hadoop    5324968 2021-09-06 07:41 /user/hive/warehouse/casedb.db/buck_salesdata/event_type=cart/000005_0
-rwxrwxrwt   1 hadoop hadoop    5342714 2021-09-06 07:41 /user/hive/warehouse/casedb.db/buck_salesdata/event_type=cart/000006_0
-rwxrwxrwt   1 hadoop hadoop    5312699 2021-09-06 07:42 /user/hive/warehouse/casedb.db/buck_salesdata/event_type=cart/000007_0
-rwxrwxrwt   1 hadoop hadoop    5339056 2021-09-06 07:42 /user/hive/warehouse/casedb.db/buck_salesdata/event_type=cart/000008_0
-rwxrwxrwt   1 hadoop hadoop    5270837 2021-09-06 07:41 /user/hive/warehouse/casedb.db/buck_salesdata/event_type=cart/000009_0
-rwxrwxrwt   1 hadoop hadoop    5394616 2021-09-06 07:41 /user/hive/warehouse/casedb.db/buck_salesdata/event_type=cart/000010_0
-rwxrwxrwt   1 hadoop hadoop    5303128 2021-09-06 07:41 /user/hive/warehouse/casedb.db/buck_salesdata/event_type=cart/000011_0
-rwxrwxrwt   1 hadoop hadoop    5357706 2021-09-06 07:42 /user/hive/warehouse/casedb.db/buck_salesdata/event_type=cart/000012_0
-rwxrwxrwt   1 hadoop hadoop    5353958 2021-09-06 07:42 /user/hive/warehouse/casedb.db/buck_salesdata/event_type=cart/000013_0
-rwxrwxrwt   1 hadoop hadoop    5334950 2021-09-06 07:41 /user/hive/warehouse/casedb.db/buck_salesdata/event_type=cart/000014_0
-rwxrwxrwt   1 hadoop hadoop    5245941 2021-09-06 07:41 /user/hive/warehouse/casedb.db/buck_salesdata/event_type=cart/000015_0
-rwxrwxrwt   1 hadoop hadoop    5382261 2021-09-06 07:41 /user/hive/warehouse/casedb.db/buck_salesdata/event_type=cart/000016_0
-rwxrwxrwt   1 hadoop hadoop    5370600 2021-09-06 07:42 /user/hive/warehouse/casedb.db/buck_salesdata/event_type=cart/000017_0
-rwxrwxrwt   1 hadoop hadoop    5335284 2021-09-06 07:42 /user/hive/warehouse/casedb.db/buck_salesdata/event_type=cart/000018_0
-rwxrwxrwt   1 hadoop hadoop    5316372 2021-09-06 07:41 /user/hive/warehouse/casedb.db/buck_salesdata/event_type=cart/000019_0
-rwxrwxrwt   1 hadoop hadoop    5366097 2021-09-06 07:41 /user/hive/warehouse/casedb.db/buck_salesdata/event_type=cart/000020_0
-rwxrwxrwt   1 hadoop hadoop    5279765 2021-09-06 07:41 /user/hive/warehouse/casedb.db/buck_salesdata/event_type=cart/000021_0
-rwxrwxrwt   1 hadoop hadoop    5305725 2021-09-06 07:42 /user/hive/warehouse/casedb.db/buck_salesdata/event_type=cart/000022_0
-rwxrwxrwt   1 hadoop hadoop    5349948 2021-09-06 07:42 /user/hive/warehouse/casedb.db/buck_salesdata/event_type=cart/000023_0
-rwxrwxrwt   1 hadoop hadoop    5288037 2021-09-06 07:41 /user/hive/warehouse/casedb.db/buck_salesdata/event_type=cart/000024_0
-rwxrwxrwt   1 hadoop hadoop    5307781 2021-09-06 07:41 /user/hive/warehouse/casedb.db/buck_salesdata/event_type=cart/000025_0
-rwxrwxrwt   1 hadoop hadoop    5338768 2021-09-06 07:41 /user/hive/warehouse/casedb.db/buck_salesdata/event_type=cart/000026_0
-rwxrwxrwt   1 hadoop hadoop    5314701 2021-09-06 07:42 /user/hive/warehouse/casedb.db/buck_salesdata/event_type=cart/000027_0
-rwxrwxrwt   1 hadoop hadoop    5395596 2021-09-06 07:42 /user/hive/warehouse/casedb.db/buck_salesdata/event_type=cart/000028_0
-rwxrwxrwt   1 hadoop hadoop    5349980 2021-09-06 07:41 /user/hive/warehouse/casedb.db/buck_salesdata/event_type=cart/000029_0
-rwxrwxrwt   1 hadoop hadoop    5339248 2021-09-06 07:41 /user/hive/warehouse/casedb.db/buck_salesdata/event_type=cart/000030_0
-rwxrwxrwt   1 hadoop hadoop    5335509 2021-09-06 07:41 /user/hive/warehouse/casedb.db/buck_salesdata/event_type=cart/000031_0
```



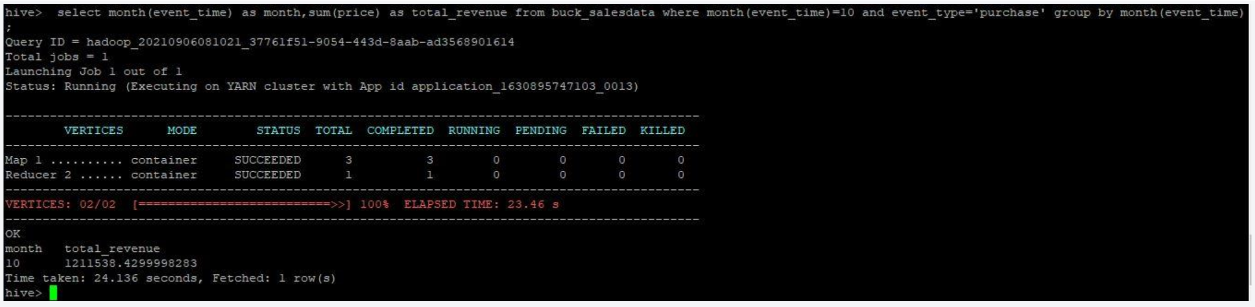
```
hive> set hive.exec.dynamic.partition.mode=nonstrict;
hive> insert into buck_salesdata partition (event_type) select event_time,product_id,catgory_id,category_code,brand,price,user_id,user_session,event_type from salesdata
;
Query ID = hadoop_20210906072243_d93c8ab5-1778-4c87-ace8-2e1a1141b0e6
Total 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_1630895747103_0011)

----------------------------------------------------------------------------------------------
        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: 168.30 s
----------------------------------------------------------------------------------------------
Loading data to table casedb.buck_salesdata partition (event_type=null)

Loaded : 4/4 partitions.
         Time taken to load dynamic partitions: 1.061 seconds
         Time taken for adding to write entity : 0.003 seconds
OK
Time taken: 180.732 seconds
hive> 
```

- ○ Show the improvement of the performance after using optimization on any single query.



