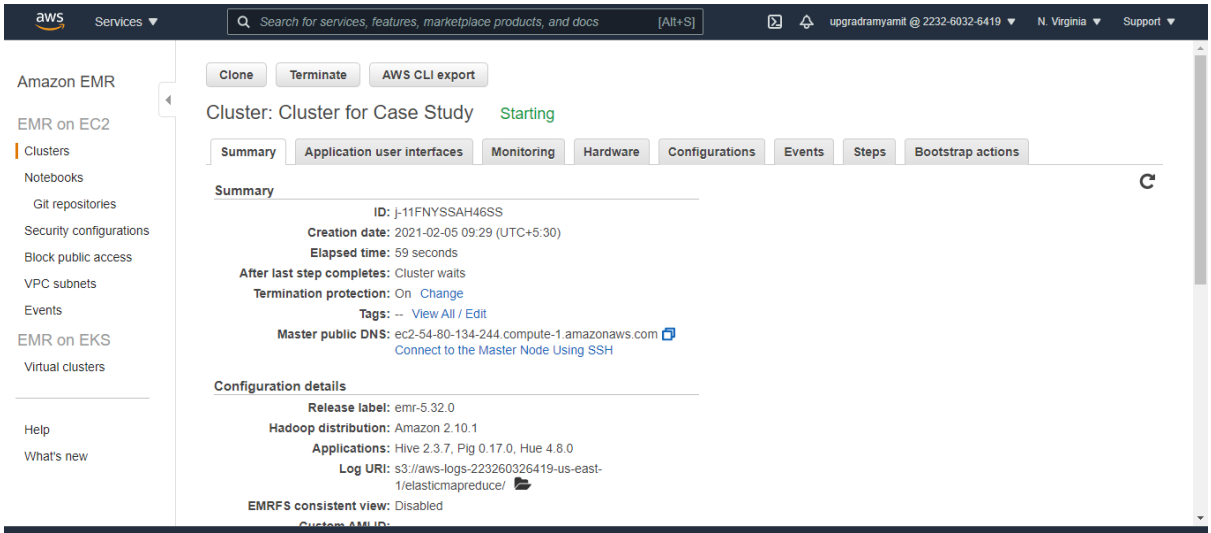
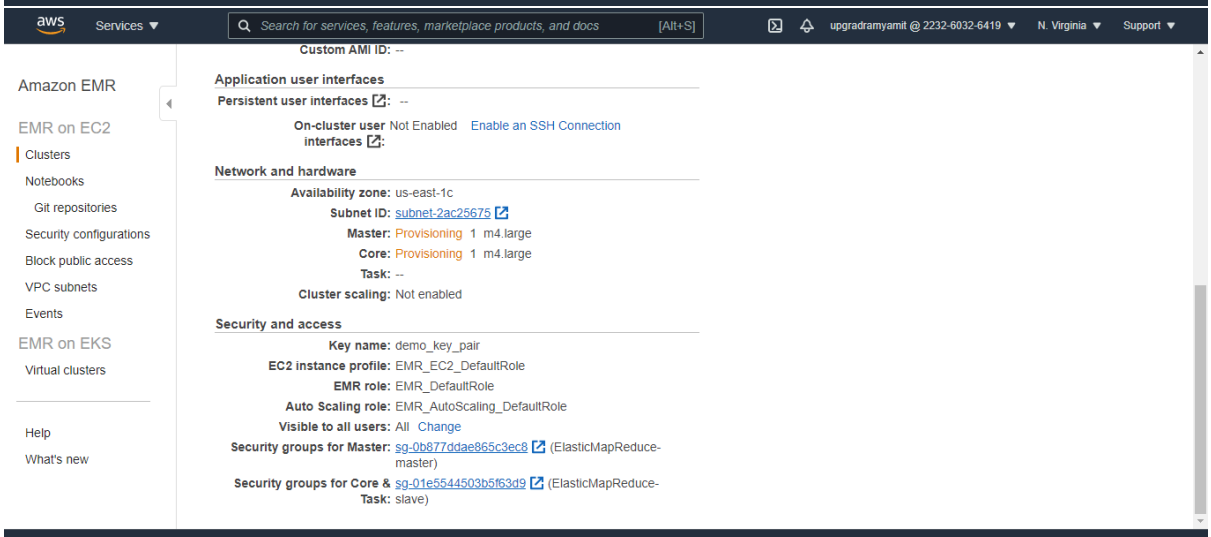
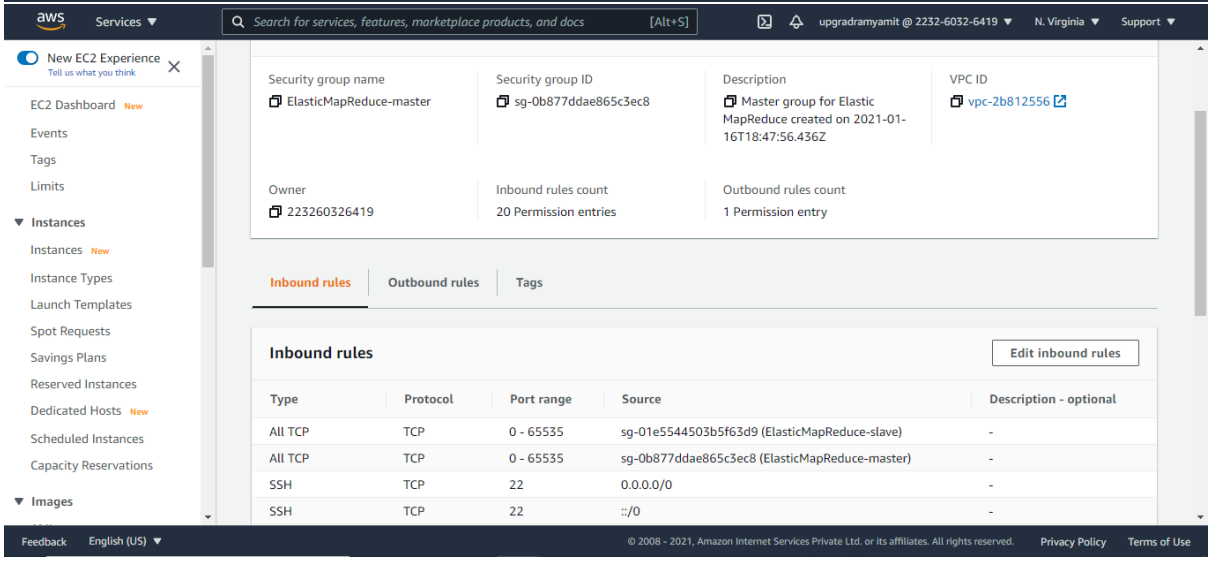
Copying the data set into the HDFS:

