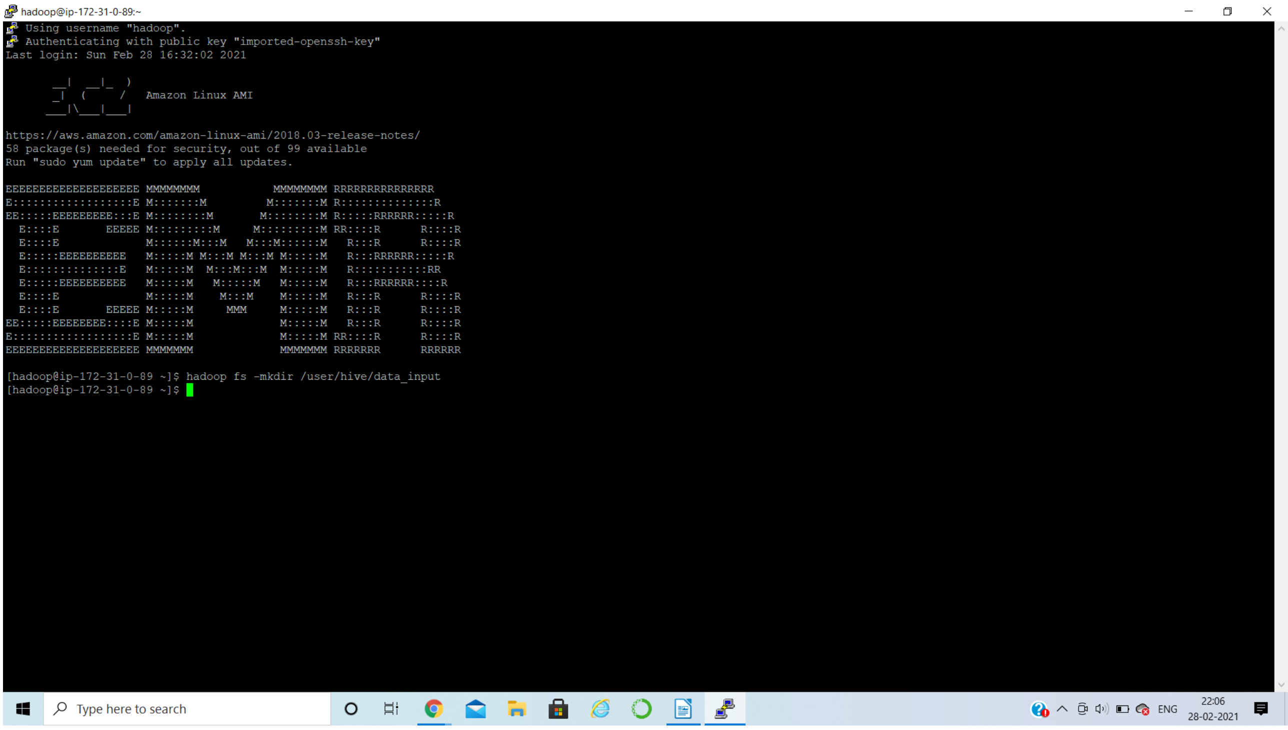
Assignment : HIVE CASE STUDY
Submitted By: Kanika Kathpalia & Gurpreet Kaur

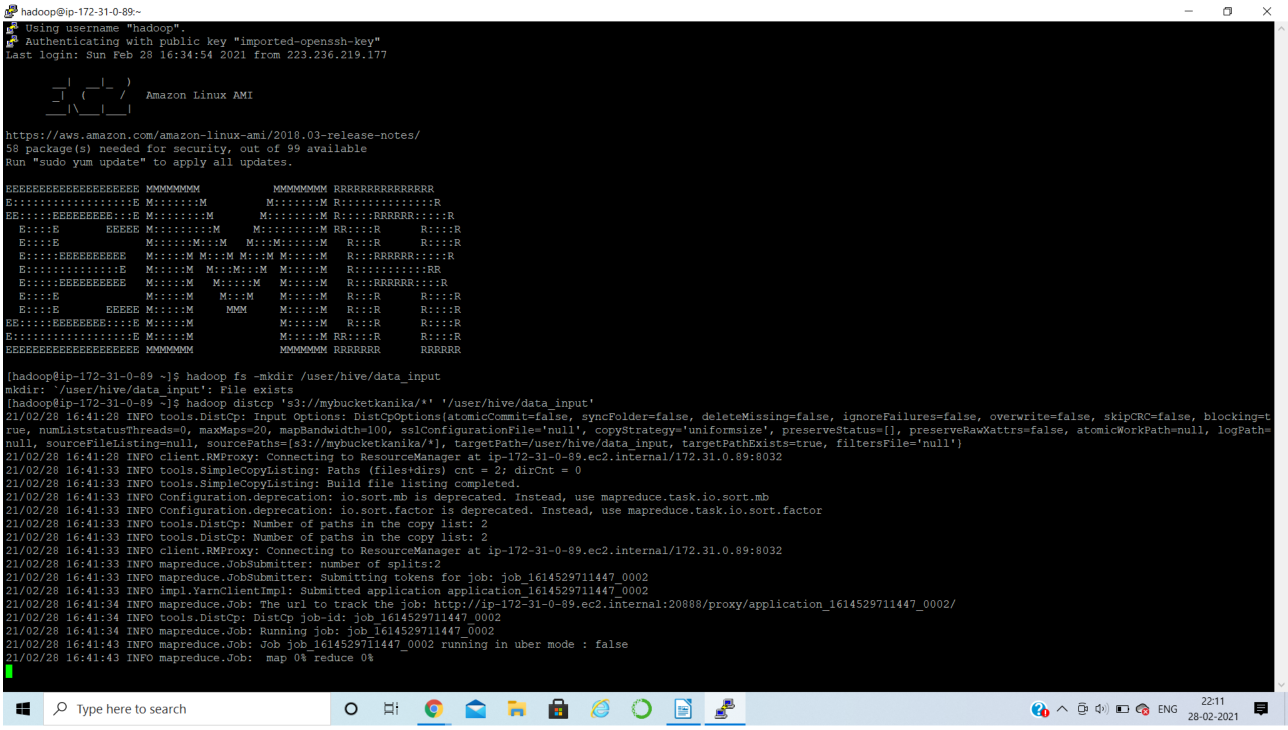
Step by Step Explanation with solved questions snapshots:

Creating a directory in HDFS to collect the input data:
hadoop fs -mkdir /user/hive/data_input

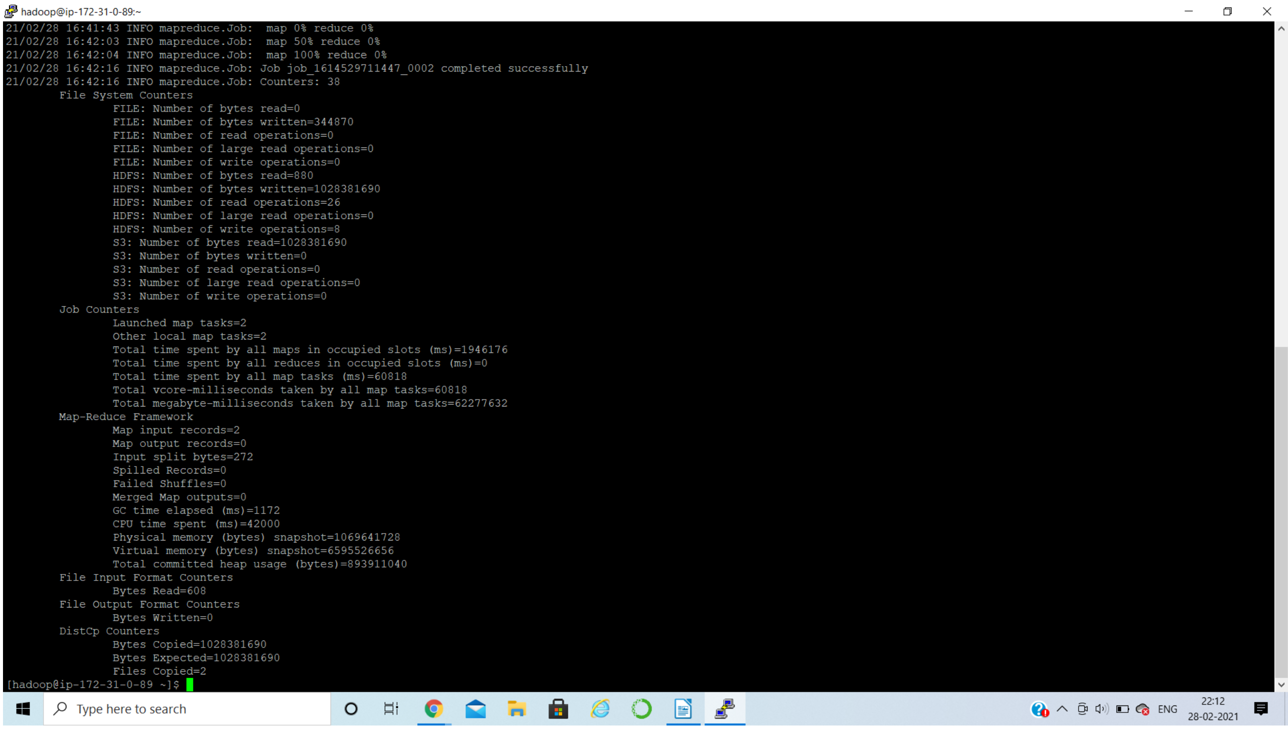


```
Using username "hadoop".
Authenticating with public key "imported-openssh-key"
Last login: Sun Feb 28 16:32:02 2021

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

https://aws.amazon.com/amazon-linux-ami/2018.03-release-notes/
58 package(s) needed for security, out of 99 available
Run "sudo yum update" to apply all updates.

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

[hadoop@ip-172-31-0-89 ~]$ hadoop fs -mkdir /user/hive/data_input
[hadoop@ip-172-31-0-89 ~]$
```

Copy data from S3 to HDFS:
hadoop distcp 's3://mybucketkanika/*' '/user/hive/data_input'