```
hive>  select month(event_time) as month,sum(price) as total_revenue from salesdata where month(event_time)=10 and event_type='purchase' group by month(event_time);
Query ID = hadoop_20210906080858_54f78d9a-e066-4756-b9f7-6e0bb14e9eb2
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1630895747103_0013)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      2          2        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      2          2        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 62.02 s
----------------------------------------------------------------------------------------------
OK
month   total_revenue
10      1211538.4299997438
Time taken: 62.577 seconds, Fetched: 1 row(s)
```

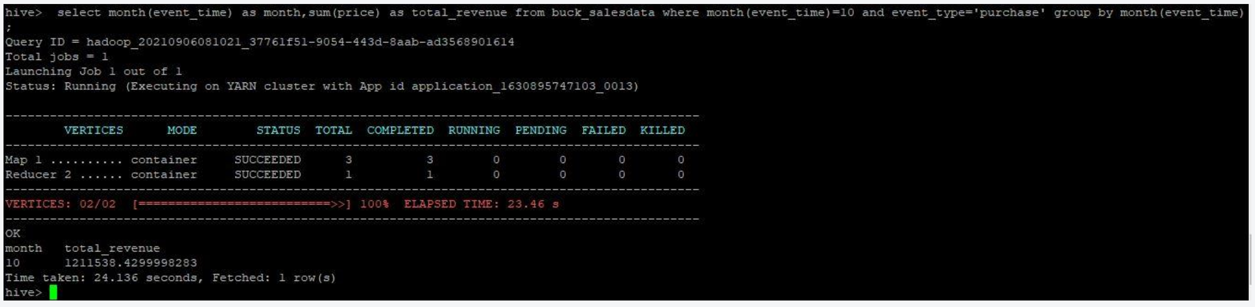
As we can see that using bucketing table the query time has improved a lot for a particular query .



```
hive>  select month(event_time) as month,sum(price) as total_revenue from buck_salesdata where month(event_time)=10 and event_type='purchase' group by month(event_time)
;
Query ID = hadoop_20210906081021_37761f51-9054-443d-8aab-ad3568901614
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1630895747103_0013)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      3          3        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 23.46 s
----------------------------------------------------------------------------------------------
OK
month   total_revenue
10      1211538.4299998283
Time taken: 24.136 seconds, Fetched: 1 row(s)
hive> 
```

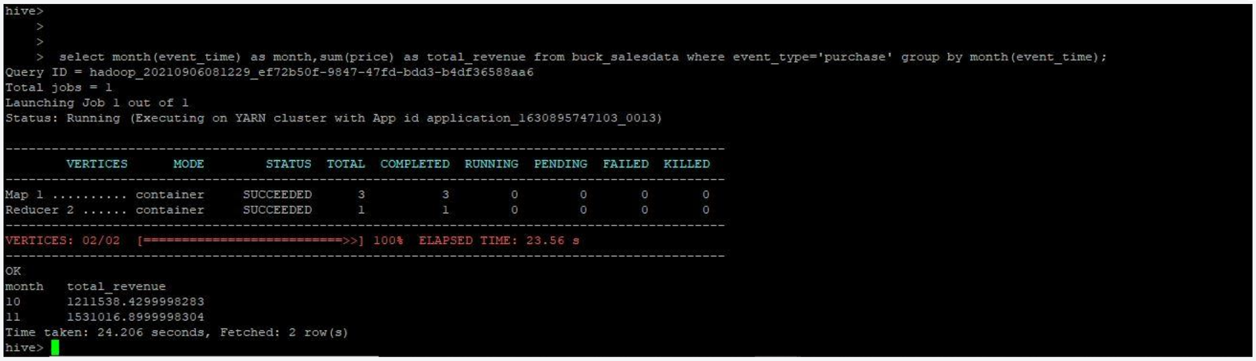
- **Run Hive queries to answer the questions given below.**

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



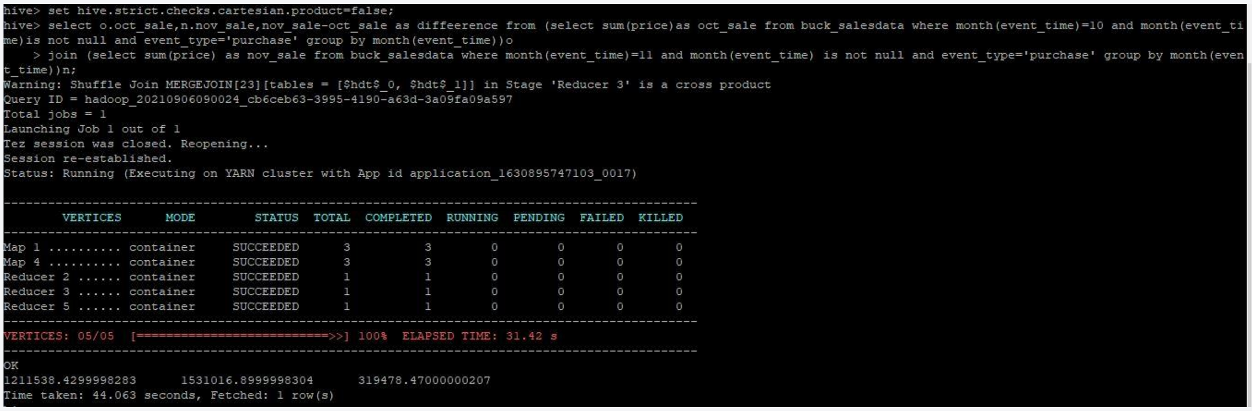
```
hive>  select month(event_time) as month,sum(price) as total_revenue from buck_salesdata where month(event_time)=10 and event_type='purchase' group by month(event_time)
;
Query ID = hadoop_20210906081021_37761f51-9054-443d-8aab-ad3568901614
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1630895747103_0013)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      3          3        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 23.46 s
----------------------------------------------------------------------------------------------
OK
month   total_revenue
10      1211538.4299998283
Time taken: 24.136 seconds, Fetched: 1 row(s)
hive> 
```

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