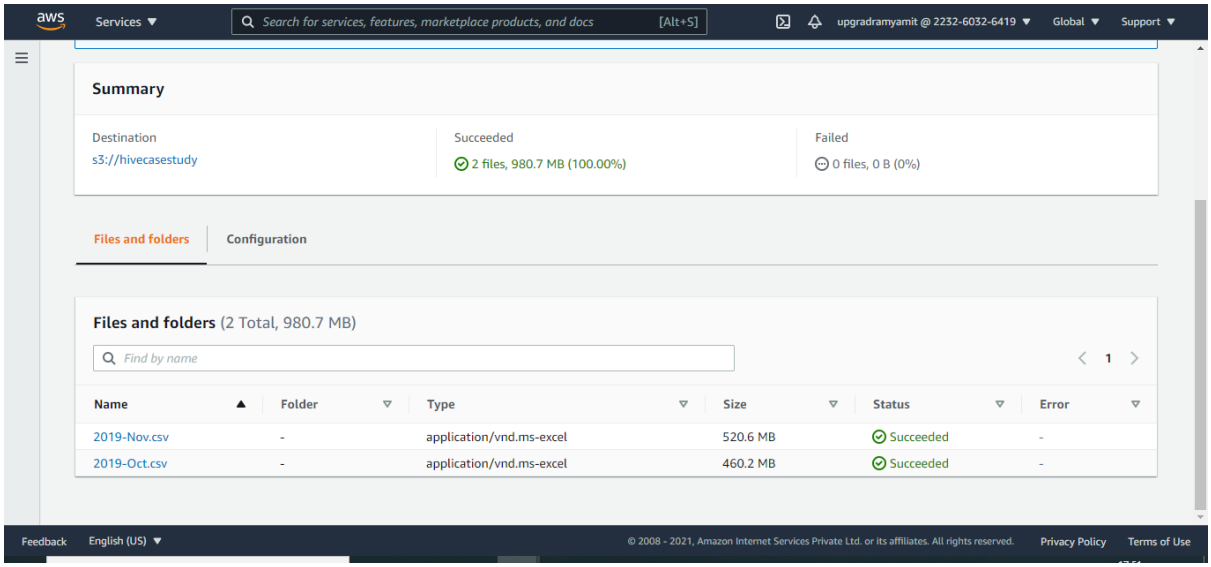
Launching an EMR cluster that utilizes the Hive services.



