

Hive Case Study on E-Commerce Data



by

Abhinav Kumar
&
Harshini V. Bhat

Hive Case Study on E-Commerce Data

Problem Statement

With online sales gaining popularity, tech companies are exploring ways to improve their sales by analyzing customer behavior and gaining insights about product trends. Furthermore, the websites make it easier for customers to find the products they require without much scavenging. Therefore, as part of this assignment, we will have to extract data and gather insights from a real-life data set of an e-commerce company.

The required preliminary setup on the AWS console to begin the analysis

1. Key Pair creation

This screenshot shows the AWS EC2 Resources page. On the left sidebar, under 'Instances', 'Key pairs' is highlighted with a red box. The main panel displays various EC2 resources: Instances (running) 0, Dedicated Hosts 0, Elastic IPs 0, Instances 0, Key pairs 5, Load balancers API Error, Placement groups 0, Security groups 7, Snapshots 0, and Volumes 0. A callout bubble provides information about easily sizing, configuring, and deploying Microsoft SQL Server Always On availability groups on AWS using the AWS Launch Wizard for SQL Server.

This screenshot shows the AWS EC2 Key pairs list page. The 'Create key pair' button is highlighted with a red box. The table lists five existing key pairs:

| Name | Type | Fingerprint | ID |
|------------------------------------|------|--|----------------|
| ShermeleAssignment | rsa | 73:83:5f:89:6e:13:4e:97:8b:7a:aef:f7:fc... | key-0194d013b0 |
| ShermeleAssignmentAgain | rsa | c2:d5:f9:5ab6:32:b6:64:6d:7ce:07:92... | key-01036f165d |
| Shermele_462248093251_2021-05-0... | rsa | b6:40:63:73:b2:1e:04:b2:91:71:1e:baf... | key-0c41501690 |
| Test | rsa | 08:04:ab:e0:67:37:f9:7ada:0bb:73:c... | key-04393c3183 |

This screenshot shows the 'Key pair' creation dialog box. The 'Name' field is filled with 'my_first_key_pair' and highlighted with a red box. The 'Key pair type' is set to 'RSA'. The 'Private key file format' is set to '.pem'. The 'Tags (Optional)' section indicates 'No tags associated with the resource.' and has an 'Add tag' button. The 'Create key pair' button is highlighted with a red box at the bottom right.

The screenshot shows the AWS EC2 Key Pairs page. At the top, a green banner says "Successfully created key pair". Below it, a table lists key pairs. One row for "my_first_key_pair" is highlighted with a red arrow pointing to it.

| Name | Type | Fingerprint | ID |
|------------------------------------|------|--|--------------------|
| ShermeleAssignment | rsa | 73:83:5f:89:6e:13:4e:97:8b:7a:aef:f... | key-0194d013b... |
| ShermeleAssignmentAgain | rsa | c2:d5:f9:5a:b6:32:b6:64:6d:7ceb:79:2... | key-01036f163d... |
| Shermele_462248093251_2021-05-0... | rsa | b6:40:63:73:b2:1e:04:b2:91:71:1ebaf... | key-0c4150169d... |
| Test | rsa | 08:04:ab:e0:67:37:f9:7a:da:0b:bb:73:c... | key-04395c3183... |
| hive_case_study_da_track | rsa | 00:fa:92:a9:a1:8e:01:6cfc:b4:eb:6a:45... | key-0fbcbcede76... |
| my_first_key_pair | rsa | 60:24:39:d8:49:baff:b7:7eb7:20:da:7... | key-027e9b72e7... |

2. Creating a role s3_access_role

The screenshot shows the AWS IAM Roles list page. A blue banner at the top says "Introducing the new Roles list experience". The main table lists existing roles, and a red box highlights the "Create role" button.

| Role name | Trusted entities | Last activity |
|---------------------------------|---|----------------|
| AWSServiceRoleForEMRCleanup | AWS Service: elasticmapreduce (Service-Linked Role) | Yesterday |
| AWSServiceRoleForOrganizations | AWS Service: organizations (Service-Linked Role) | 204 days ago |
| AWSServiceRoleForSSO | AWS Service: sso (Service-Linked Role) | - |
| AWSServiceRoleForSupport | AWS Service: support (Service-Linked Role) | - |
| AWSServiceRoleForTrustedAdvisor | AWS Service: trustedadvisor (Service-Linked Role) | - |
| EMR_AutoScaling_DefaultRole | AWS Service: elasticmapreduce, and 1 more. Edit | - |
| EMR_DefaultRole | AWS Service: elasticmapreduce | 20 minutes ago |

The screenshot shows the "Create role" wizard, step 1: "Select type of trusted entity". It has four tabs: "AWS service" (selected), "Another AWS account", "Web Identity", and "SAML 2.0 federation". A red box highlights the "AWS service" tab. Below it, a note says "Allows AWS services to perform actions on your behalf. [Learn more](#)".

▼ Attach permissions policies

Choose one or more policies to attach to your new role.

[Create policy](#)

[Filter policies](#) Showing 9 results

| Policy name | Used as |
|--|---------|
| <input type="checkbox"/> AmazonDMSRedshiftS3Role | None |
| <input checked="" type="checkbox"/> AmazonS3FullAccess | None |
| <input type="checkbox"/> AmazonS3ObjectLambdaExecutionRolePolicy | None |
| <input type="checkbox"/> AmazonS3OutpostsFullAccess | None |
| <input type="checkbox"/> AmazonS3OutpostsReadOnlyAccess | None |
| <input type="checkbox"/> AmazonS3ReadOnlyAccess | None |
| <input type="checkbox"/> IVSRecordToS3 | None |
| <input type="checkbox"/> QuickSightAccessForS3StorageManagementAnalyticsReadOnly | None |

► Set permissions boundary

* Required [Cancel](#) [Previous](#) [Next: Tags](#)

AWS Services Search for services, features, blogs, docs, and more [Option+S] Global upgradabhinavkumar

EC2 EMR

Create role

Review

Provide the required information below and review this role before you create it.

Role name* Use alphanumeric and '+,-, @,-' characters. Maximum 64 characters.

Role description Allows EC2 instances to call AWS services on your behalf.

Maximum 1000 characters. Use alphanumeric and '+,-, @,-' characters.

Trusted entities AWS service: ec2.amazonaws.com

Policies AmazonS3FullAccess

Permissions boundary Permissions boundary is not set

No tags were added.

* Required [Cancel](#) [Previous](#) [Create role](#)

Navigate back to EC2 service to apply the role created

AWS Services Search for services, features, blogs, docs, and more [Option+S] N. Virginia upgradabhinavkumar @ 4622-4809-3251

EC2 EMR

New EC2 Experience Tell us what you think

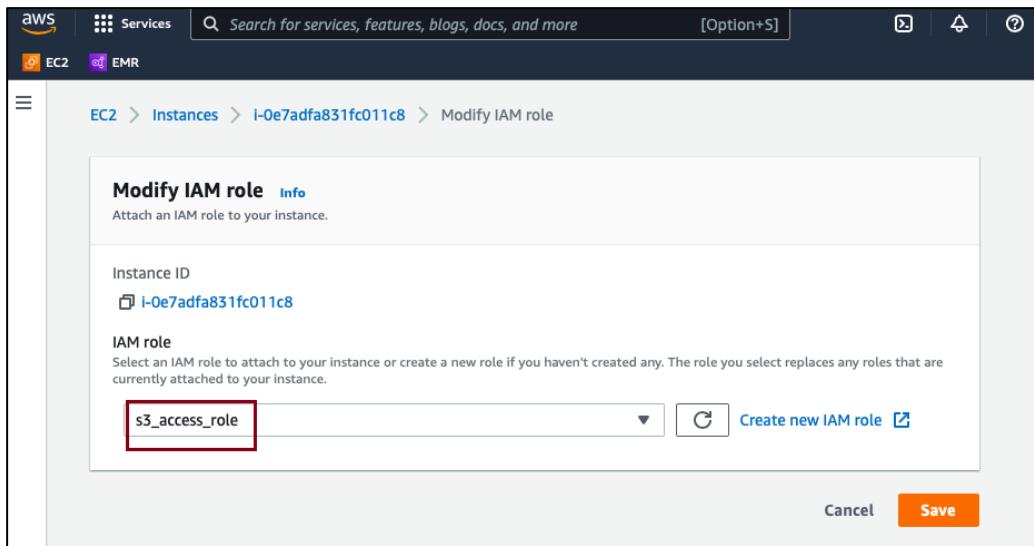
EC2 Dashboard

Instances (1/4) Info Actions Launch instances

Instances (1/4) Filter instances

| Public IPv4 ... | Elastic IP | IPv6 IPs | Monitoring | Security group name | Key name | Launch time |
|-----------------|------------|----------|------------|-------------------------|------------------|-------------|
| - | - | - | disabled | - | hive_case_stu... | 2021/11/25 |
| - | - | - | disabled | - | hive_case_stu... | 2021/11/25 |
| co... | - | - | disabled | ElasticMapReduce-slave | my_first_key_... | 2021/11/25 |
| co... | - | - | disabled | ElasticMapReduce-master | my_first_key_... | 2021/11/25 |

Select the master-reducer > go to actions > security > modify IAM role



Now, all the required preliminary setup is completed. Hence, we will move for the cluster creation.