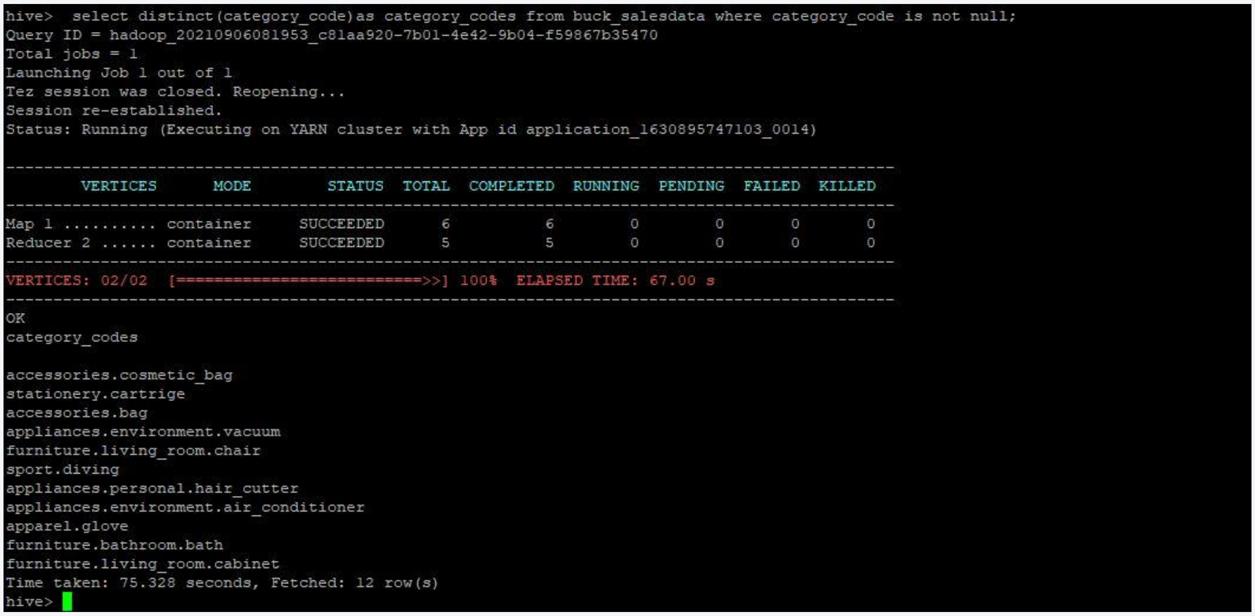
```
hive>
    >
    >
    >  select month(event_time) as month,sum(price) as total_revenue from buck_salesdata where event_type='purchase' group by month(event_time);
Query ID = hadoop_20210906081229_ef72b50f-9847-47fd-bdd3-b4df36588aa6
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1630895747103_0013)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      3          3        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 23.56 s
----------------------------------------------------------------------------------------------
OK
month   total_revenue
10      1211538.4299998283
11      1531016.8999998304
Time taken: 24.206 seconds, Fetched: 2 row(s)
hive>
```

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



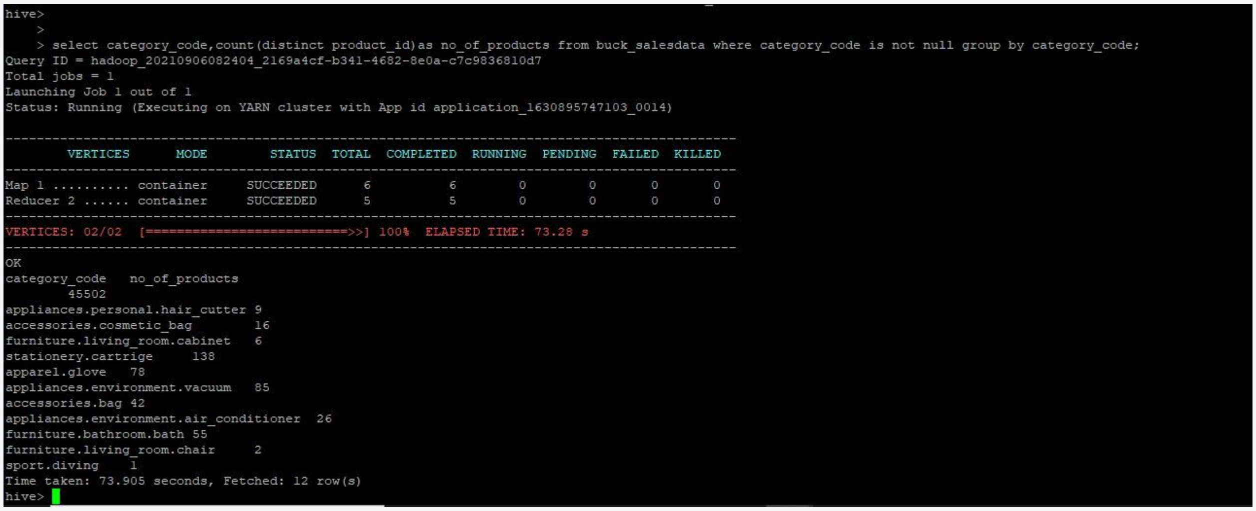
```
hive> set hive.strict.checks.cartesian.product=false;
hive> select o.oct_sale,n.nov_sale,nov_sale-oct_sale as diffeerence from (select sum(price)as oct_sale from buck_salesdata where month(event_time)=10 and month(event_ti
me)is not null and event_type='purchase' group by month(event_time))o
    > join (select sum(price) as nov_sale from buck_salesdata where month(event_time)=11 and month(event_time) is not null and event_type='purchase' group by month(even
t_time))n;
Warning: Shuffle Join MERGEJOIN[23][tables = [$hdt$_0, $hdt$_1]] in Stage 'Reducer 3' is a cross product
Query ID = hadoop_20210906090024_cb6ceb63-3995-4190-a63d-3a09fa09a597
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_1630895747103_0017)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      3          3        0        0       0       0
Map 4 .......... container     SUCCEEDED      3          3        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
Reducer 5 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 05/05  [==========================>>] 100%  ELAPSED TIME: 31.42 s
----------------------------------------------------------------------------------------------
OK
1211538.4299998283      1531016.8999998304      319478.47000000207
Time taken: 44.063 seconds, Fetched: 1 row(s)
```

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



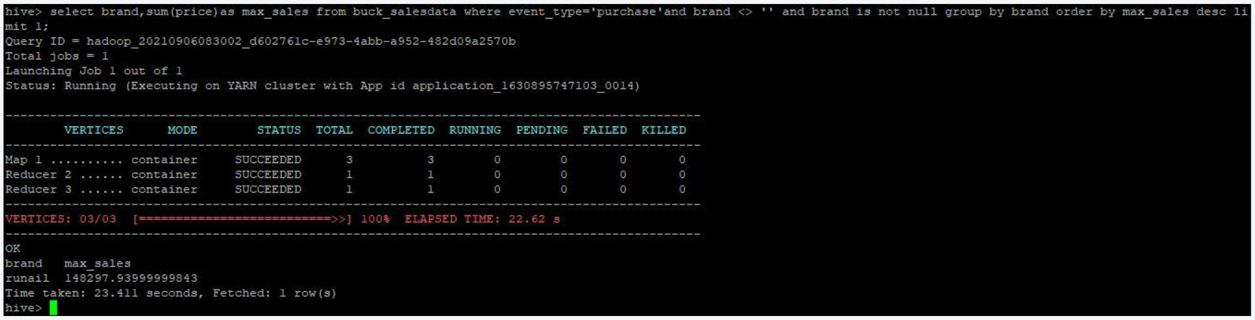
```
hive>  select distinct(category_code)as category_codes from buck_salesdata where category_code is not null;
Query ID = hadoop_20210906081953_c81aa920-7b01-4e42-9b04-f59867b35470
Total jobs = 1
Launching Job 1 out of 1
Tez session was closed. Reopening...