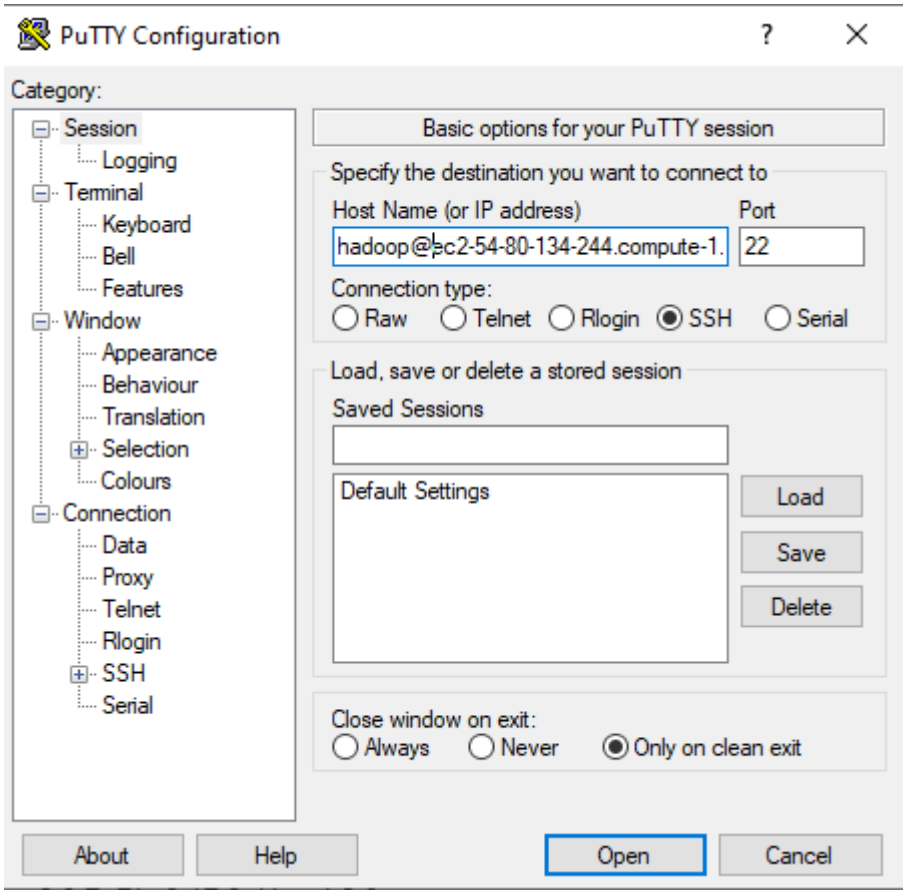


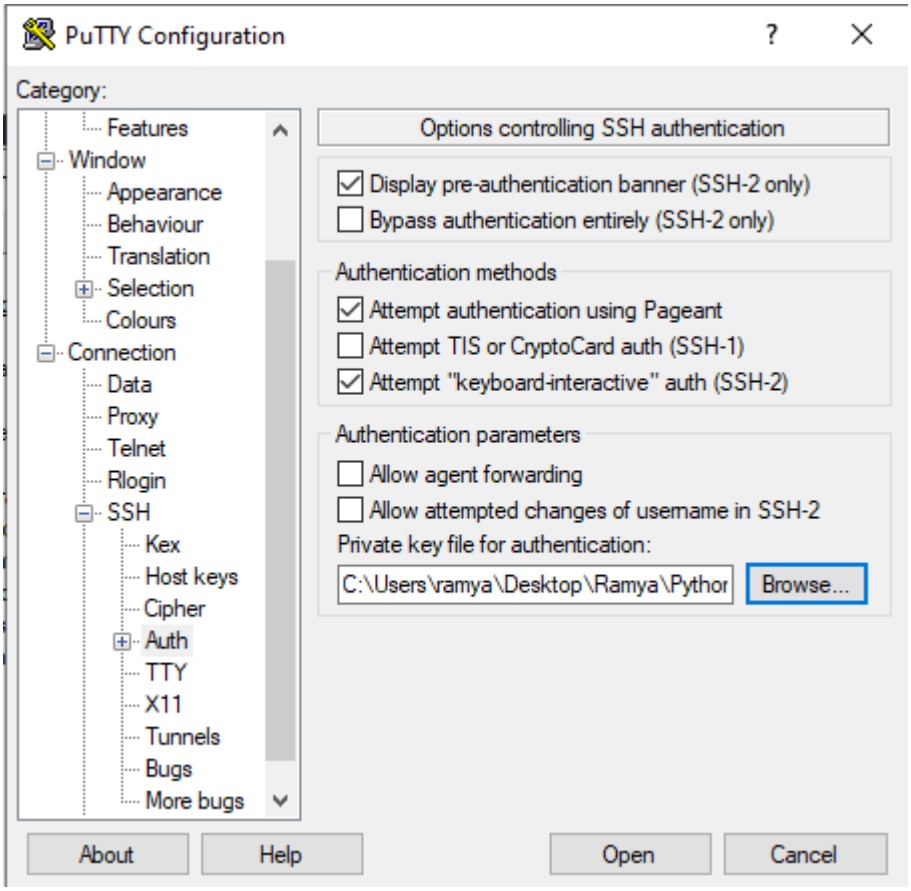


Files uploaded in s3



Launching Putty





Moving the data from the S3 bucket into the HDFS.

```
[hadoop@ip-172-31-41-129 ~]$ hadoop fs -ls /user/hive
Found 1 items
drwxrwxrwt   - hdfs hadoop          0 2021-02-05 04:05 /user/hive/warehouse
[hadoop@ip-172-31-41-129 ~]$ hadoop fs -mkdir /user/hive/casestudy
[hadoop@ip-172-31-41-129 ~]$ hadoop fs -ls /user/hive
Found 2 items
drwxr-xr-x   - hadoop hadoop          0 2021-02-05 04:11 /user/hive/casestudy
drwxrwxrwt   - hdfs   hadoop          0 2021-02-05 04:05 /user/hive/warehouse
[hadoop@ip-172-31-41-129 ~]$ aws s3 ls hivecasestudy
2021-02-02 09:39:41  545839412 2019-Nov.csv
2021-02-02 09:39:41  482542278 2019-Oct.csv
[hadoop@ip-172-31-41-129 ~]$ hadoop distcp 's3://hivecasestudy/*' '/user/hive/casestudy/'
21/02/05 04:16:00 INFO tools.OptionsParser: parseChunkSize: blocksperchunk false
21/02/05 04:16:01 INFO tools.DistCp: Input Options: DistCpOptions{atomicCommit=false, syncFolder=false, deleteMissing=false, ignoreFailures=false, overwrite=false, appe
nd=false, useDiff=false, useRdiff=false, fromSnapshot=null, toSnapshot=null, skipCRC=false, blocking=true, numListstatusThreads=0, maxMaps=20, mapBandwidth=100, sslConf
igurationFile='null', copyStrategy='uniformsize', preserveStatus=[], preserveRawXattrs=false, atomicWorkPath=null, logPath=null, sourceFileListing=null, sourcePaths=[s3
://hivecasestudy/*], targetPath=/user/hive/casestudy, targetPathExists=true, filtersFile='null', blocksPerChunk=0, copyBufferSize=8192, verboseLog=false}
21/02/05 04:16:01 INFO client.RMProxy: Connecting to ResourceManager at ip-172-31-41-129.ec2.internal/172.31.41.129:8032
21/02/05 04:16:01 INFO client.AHSProxy: Connecting to Application History server at ip-172-31-41-129.ec2.internal/172.31.41.129:10200
21/02/05 04:16:06 INFO tools.SimpleCopyListing: Paths (files+dirs) cnt = 2; dirCnt = 0
21/02/05 04:16:06 INFO tools.SimpleCopyListing: Build file listing completed.
21/02/05 04:16:06 INFO Configuration.deprecation: io.sort.mb is deprecated. Instead, use mapreduce.task.io.sort.mb
21/02/05 04:16:06 INFO Configuration.deprecation: io.sort.factor is deprecated. Instead, use mapreduce.task.io.sort.factor
21/02/05 04:16:06 INFO tools.DistCp: Number of paths in the copy list: 2
21/02/05 04:16:06 INFO tools.DistCp: Number of paths in the copy list: 2
21/02/05 04:16:06 INFO client.RMProxy: Connecting to ResourceManager at ip-172-31-41-129.ec2.internal/172.31.41.129:8032
21/02/05 04:16:06 INFO client.AHSProxy: Connecting to Application History server at ip-172-31-41-129.ec2.internal/172.31.41.129:10200
21/02/05 04:16:06 INFO mapreduce.JobSubmitter: number of splits:2
21/02/05 04:16:06 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1612497986037_0002
21/02/05 04:16:07 INFO conf.Configuration: resource-types.xml not found
21/02/05 04:16:07 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
21/02/05 04:16:07 INFO resource.ResourceUtils: Adding resource type - name = memory-mb, units = Mi, type = COUNTABLE
21/02/05 04:16:07 INFO resource.ResourceUtils: Adding resource type - name = vcores, units = , type = COUNTABLE
21/02/05 04:16:07 INFO impl.YarnClientImpl: Submitted application application_1612497986037_0002