3. Creating an EMR Cluster

| Name | ID | Status | Creation time (UTC+5:30) | Elapsed time | Normal instances |
|--------------------------|-----------------|--|-----------------------------|---------------------|------------------|
| UpGrad_Hive_Querying_mod | j-2LXNJQNQJ2JL1 | Terminated | 2021-11-24 14:59 (UTC+5:30) | 5 hours, 38 minutes | 48 |
| UpGrad_Hive_Querying_mod | j-16FEU63E1PMNL | Terminated | 2021-11-23 19:53 (UTC+5:30) | 7 hours, 28 minutes | 64 |
| DSC_30_Lab_Slot_2 | j-23N6X2GGYBYSB | Terminated | 2021-11-20 15:57 (UTC+5:30) | 7 hours, 34 minutes | 96 |
| Demo_Cluster | j-1DHKRJMJIFBVZ | Terminated | 2021-11-19 19:23 (UTC+5:30) | 1 hour, 9 minutes | 8 |
| My_First_Demo_Cluster | j-FKECJ18IEY3C | Terminated with errors Instance failure | 2021-11-10 20:32 (UTC+5:30) | 2 hours, 25 minutes | 36 |

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-5.29.0**

| | | |
|--|---|--|
| <input checked="" type="checkbox"/> Hadoop 2.8.5 | <input type="checkbox"/> Zeppelin 0.8.2 | <input type="checkbox"/> Livy 0.6.0 |
| <input type="checkbox"/> JupyterHub 1.0.0 | <input checked="" type="checkbox"/> Tez 0.9.2 | <input type="checkbox"/> Flink 1.9.1 |
| <input type="checkbox"/> Ganglia 3.7.2 | <input type="checkbox"/> HBase 1.4.10 | <input type="checkbox"/> Pig 0.17.0 |
| <input checked="" type="checkbox"/> Hive 2.3.6 | <input type="checkbox"/> Presto 0.227 | <input type="checkbox"/> ZooKeeper 3.4.14 |
| <input type="checkbox"/> MXNet 1.5.1 | <input type="checkbox"/> Sqoop 1.4.7 | <input type="checkbox"/> Mahout 0.13.0 |
| <input checked="" type="checkbox"/> Hue 4.4.0 | <input type="checkbox"/> Phoenix 4.14.3 | <input type="checkbox"/> Oozie 5.1.0 |
| <input checked="" type="checkbox"/> Spark 2.4.4 | <input type="checkbox"/> HCatalog 2.3.6 | <input type="checkbox"/> TensorFlow 1.14.0 |

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 Use for Spark table metadata

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](#)

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 <input type="checkbox"/> | m4.large <input checked="" type="checkbox"/> 2 vCore, 8 GiB memory, EBS only storage EBS Storage: 32 GiB <input type="checkbox"/> Add configuration settings <input type="checkbox"/> | 1 Instances <input checked="" type="checkbox"/> | <input checked="" type="radio"/> On-demand <input type="checkbox"/> Spot <input type="checkbox"/> Use on-demand as max price |
| Core Core - 2 <input type="checkbox"/> | m4.large <input checked="" type="checkbox"/> 2 vCore, 8 GiB memory, EBS only storage EBS Storage: 32 GiB <input type="checkbox"/> Add configuration settings <input type="checkbox"/> | 1 Instances <input checked="" type="checkbox"/> | <input checked="" type="radio"/> On-demand <input type="checkbox"/> Spot <input type="checkbox"/> Use on-demand as max price |
| Task Task - 3 <input type="checkbox"/> | m4.large <input checked="" type="checkbox"/> 2 vCore, 8 GiB memory, EBS only storage EBS Storage: 32 GiB <input type="checkbox"/> Add configuration settings <input type="checkbox"/> | 0 Instances <input type="checkbox"/> | <input checked="" type="radio"/> On-demand <input type="checkbox"/> Spot <input type="checkbox"/> Use on-demand as max price |

[+ Add task instance group](#)

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 **Hive_Case_Study_DA_Track**

Logging S3 folder `s3://aws-logs-462248093251-us-east-1/elasticma/`

Debugging Termination protection

Tags

| | |
|--|-------------------------|
| Key | Value (optional) |
| <input type="text"/> Add a key to create a tag | <input type="text"/> |

Additional Options

EMRFS consistent view Custom AMI ID `None`

Create Cluster - Advanced Options [Go to quick options](#)

Step 1: Software and Steps

Step 2: Hardware

Step 3: General Cluster Settings

Step 4: Security

Security Options

EC2 key pair: my_first_key_pair [Edit](#)

Cluster visible to all IAM users in account [Edit](#)

Permissions

Default Custom [Edit](#)

Select custom roles to tailor permissions for your cluster.

EMR role: EMR_DefaultRole [Edit](#)

EC2 instance profile: s3_access_role [Edit](#)

Auto Scaling role: EMR_AutoScaling_DefaultRole [Edit](#)

Security Configuration

EC2 security groups

An EC2 security group acts as a virtual firewall for your cluster nodes to control inbound and outbound traffic. There are two types of security groups you can configure, [EMR managed security groups](#) and [additional security groups](#). EMR will [automatically update](#) the rules in the EMR managed security groups in order to launch a cluster. [Learn more](#).

| Type | EMR managed security groups | Additional security groups |
|-------------|--|---|
| Master | Default: sg-0305c7d2fe95e00f0 (ElasticMapReduce-mast) Edit No security groups selected Edit | EMR will automatically update the selected group Edit |
| Core & Task | Default: sg-0fd14326d7497a526 (ElasticMapReduce-slave) Edit No security groups selected Edit | EMR will not modify the selected groups Edit |

[Create a security group](#)

While creating the EMR cluster, we have configured permissions for IAM roles to custom where we chose s3_access_role for EC2 instance profile.

.compute-1.amazonaws.com, and Configuration details like Release label: emr-5.29.0, Hadoop distribution: Amazon 2.8.5, Applications: Hive 2.3.6, Tez 0.9.2, Hue 4.4.0, Spark 2.4.4, Log URI: s3://aws-logs-462248093251-us-east-1/elasticsearch/, EMRFS consistent view: Disabled, and Custom AMI ID: --. The Application user interfaces section lists Persistent user interfaces: Spark history server and On-cluster user: HDFS Name Node, Hue, Spark History Server, Tez interfaces: UI, Resource Manager. Network and hardware details include Availability zone: us-east-1, Subnet ID: subnet-[, and Master: Running 1 m4.large."/>](#)

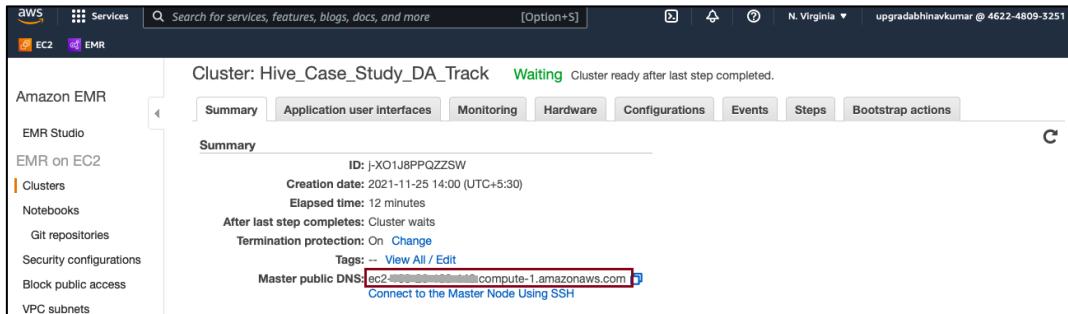
To connect the SSH to the master node, we need to change the security group settings of the master node.

| Name | Security group ID | Security group name | VPC ID | Description |
|----------------------|-----------------------|--------------------------|-----------------------|----------------------------|
| sg-0305c7d2fe95e00f0 | vpc-0efaaa2929885a3eb | ElasticMapReduce-mast... | vpc-0efaaa2929885a3eb | Master group for Elasti... |
| sg-0fd14326d7497a526 | vpc-0efaaa2929885a3eb | ElasticMapReduce-slave | vpc-0efaaa2929885a3eb | Slave group for Elastic... |

Following steps are required to change the security group setting of the master node:

- Navigate to the link of the master **Security Group id** as shown in the above illustration.
- Click on **Edit inbound rules** to add a new rule.
- After clicking the **Add Rule** button, select **All TCP** and here we have chosen IP as **My IP**.
- Click on **Save Rules** button to save rule.

4. Connecting to the terminal



Steps involved in connecting SSH to the master node in command-line interface (CLI) i.e. Terminal in our case.

- Open the terminal and change directory to the location where key-pair.pem file is located.
 - Copy **Master public DNS** from cluster summary.
 - We have executed a command to establish SSH connection to the master node according to the following syntax:
`ssh -i mykeypair.pem hadoop@ec2-###-##-##-##.compute-1.amazonaws.com`

Successfully connected SSH to the master node and then accessing the data from the AWS S3.