Session re-established.
Status: Running (Executing on YARN cluster with App id application_1630895747103_0014)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      6          6        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      5          5        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 67.00 s
----------------------------------------------------------------------------------------------
OK
category_codes

accessories.cosmetic_bag
stationery.cartrige
accessories.bag
appliances.environment.vacuum
furniture.living_room.chair
sport.diving
appliances.personal.hair_cutter
appliances.environment.air_conditioner
apparel.glove
furniture.bathroom.bath
furniture.living_room.cabinet
Time taken: 75.328 seconds, Fetched: 12 row(s)
hive>
```

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



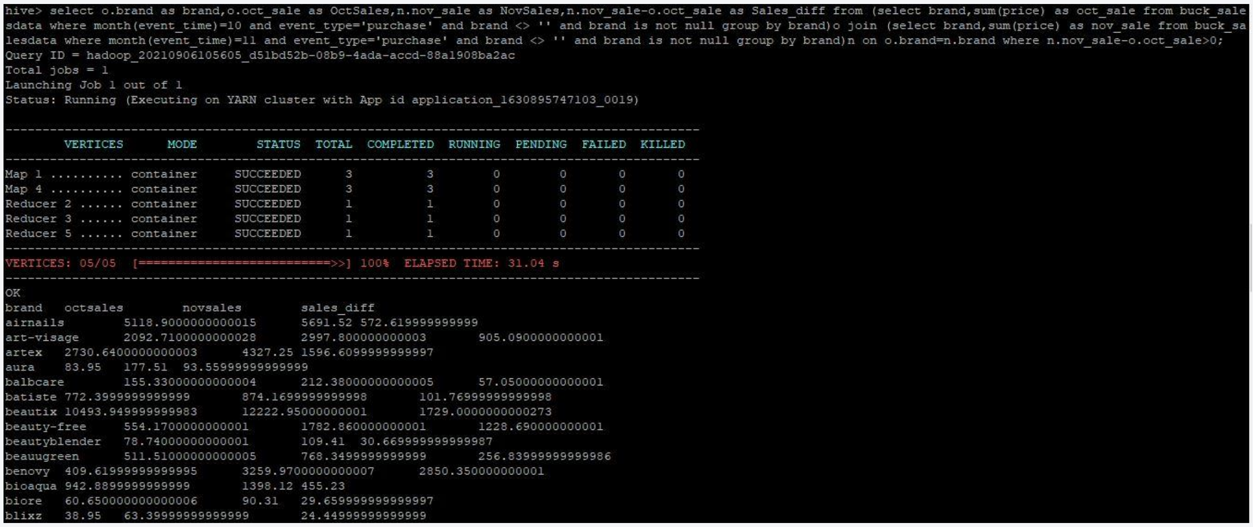
```
hive>
    >
    > select category_code,count(distinct product_id)as no_of_products from buck_salesdata where category_code is not null group by category_code;
Query ID = hadoop_20210906082404_2169a4cf-b341-4682-8e0a-c7c9836810d7
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1630895747103_0014)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      6          6        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      5          5        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 73.28 s
----------------------------------------------------------------------------------------------
OK
category_code   no_of_products
        45502
appliances.personal.hair_cutter 9
accessories.cosmetic_bag        16
furniture.living_room.cabinet   6
stationery.cartrige     138
apparel.glove   78
appliances.environment.vacuum   85
accessories.bag 42
appliances.environment.air_conditioner  26
furniture.bathroom.bath 55
furniture.living_room.chair     2
sport.diving    1
Time taken: 73.905 seconds, Fetched: 12 row(s)
hive>
```

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

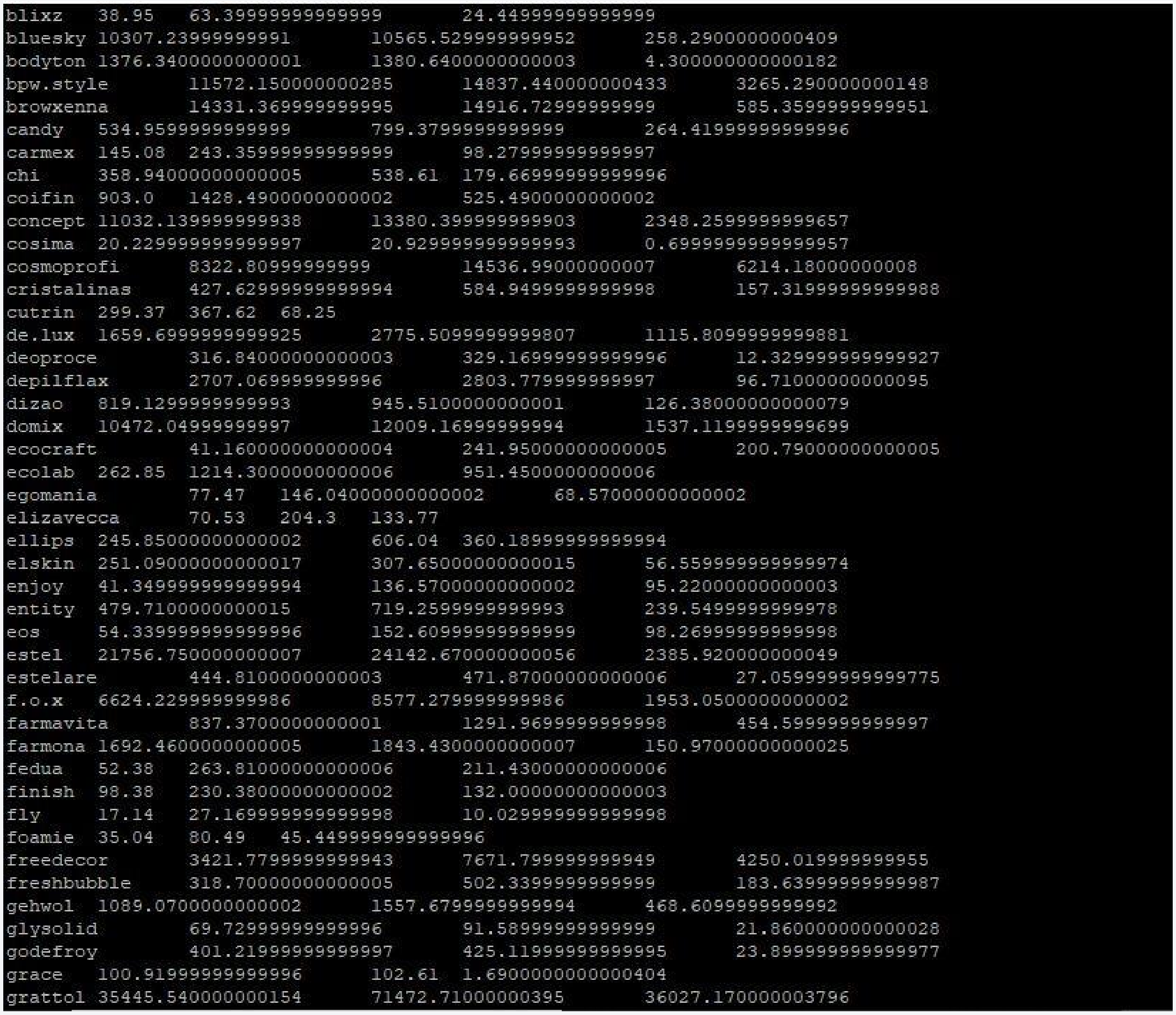


```
hive> select brand,sum(price)as max_sales from buck_salesdata where event_type='purchase'and brand <> '' and brand is not null group by brand order by max_sales desc li
mit 1;
Query ID = hadoop_20210906083002_d602761c-e973-4abb-a952-482d09a2570b
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1630895747103_0014)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      3          3        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 22.62 s
----------------------------------------------------------------------------------------------
OK
brand   max_sales
runail  148297.93999999843
Time taken: 23.411 seconds, Fetched: 1 row(s)
hive>
```