21/02/05 04:16:07 INFO mapreduce.Job: The url to track the job: http://ip-172-31-41-129.ec2.internal:20888/proxy/application_1612497986037_0002/
21/02/05 04:16:07 INFO tools.DistCp: DistCp job-id: job_1612497986037_0002
21/02/05 04:16:07 INFO mapreduce.Job: Running job: job_1612497986037_0002
21/02/05 04:16:30 INFO mapreduce.Job: Job job_1612497986037_0002 running in uber mode : false
21/02/05 04:16:30 INFO mapreduce.Job:  map 0% reduce 0%
21/02/05 04:16:52 INFO mapreduce.Job:  map 100% reduce 0%
21/02/05 04:17:07 INFO mapreduce.Job: Job job_1612497986037_0002 completed successfully
21/02/05 04:17:08 INFO mapreduce.Job: Counters: 38
        File System Counters
                FILE: Number of bytes read=0
                FILE: Number of bytes written=445734
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=922
                HDFS: Number of bytes written=1028381690
                HDFS: Number of read operations=22
```

```
                S3: Number of bytes written=0
                S3: Number of read operations=0
                S3: Number of large read operations=0
                S3: Number of write operations=0
        Job Counters
                Launched map tasks=2
                Other local map tasks=2
                Total time spent by all maps in occupied slots (ms)=2131520
                Total time spent by all reduces in occupied slots (ms)=0
                Total time spent by all map tasks (ms)=66610
                Total vcore-milliseconds taken by all map tasks=66610
                Total megabyte-milliseconds taken by all map tasks=68208640
        Map-Reduce Framework
                Map input records=2
                Map output records=0
                Input split bytes=272
                Spilled Records=0
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=1354
                CPU time spent (ms)=43800
                Physical memory (bytes) snapshot=1258647552
                Virtual memory (bytes) snapshot=6672437248
                Total committed heap usage (bytes)=1000341504
        File Input Format Counters
                Bytes Read=650
        File Output Format Counters
                Bytes Written=0
        DistCp Counters
                Bytes Copied=1028381690
                Bytes Expected=1028381690
                Files Copied=2
[hadoop@ip-172-31-41-129 ~]$ hadoop fs -ls /user/hive/casestudy/*
-rw-r--r--   1 hadoop hadoop  545839412 2021-02-05 04:17 /user/hive/casestudy/2019-Nov.csv
-rw-r--r--   1 hadoop hadoop  482542278 2021-02-05 04:17 /user/hive/casestudy/2019-Oct.csv
[hadoop@ip-172-31-41-129 ~]$ hadoop fs -cat /user/hive/casestudy/2019* | head -5
event_time,event_type,product_id,category_id,category_code,brand,price,user_id,user_session
2019-11-01 00:00:02 UTC,view,5802432,1487580009286598681,,,0.32,562076640,09fafd6c-6c99-46b1-834f-33527f4de241
2019-11-01 00:00:09 UTC,cart,5844397,1487580006317032337,,,2.38,553329724,2067216c-31b5-455d-a1cc-af0575a34ffb
2019-11-01 00:00:10 UTC,view,5837166,1783999064103190764,,pnb,22.22,556138645,57ed222e-a54a-4907-9944-5a875c2d7f4f
2019-11-01 00:00:11 UTC,cart,5876812,1487580010100293687,,jessnail,3.16,564506666,186c1951-8052-4b37-adce-dd9644b1d5f7
cat: Unable to write to output stream.
cat: Unable to write to output stream.
[hadoop@ip-172-31-41-129 ~]$
```