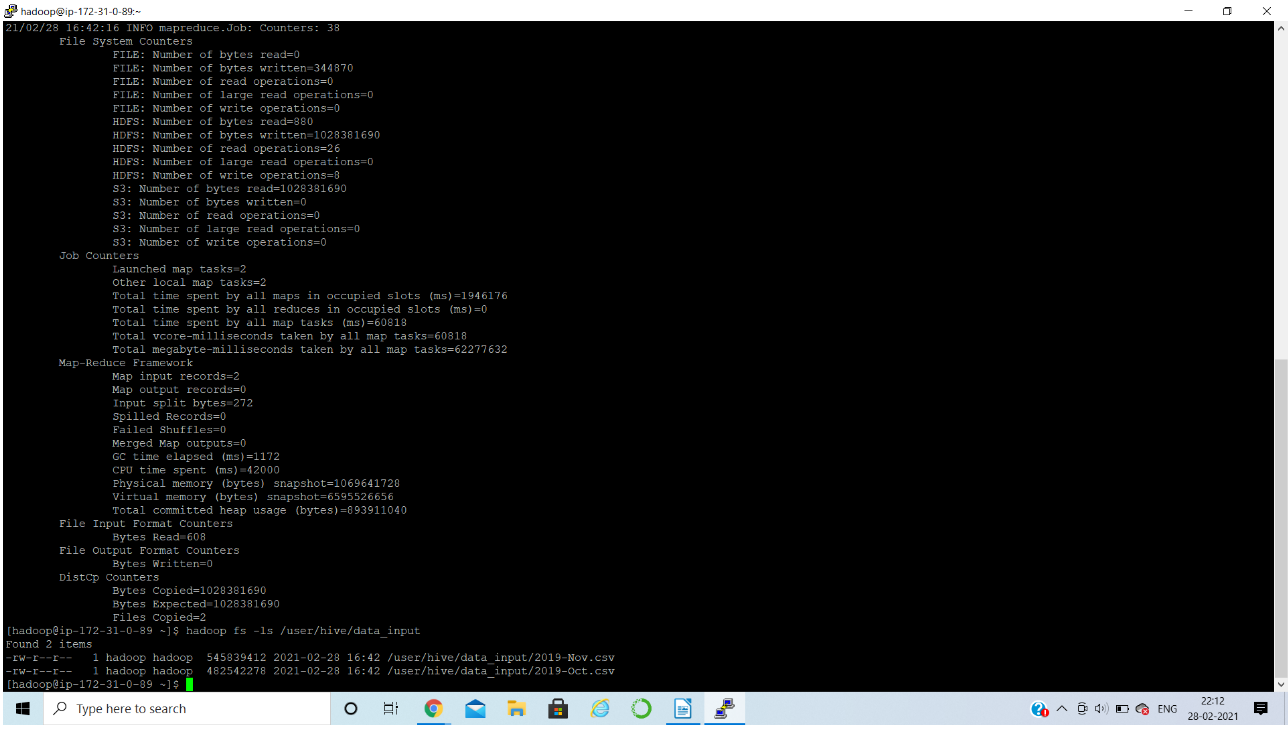
```
Using username "hadoop".
Authenticating with public key "imported-openssh-key"
Last login: Sun Feb 28 16:34:54 2021 from 223.236.219.177

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

https://aws.amazon.com/amazon-linux-ami/2018.03-release-notes/
58 package(s) needed for security, out of 99 available
Run "sudo yum update" to apply all updates.

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

[hadoop@ip-172-31-0-89 ~]$ hadoop fs -mkdir /user/hive/data_input
mkdir: `/user/hive/data_input': File exists
[hadoop@ip-172-31-0-89 ~]$ hadoop distcp 's3://mybucketkanika/*' '/user/hive/data_input'
21/02/28 16:41:28 INFO tools.DistCp: Input Options: DistCpOptions{atomicCommit=false, syncFolder=false, deleteMissing=false, ignoreFailures=false, overwrite=false, skipCRC=false, blocking=t
rue, numListstatusThreads=0, maxMaps=20, mapBandwidth=100, sslConfigurationFile='null', copyStrategy='uniformsize', preserveStatus=[], preserveRawXattrs=false, atomicWorkPath=null, logPath=
null, sourceFileListing=null, sourcePaths=[s3://mybucketkanika/*], targetPath=/user/hive/data_input, targetPathExists=true, filtersFile='null'}
21/02/28 16:41:28 INFO client.RMProxy: Connecting to ResourceManager at ip-172-31-0-89.ec2.internal/172.31.0.89:8032
21/02/28 16:41:33 INFO tools.SimpleCopyListing: Paths (files+dirs) cnt = 2; dirCnt = 0
21/02/28 16:41:33 INFO tools.SimpleCopyListing: Build file listing completed.
21/02/28 16:41:33 INFO Configuration.deprecation: io.sort.mb is deprecated. Instead, use mapreduce.task.io.sort.mb
21/02/28 16:41:33 INFO Configuration.deprecation: io.sort.factor is deprecated. Instead, use mapreduce.task.io.sort.factor
21/02/28 16:41:33 INFO tools.DistCp: Number of paths in the copy list: 2
21/02/28 16:41:33 INFO tools.DistCp: Number of paths in the copy list: 2
21/02/28 16:41:33 INFO client.RMProxy: Connecting to ResourceManager at ip-172-31-0-89.ec2.internal/172.31.0.89:8032
21/02/28 16:41:33 INFO mapreduce.JobSubmitter: number of splits:2
21/02/28 16:41:33 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1614529711447_0002
21/02/28 16:41:33 INFO impl.YarnClientImpl: Submitted application application_1614529711447_0002
21/02/28 16:41:34 INFO mapreduce.Job: The url to track the job: http://ip-172-31-0-89.ec2.internal:20888/proxy/application_1614529711447_0002/
21/02/28 16:41:34 INFO tools.DistCp: DistCp job-id: job_1614529711447_0002
21/02/28 16:41:34 INFO mapreduce.Job: Running job: job_1614529711447_0002
21/02/28 16:41:43 INFO mapreduce.Job: Job job_1614529711447_0002 running in uber mode : false
21/02/28 16:41:43 INFO mapreduce.Job:  map 0% reduce 0%
```



```
21/02/28 16:41:43 INFO mapreduce.Job:  map 0% reduce 0%
21/02/28 16:42:03 INFO mapreduce.Job:  map 50% reduce 0%
21/02/28 16:42:04 INFO mapreduce.Job:  map 100% reduce 0%
21/02/28 16:42:16 INFO mapreduce.Job: Job job_1614529711447_0002 completed successfully
21/02/28 16:42:16 INFO mapreduce.Job: Counters: 38
        File System Counters
                FILE: Number of bytes read=0
                FILE: Number of bytes written=344870
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=880
                HDFS: Number of bytes written=1028381690
                HDFS: Number of read operations=26
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=8
                S3: Number of bytes read=1028381690
                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)=1946176
                Total time spent by all reduces in occupied slots (ms)=0
                Total time spent by all map tasks (ms)=60818
                Total vcore-milliseconds taken by all map tasks=60818
                Total megabyte-milliseconds taken by all map tasks=62277632
        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)=1172
                CPU time spent (ms)=42000
                Physical memory (bytes) snapshot=1069641728
                Virtual memory (bytes) snapshot=6595526656
                Total committed heap usage (bytes)=893911040
        File Input Format Counters
                Bytes Read=608
        File Output Format Counters
                Bytes Written=0
        DistCp Counters
                Bytes Copied=1028381690
                Bytes Expected=1028381690
                Files Copied=2
[hadoop@ip-172-31-0-89 ~]$
```