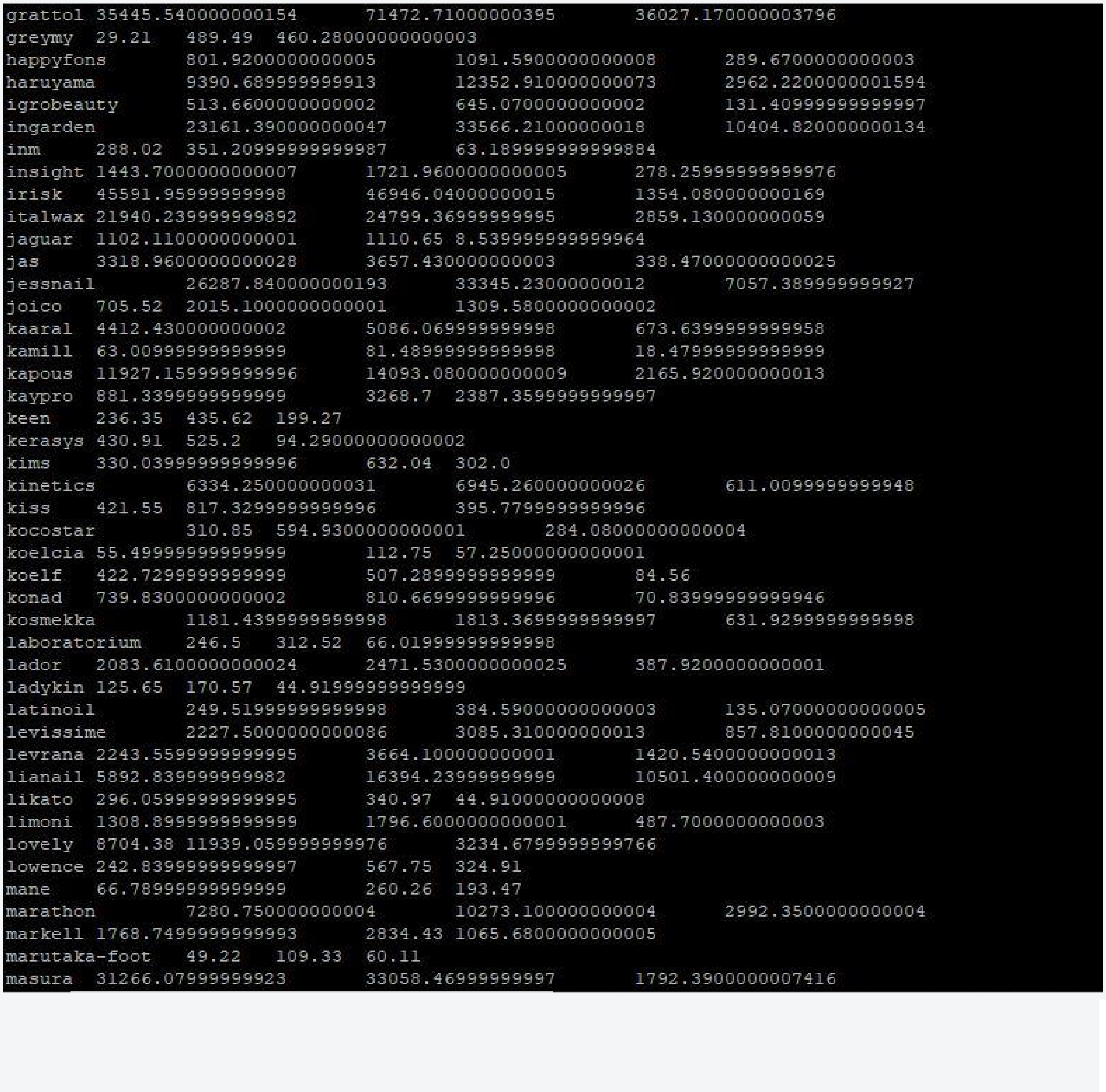
7. Which brands increased their sales from October to November?



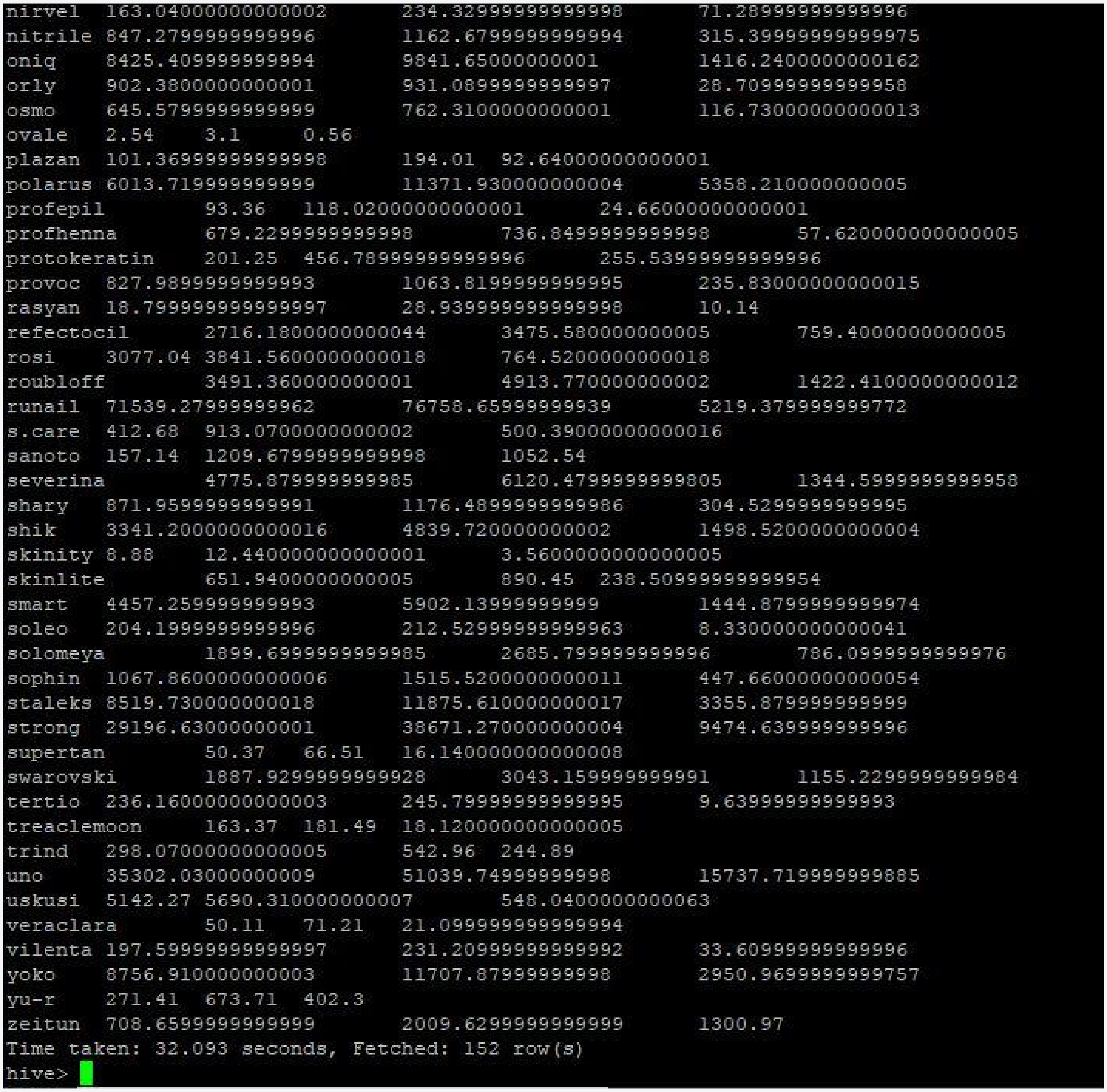
```
hive> select o.brand as brand,o.oct_sale as OctSales,n.nov_sale as NovSales,n.nov_sale-o.oct_sale as Sales_diff from (select brand,sum(price) as oct_sale from buck_sale
sdata where month(event_time)=10 and event_type='purchase' and brand <> '' and brand is not null group by brand)o join (select brand,sum(price) as nov_sale from buck_sa
lesdata where month(event_time)=11 and event_type='purchase' and brand <> '' and brand is not null group by brand)n on o.brand=n.brand where n.nov_sale-o.oct_sale>0;
Query ID = hadoop_20210906105605_d51bd52b-08b9-4ada-accd-88a1908ba2ac
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1630895747103_0019)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      3          3        0        0       0       0
Map 4 .......... container     SUCCEEDED      3          3        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
Reducer 5 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 05/05  [==========================>>] 100%  ELAPSED TIME: 31.04 s
----------------------------------------------------------------------------------------------
OK
brand   octsales        novsales        sales_diff
airnails        5118.9000000000015      5691.52 572.619999999999
art-visage      2092.7100000000028      2997.800000000003       905.0900000000001
artex   2730.6400000000003      4327.25 1596.6099999999997
aura    83.95   177.51  93.55999999999999
balbcare        155.33000000000004      212.38000000000005      57.05000000000001
batiste 772.3999999999999       874.1699999999998       101.76999999999998
beautix 10493.949999999983      12222.95000000001       1729.0000000000273
beauty-free     554.1700000000001       1782.860000000001       1228.690000000001
beautyblender   78.74000000000001       109.41  30.669999999999987
beauugreen      511.51000000000005      768.3499999999999       256.83999999999986
benovy  409.61999999999995      3259.9700000000007      2850.350000000001
bioaqua 942.8899999999999       1398.12 455.23
biore   60.650000000000006      90.31   29.659999999999997
blixz   38.95   63.39999999999999       24.44999999999999
```



```
blixz   38.95   63.39999999999999       24.44999999999999
bluesky 10307.23999999991       10565.529999999952      258.2900000000409
bodyton 1376.3400000000001      1380.6400000000003      4.300000000000182
bpw.style       11572.150000000285      14837.440000000433      3265.290000000148
browxenna       14331.369999999995      14916.72999999999       585.3599999999951
candy   534.9599999999999       799.3799999999999       264.41999999999996
carmex  145.08  243.35999999999999      98.27999999999997
chi     358.94000000000005      538.61  179.66999999999996
coifin  903.0   1428.4900000000002      525.4900000000002
concept 11032.139999999938      13380.399999999903      2348.259999999657
cosima  20.229999999999997      20.929999999999993      0.6999999999999957
cosmoprofi      8322.80999999999        14536.99000000007       6214.18000000008
cristalinas     427.62999999999994      584.9499999999998       157.31999999999988
cutrin  299.37  367.62  68.25
de.lux  1659.6999999999925      2775.5099999999807      1115.8099999999881
deoproce        316.84000000000003      329.16999999999996      12.329999999999927