Creating the database and launching Hive queries on your EMR cluster:

Creating the structure of database

```
hive> show databases;
OK
default
Time taken: 0.703 seconds, Fetched: 1 row(s)
hive> create database casestudy;
OK
Time taken: 0.324 seconds
hive> show databases;
OK
casestudy
default
Time taken: 0.058 seconds, Fetched: 2 row(s)
hive> use casestudy;
OK
Time taken: 0.05 seconds
hive> describe database extended casestudy;
OK
casestudy               hdfs://ip-172-31-35-204.ec2.internal:8020/user/hive/warehouse/casestudy.db      hadoop  USER
Time taken: 0.037 seconds, Fetched: 1 row(s)
```

```
hive> CREATE EXTERNAL TABLE IF NOT EXISTS ecommerce_data(event_time timestamp, event_type string, product_id string,
    > category_id string, category_code string, brand string, price float, user_id bigint, user_session string)
    > ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.OpenCSVSerde'
    > STORED AS TEXTFILE
    > LOCATION '/user/hive/casestudy/'
    > tblproperties('skip.header.line.count'='1');
OK
Time taken: 0.943 seconds
hive> describe extended ecommerce_data;
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                   string                  from deserializer
user_id                 string                  from deserializer
user_session            string                  from deserializer

Detailed Table Information      Table(tableName:ecommerce_data, dbName:casestudy, owner:hadoop, createTime:1612791153, lastAccessTime:0, retention:0, sd:StorageDescript
or(cols:[FieldSchema(name:event_time, type:timestamp, comment:null), FieldSchema(name:event_type, type:string, comment:null), FieldSchema(name:product_id, type:string,
comment:null), FieldSchema(name:category_id, type:string, comment:null), FieldSchema(name:category_code, type:string, comment:null), FieldSchema(name:brand, type:string
, comment:null), FieldSchema(name:price, type:float, comment:null), FieldSchema(name:user_id, type:bigint, comment:null), FieldSchema(name:user_session, type:string, co
mment:null)], location:hdfs://ip-172-31-35-204.ec2.internal:8020/user/hive/casestudy, inputFormat:org.apache.hadoop.mapred.TextInputFormat, outputFormat:org.apache.hado
op.hive.ql.io.HiveIgnoreKeyTextOutputFormat, compressed:false, numBuckets:-1, serdeInfo:SerDeInfo(name:null, serializationLib:org.apache.hadoop.hive.serde2.OpenCSVSerde
, parameters:{serialization.format=1}), bucketCols:[], sortCols:[], parameters:{}, skewedInfo:SkewedInfo(skewedColNames:[], skewedColValues:[], skewedColValueLocationMa
ps:{}), storedAsSubDirectories:false), partitionKeys:[], parameters:{skip.header.line.count=1, transient_lastDdlTime=1612791153, totalSize=1028381690, EXTERNAL=TRUE, nu
mFiles=2}, viewOriginalText:null, viewExpandedText:null, tableType:EXTERNAL_TABLE, rewriteEnabled:false)
Time taken: 0.134 seconds, Fetched: 11 row(s)
hive> SELECT * FROM ecommerce_data 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: 2.248 seconds, Fetched: 5 row(s)
```

```
hive> set hive.exec.dynamic.partition.mode=nonstrict;
hive> set hive.exec.dynamic.partition=true;
hive> set hive.enforce.bucketing=true;
hive>
    > CREATE TABLE IF NOT EXISTS ecommerce_table1(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 50 buckets
    > ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.OpenCSVSerde'
    > STORED AS TEXTFILE;
OK
Time taken: 0.258 seconds
hive> insert into table ecommerce_table1 partition (event_type)
    > select
    > from_unixtime(unix_timestamp(event_time , 'yyyy-MM-dd HH:mm:ss')) as event_time,
    > product_id,
    > category_id,
    > category_code,
    > brand,
    > CAST(price AS float) as price,
    > CAST(user_id AS bigint) as user_id,
    > user_session,
    > event_type
    > from ecommerce_data;
Query ID = hadoop_20210208143240_1f0aed75-6db2-472d-87ef-f2ee35e95876
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1612790815366_0005)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      4          4        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      9          9        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 420.74 s
----------------------------------------------------------------------------------------------
Loading data to table casestudy.ecommerce_table1 partition (event_type=null)

Loaded : 4/4 partitions.
        Time taken to load dynamic partitions: 0.777 seconds
        Time taken for adding to write entity : 0.006 seconds
OK
Time taken: 427.459 seconds
```