5. Loading the data from S3 into HDFS

- Checking the Hadoop directory

```
[hadoop@ip-10-0-0-0 ~]$ hadoop fs -ls /
Found 4 items
drwxr-xr-x  - hdfs hadoop          0 2021-11-25 09:13 /apps
drwxrwxrwt  - hdfs hadoop          0 2021-11-25 09:15 /tmp
drwxr-xr-x  - hdfs hadoop          0 2021-11-25 09:13 /user
drwxr-xr-x  - hdfs hadoop          0 2021-11-25 09:13 /var
[hadoop@ip-10-0-0-0 ~]$
[hadoop@ip-10-0-0-0 ~]$ aws s3 ls e-commerce-events-ml
2020-03-17 11:47:09 545839412 2019-Nov.csv
2020-03-17 11:37:31 482542278 2019-Oct.csv
```

- Making directory 'hive-case-study' in Hadoop

```
[hadoop@ip-10-0-0-0 ~]$ hadoop fs -mkdir /hive-case-study
[hadoop@ip-10-0-0-0 ~]$ hadoop fs -ls /
Found 5 items
drwxr-xr-x  - hdfs hadoop          0 2021-11-25 09:13 /apps
drwxr-xr-x  - hdfs hadoop          0 2021-11-25 09:33 /hive-case-study
drwxrwxrwt  - hdfs hadoop          0 2021-11-25 09:15 /tmp
drwxr-xr-x  - hdfs hadoop          0 2021-11-25 09:13 /user
drwxr-xr-x  - hdfs hadoop          0 2021-11-25 09:13 /var
```

- Loading the E-Commerce data from S3 to HDFS
 - Loading 2019-Oct.csv into 'hive-case-study'

```
[hadoop@ip-10-0-0-0 ~]$ hadoop distcp s3n://e-commerce-events-ml/2019-Oct.csv /hive-case-study/2019-oct.csv
21/11/25 09:36:31 INFO tools.DistCp: Input Options: DistCpOptions{atomicCommit=false, deleteMissing=false, ignoreFailures=false, overwrite=false, skipCRC=false, blocking=true, numListThreads=0, maxMaps=20, mapBandwidth=100, sslConfigurationFile='null', copyStrategy='uniformsize', preserveStatus=[], preserveRawXattrs=false, atomicWorkPath=null, logPath=null, sourceFileListing=null, sourcePaths=[s3n://e-commerce-events-ml/2019-Oct.csv], targetPath=/hive-case-study/2019-oct.csv, targetPathExists=false, filtersFile='null'}
21/11/25 09:36:31 INFO client.RMProxy: Connecting to ResourceManager at ip-10-0-0-0.ec2.internal/10.0.0.1:8032
21/11/25 09:36:35 INFO tools.SimpleCopyListing: Paths (files+dirs) cnt = 1; dirCnt = 0
21/11/25 09:36:35 INFO tools.SimpleCopyListing: Build file listing completed.
21/11/25 09:36:35 INFO Configuration deprecation: io.sort.mb is deprecated. Instead, use mapreduce.task.io.sort.mb
21/11/25 09:36:35 INFO Configuration deprecation: io.sort.factor is deprecated. Instead, use mapreduce.task.io.sort.factor
21/11/25 09:36:35 INFO tools.DistCp: Number of paths in the copy list: 1
21/11/25 09:36:36 INFO tools.DistCp: Number of paths in the copy list: 1
21/11/25 09:36:36 INFO client.RMProxy: Connecting to ResourceManager at ip-10-0-0-0.ec2.internal/10.0.0.1:8032
21/11/25 09:36:36 INFO mapreduce.JobSubmitter: number of splits=1
21/11/25 09:36:36 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1637831653255_0001
21/11/25 09:36:37 INFO impl.YarnClientImpl: Submitted application application_1637831653255_0001
21/11/25 09:36:37 INFO mapreduce.Job: The url to track the job: http://ip-10-0-0-0.intern:20888/proxy/application_1637831653255_0001/
21/11/25 09:36:37 INFO tools.DistCp: DistCp job-id: job_1637831653255_0001
21/11/25 09:36:37 INFO mapreduce.Job: Running job: job_1637831653255_0001
21/11/25 09:36:48 INFO mapreduce.Job: Job job_1637831653255_0001 running in uber mode : false
21/11/25 09:36:48 INFO mapreduce.Job: map 10% reduce 0%
21/11/25 09:37:06 INFO mapreduce.Job: map 100% reduce 0%
21/11/25 09:37:08 INFO mapreduce.Job: Job job_1637831653255_0001 completed successfully
21/11/25 09:37:08 INFO mapreduce.Job: Counters: 38
File System Counters
FILE: Number of bytes read=0
FILE: Number of bytes written=172777
FILE: Number of read operations=0
FILE: Number of large read operations=0
FILE: Number of write operations=0
HDFS: Number of bytes read=361
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
S3N: Number of bytes read=482542278
S3N: Number of bytes written=0
S3N: Number of read operations=0
S3N: Number of large read operations=0
S3N: 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)=567648
Total time spent by all reduces in occupied slots (ms)=0
Total time spent by all map tasks (ms)=17739
Total vcore-milliseconds taken by all map tasks=17739
Total megabyte-milliseconds taken by all map tasks=18164736
Map-Reduce Framework
Map input records=1
Map output records=0
Input split bytes=137
Spilled Records=0
Failed Shuffles=0
Merged Map outputs=0
GC time elapsed (ms)=377
CPU time spent (ms)=19730
Physical memory (bytes) snapshot=531374080
Virtual memory (bytes) snapshot=3308642304
Total committed heap usage (bytes)=448266240
File Input Format Counters
Bytes Read=224
File Output Format Counters
Bytes Written=0
DistCp Counters
Bytes Copied=482542278
Bytes Expected=482542278
Files Copied=1
```

- Loading 2019-Nov.csv into ‘hive-case-study’

```
[hadoop@ip-10-0-1-10 ~]$ hadoop distcp s3n://e-commerce-events-ml/2019-Nov.csv /hive-case-study/2019-nov.csv
21/11/25 09:38:40 INFO tools.DistCp: Input Options: DistCpOptionsAtomicCommit=false, syncOrder=false, deleteMissing=false, ignoreFailures=false, overwrite=false, skipCRC=false, blocking=true, sourceListStatusThreads=0, maxMaps=20, mapBandwidth=100, sslConfigurationFile='null', copyStrategy='uniformsize', preserveStatus=[], preserveRawXattrs=false, atomicWorkPath=null, logPath=null,
21/11/25 09:38:40 INFO client.RMProxy: Connecting to ResourceManager at ip-10-0-1-10.ec2.internal/10.0.1.10:8032
21/11/25 09:38:44 INFO tools.RMProxy: Connecting to ResourceManager at ip-10-0-1-10.ec2.internal/10.0.1.10:8032
21/11/25 09:38:44 INFO tools.SimpleCopyListing: Paths: (files dirs) cnt = 1; dirCnt = 0
21/11/25 09:38:44 INFO tools.SimpleCopyListing: Build file listing completed.
21/11/25 09:38:44 INFO Configuration.deprecation: io.sort.mb is deprecated. Instead, use mapreduce.task.io.sort.mb
21/11/25 09:38:44 INFO Configuration.deprecation: io.sort.factor is deprecated. Instead, use mapreduce.task.io.sort.factor
21/11/25 09:38:44 INFO tools.DistCp: Number of paths in the copy list: 1
21/11/25 09:38:44 INFO tools.DistCp: Number of paths in the copy list: 1
21/11/25 09:38:44 INFO client.RMProxy: Connecting to ResourceManager at ip-10-0-1-10.ec2.internal/10.0.1.10:8032
21/11/25 09:38:45 INFO mapreduce.JobSubmitter: number of splits:1
21/11/25 09:38:45 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1637831653255_0002
21/11/25 09:38:45 INFO impl.YarnClientImpl: Submitted application application_1637831653255_0002
21/11/25 09:38:45 INFO mapreduce.Job: The url to track the job: http://ip-10-0-1-10.ec2.internal:20888/proxy/application_1637831653255_0002/
21/11/25 09:38:45 INFO tools.DistCp: JobId: job_1637831653255_0002
21/11/25 09:38:45 INFO tools.DistCp: JobId: job_1637831653255_0002
21/11/25 09:38:53 INFO mapreduce.Job: Job job_1637831653255_0002 running in uber mode : false
21/11/25 09:38:53 INFO mapreduce.Job: map 0% reduce 0%
21/11/25 09:39:12 INFO mapreduce.Job: map 100% reduce 0%
21/11/25 09:39:14 INFO mapreduce.Job: Job job_1637831653255_0002 completed successfully
21/11/25 09:39:14 INFO mapreduce.Job: Counters: 38
  File System Counters
    FILE: Number of bytes read=0
    FILE: Number of bytes written=172777
    FILE: Number of read operations=0
    FILE: Number of large read operations=0
    FILE: Number of write operations=0
    HDFS: Number of bytes read=361
    HDFS: Number of bytes written=545839412
    HDFS: Number of read operations=42
    HDFS: Number of large read operations=0
    HDFS: Number of write operations=4
    S3N: Number of bytes read=545839412
    S3N: Number of bytes written=0
    S3N: Number of read operations=0
    S3N: Number of large read operations=0
    S3N: 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)=546816
    Total time spent by all reduces in occupied slots (ms)=0
    Total time spent by all map tasks (ms)=17088
    Total time spent in milli-seconds taken by all map tasks=17088
    Total megabyte-milliseconds taken by all map tasks=17498112
  Map-Reduce Framework
    Map input records=1
    Map output records=0
    Input split bytes=137
    Spilled Records=0
    Failed Shuffles=0
    Merged Map outputs=0
    GC time elapsed (ms)=299
    CPU time spent (ms)=20630
    Physical memory (bytes) snapshot=563523584
    Virtual memory (bytes) snapshot=3304165376
    Total resident heap usage (bytes)=458227712
  File Input Format Counters
    Bytes Read=224
  File Output Format Counters
    Bytes Written=0
  DistCp Counters
    Bytes Copied=545839412
    Bytes Expected=545839412
    Files Copied=1
```