depilflax       2707.069999999996       2803.779999999997       96.71000000000095
dizao   819.1299999999993       945.5100000000001       126.38000000000079
domix   10472.04999999997       12009.16999999994       1537.1199999999699
ecocraft        41.160000000000004      241.95000000000005      200.79000000000005
ecolab  262.85  1214.3000000000006      951.4500000000006
egomania        77.47   146.04000000000002      68.57000000000002
elizavecca      70.53   204.3   133.77
ellips  245.85000000000002      606.04  360.18999999999994
elskin  251.09000000000017      307.65000000000015      56.559999999999974
enjoy   41.349999999999994      136.57000000000002      95.22000000000003
entity  479.7100000000015       719.2599999999993       239.5499999999978
eos     54.339999999999996      152.60999999999999      98.26999999999998
estel   21756.750000000007      24142.670000000056      2385.920000000049
estelare        444.8100000000003       471.87000000000006      27.059999999999775
f.o.x   6624.229999999986       8577.279999999986       1953.0500000000002
farmavita       837.3700000000001       1291.9699999999998      454.5999999999997
farmona 1692.4600000000005      1843.4300000000007      150.97000000000025
fedua   52.38   263.81000000000006      211.43000000000006
finish  98.38   230.38000000000002      132.00000000000003
fly     17.14   27.169999999999998      10.029999999999998
foamie  35.04   80.49   45.449999999999996
freedecor       3421.7799999999943      7671.799999999949       4250.019999999955
freshbubble     318.70000000000005      502.3399999999999       183.63999999999987
gehwol  1089.0700000000002      1557.6799999999994      468.6099999999992
glysolid        69.72999999999996       91.58999999999999       21.860000000000028
godefroy        401.21999999999997      425.11999999999995      23.899999999999977
grace   100.91999999999996      102.61  1.6900000000000404
grattol 35445.540000000154      71472.71000000395       36027.170000003796
```

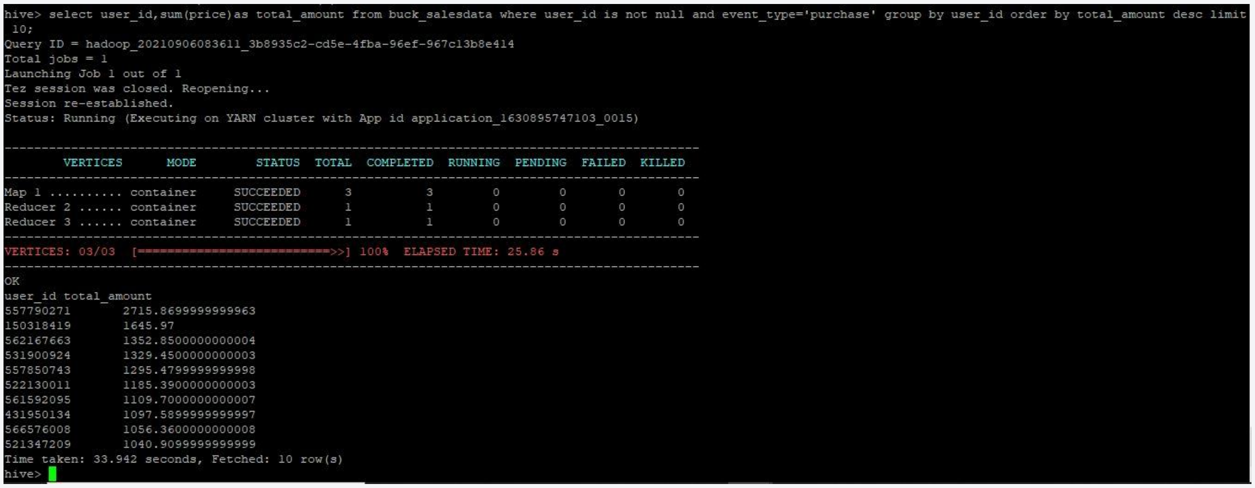


```
grattol 35445.540000000154      71472.71000000395       36027.170000003796
greymy  29.21   489.49  460.28000000000003
happyfons       801.9200000000005       1091.5900000000008      289.6700000000003
haruyama        9390.689999999913       12352.910000000073      2962.2200000001594
igrobeauty      513.6600000000002       645.0700000000002       131.40999999999997
ingarden        23161.390000000047      33566.21000000018       10404.820000000134
inm     288.02  351.20999999999987      63.189999999999884
insight 1443.7000000000007      1721.9600000000005      278.25999999999976
irisk   45591.95999999998       46946.04000000015       1354.080000000169
italwax 21940.239999999892      24799.36999999995       2859.130000000059
jaguar  1102.1100000000001      1110.65 8.539999999999964
jas     3318.9600000000028      3657.430000000003       338.47000000000025
jessnail        26287.840000000193      33345.23000000012       7057.389999999927
joico   705.52  2015.1000000000001      1309.5800000000002
kaaral  4412.430000000002       5086.069999999998       673.6399999999958
kamill  63.00999999999999       81.48999999999998       18.47999999999999
kapous  11927.159999999996      14093.080000000009      2165.920000000013
kaypro  881.3399999999999       3268.7  2387.3599999999997
keen    236.35  435.62  199.27
kerasys 430.91  525.2   94.29000000000002
kims    330.03999999999996      632.04  302.0