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

```
hive> select month(event_time) ,sum(price)  from ecommerce_data
    > where month(event_time)= 10 and event_type = 'purchase' group by month(event_time);
Query ID = hadoop_20210208144614_6959bc6e-b932-4da2-b389-0292b58f28a4
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1612790815366_0006)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      2          2        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      6          6        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 70.44 s
----------------------------------------------------------------------------------------------
OK
10      1211538.4299997438
Time taken: 75.477 seconds, Fetched: 1 row(s)
```

```
hive> select month(event_time) ,sum(price) from ecommerce_table1
    > where month(event_time)= 10 and event_type = 'purchase' group by month(event_time);
Query ID = hadoop_20210208145652_c824b9f8-888d-46ae-abbd-d811eca8aeb1
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1612790815366_0006)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      6          6        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      2          2        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 25.47 s
----------------------------------------------------------------------------------------------
OK
10      1211538.430000022
Time taken: 26.186 seconds, Fetched: 1 row(s)
```

Running Hive queries to answer the questions given below.

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


```
hive> select month(event_time) ,sum(price) from ecommerce_table1
    > where month(event_time)= 10 and event_type = 'purchase' group by month(event_time);
Query ID = hadoop_20210208145652_c824b9f8-888d-46ae-abbd-d811eca8aeb1
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1612790815366_0006)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      6          6        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      2          2        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 25.47 s
----------------------------------------------------------------------------------------------
OK
10      1211538.430000022
Time taken: 26.186 seconds, Fetched: 1 row(s)
```


Finding the total sum of purchases per month in a single output.

```
hive> select month(event_time) ,sum(price)  from ecommerce_table1 where event_type = 'purchase' group by month(event_time);
Query ID = hadoop_20210207063430_1656da38-c8c8-42bd-bffa-4234b373846f
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1612674840512_0004)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      5          5        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      2          2        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 21.64 s
--------------------------------------------------------------------------------
OK
10      1211166.80953072
11      1531016.8991247676
Time taken: 22.245 seconds, Fetched: 2 row(s)
```

Finding the change in revenue generated due to purchases from October to November.

```
hive> set hive.strict.checks.cartesian.product=false;
hive> select o.Osale ,n.Nsale,n.Nsale - o.Osale from
    > (Select sum(price) as Osale from ecommerce_table1 where month(event_time) = 10 and month(event_time) is not NULL and event_type = 'purchase' group by month(event_time)) o
    > cross join
    > (Select sum(price) as Nsale from ecommerce_table1 where month(event_time) = 11 and month(event_time) is not NULL and event_type = 'purchase' group by month(event_time)) n;
Warning: Map Join MAPJOIN[23][bigTable=?] in task 'Reducer 2' is a cross product
Query ID = hadoop_20210207094916_746d1282-ca80-46e9-b09a-6c7f17a8f6c4
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1612674840512_0009)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      5          5        0        0       0       0
Map 3 .......... container     SUCCEEDED      5          5        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      2          2        0        0       0       0
Reducer 4 ...... container     SUCCEEDED      2          2        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 04/04  [==========================>>] 100%  ELAPSED TIME: 23.92 s
--------------------------------------------------------------------------------
OK
1211166.80953072        1531016.8991247676      319850.0895940475
Time taken: 24.957 seconds, Fetched: 1 row(s)
```

Finding distinct categories of products ignoring the categories with null category code.

```
hive> select distinct(category_code) from ecommerce_table1 where category_code is not NULL;
Query ID = hadoop_20210208154955_fb0f5c1a-cadd-48d5-9d1d-0c734d25f737
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1612798691960_0003)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      6          6        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 64.93 s
--------------------------------------------------------------------------------
OK

accessories.bag
accessories.cosmetic_bag
apparel.glove
appliances.environment.air_conditioner
appliances.environment.vacuum
appliances.personal.hair_cutter
furniture.bathroom.bath
furniture.living_room.cabinet
furniture.living_room.chair
sport.diving
stationery.cartrige
Time taken: 65.806 seconds, Fetched: 12 row(s)
```

Finding the total number of products available under each category.

```
hive> Select category_code,count(distinct product_id) from ecommerce_table1
    > where category_code is not NULL
    > Group by category_code;
Query ID = hadoop_20210208155104_78c06e46-1b11-40ee-b1f8-a8637422f657
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1612798691960_0003)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      6          6        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 57.20 s
--------------------------------------------------------------------------------
OK
        45502
accessories.bag 42
accessories.cosmetic_bag        16
apparel.glove   78
appliances.environment.air_conditioner  26
appliances.environment.vacuum   85
appliances.personal.hair_cutter 9
furniture.bathroom.bath 55
furniture.living_room.cabinet   6
furniture.living_room.chair     2
sport.diving    1
stationery.cartrige     138
Time taken: 58.041 seconds, Fetched: 12 row(s)
```

Finding the brand that had the maximum sales in October and November combined.

```
hive> select brand,sum(price) as sales
    > from ecommerce_table1 where event_type='purchase' and brand <> '' and brand is not NULL
    > Group by brand
    > Order by sales desc
    > Limit 1;
Query ID = hadoop_20210208145732_2a7e8911-6e9b-43c6-86e1-1dc2ee848c64
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1612790815366_0006)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      6          6        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      2          2        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 25.57 s
--------------------------------------------------------------------------------
OK
runail  148297.93999999904
Time taken: 26.329 seconds, Fetched: 1 row(s)
```