6. Launching Hive interface in hadoop

```
[hadoop@ip-10-0-1-10 ~]$ hive
Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j2.properties Async: false
hive>
```

7. Creating the database ‘ecom_db’

```
hive> show databases;
OK
default
Time taken: 0.627 seconds, Fetched: 1 row(s)
hive> create database if not exists ecom_db ;
OK
Time taken: 0.411 seconds
hive> describe database ecom_db;
OK
ecom_db          hdfs://ip-10-0-1-10.ec2.internal:8020/user/hive/warehouse/com_db.db  hadoop USER
Time taken: 0.064 seconds, Fetched: 1 row(s)
```

8. Creating the table 'ecom_table' in our database 'ecom_db'

```
hive> -- Creating table in ecom_db
hive> create external table if not exists ecom_table ( event_time timestamp, event_type string, product_id string, category_id string, category_code string, brand string, price float, user_id bigint, user_session string)
> ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.OpenCSVSerde'
> WITH SERDEPROPERTIES
> ( "separatorChar" = ",", "quoteChar" = "\"", "escapeChar" = "\\\"")
> stored as textfile LOCATION '/hive-case-study'
> TBLPROPERTIES ('skip.header.line.count'=1");
OK
Time taken: 0.168 seconds
hive> describe ecom_table ;
OK
event_time          string          from deserializer
event_type          string          from deserializer
product_id          string          from deserializer
category_id         string          from deserializer
category_code       string          from deserializer
brand               string          from deserializer
price               float           from deserializer
user_id              string          from deserializer
user_session         string          from deserializer
Time taken: 0.069 seconds, Fetched: 9 row(s)
```

Here, we can observe that the data type of the attributes is not of the required type. So, after loading the data into ecom_table, we have to create another external table and insert data using ecom_table to that newly created external table to ensure that all the attributes are of the correct datatypes.

- Loading the data into the 'ecom_table'

```
hive> -- Loading data in the ecom_table from both the datasets i.e. 2019-Oct & 2019-Nov
hive>
> LOAD DATA INPATH '/hive-case-study/2019-oct.csv' into table ecom_table;
Loading data to table ecom_db.ecom_table
OK
Time taken: 1.42 seconds
hive> LOAD DATA INPATH '/hive-case-study/2019-nov.csv' into table ecom_table;
Loading data to table ecom_db.ecom_table
OK
Time taken: 0.898 seconds
```

- Viewing the data in 'ecom_table'

```
hive> -- Checking the data in the ecom_table after loading
hive> select * from ecom_table limit 5;
OK
2019-11-01 00:00:02 UTC view      5802432 1487580009286598681          0.32    562076640      09fafd6c-6c99-46b1-834f-33
527f4de241
2019-11-01 00:00:09 UTC cart     5844397 1487580006317032337          2.38    553329724      2067216c-31b5-455d-a1cc-af
0575a34ffb
2019-11-01 00:00:10 UTC view      5837166 1783999064103190764          pnb    22.22    556138645      57ed222e-a54a-4907-9944-5a
875c2d7f4f
2019-11-01 00:00:11 UTC cart     5876812 1487580010100293687          jessnail   3.16    564506666      186c1951-8052-4b37
-adce-dd9644bd5f7
2019-11-01 00:00:24 UTC remove_from_cart 5826182 1487580007483048900          3.33    553329724      2067216c-3
1b5-455d-a1cc-af0575a34ffb
Time taken: 3.012 seconds, Fetched: 5 row(s)
```

9. Creating another table ‘new_ecom_table’ as aforementioned

```
hive> -- Creating another table for data analysis with correct datatype of attributes
hive> create external table if not exists new_ecom_table (
    > 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;
OK
Time taken: 0.103 seconds
```

```
hive> describe new_ecom_table;
OK
+-----+-----+
| 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  |
+-----+-----+
Time taken: 0.316 seconds, Fetched: 9 row(s)
```

In the new_ecom_table, we can observe that the data type of the attributes is of the required type.

- Inserting data into ‘new_ecom_table’ using ‘ecom_table’.

```
hive> -- Inserting data from ecom_table to new_ecom_table
hive> INSERT INTO new_ecom_table
    > select
    > cast(replace(event_time,'UTC','') as timestamp),
    > event_type,
    > product_id,
    > category_id,
    > category_code,
    > brand,
    > cast(price as float),
    > cast(user_id as bigint),
    > user_session
    > from ecom_table;
Query ID = hadoop_20211125125346_db70616a-c10c-4960-a42e-8083b3d5cf3d
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_1637831653255_0008)

-----  

      VERTICES      MODE      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED  

-----  

Map 1 ..... container      SUCCEEDED      2        2        0        0        0        0        0
-----  

VERTICES: 01/01  [=====>>>] 100%  ELAPSED TIME: 121.10 s  

-----  

Loading data to table ecom_db.new_ecom_table
OK
Time taken: 131.658 seconds
```

- Viewing the data in ‘new_ecom_table’

```
hive> select * from new_ecom_table limit 5;
OK
2019-11-01 00:00:02    view    5802432 1487580009286598681          0.32    562076640    09fafd6c-6c99-46b1-834f-33
527f4de241
2019-11-01 00:00:09    cart    5844397 1487580006317032337          2.38    553329724    2067216c-31b5-455d-a1cc-af
0575a34ffb
2019-11-01 00:00:10    view    5837166 1783999064103190764      pnb    22.22    556138645    57ed222e-a54a-4907-9944-5a
875c2d7f4f
2019-11-01 00:00:11    cart    5876812 1487580010100293687      jessnail   3.16    564506666    186c1951-8052-4b37
-adce-dd9644b1d5f7
2019-11-01 00:00:24    remove_from_cart    5826182 1487580007483048900          3.33    553329724    2067216c-3
1b5-455d-a1cc-af0575a34ffb
Time taken: 0.162 seconds, Fetched: 5 row(s)
```

Now, we will make a query on our new_ecom_table to check the time taken to run a query.

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

```
select sum(price)
from new_ecom_table
where MONTH(event_time)= 10 and event_type = 'purchase';
```

```
hive> select sum(price)
> from new_ecom_table
> where MONTH(event_time)= 10 and event_type = 'purchase';
Query ID = hadoop_20211125181851_03579fd1-7c81-4200-921b-cfef3ce51012
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1637831653255_0024)
```

| VERTICES | MODE | STATUS | TOTAL | COMPLETED | RUNNING | PENDING | FAILED | KILLED |
|-----------------|-----------|-----------|-------|-----------|---------|---------|--------|--------|
| Map 1 | container | SUCCEEDED | 7 | 7 | 0 | 0 | 0 | 0 |
| Reducer 2 | container | SUCCEEDED | 1 | 1 | 0 | 0 | 0 | 0 |

VERTICES: 02/02 [=====>>] 100% ELAPSED TIME: 33.91 s

```
OK
1211538.4295325726
Time taken: 37.511 seconds, Fetched: 1 row(s)
```

Here, in the above query, the time taken is 37.511 seconds.

In order to optimise our query, we can use optimisation techniques such as **partitioning** and **bucketing**.

10. Creating a partitioned table 'part_new_ecom_table'

```
hive> -- Creating a partitioned table for better optimisation of query
hive> create table if not exists part_new_ecom_table (
    > 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)
    > row format delimited fields terminated by ','
    > lines terminated by '\n' ;
OK
Time taken: 0.136 seconds
hive> describe part_new_ecom_table;
OK
event_time          timestamp
product_id         string
category_id        string
category_code      string
brand              string
price              float
user_id             bigint
user_session       string
event_type          string

# Partition Information
# col_name           data_type            comment
event_type          string
Time taken: 0.316 seconds, Fetched: 14 row(s)
```

We are creating partition based on an attribute 'event_type' using the static partitioning method.

- Checking the distinct values in event_type.

```
hive> select distinct event_type
    > from new_ecom_table;
Query ID = hadoop_20211125134627_2011087c-5dd4-4514-99ef-1d550777ea20
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_1637831653255_0011)

-----  

     VERTICES      MODE      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED  

-----  

Map 1 ..... container  SUCCEEDED    7      7      0      0      0      0      0  

Reducer 2 ..... container  SUCCEEDED    4      4      0      0      0      0      0  

-----  

VERTICES: 02/02  [=====>>>] 100%  ELAPSED TIME: 31.76 s  

-----  

OK
view
purchase
cart
remove_from_cart
Time taken: 40.156 seconds, Fetched: 4 row(s)
```

We can observe that there are four distinct values in the attribute 'event_type'. They are view, purchase, cart and remove_from_cart.