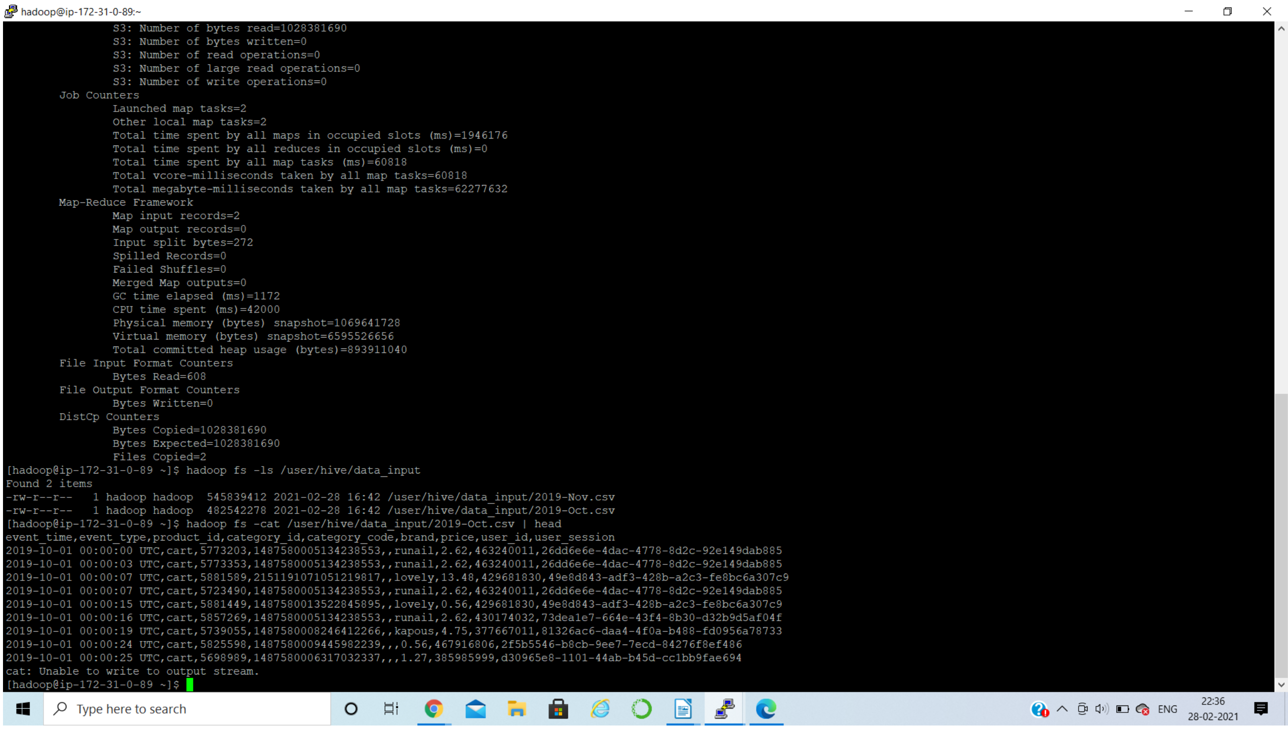
Checking that the data was copied:

hadoop fs -ls /user/hive/data_input



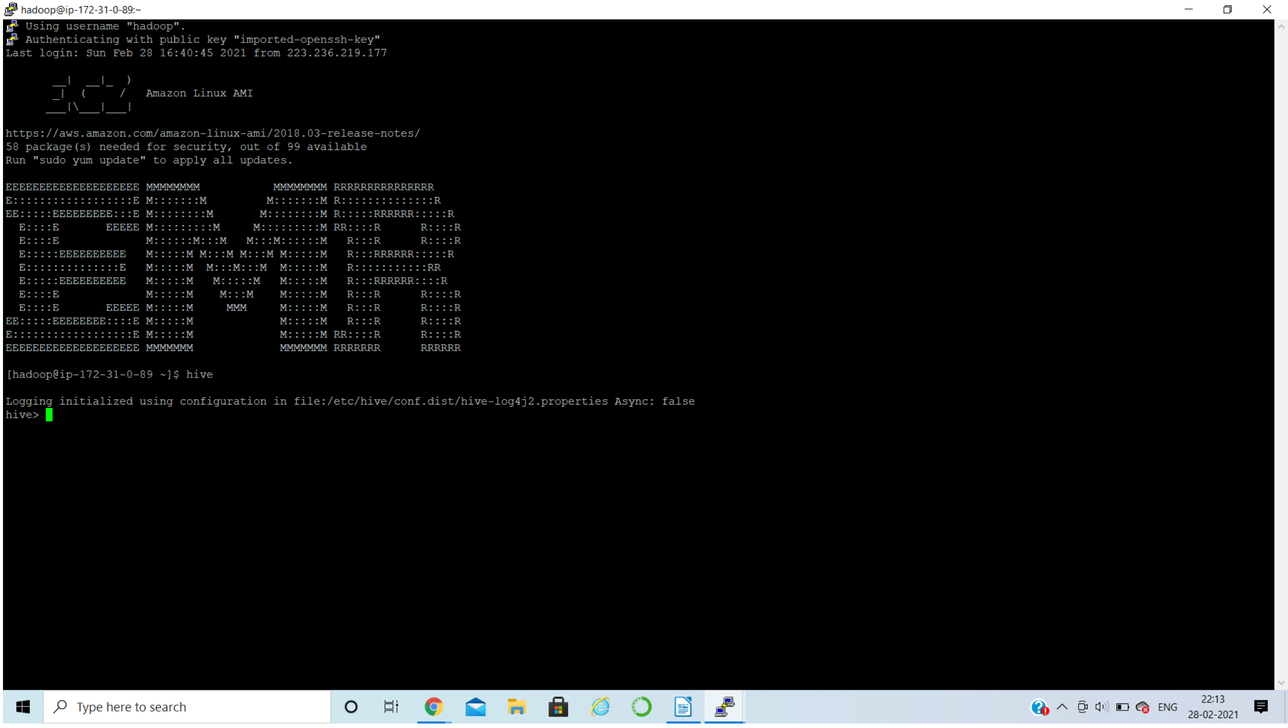
```
21/02/28 16:42:16 INFO mapreduce.Job: Counters: 38
        File System Counters
                FILE: Number of bytes read=0
                FILE: Number of bytes written=344870
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=880
                HDFS: Number of bytes written=1028381690
                HDFS: Number of read operations=26
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=8
                S3: Number of bytes read=1028381690
                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)=1946176
                Total time spent by all reduces in occupied slots (ms)=0
                Total time spent by all map tasks (ms)=60818
                Total vcore-milliseconds taken by all map tasks=60818
                Total megabyte-milliseconds taken by all map tasks=62277632
        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)=1172
                CPU time spent (ms)=42000
                Physical memory (bytes) snapshot=1069641728
                Virtual memory (bytes) snapshot=6595526656
                Total committed heap usage (bytes)=893911040
        File Input Format Counters
                Bytes Read=608
        File Output Format Counters
                Bytes Written=0
        DistCp Counters
                Bytes Copied=1028381690
                Bytes Expected=1028381690
                Files Copied=2
[hadoop@ip-172-31-0-89 ~]$ hadoop fs -ls /user/hive/data_input
Found 2 items
-rw-r--r--   1 hadoop hadoop  545839412 2021-02-28 16:42 /user/hive/data_input/2019-Nov.csv
-rw-r--r--   1 hadoop hadoop  482542278 2021-02-28 16:42 /user/hive/data_input/2019-Oct.csv
[hadoop@ip-172-31-0-89 ~]$
```