Finding the brands that increased their sales from October to November.

```
hive> select o.brand as brand,o.Osale as OctSales,n.Nsale as NovSales,n.Nsale - o.Osale as Sales_diff from
    > (Select brand,sum(price) as Osale
    > from ecommerce_table1
    > 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 Nsale
    > from ecommerce_table1
    > 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.Nsale - o.Osale > 0;
Query ID = hadoop_20210208145802_6df44905-5877-494c-8620-8717308797a1
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1612790815366_0006)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      6          6        0        0       0       0
Map 3 .......... container     SUCCEEDED      6          6        0        0       0       1
Reducer 2 ...... container     SUCCEEDED      2          2        0        0       0       0
Reducer 4 ...... container     SUCCEEDED      2          2        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 04/04  [==========================>>] 100%  ELAPSED TIME: 21.04 s
--------------------------------------------------------------------------------
OK
art-visage      2092.7099999999996      2997.8000000000001      905.0900000000015
artex   2730.6399999999994      4327.249999999998       1596.6099999999988
batiste 772.4   874.1699999999998       101.76999999999987
beautix 10493.949999999993      12222.95        1729.0000000000073
beautyblender   78.74000000000001       109.41  30.669999999999987
bioaqua 942.8899999999999       1398.1200000000003      455.2300000000005
biore   60.65   90.30999999999999       29.65999999999999
blixz   38.95   63.400000000000006      24.450000000000003
browxenna       14331.369999999999      14916.73        585.3600000000006
carmex  145.07999999999998      243.36  98.28000000000003
concept 11032.140000000014      13380.400000000012      2348.2599999999984
cutrin  299.37  367.62  68.25
deoproce        316.84000000000003      329.16999999999996      12.329999999999927
domix   10472.050000000016      12009.170000000013      1537.1199999999972
ecocraft        41.160000000000004      241.95000000000002      200.79000000000002
ecolab  262.85  1214.3000000000002       951.4500000000002
```

```
egomania        77.47   146.04  68.57
ellips  245.85  606.04  360.18999999999994
elskin  251.08999999999997      307.6499999999999       56.559999999999945
entity  479.7100000000005       719.2599999999998       239.54999999999927
eos     54.339999999999999      152.61  98.27000000000002
f.o.x   6624.229999999999       8577.279999999992       1953.049999999993
farmavita       837.37  1291.97 454.6
fedua   52.38   263.81  211.43
fly     17.14   27.169999999999998      10.029999999999998
freshbubble     318.70000000000005      502.34  183.63999999999993
gehwol  1089.0700000000002      1557.6799999999998      468.6099999999997
glysolid        69.73   91.59   21.86
greymy  29.21   489.49  460.28000000000003
happyfons       801.9200000000003       1091.5900000000004      289.6700000000001
haruyama        9390.690000000005       12352.909999999983      2962.219999999934
inm     288.02  351.21000000000004      63.190000000000055
insight 1443.6999999999998      1721.96 278.2600000000002
jaguar  1102.11 1110.65 8.540000000000191
joico   705.52  2015.1  1309.58
kaaral  4412.43 5086.069999999999       673.6399999999985
kamill  63.01   81.49000000000001       18.48000000000001
kaypro  881.3399999999999       3268.7  2387.3599999999997
keen    236.34999999999997      435.61999999999995      199.26999999999998
konad   739.8299999999999       810.6700000000001       70.84000000000015
laboratorium    246.5   312.52  66.01999999999998
levissime       2227.4999999999977      3085.3100000000013      857.8100000000036
lianail 5892.839999999998       16394.23999999994       10501.399999999941
likato  296.06  340.96999999999997      44.90999999999997
limoni  1308.9  1796.6  487.6999999999998
lovely  8704.379999999994       11939.060000000001      3234.6800000000076
mane    66.78999999999999       260.26  193.47
marathon        7280.75 10273.1 2992.3500000000004
markell 1768.7499999999999      2834.429999999999       1065.6799999999998
masura  31266.07999999908       33058.46999999917       1792.3900000000904
mavala  409.04  446.32000000000005      37.28000000000003
milv    3904.9399999999914      5642.009999999988       1737.069999999997
miskin  158.04000000000002      293.07000000000005      135.03000000000003
missha  1293.8300000000002      2150.28 856.45
moyou   5.71    10.280000000000001      4.570000000000001
nagaraku        4369.739999999996       5327.679999999993       957.9399999999969
nefertiti       233.52000000000004      366.64  133.11999999999995
nirvel  163.04000000000002      234.32999999999998      71.28999999999996
nitrile 847.28  1162.6800000000003      315.4000000000003
orly    902.3799999999999       931.09  28.710000000000015
```

```
osmo    645.5800000000002       762.31  116.72999999999979
ovale   2.54    3.1     0.56
plazan  101.37  194.01  92.63999999999999
profhenna       679.2299999999999       736.8499999999998       57.61999999999989
protokeratin    201.25  456.79  255.54000000000002
provoc  827.9899999999997       1063.8199999999997      235.83000000000004
runail  71539.28000000057       76758.66000000067       5219.3800000001065
shik    3341.2000000000007      4839.72 1498.5199999999995
skinlite        651.9400000000003       890.4500000000003       238.51
smart   4457.260000000002       5902.14 1444.8799999999983
soleo   204.20000000000001      212.53000000000011      8.330000000000013
sophin  1067.8600000000001      1515.52 447.65999999999985
strong  29196.629999999997      38671.26999999999       9474.639999999992
trind   298.07  542.96  244.89000000000004
uno     35302.02999999982       51039.74999999995       15737.720000000132
uskusi  5142.27000000001        5690.310000000007       548.0399999999963
yoko    8756.909999999993       11707.879999999997      2950.970000000005
airnails        5118.900000000005       5691.520000000006       572.6200000000008
aura    83.95   177.51  93.55999999999999
balbcare        155.33000000000004      212.38000000000002      57.04999999999998
beauty-free     554.1700000000001       1782.8599999999988      1228.6899999999987
beauugreen      511.51000000000005      768.3500000000001       256.8400000000001
benovy  409.62  3259.970000000001       2850.3500000000013
bluesky 10307.240000000029      10565.530000000006      258.2899999999772
bodyton 1376.3400000000001      1380.6400000000003      4.300000000000182
bpw.style       11572.149999999874      14837.439999999862      3265.289999999988
candy   534.96  799.38  264.41999999999996
chi     358.94  538.6099999999999       179.6699999999999
coifin  903.0   1428.4900000000002      525.4900000000002
cosima  20.229999999999997      20.93   0.7000000000000028
cosmoprofi      8322.81 14536.989999999998      6214.1799999999985
cristalinas     427.63  584.95  157.32000000000005
de.lux  1659.7000000000001      2775.5100000000001      1115.8100000000002
depilflax       2707.0699999999993      2803.7799999999997      96.71000000000049
dizao   819.13  945.5100000000009       126.3800000000009
elizavecca      70.53   204.3   133.77
enjoy   41.35   136.57000000000002      95.22000000000003
estel   21756.749999999985      24142.669999999976      2385.919999999991
estelare        444.81000000000006      471.87  27.059999999999945
farmona 1692.4599999999996      1843.4299999999998      150.97000000000025
finish  98.38   230.38  132.0
foamie  35.04   80.49   45.449999999999996
freedecor       3421.7799999999947      7671.799999999997       4250.020000000002
godefroy        401.21999999999997      425.11999999999995      23.899999999999977
grace   100.92  102.61  1.6899999999999977
grattol 35445.53999999989       71472.70999999961       36027.16999999972
igrobeauty      513.6599999999999       645.07  131.4100000000002
ingarden        23161.390000000014      33566.209999999875      10404.819999999861
irisk   45591.95999999925       46946.03999999975       1354.0800000005038
italwax 21940.23999999997       24799.369999999948      2859.129999999979
jas     3318.96 3657.430000000001       338.47000000000116
jessnail        26287.839999999942      33345.23000000004       7057.390000000098
kapous  11927.159999999974      14093.079999999973      2165.9199999999983
kerasys 430.91000000000001      525.1999999999999       94.28999999999985
kims    330.03999999999996      632.04  302.0
kinetics        6334.250000000003       6945.259999999995       611.009999999992
kiss    421.55000000000007      817.3300000000002       395.7800000000001
kocostar        310.85  594.93  284.0799999999999
koelcia 55.5    112.75  57.25
koelf   422.73  507.28999999999996      84.55999999999995
kosmekka        1181.44 1813.37 631.9299999999998
lador   2083.6100000000015      2471.530000000001       387.9199999999996
ladykin 125.65  170.57  44.91999999999999
latinoil        249.52000000000004      384.59000000000003      135.07
levrana 2243.5600000000004      3664.099999999999       1420.5399999999986
lowence 242.84  567.75  324.90999999999997
marutaka-foot   49.22   109.33000000000001      60.110000000000014
matrix  3243.25 3726.7400000000002      483.49000000000024
metzger 5373.450000000002       6457.16 1083.7099999999982
neoleor 43.41   51.7    8.290000000000006
oniq    8425.410000000001       9841.650000000005       1416.2399999999943
polarus 6013.719999999999       11371.929999999998      5358.209999999999
profepil        93.36000000000001       118.02000000000001      24.659999999999997
rasyan  18.799999999999997      28.939999999999998      10.14
refectocil      2716.1799999999994      3475.5800000000017      759.4000000000024
rosi    3077.04 3841.5599999999986      764.5199999999986
roubloff        3491.3599999999983      4913.769999999997       1422.4099999999985
s.care  412.68  913.07  500.39000000000004
sanoto  157.14  1209.6799999999998      1052.54
severina        4775.879999999993       6120.479999999992       1344.5999999999995
shary   871.9600000000003       1176.49 304.52999999999975
skinity 8.88    12.440000000000001      3.5600000000000005
solomeya        1899.6999999999996      2685.7999999999997      786.1000000000001
staleks 8519.730000000005       11875.61        3355.8799999999956
supertan        50.36999999999999       66.50999999999999       16.14
swarovski       1887.9300000000007      3043.159999999998       1155.2299999999973
tertio  236.16000000000003      245.8   9.639999999999986
treaclemoon     163.37  181.49  18.120000000000005
veraclara       50.11   71.21000000000001       21.10000000000001
vilenta 197.6   231.20999999999998      33.60999999999985
yu-r    271.40999999999997      673.7099999999999       402.29999999999995
zeitun  708.66  2009.6299999999999       1300.9699999999998
Time taken: 21.927 seconds, Fetched: 152 row(s)
```