- Inserting data into the 'part_new_ecom_table' based on distinct values of event_type attribute

```

hive> -- Inserting data from new_ecom_table to part_new_ecom_table
hive>
> insert into table part_new_ecom_table partition (event_type = 'view')
> select event_time, product_id, category_id, category_code, brand, price, user_id, user_session
| > from new_ecom_table where event_type = 'view';
Query ID = hadoop_20211125140714_c52fa5fe-0abe-4307-a22c-30c5efd25ce5
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_1637831653255_0012)

-----  

      VERTICES      MODE      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED  

-----  

Map 1 ..... container  SUCCEEDED    7        7        0        0        0        0        0
-----  

VERTICES: 01/01 [=====>>] 100% ELAPSED TIME: 53.51 s  

-----  

Loading data to table ecom_db.part_new_ecom_table partition (event_type=view)
OK
Time taken: 62.756 seconds

```

```

hive> insert into table part_new_ecom_table partition (event_type = 'purchase')
> select event_time, product_id, category_id, category_code, brand, price, user_id, user_session
> from new_ecom_table where event_type = 'purchase';
Query ID = hadoop_20211125140913_55111259-1708-417a-8a84-6c4c43117940
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1637831653255_0012)

-----  

      VERTICES      MODE      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED  

-----  

Map 1 ..... container  SUCCEEDED    7        7        0        0        0        0        0
-----  

VERTICES: 01/01 [=====>>] 100% ELAPSED TIME: 33.03 s  

-----  

Loading data to table ecom_db.part_new_ecom_table partition (event_type=purchase)
OK
Time taken: 34.239 seconds

```

```

hive> insert into table part_new_ecom_table partition (event_type = 'remove_from_cart')
> select event_time, product_id, category_id, category_code, brand, price, user_id, user_session
| > from new_ecom_table where event_type = 'remove_from_cart';
Query ID = hadoop_20211125141130_83eada57-3582-46a8-ae60-5ec495aba884
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1637831653255_0012)

-----  

      VERTICES      MODE      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED  

-----  

Map 1 ..... container  SUCCEEDED    7        7        0        0        0        0        0
-----  

VERTICES: 01/01 [=====>>] 100% ELAPSED TIME: 40.70 s  

-----  

Loading data to table ecom_db.part_new_ecom_table partition (event_type=remove_from_cart)
OK
Time taken: 41.899 seconds

```

```

hive> insert into table part_new_ecom_table partition (event_type = 'cart')
> select event_time, product_id, category_id, category_code, brand, price, user_id, user_session
> from new_ecom_table where event_type = 'cart';
Query ID = hadoop_20211125141000_c64cf84b-b9f2-475d-9c62-a5727cc877b5
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1637831653255_0012)

-----  

  VERTICES      MODE      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED  

Map 1 ..... container    SUCCEEDED     7       7       0       0       0       0       0
-----  

VERTICES: 01/01 [=====>] 100% ELAPSED TIME: 43.29 s  

-----  

Loading data to table ecom_db.part_new_ecom_table partition (event_type=cart)
OK
Time taken: 44.569 seconds

```

- Checking the partitions created in the ‘part_new_ecom_table’.

```

hive> show partitions part_new_ecom_table;
OK
event_type=cart
event_type=purchase
event_type=remove_from_cart
event_type=view
Time taken: 0.056 seconds, Fetched: 4 row(s)

```

We can observe that the four partitions in the part_new_ecom_table are created successfully.

- Querying using the partitioned table ‘part_new_ecom_table’.

```

select sum(price)

from part_new_ecom_table

where MONTH(event_time)= 10 and event_type = 'purchase';

```

```

hive> select sum(price)
> from part_new_ecom_table
> where MONTH(event_time)= 10 and event_type = 'purchase';
Query ID = hadoop_20211125200414_b6ab84f0-cc76-4526-b1ba-e392de11f375
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1637831653255_0032)

-----  

  VERTICES      MODE      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED  

Map 1 ..... container    SUCCEEDED     3       3       0       0       0       0       0
Reducer 2 ..... container    SUCCEEDED     1       1       0       0       0       0       0
-----  

VERTICES: 02/02 [=====>] 100% ELAPSED TIME: 20.77 s  

-----  

OK
1211538.4295325726
Time taken: 22.5 seconds, Fetched: 1 row(s)

```

The time taken in the given query is 22.5 seconds, which is 15.011 seconds shorter than the time taken without employing the partitioning optimisation technique.

Now, we are going to use another optimisation technique called bucketing.

Before bucketing, we did some settings in the hive environment.

```
set hive.cli.print.header=true;
set hive.exec.dynamic.partition=true;
set hive.exec.dynamic.partition.mode=nonstrict;
set hive.enforce.bucketing = true;
```

11. Creating a bucketed table 'buck_new_ecom_table'.

```
hive> -- Creating a bucketed table for better optimisation of query
hive>
> create table if not exists buck_new_ecom_table (
> event_time timestamp,
> product_id string,
> category_id string,
> category_code string,
> brand string,
> price float,
> user_id bigint,
> user_session string )
> partitioned by (event_type string)
> clustered by (category_code) into 6 buckets
> row format delimited fields terminated by ','
> lines terminated by '\n' stored as textfile;
OK
Time taken: 0.063 seconds
hive> describe buck_new_ecom_table;
OK
event_time          timestamp
product_id         string
category_id        string
category_code      string
brand              string
price              float
user_id             bigint
user_session       string
event_type          string
# Partition Information
# col_name           data_type           comment
event_type          string
Time taken: 0.059 seconds, Fetched: 14 row(s)
```

- Inserting data into the bucketed table buck_new_ecom_table'.

```
hive> -- Inserting data from new_ecom_table to buck_new_ecom_table
hive>
> insert into table buck_new_ecom_table partition (event_type)
> select event_time, product_id, category_id, category_code, brand, price, user_id, user_session, event_type
> from new_ecom_table;
Query ID = hadoop_20211125202727_8316b546-8e52-4f41-b891-e957b21e5a72
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_1637831653255_0033)

-----  

 VERTICES    MODE     STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED  

Map 1 ..... container    SUCCEEDED    7        7        0        0        0        0  

Reducer 2 ..... container    SUCCEEDED    4        4        0        0        0        0  

-----  

VERTICES: 02/02 [=====>>>] 100% ELAPSED TIME: 152.23 s  

-----  

Loading data to table ecom_db.buck_new_ecom_table partition (event_type=null)  

Loaded : 4/4 partitions.  

    Time taken to load dynamic partitions: 0.266 seconds  

    Time taken for adding to write entity : 0.003 seconds  

OK
event_time      product_id      category_id      category_code      brand      price      user_id      user_session      event_type
Time taken: 161.366 seconds
```

- Querying using the bucketed table ‘buck_new_ecom_table’

```
select sum(price)

from buck_new_ecom_table

where MONTH(event_time)= 10 and event_type = 'purchase';
```

```
hive> select sum(price)
> from buck_new_ecom_table
> where MONTH(event_time)= 10 and event_type = 'purchase';
Query ID = hadoop_20211125203246_be158112-d0f3-47c3-a4b9-9049dcb746c2
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1637831653255_0033)

-----  

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: 14.08 s  

-----  

OK  

_c0  

1211538.4295325726  

Time taken: 14.78 seconds, Fetched: 1 row(s)
```

We observe that the time taken to run the query by using the

- new_ecom_table is **37.5 s**
- part_new_ecom_table is **22.5 s**
- buck_new_ecom_table is **14.78 s**

Therefore, we can observe that bucketing optimisation technique is taking less time than partitioning and normal querying on tables in the ecom_db database. Hence, we will be using bucketing optimisation technique for querying on our bucketed table i.e. ‘buck_new_ecom_table’.

12. Queries

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

```
select sum(price)

from buck_new_ecom_table

where MONTH(event_time)= 10 and event_type = 'purchase';
```

```
hive> select sum(price)
> from buck_new_ecom_table
> where MONTH(event_time)= 10 and event_type = 'purchase';
Query ID = hadoop_20211125203246_be158112-d0f3-47c3-a4b9-9049dcb746c2
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1637831653255_0033)

-----  

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: 14.08 s  

-----  

OK  

_c0  

1211538.4295325726  

Time taken: 14.78 seconds, Fetched: 1 row(s)
```

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

```
select MONTH(event_time) as month , sum(price)
from buck_new_ecom_table
where event_type = 'purchase'
group by MONTH(event_time);
```

```
hive> select MONTH(event_time) as month , sum(price)
> from buck_new_ecom_table
> where event_type = 'purchase'
> group by MONTH(event_time);
Query ID = hadoop_20211125203329_c7aedf1c-8753-4cd7-af06-b06a17f203bf
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1637831653255_0033)

-----  
 VERTICES      MODE      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED  
-----  
Map 1 ..... container  SUCCEEDED    2        2        0        0        0        0        0  
Reducer 2 ..... container  SUCCEEDED    1        1        0        0        0        0        0  
-----  
VERTICES: 02/02  [=====>>>] 100%  ELAPSED TIME: 13.51 s  
-----  
OK  
month _c1  
10    1211538.4295325726  
11    1531016.8991247676  
Time taken: 14.078 seconds, Fetched: 2 row(s)
```

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

```
select (sum(case when month(event_time)=11 then price else 0 end) -
sum(case when
month(event_time)=10 then price else 0 end)) as total_revenue_diff
from buck_new_ecom_table
where event_type='purchase' ;
```

```
hive> select (sum(case when month(event_time)=11 then price else 0 end) - sum(case when
> month(event_time)=10 then price else 0 end)) as total_revenue_diff
> from buck_new_ecom_table
> where event_type='purchase' ;
Query ID = hadoop_20211125203347_9d7b4eb3-ba28-4e47-aac0-f10ef55f0275
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1637831653255_0033)

-----  
 VERTICES      MODE      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED  
-----  
Map 1 ..... container  SUCCEEDED    2        2        0        0        0        0        0  
Reducer 2 ..... container  SUCCEEDED    1        1        0        0        0        0        0  
-----  
VERTICES: 02/02  [=====>>>] 100%  ELAPSED TIME: 3.31 s  
-----  
OK
total_revenue_diff
319478.469592195
Time taken: 3.838 seconds, Fetched: 1 row(s)
```