hadoop fs -cat /user/hive/data_input/2019-Oct.csv | head



```
                S3: Number of bytes read=1028381690
                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)=1946176
                Total time spent by all reduces in occupied slots (ms)=0
                Total time spent by all map tasks (ms)=60818
                Total vcore-milliseconds taken by all map tasks=60818
                Total megabyte-milliseconds taken by all map tasks=62277632
        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)=1172
                CPU time spent (ms)=42000
                Physical memory (bytes) snapshot=1069641728
                Virtual memory (bytes) snapshot=6595526656
                Total committed heap usage (bytes)=893911040
        File Input Format Counters
                Bytes Read=608
        File Output Format Counters
                Bytes Written=0
        DistCp Counters
                Bytes Copied=1028381690
                Bytes Expected=1028381690
                Files Copied=2
[hadoop@ip-172-31-0-89 ~]$ hadoop fs -ls /user/hive/data_input
Found 2 items
-rw-r--r--   1 hadoop hadoop  545839412 2021-02-28 16:42 /user/hive/data_input/2019-Nov.csv
-rw-r--r--   1 hadoop hadoop  482542278 2021-02-28 16:42 /user/hive/data_input/2019-Oct.csv
[hadoop@ip-172-31-0-89 ~]$ hadoop fs -cat /user/hive/data_input/2019-Oct.csv | head
event_time,event_type,product_id,category_id,category_code,brand,price,user_id,user_session
2019-10-01 00:00:00 UTC,cart,5773203,1487580005134238553,,runail,2.62,463240011,26dd6e6e-4dac-4778-8d2c-92e149dab885
2019-10-01 00:00:03 UTC,cart,5773353,1487580005134238553,,runail,2.62,463240011,26dd6e6e-4dac-4778-8d2c-92e149dab885
2019-10-01 00:00:07 UTC,cart,5881589,2151191071051219817,,lovely,13.48,429681830,49e8d843-adf3-428b-a2c3-fe8bc6a307c9
2019-10-01 00:00:07 UTC,cart,5723490,1487580005134238553,,runail,2.62,463240011,26dd6e6e-4dac-4778-8d2c-92e149dab885
2019-10-01 00:00:15 UTC,cart,5881449,1487580013522845895,,lovely,0.56,429681830,49e8d843-adf3-428b-a2c3-fe8bc6a307c9
2019-10-01 00:00:16 UTC,cart,5857269,1487580005134238553,,runail,2.62,430174032,73dea1e7-664e-43f4-8b30-d32b9d5af04f
2019-10-01 00:00:19 UTC,cart,5739055,1487580008246412266,,kapous,4.75,377667011,81326ac6-daa4-4f0a-b488-fd0956a78733
2019-10-01 00:00:24 UTC,cart,5825598,1487580009445982239,,,0.56,467916806,2f5b5546-b8cb-9ee7-7ecd-84276f8ef486
2019-10-01 00:00:25 UTC,cart,5698989,1487580006317032337,,,1.27,385985999,d30965e8-1101-44ab-b45d-cc1bb9fae694
cat: Unable to write to output stream.
[hadoop@ip-172-31-0-89 ~]$
```