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

```
hive> SELECT user_id, sum(price)as sales from ecommerce_table1
    > where user_id is not NULL and event_type = 'purchase'
    > Group by user_id
    > Order by sales Desc
    > LIMIT 10;
Query ID = hadoop_20210208145851_b889967b-d279-4b70-9d21-23ee1b0e934f
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1612790815366_0006)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 ..........   container     SUCCEEDED      6          6        0        0       0       0
Reducer 2 ......   container     SUCCEEDED      2          2        0        0       0       0
Reducer 3 ......   container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 27.98 s
--------------------------------------------------------------------------------
OK
557790271       2715.8699999999944
150318419       1645.9699999999998
562167663       1352.85
531900924       1329.4499999999998
557850743       1295.48
522130011       1185.3900000000003
561592095       1109.7
431950134       1097.5899999999995
566576008       1056.3600000000015
521347209       1040.9099999999999
Time taken: 28.639 seconds, Fetched: 10 row(s)
```

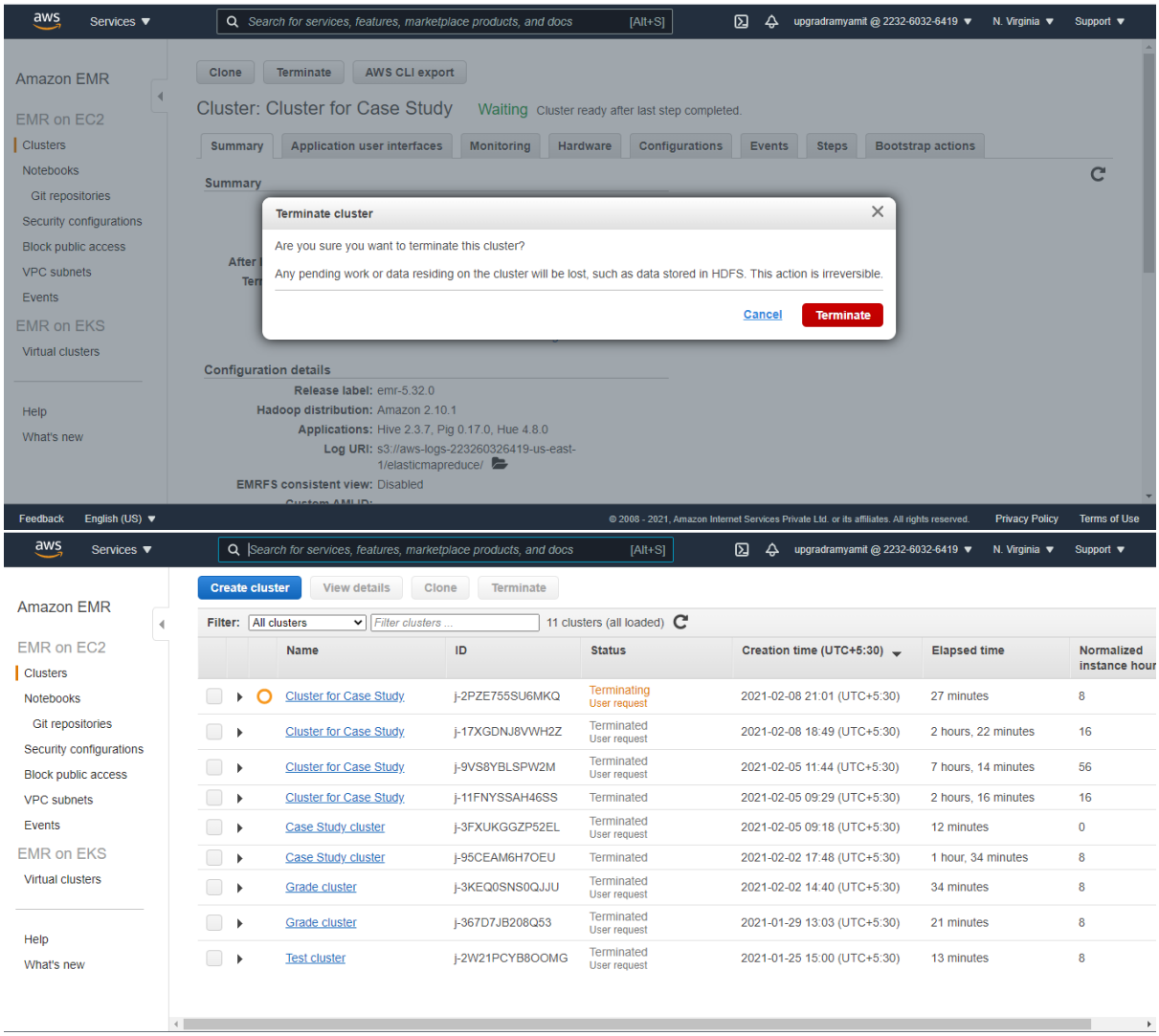
Cleaning up

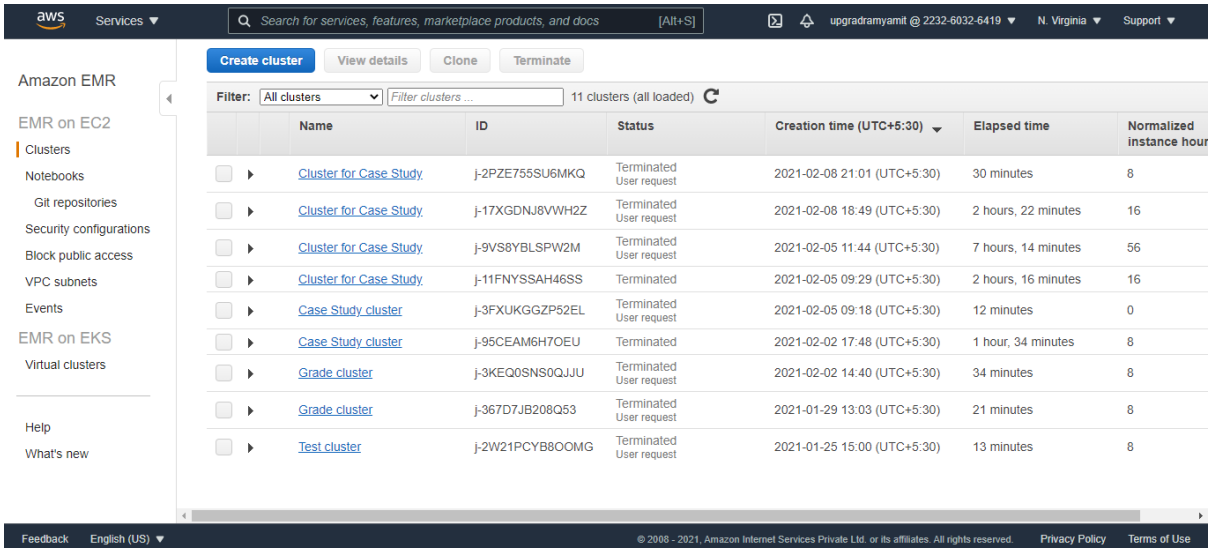
Dropping the database

```
hive> drop table ecommerce_table1;
OK
Time taken: 0.258 seconds
hive> drop table ecommerce_data;
OK
Time taken: 0.357 seconds
hive> drop database casestudy;
OK
Time taken: 0.075 seconds
```

```
[hadoop@ip-172-31-47-49 ~]$ hadoop fs -ls /user/hive/casestudy/*
-rw-r--r--   1 hadoop hadoop  545839412 2021-02-08 15:41 /user/hive/casestudy/2019-Nov.csv
-rw-r--r--   1 hadoop hadoop  482542278 2021-02-08 15:41 /user/hive/casestudy/2019-Oct.csv
[hadoop@ip-172-31-47-49 ~]$ hadoop fs -rm -r -f /user/hive/casestudy
Deleted /user/hive/casestudy
[hadoop@ip-172-31-47-49 ~]$ hadoop fs -ls /user/hive/casestudy/*
ls: `/user/hive/casestudy/*': No such file or directory
```

Terminating the cluster




aws
Services
Search for services, features, marketplace products, and docs [Alt+S]
upgradramyamit @ 2232-6032-6419
N. Virginia
Support
Amazon EMR
EMR on EC2
Clusters
Notebooks
Git repositories
Security configurations
Block public access
VPC subnets
Events
EMR on EKS
Virtual clusters
Help
What's new
Create cluster
View details
Clone
Terminate
Filter: All clusters
Filter clusters ...
11 clusters (all loaded)
Name
ID
Status
Creation time (UTC+5:30)
Elapsed time
Normalized instance hour
Cluster for Case Study j-2PZE755SU6MKQ Terminated User request 2021-02-08 21:01 (UTC+5:30) 30 minutes 8
Cluster for Case Study j-17XGDNJ8VWH2Z Terminated User request 2021-02-08 18:49 (UTC+5:30) 2 hours, 22 minutes 16
Cluster for Case Study j-9VS8YBLSPW2M Terminated User request 2021-02-05 11:44 (UTC+5:30) 7 hours, 14 minutes 56
Cluster for Case Study j-11FNYSSAH46SS Terminated 2021-02-05 09:29 (UTC+5:30) 2 hours, 16 minutes 16
Case Study cluster j-3FXUKGGZP52EL Terminated User request 2021-02-05 09:18 (UTC+5:30) 12 minutes 0
Case Study cluster j-95CEAM6H7OEU Terminated 2021-02-02 17:48 (UTC+5:30) 1 hour, 34 minutes 8
Grade cluster j-3KEQ0SNS0QJJU Terminated User request 2021-02-02 14:40 (UTC+5:30) 34 minutes 8
Grade cluster j-367D7JB208Q53 Terminated User request 2021-01-29 13:03 (UTC+5:30) 21 minutes 8
Test cluster j-2W21PCYB8OOMG Terminated User request 2021-01-25 15:00 (UTC+5:30) 13 minutes 8
Feedback
English (US)
© 2008 - 2021, Amazon Internet Services Private Ltd. or its affiliates. All rights reserved.
Privacy Policy
Terms of Use