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

```
select distinct(category_code)
from buck_new_ecom_table
where category_code != '';
```

```
hive> select distinct(category_code)
> from buck_new_ecom_table
> where category_code != '';
Query ID = hadoop_20211125203402_9afbf069-2f68-4dbe-9b64-34f0971017b0
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1637831653255_0033)

-----  
 VERTICES      MODE      STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED  
-----  
Map 1 ..... container    SUCCEEDED    7      7      0      0      0      0  
Reducer 2 ..... container    SUCCEEDED    4      4      0      0      0      0  
-----  
VERTICES: 02/02  [=====>>] 100% ELAPSED TIME: 26.08 s  
-----  
OK  
category_code  
accessories.bag  
appliances.environment.vacuum  
appliances.personal.hair_cutter  
sport.diving  
apparel.glove  
furniture.bathroom.bath  
furniture.living_room.cabinet  
stationery.cartrige  
accessories.cosmetic_bag  
appliances.environment.air_conditioner  
furniture.living_room.chair  
Time taken: 26.633 seconds, Fetched: 11 row(s)
```

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

```
select category_code, count(product_id) as total_products
from buck_new_ecom_table
group by category_code;
```

```
hive> select category_code, count(product_id) as total_products
> from buck_new_ecom_table
> group by category_code;
Query ID = hadoop_20211125203432_40974583-4a39-45a5-91f8-9ef36a246703
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1637831653255_0033)

-----  
 VERTICES      MODE      STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED  
-----  
Map 1 ..... container    SUCCEEDED    7      7      0      0      0      0  
Reducer 2 ..... container    SUCCEEDED    4      4      0      0      0      0  
-----  
VERTICES: 02/02  [=====>>] 100% ELAPSED TIME: 11.47 s  
-----  
OK  
category_code  total_products  
accessories.bag 11681  
appliances.environment.vacuum 59761  
appliances.personal.hair_cutter 1643  
sport.diving 2  
          8594895  
apparel.glove 18232  
furniture.bathroom.bath 9857  
furniture.living_room.cabinet 13439  
stationery.cartrige 26722  
accessories.cosmetic_bag 1248  
appliances.environment.air_conditioner 332  
furniture.living_room.chair 308  
Time taken: 12.018 seconds, Fetched: 12 row(s)
```

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

```
select brand, sum(price) as total_sales  
from buck_new_ecom_table  
where event_type = 'purchase' and brand != ''  
group by brand  
order by total_sales desc  
limit 1;
```

```
hive> select brand, sum(price) as total_sales  
> from buck_new_ecom_table  
> where event_type = 'purchase' and brand != ''  
> group by brand  
> order by total_sales desc  
> limit 1;  
Query ID = hadoop_20211125203447_b474a8cf-8d3b-409d-a1b6-615e05af850d  
Total jobs = 1  
Launching Job 1 out of 1  
Status: Running (Executing on YARN cluster with App id application_1637831653255_0033)  
  
-----  
 VERTICES      MODE      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED  
-----  
 Map 1 ..... container  SUCCEEDED   2       2        0        0        0        0        0  
 Reducer 2 ..... container  SUCCEEDED   1       1        0        0        0        0        0  
 Reducer 3 ..... container  SUCCEEDED   1       1        0        0        0        0        0  
-----  
VERTICES: 03/03  [=====>>>] 100%  ELAPSED TIME: 2.70 s  
-----  
OK  
brand    total_sales  
runail  148297.93996394053  
Time taken: 3.195 seconds, Fetched: 1 row(s)
```

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

```
with brand_sales_details as  
(select brand,  
     sum(case when month(event_time)=11 then price else 0 end) as  
          nov_sales,  
     sum(case when month(event_time)=10 then price else 0 end) as  
          oct_sales  
  from buck_new_ecom_table  
 where event_type='purchase'  
 group by brand)  
  
select brand, nov_sales, oct_sales, (nov_sales - oct_sales) as  
      nov_oct_sales_diff  
  from brand_sales_details  
 where (nov_sales-oct_sales)>0;
```

```

hive> with brand_sales_details as
  > ( select brand,
  > sum(case when month(event_time)=11 then price else 0 end) as nov_sales,
  > sum(case when month(event_time)=10 then price else 0 end) as oct_sales
  > from buck_new_ecom_table
  > where event_type='purchase'
  > group by brand )
  >
  > select brand, nov_sales, oct_sales, (nov_sales - oct_sales) as nov_oct_sales_diff
  > from brand_sales_details
  > where (nov_sales-oct_sales)>0 ;

```

Query ID = hadoop_20211125203509_6cff4120-508f-45bd-b388-7d06237dd4a0

Total jobs = 1

Launching Job 1 out of 1

Status: Running (Executing on YARN cluster with App id application_1637831653255_0033)

| 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: 14.07 s

OK

| brand | nov_sales | oct_sales | nov_oct_sales_diff |
|-------------------|--------------------|--------------------|--------------------|
| 619509.2397020273 | 474679.05964545906 | 144830.18005656824 | |
| airnails | 5691.519957095385 | 5118.899943232536 | 572.6200138628483 |
| art-visage | 2997.8000057935715 | 2092.7100064754486 | 905.0899993181229 |
| artex | 4327.249953508377 | 2730.6399517059326 | 1596.6100018024445 |
| aura | 177.5100040435791 | 83.95000076293945 | 93.56000328063965 |
| balbcare | 212.3800015449524 | 155.33000373840332 | 57.04999780654907 |
| barbie | 12.390000343322754 | 0.0 | 12.390000343322754 |
| batiste | 874.1700088977814 | 772.400013923645 | 101.76999497413635 |
| beautix | 12222.95004272461 | 10493.949965000153 | 1729.0000777244568 |
| beauty-free | 1782.8599914312363 | 554.1699986457825 | 1228.6899927854538 |
| beautyblender | 109.41000175476074 | 78.73999977111816 | 30.670001983642578 |
| beauugreen | 768.3499994277954 | 511.5099878311157 | 256.8400115966797 |
| benovy | 3259.969982147217 | 409.619996547699 | 2850.349985599518 |
| binacil | 24.260000228881836 | 0.0 | 24.260000228881836 |
| bioaqua | 1398.1200065612793 | 942.8900030851364 | 455.2300034761429 |
| biore | 90.30999946594238 | 60.650001525878906 | 29.659997940063477 |
| blixz | 63.400001764297485 | 38.94999921321869 | 24.450002551078796 |
| bluesky | 10565.529949843884 | 10307.239978790283 | 258.2899710536003 |
| bodyton | 1380.639987230301 | 1376.3399817943573 | 4.3000054359436035 |
| bpw.style | 14837.440190911293 | 11572.1500659585 | 3265.290124952793 |
| browxenna | 14916.730226278305 | 14331.370284080505 | 585.3599421977997 |
| candy | 799.3799992799759 | 534.9600057601929 | 264.419993519783 |

| | | | |
|-------------|--------------------|--------------------|--------------------|
| carmex | 243.3599967956543 | 145.07999897003174 | 98.27999782562256 |
| chi | 538.6099972724915 | 358.93999576568604 | 179.67000150680542 |
| coifin | 1428.4900131225586 | 902.9999961853027 | 525.4900169372559 |
| concept | 13380.400002479553 | 11032.14000660181 | 2348.2599958777428 |
| cosima | 20.930000603199005 | 20.230000972747803 | 0.6999996304512024 |
| cosmoprofi | 14536.989881515503 | 8322.80991601944 | 6214.179965496063 |
| cristalinas | 584.950008392334 | 427.63000297546387 | 157.32000541687012 |
| cutrin | 367.6199998855591 | 299.3700017929077 | 68.24999809265137 |
| de.lux | 2775.510024756193 | 1659.7000161707401 | 1115.810008585453 |
| deoproce | 329.1699993610382 | 316.8399999141693 | 12.329999446868896 |
| depilflax | 2803.7799961566925 | 2707.0699973106384 | 96.70999884605408 |
| dewal | 61.28999876976013 | 0.0 | 61.28999876976013 |
| dizao | 945.5100176334381 | 819.1300112009048 | 126.38000643253326 |
| domix | 12009.170008182526 | 10472.05003106594 | 1537.1199771165848 |
| ecocraft | 241.9500012397766 | 41.15999937057495 | 200.79000186920166 |
| ecolab | 1214.30000436306 | 262.84999895095825 | 951.4500054121017 |
| egomania | 146.04000091552734 | 77.46999835968018 | 68.57000255584717 |
| elizavecca | 204.29999923706055 | 70.52999973297119 | 133.76999950408936 |
| ellips | 606.0399996042252 | 245.8499938249588 | 360.19000577926636 |
| elskin | 307.6499996781349 | 251.0900001525879 | 56.55999952554703 |
| enjoy | 136.57000184059143 | 41.34999966621399 | 95.22000217437744 |
| entity | 719.2599903345108 | 479.70999866724014 | 239.54999166727066 |
| eos | 152.60999727249146 | 54.34000015258789 | 98.26999711990356 |
| estel | 24142.66994935274 | 21756.749947547913 | 2385.9200018048286 |
| estelare | 471.86999905109406 | 444.80999556183815 | 27.060003489255905 |
| f.o.x | 8577.279987692833 | 6624.229980587959 | 1953.0500071048737 |
| farmavita | 1291.969996213913 | 837.3699972629547 | 454.59999895095825 |
| farmona | 1843.4299907684326 | 1692.46000289917 | 150.9699878692627 |
| fedua | 263.80999755859375 | 52.3799991607666 | 211.42999839782715 |
| finish | 230.37999820709229 | 98.37999773025513 | 132.00000047683716 |
| fly | 27.170000553131104 | 17.139999389648438 | 10.030001163482666 |
| foamie | 80.48999977111816 | 35.03999996185303 | 45.44999980926514 |
| freedecor | 7671.800070524216 | 3421.7800273299217 | 4250.020043194294 |
| freshbubble | 502.3399975299835 | 318.6999980926514 | 183.63999772071838 |
| gehwol | 1557.6799898147583 | 1089.0699853897095 | 468.6100044250488 |
| glysolid | 91.59000062942505 | 69.73000013828278 | 21.860000491142273 |
| godefroy | 425.1200022697449 | 401.22000312805176 | 23.899999141693115 |
| grace | 102.61000108718872 | 100.9200005531311 | 1.6900005340576172 |
| grattol | 71472.70888674259 | 35445.53947067261 | 36027.169416069984 |
| greymy | 489.48999214172363 | 29.209999084472656 | 460.279993057251 |
| happyfons | 1091.5899834632874 | 801.9199857711792 | 289.66999769210815 |
| haruyama | 12352.910059452057 | 9390.690077126026 | 2962.2199823260307 |
| helloganic | 3.0999999046325684 | 0.0 | 3.0999999046325684 |
| igrobeauty | 645.0699995160103 | 513.6600003838539 | 131.40999913215637 |
| ingarden | 33566.209977939725 | 23161.38997283578 | 10404.820005103946 |
| inm | 351.2099983692169 | 288.01999855041504 | 63.18999981880188 |
| insight | 1721.9600095748901 | 1443.7000050544739 | 278.26000452041626 |
| irisk | 46946.04018642008 | 45591.96021157503 | 1354.0799748450518 |