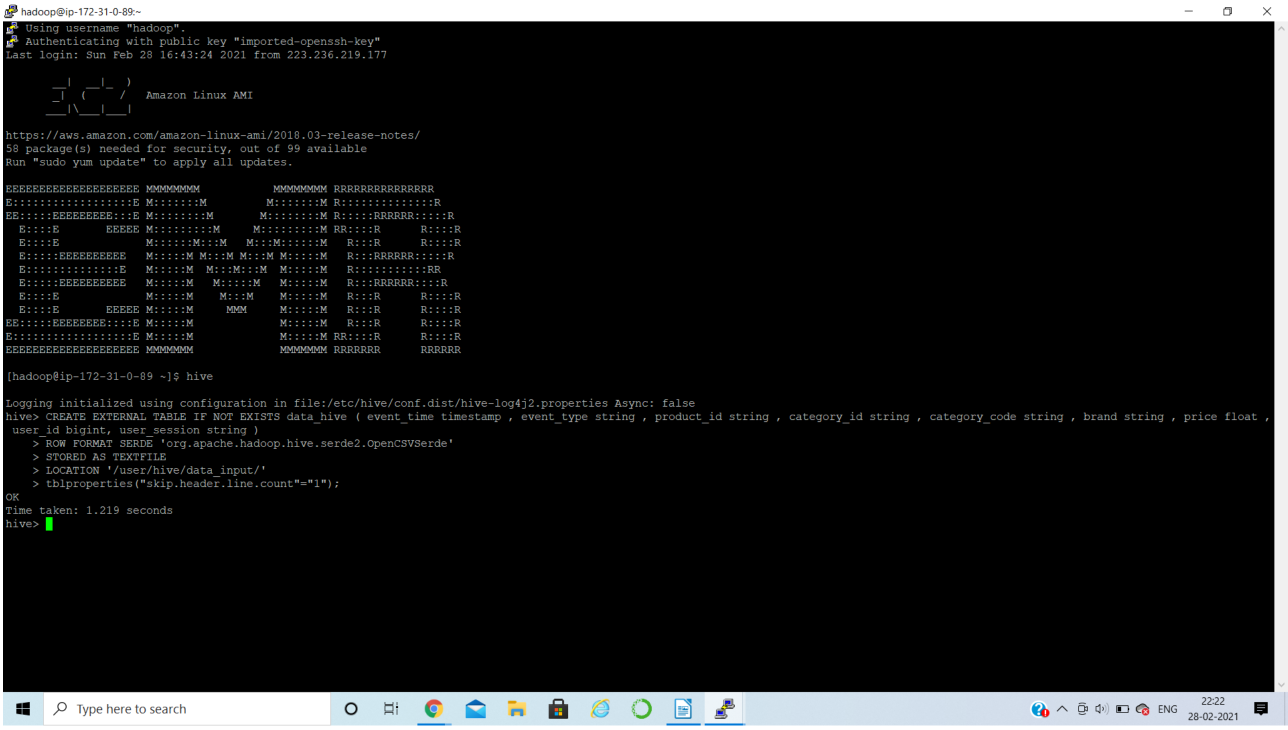
Open a Duplicate session: To work on hive



```
 Using username "hadoop".
 Authenticating with public key "imported-openssh-key"
Last login: Sun Feb 28 16:40:45 2021 from 223.236.219.177

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

https://aws.amazon.com/amazon-linux-ami/2018.03-release-notes/
58 package(s) needed for security, out of 99 available
Run "sudo yum update" to apply all updates.

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

[hadoop@ip-172-31-0-89 ~]$ hive

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

Creating hive table:

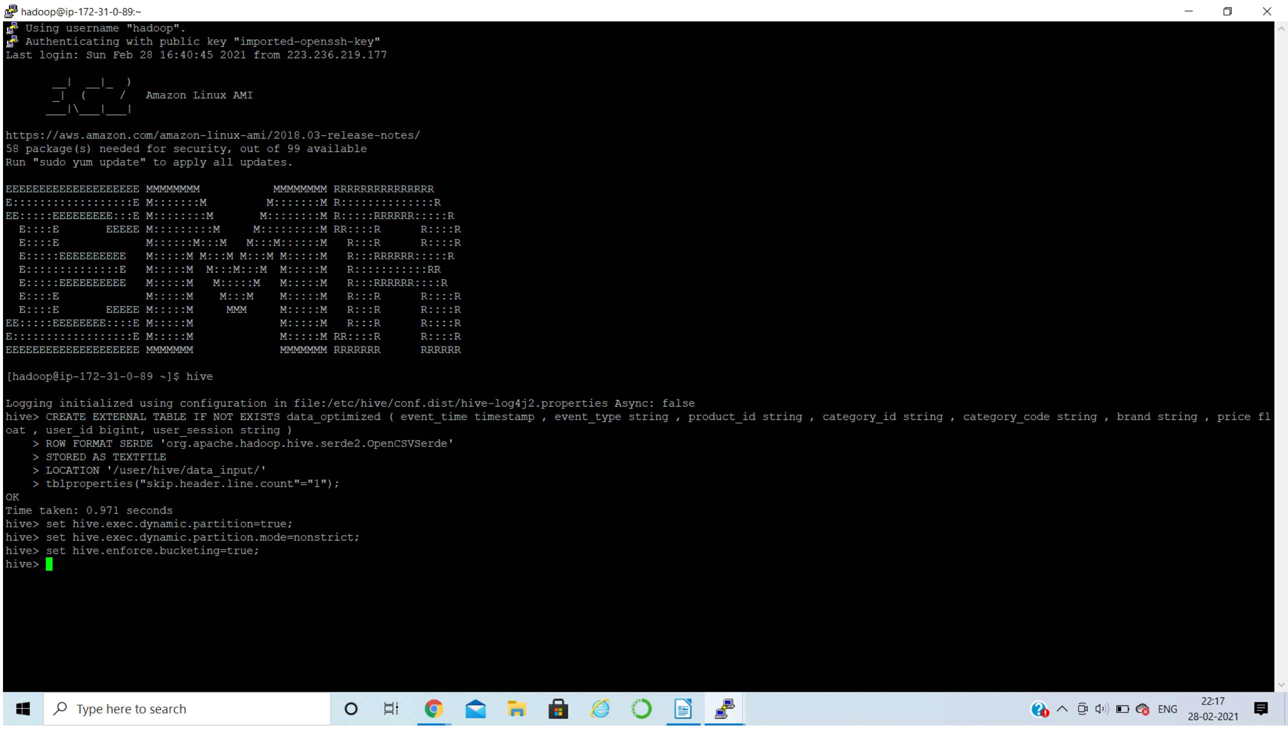
CREATE EXTERNAL TABLE IF NOT EXISTS data_hive (event_time timestamp , event_type string , product_id string , category_id string , category_code string , brand string , price float , user_id bigint, user_session string)
ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.OpenCSVSerde'
STORED AS TEXTFILE
LOCATION '/user/hive/data_input/'
tblproperties("skip.header.line.count"="1");



```
Using username "hadoop".
Authenticating with public key "imported-openssh-key"
Last login: Sun Feb 28 16:43:24 2021 from 223.236.219.177

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

https://aws.amazon.com/amazon-linux-ami/2018.03-release-notes/
58 package(s) needed for security, out of 99 available
Run "sudo yum update" to apply all updates.