kinetics        6334.250000000031       6945.260000000026       611.0099999999948
kiss    421.55  817.3299999999996       395.7799999999996
kocostar        310.85  594.9300000000001       284.08000000000004
koelcia 55.49999999999999       112.75  57.25000000000001
koelf   422.7299999999999       507.2899999999999       84.56
konad   739.8300000000002       810.6699999999996       70.83999999999946
kosmekka        1181.4399999999998      1813.3699999999997      631.9299999999998
laboratorium    246.5   312.52  66.01999999999998
lador   2083.6100000000024      2471.5300000000025      387.9200000000001
ladykin 125.65  170.57  44.91999999999999
latinoil        249.51999999999998      384.59000000000003      135.07000000000005
levissime       2227.5000000000086      3085.310000000013       857.8100000000045
levrana 2243.5599999999995      3664.100000000001       1420.540000000013
lianail 5892.839999999982       16394.23999999999       10501.400000000009
likato  296.05999999999995      340.97  44.91000000000008
limoni  1308.8999999999999      1796.6000000000001      487.7000000000003
lovely  8704.38 11939.059999999976      3234.6799999999766
lowence 242.83999999999997      567.75  324.91
mane    66.78999999999999       260.26  193.47
marathon        7280.750000000004       10273.100000000004      2992.3500000000004
markell 1768.7499999999993      2834.43 1065.6800000000005
marutaka-foot   49.22   109.33  60.11
masura  31266.07999999923       33058.46999999997       1792.3900000007416
```



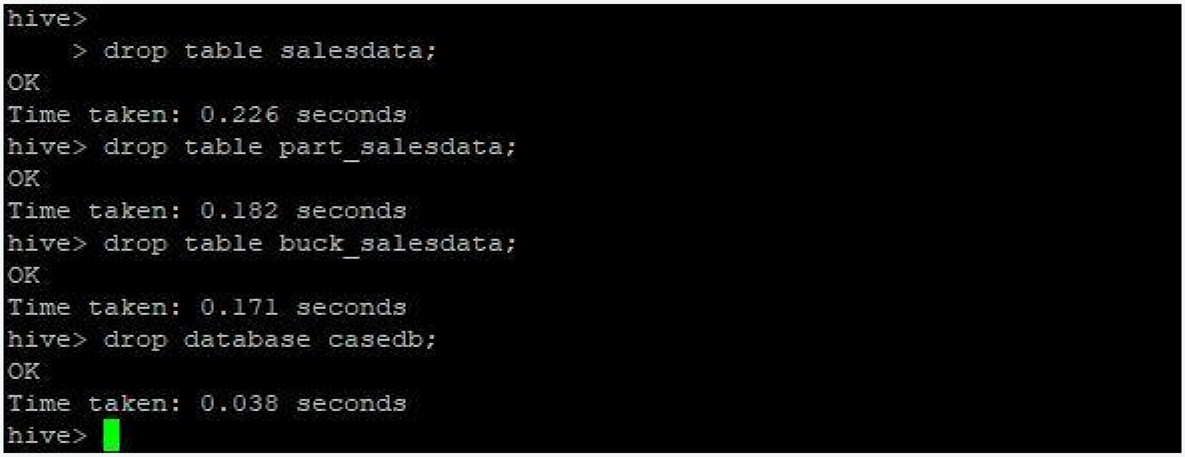
```
nirvel  163.04000000000002      234.32999999999998      71.28999999999996
nitrile 847.2799999999996       1162.6799999999994      315.39999999999975
oniq    8425.409999999994       9841.65000000001        1416.2400000000162
orly    902.3800000000001       931.0899999999997       28.709999999999958
osmo    645.5799999999999       762.3100000000001       116.73000000000013
ovale   2.54    3.1     0.56
plazan  101.36999999999998      194.01  92.64000000000001
polarus 6013.719999999999       11371.930000000004      5358.210000000005
profepil        93.36   118.02000000000001      24.66000000000001
profhenna       679.2299999999998       736.8499999999998       57.620000000000005
protokeratin    201.25  456.78999999999996      255.53999999999996
provoc  827.9899999999993       1063.8199999999995      235.83000000000015
rasyan  18.799999999999997      28.939999999999998      10.14
refectocil      2716.1800000000044      3475.580000000005       759.4000000000005
rosi    3077.04 3841.5600000000018      764.5200000000018
roubloff        3491.360000000001       4913.770000000002       1422.4100000000012
runail  71539.27999999962       76758.65999999939       5219.379999999772
s.care  412.68  913.0700000000002       500.39000000000016
sanoto  157.14  1209.6799999999998      1052.54
severina        4775.879999999985       6120.4799999999805      1344.5999999999958
shary   871.9599999999991       1176.4899999999986      304.5299999999995
shik    3341.2000000000016      4839.720000000002       1498.5200000000004
skinity 8.88    12.440000000000001      3.5600000000000005
skinlite        651.9400000000005       890.45  238.50999999999954
smart   4457.259999999993       5902.13999999999        1444.8799999999974
soleo   204.1999999999996       212.52999999999963      8.330000000000041
solomeya        1899.6999999999985      2685.799999999996       786.0999999999976