| | | | |
|---------------|--------------------|--------------------|--------------------|
| italwax | 24799.37004429102 | 21940.239994883537 | 2859.130049407482 |
| jaguar | 1110.6500117778778 | 1102.110021829605 | 8.539989948272705 |
| jas | 3657.4299937039614 | 3318.9600024223328 | 338.4699912816286 |
| jessnail | 33345.23023867607 | 26287.840348243713 | 7057.389890432358 |
| joico | 2015.1000146865845 | 705.5200037956238 | 1309.5800108909607 |
| juno | 21.079999923706055 | 0.0 | 21.079999923706055 |
| kaaral | 5086.069996476173 | 4412.429983615875 | 673.6400128602982 |
| kamill | 81.48999953269958 | 63.010000228881836 | 18.47999930381775 |
| kapous | 14093.079938054085 | 11927.159952402115 | 2165.91998565197 |
| kares | 59.45000076293945 | 0.0 | 59.45000076293945 |
| kaypro | 3268.700007915497 | 881.3400187492371 | 2387.3599891662598 |
| keen | 435.6199960708618 | 236.34999418258667 | 199.27000188827515 |
| kerasys | 525.2000050544739 | 430.9100044965744 | 94.29000055789948 |
| kims | 632.0399990081787 | 330.0399923324585 | 302.0000066757202 |
| kinetics | 6945.260000705719 | 6334.249932765961 | 611.0100679397583 |
| kiss | 817.3299901485443 | 421.54999327659607 | 395.77999687194824 |
| kocostar | 594.9300029277802 | 310.8499982357025 | 284.08000469207764 |
| koelcia | 112.75 | 55.5 | 57.25 |
| koelf | 507.29000186920166 | 422.7300081253052 | 84.55999374389648 |
| konad | 810.6699978709221 | 739.8300001621246 | 70.83999770879745 |
| kosmekka | 1813.3700094223022 | 1181.4400033950806 | 631.9300060272217 |
| laboratorium | 312.5199975967407 | 246.4999952316284 | 66.01999807357788 |
| lador | 2471.5300159454346 | 2083.610013961792 | 387.9200019836426 |
| ladykin | 170.56999969482422 | 125.64999961853027 | 44.920000076293945 |
| latinoil | 384.5899987220764 | 249.5199966430664 | 135.07000207901 |
| levissime | 3085.3099098205566 | 2227.4999141693115 | 857.8099956512451 |
| levrana | 3664.0999879837036 | 2243.5599967837334 | 1420.5399911999702 |
| lianail | 16394.239884018898 | 5892.839952707291 | 10501.399931311607 |
| likato | 340.9699954986572 | 296.0599980354309 | 44.90999746322632 |
| limoni | 1796.6000032424927 | 1308.9000149965286 | 487.69998824596405 |
| lovely | 11939.059989094734 | 8704.380010932684 | 3234.6799781620502 |
| lowence | 567.7499952316284 | 242.83999252319336 | 324.91000270843506 |
| mane | 260.26000118255615 | 66.79000186920166 | 193.4699993133545 |
| marathon | 10273.099990844727 | 7280.749939441681 | 2992.3500514030457 |
| markell | 2834.43000292778 | 1768.7500059604645 | 1065.6799969673157 |
| marutaka-foot | 109.33000040054321 | 49.21999979019165 | 60.11000061035156 |
| masura | 33058.469878435135 | 31266.079910814762 | 1792.3899676203728 |
| matreshka | 182.66999757289886 | 0.0 | 182.66999757289886 |
| matrix | 3726.739989757538 | 3243.249990463257 | 483.489999294281 |
| mavala | 446.32000255584717 | 409.0400023460388 | 37.28000020980835 |
| metzger | 6457.159960865974 | 5373.4499744176865 | 1083.709986448288 |
| milv | 5642.01002573967 | 3904.940046072006 | 1737.0699796676636 |
| miskin | 293.0700011253357 | 158.04000186920166 | 135.02999925613403 |
| misssha | 2150.2800248861313 | 1293.830022573471 | 856.4500023126602 |
| moyou | 10.28000020980835 | 5.710000038146973 | 4.570000171661377 |
| nagaraku | 5327.680045571178 | 4369.7400778234005 | 957.9399677477777 |
| naomi | 389.0000011920929 | 0.0 | 389.0000011920929 |
| nefertiti | 366.64000034332275 | 233.51999759674072 | 133.12000274658203 |

| | | | |
|--------------|--------------------|--------------------|--------------------|
| neoleor | 51.7000076293945 | 43.4099984741211 | 8.29000915527344 |
| nirvel | 234.33000826835632 | 163.04000329971313 | 71.29000496864319 |
| nitrile | 1162.6800317764282 | 847.2800407409668 | 315.3999910354614 |
| oniq | 9841.649902820587 | 8425.409879207611 | 1416.240023612976 |
| orly | 931.0899903774261 | 902.3799939155579 | 28.709996461868286 |
| osmo | 762.3100028038025 | 645.5800037384033 | 116.72999906539917 |
| ovale | 3.099999046325684 | 2.5399999618530273 | 0.559999942779541 |
| plazan | 194.010000705719 | 101.37000036239624 | 92.64000034332275 |
| polarus | 11371.930022716522 | 6013.720007181168 | 5358.210015535355 |
| profepil | 118.01999974250793 | 93.36000156402588 | 24.659998178482056 |
| profhenna | 736.8500001430511 | 679.2300038337708 | 57.619996309280396 |
| protokeratin | 456.790002822876 | 201.24999809265137 | 255.5400047302246 |
| provoc | 1063.8200211524963 | 827.9900186061859 | 235.83000254631042 |
| rasyan | 28.940000295639038 | 18.799999952316284 | 10.140000343322754 |
| refectocil | 3475.57999587059 | 2716.1799943447113 | 759.4000015258789 |
| rosi | 3841.5600021481514 | 3077.0399764180183 | 764.520025730133 |
| rouabloff | 4913.770027637482 | 3491.3600150346756 | 1422.410012602806 |
| runail | 76758.65991047397 | 71539.28005346656 | 5219.379857007414 |
| s.care | 913.0699844360352 | 412.67999267578125 | 500.3899917602539 |
| sanoto | 1209.6799850463867 | 157.13999938964844 | 1052.5399856567383 |
| severina | 6120.479953020811 | 4775.8799668848515 | 1344.5999861359596 |
| shary | 1176.4899995326996 | 871.9600003957748 | 304.52999913692474 |
| shik | 4839.720018148422 | 3341.199989080429 | 1498.5200290679932 |
| skinity | 12.440000057220459 | 8.880000114440918 | 3.559999942779541 |
| skinlite | 890.4499936699867 | 651.9399995803833 | 238.50999408960342 |
| smart | 5902.139976501465 | 4457.259982824326 | 1444.8799936771393 |
| soleo | 212.52999597787857 | 204.1999952197075 | 8.330000758171082 |
| solomeya | 2685.8000009655952 | 1899.6999986171722 | 786.100002348423 |
| sophin | 1515.5200046300888 | 1067.8599869012833 | 447.66001772880554 |
| staleks | 11875.610019385815 | 8519.730030417442 | 3355.8799889683723 |
| strong | 38671.27037525177 | 29196.63009786606 | 9474.640277385712 |
| supertan | 66.51000016927719 | 50.37000048160553 | 16.13999968767166 |
| swarovski | 3043.1599831581116 | 1887.9299856424332 | 1155.2299975156784 |
| tertio | 245.80000019073486 | 236.15999841690063 | 9.640001773834229 |
| treaclemoon | 181.48999691009521 | 163.36999654769897 | 18.12000036239624 |
| trind | 542.9599976539612 | 298.0699954032898 | 244.8900022506714 |
| uno | 51039.74947929382 | 35302.029363155365 | 15737.720116138458 |
| uskusi | 5690.31001329422 | 5142.270027637482 | 548.0399856567383 |
| veraclara | 71.21000015735626 | 50.11000084877014 | 21.09999930858612 |
| vilenta | 231.20999908447266 | 197.59999787807465 | 33.61000120639801 |
| yoko | 11707.88005465269 | 8756.910053431988 | 2950.970001220703 |
| yu-r | 673.710018157959 | 271.4100036621094 | 402.3000144958496 |
| zeitun | 2009.6300013065338 | 708.6600031852722 | 1300.9699981212616 |