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

[hadoop@ip-172-31-0-89 ~]$ hive

Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j2.properties Async: false
hive> CREATE EXTERNAL TABLE IF NOT EXISTS data_hive ( 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/data_input/'
    > tblproperties("skip.header.line.count"="1");
OK
Time taken: 1.219 seconds
hive> 
```

Now enabling dynamic partitioning:

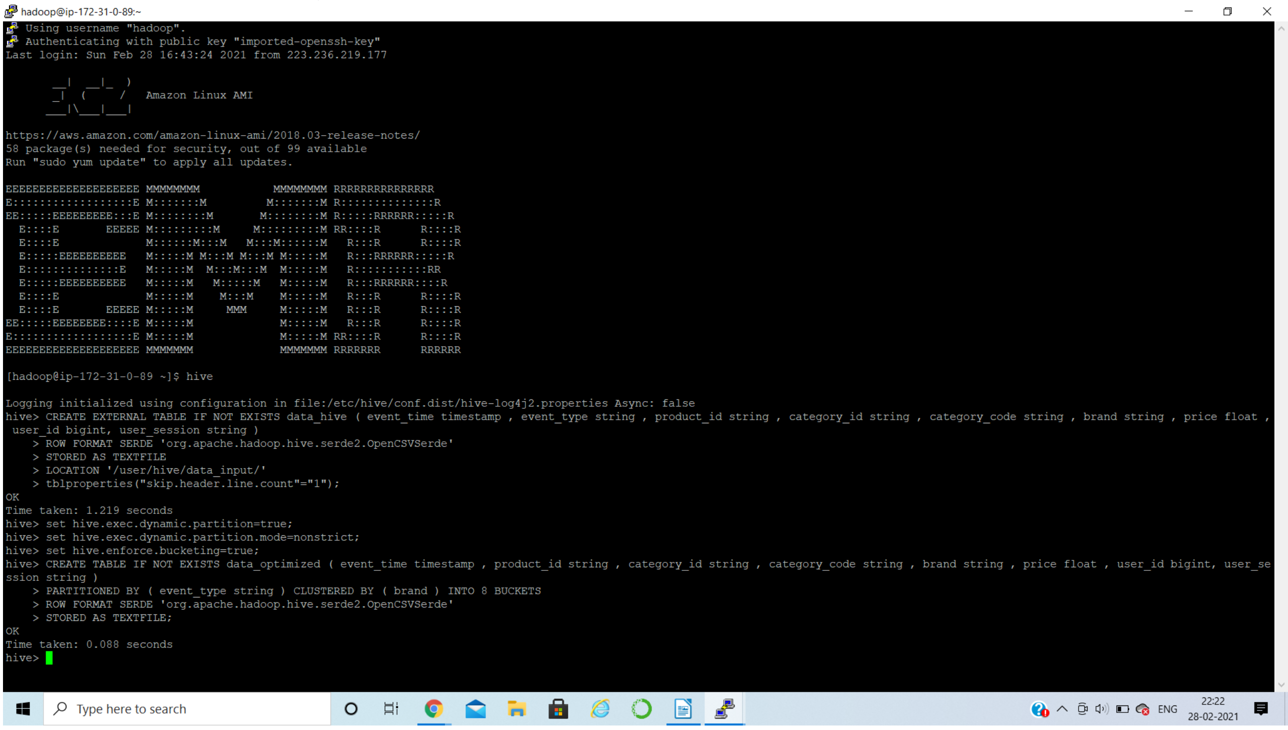
set hive.exec.dynamic.partition=true;
set hive.exec.dynamic.partition.mode=nonstrict;
set hive.enforce.bucketing=true;



```
Using username "hadoop".
Authenticating with public key "imported-openssh-key"
Last login: Sun Feb 28 16:40:45 2021 from 223.236.219.177

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

https://aws.amazon.com/amazon-linux-ami/2018.03-release-notes/
58 package(s) needed for security, out of 99 available
Run "sudo yum update" to apply all updates.

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

[hadoop@ip-172-31-0-89 ~]$ hive

Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j2.properties Async: false
hive> CREATE EXTERNAL TABLE IF NOT EXISTS data_optimized ( event_time timestamp , event_type string , product_id string , category_id string , category_code string , brand string , price fl
oat , user_id bigint, user_session string )
    > ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.OpenCSVSerde'
    > STORED AS TEXTFILE
    > LOCATION '/user/hive/data_input/'
    > tblproperties("skip.header.line.count"="1");
OK
Time taken: 0.971 seconds
hive> set hive.exec.dynamic.partition=true;
hive> set hive.exec.dynamic.partition.mode=nonstrict;
hive> set hive.enforce.bucketing=true;
hive> 
```

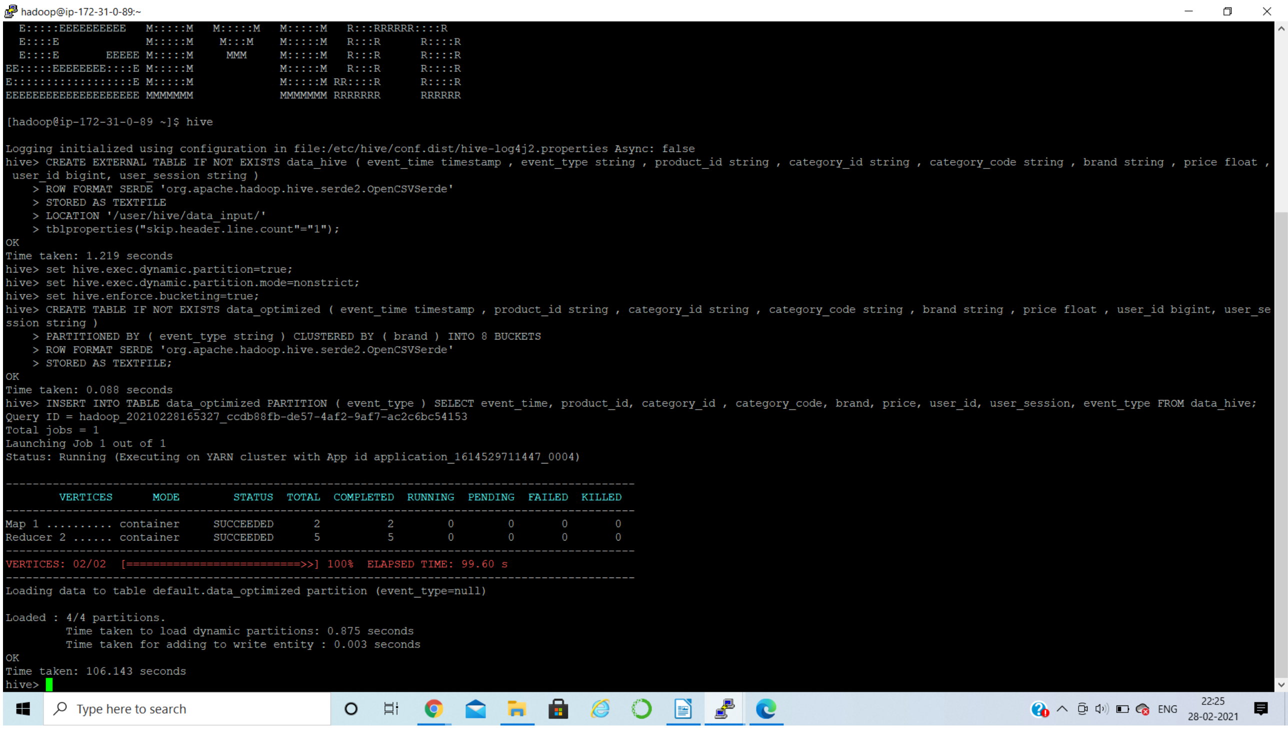
Creating optimized table:

CREATE TABLE IF NOT EXISTS data_optimized (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 (brand) INTO 8 BUCKETS
ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.OpenCSVSerde'
STORED AS TEXTFILE;



```
 Using username "hadoop".
 Authenticating with public key "imported-openssh-key"
Last login: Sun Feb 28 16:43:24 2021 from 223.236.219.177

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

https://aws.amazon.com/amazon-linux-ami/2018.03-release-notes/
58 package(s) needed for security, out of 99 available
Run "sudo yum update" to apply all updates.