sophin  1067.8600000000006      1515.5200000000011      447.66000000000054
staleks 8519.730000000018       11875.610000000017      3355.879999999999
strong  29196.63000000001       38671.270000000004      9474.639999999996
supertan        50.37   66.51   16.140000000000008
swarovski       1887.9299999999928      3043.159999999991       1155.2299999999984
tertio  236.16000000000003      245.79999999999995      9.639999999999993
treaclemoon     163.37  181.49  18.120000000000005
trind   298.07000000000005      542.96  244.89
uno     35302.03000000009       51039.74999999998       15737.719999999885
uskusi  5142.27 5690.310000000007       548.0400000000063
veraclara       50.11   71.21   21.099999999999994
vilenta 197.59999999999997      231.20999999999992      33.60999999999996
yoko    8756.910000000003       11707.87999999998       2950.9699999999757
yu-r    271.41  673.71  402.3
zeitun  708.6599999999999       2009.6299999999999      1300.97
Time taken: 32.093 seconds, Fetched: 152 row(s)
hive>
```

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.



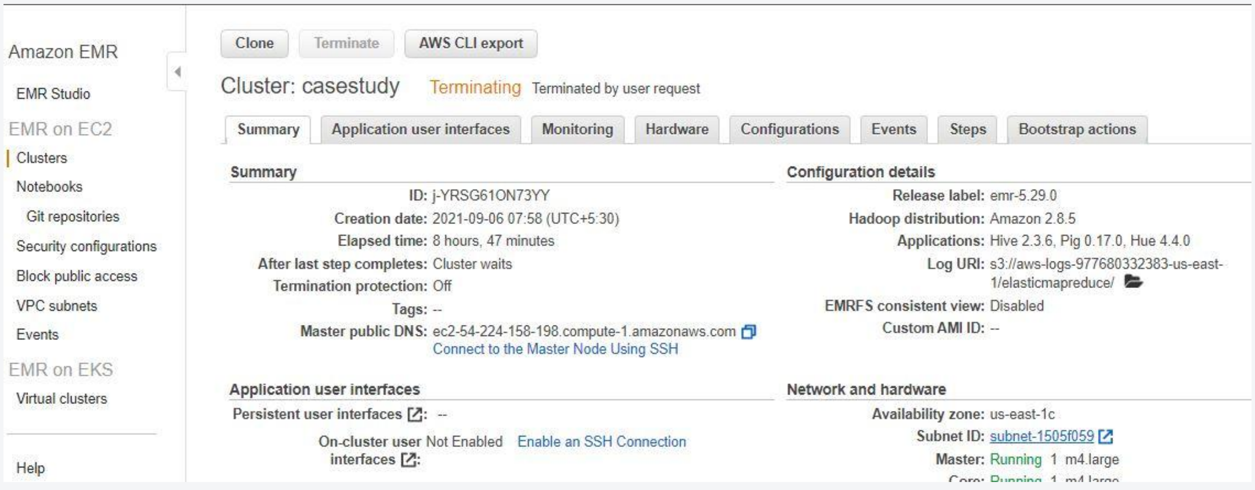
```
hive> select user_id,sum(price)as total_amount from buck_salesdata where user_id is not null and event_type='purchase' group by user_id order by total_amount desc limit
 10;
Query ID = hadoop_20210906083611_3b8935c2-cd5e-4fba-96ef-967c13b8e414
Total jobs = 1
Launching Job 1 out of 1
Tez session was closed. Reopening...
Session re-established.
Status: Running (Executing on YARN cluster with App id application_1630895747103_0015)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      3          3        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 25.86 s
----------------------------------------------------------------------------------------------
OK
user_id total_amount
557790271       2715.8699999999963
150318419       1645.97
562167663       1352.8500000000004
531900924       1329.4500000000003
557850743       1295.4799999999998
522130011       1185.3900000000003
561592095       1109.7000000000007
431950134       1097.5899999999997
566576008       1056.3600000000008
521347209       1040.9099999999999
Time taken: 33.942 seconds, Fetched: 10 row(s)
hive>
```

- **Cleaning up**
  - Drop your database



```
hive>
    > drop table salesdata;
OK
Time taken: 0.226 seconds
hive> drop table part_salesdata;
OK
Time taken: 0.182 seconds
hive> drop table buck_salesdata;
OK
Time taken: 0.171 seconds
hive> drop database casedb;
OK
Time taken: 0.038 seconds
hive>
```

  - Terminate your cluster



Cluster: casestudy Terminating Terminated by user request

Summary

ID: j-YRSG61ON73YY
Creation date: 2021-09-06 07:58 (UTC+5:30)
Elapsed time: 8 hours, 47 minutes
After last step completes: Cluster waits
Termination protection: Off
Tags: --
Master public DNS: ec2-54-224-158-198.compute-1.amazonaws.com
Connect to the Master Node Using SSH

Configuration details

Release label: emr-5.29.0
Hadoop distribution: Amazon 2.8.5
Applications: Hive 2.3.6, Pig 0.17.0, Hue 4.4.0
Log URI: s3://aws-logs-977680332383-us-east-1/elasticmapreduce/
EMRFS consistent view: Disabled
Custom AMI ID: --

Application user interfaces

Persistent user interfaces: --
On-cluster user interfaces: Not Enabled Enable an SSH Connection

Network and hardware

Availability zone: us-east-1c
Subnet ID: subnet-1505f059
Master: Running 1 m4.large
Core: Running 1 m4.large

Case Study by:

Raj Patel & Sachin Prajapati .