Time taken: 14.611 seconds, Fetched: 161 row(s)

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.

```
select user_id, sum(price) as total_amt_spent
from buck_new_ecom_table
where event_type='purchase'
group by user_id
order by total_amt_spent desc
limit 10;
```

```
hive> select user_id, sum(price) as total_amt_spent
> from buck_new_ecom_table
> where event_type='purchase'
> group by user_id
> order by total_amt_spent desc
> limit 10 ;
Query ID = hadoop_20211125203528_6fabaf08-c414-4a63-8776-8ea83ba908e7
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1637831653255_0033)
```

| 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: 4.82 s
```

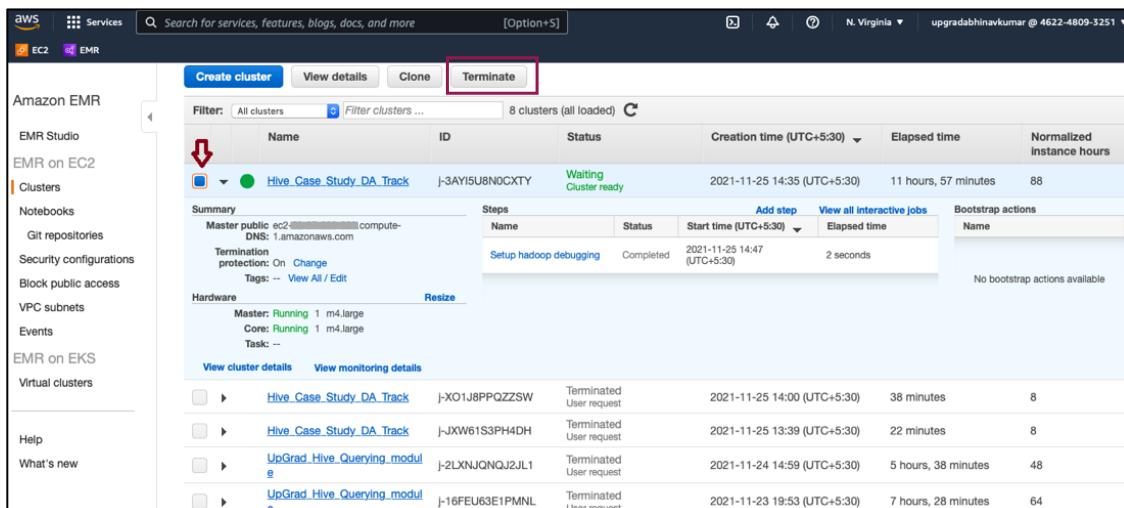
```
OK
```

```
user_id total_amt_spent
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: 5.573 seconds, Fetched: 10 row(s)
```

13. Dropping the database 'ecom_db'

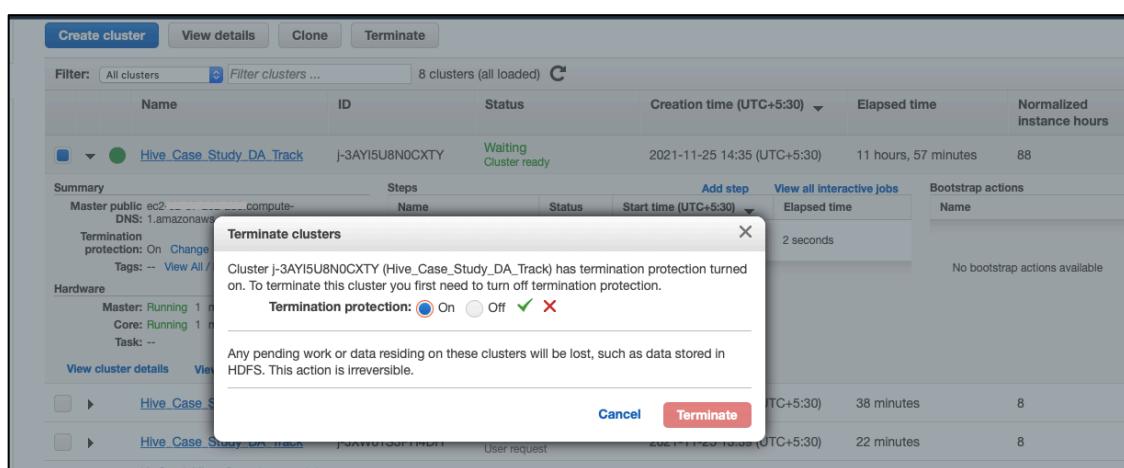
```
[hive> show databases;
OK
database_name
default
ecom_db
Time taken: 0.016 seconds, Fetched: 2 row(s)
[hive> drop database ecom_db cascade;
OK
Time taken: 0.448 seconds
[hive> show databases;
OK
database_name
default
Time taken: 0.008 seconds, Fetched: 1 row(s)
```

14. Terminating the cluster



The screenshot shows the AWS EMR console with the 'Clusters' section selected. A list of 8 clusters is displayed. The first cluster, 'Hive_Case_Study_DA_Track', has its 'Terminate' button highlighted with a red box. The cluster details show it is in a 'Waiting Cluster ready' state.

Select the cluster to be terminated and click on Terminate button.



The screenshot shows the 'Terminate clusters' dialog box. It contains a warning message: 'Cluster j-3AYI5U8N0CXTY (Hive_Case_Study_DA_Track) has termination protection turned on. To terminate this cluster you first need to turn off termination protection.' Below this, there is a radio button group for 'Termination protection' with two options: 'On' (selected) and 'Off'. At the bottom of the dialog are 'Cancel' and 'Terminate' buttons.

Terminate clusters

Cluster j-3AYI5U8N0CXTY (Hive_Case_Study_DA_Track) has termination protection turned on. To terminate this cluster you first need to turn off termination protection.

Termination protection: On Off ✓ ✗

Any pending work or data residing on these clusters will be lost, such as data stored in HDFS. This action is irreversible.

Cancel **Terminate**

Terminate clusters

Cluster j-3AYI5U8N0CXTY (Hive_Case_Study_DA_Track) has termination protection turned on. To terminate this cluster you first need to turn off termination protection.

Termination protection: Off Change

Any pending work or data residing on these clusters will be lost, such as data stored in HDFS. This action is irreversible.

Cancel **Terminate**

| Name | ID | Status | Creation time (UTC+5:30) | Elapsed time | Normalized instance hours |
|--|-----------------|--------------------------|-----------------------------|---------------------|---------------------------|
| <input checked="" type="checkbox"/> Hive_Case_Study_DA_Track | j-3AYI5U8N0CXTY | Terminating User request | 2021-11-25 14:35 (UTC+5:30) | 12 hours | 88 |
| <input type="checkbox"/> Hive_Case_Study_DA_Track | j-XO1J8PPQZZSW | Terminated User request | 2021-11-25 14:00 (UTC+5:30) | 38 minutes | 8 |
| <input type="checkbox"/> Hive_Case_Study_DA_Track | j-JXW61S3PH4DH | Terminated User request | 2021-11-25 13:39 (UTC+5:30) | 22 minutes | 8 |
| <input type="checkbox"/> UpGrad_Hive_Querying_module | j-2LXNJQNQJ2JL1 | Terminated User request | 2021-11-24 14:59 (UTC+5:30) | 5 hours, 38 minutes | 48 |
| <input type="checkbox"/> UpGrad_Hive_Querying_module | j-16FEU63E1PMNL | Terminated User request | 2021-11-23 19:53 (UTC+5:30) | 7 hours, 28 minutes | 64 |
| <input type="checkbox"/> DSC_30_Lab_Slot_2 | j-23N6X2GGYBYSB | Terminated User request | 2021-11-20 15:57 (UTC+5:30) | 7 hours, 34 minutes | 96 |
| <input type="checkbox"/> Demo_Cluster | j-1DHKRIJMIFBVZ | Terminated User request | 2021-11-19 19:23 (UTC+5:30) | 1 hour, 9 minutes | 8 |

| Name | ID | Status | Creation time (UTC+5:30) | Elapsed time | Normalized instance hours |
|--|-----------------|-------------------------|-----------------------------|--------------|---------------------------|
| <input checked="" type="checkbox"/> Hive_Case_Study_DA_Track | j-3AYI5U8N0CXTY | Terminated User request | 2021-11-25 14:35 (UTC+5:30) | 12 hours | 96 |

Successfully terminated the cluster