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

[hadoop@ip-172-31-0-89 ~]$ hive

Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j2.properties Async: false
hive> CREATE EXTERNAL TABLE IF NOT EXISTS data_hive ( 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/data_input/'
    > tblproperties("skip.header.line.count"="1");
OK
Time taken: 1.219 seconds
hive> set hive.exec.dynamic.partition=true;
hive> set hive.exec.dynamic.partition.mode=nonstrict;
hive> set hive.enforce.bucketing=true;
hive> CREATE TABLE IF NOT EXISTS data_optimized ( event_time timestamp , product_id string , category_id string , category_code string , brand string , price float , user_id bigint, user_se
ssion string )
    > PARTITIONED BY ( event_type string ) CLUSTERED BY ( brand ) INTO 8 BUCKETS
    > ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.OpenCSVSerde'
    > STORED AS TEXTFILE;
OK
Time taken: 0.088 seconds
hive>
```

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



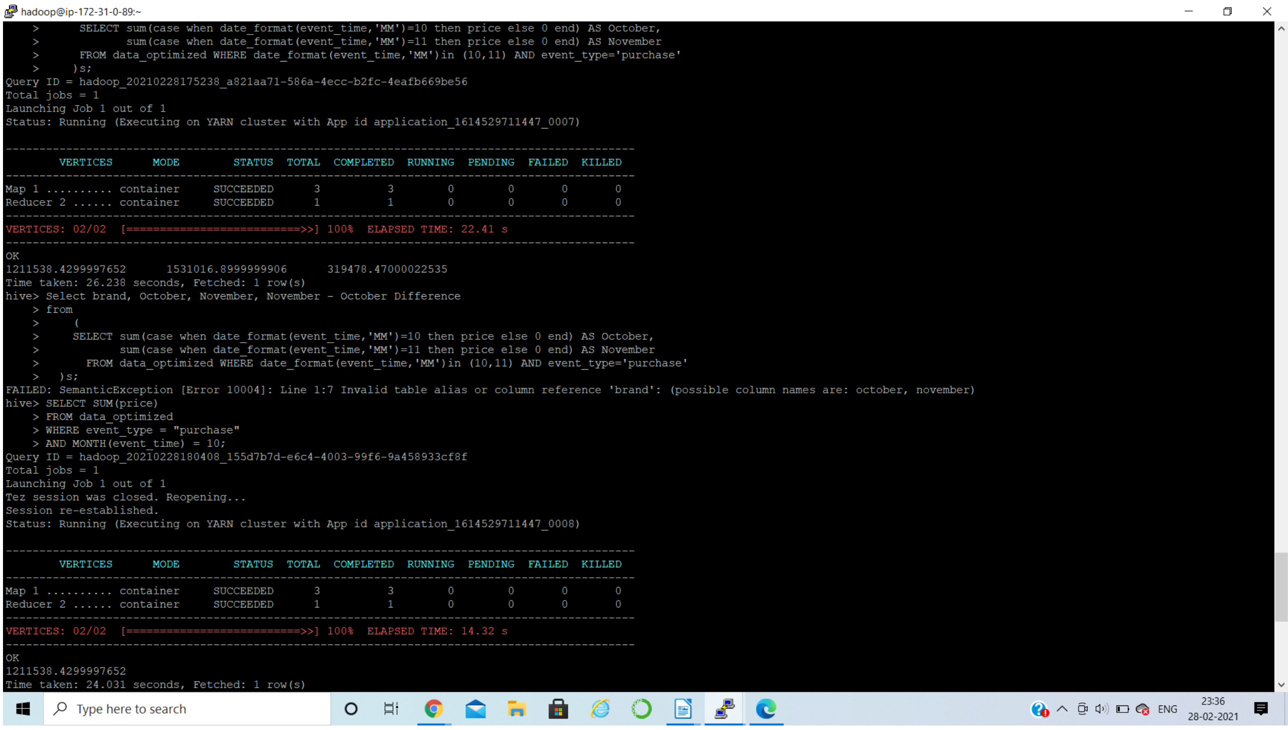
```
  E:::::EEEEEEEEEE   M:::::M   M:::::M   M:::::M   R:::RRRRRR::::R
  E::::E             M:::::M    M:::M    M:::::M   R:::R      R::::R
  E::::E       EEEEE M:::::M     MMM     M:::::M   R:::R      R::::R
EE:::::EEEEEEEE::::E M:::::M             M:::::M   R:::R      R::::R
E::::::::::::::::::E M:::::M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

[hadoop@ip-172-31-0-89 ~]$ hive

Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j2.properties Async: false
hive> CREATE EXTERNAL TABLE IF NOT EXISTS data_hive ( 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/data_input/'
    > tblproperties("skip.header.line.count"="1");
OK
Time taken: 1.219 seconds
hive> set hive.exec.dynamic.partition=true;
hive> set hive.exec.dynamic.partition.mode=nonstrict;
hive> set hive.enforce.bucketing=true;
hive> CREATE TABLE IF NOT EXISTS data_optimized ( event_time timestamp , product_id string , category_id string , category_code string , brand string , price float , user_id bigint, user_se
ssion string )
    > PARTITIONED BY ( event_type string ) CLUSTERED BY ( brand ) INTO 8 BUCKETS
    > ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.OpenCSVSerde'
    > STORED AS TEXTFILE;
OK
Time taken: 0.088 seconds
hive> INSERT INTO TABLE data_optimized PARTITION ( event_type ) SELECT event_time, product_id, category_id , category_code, brand, price, user_id, user_session, event_type FROM data_hive;
Query ID = hadoop_20210228165327_ccdb88fb-de57-4af2-9af7-ac2c6bc54153
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1614529711447_0004)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      2          2        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      5          5        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 99.60 s
----------------------------------------------------------------------------------------------
Loading data to table default.data_optimized partition (event_type=null)

Loaded : 4/4 partitions.
         Time taken to load dynamic partitions: 0.875 seconds
         Time taken for adding to write entity : 0.003 seconds
OK
Time taken: 106.143 seconds
hive>
```

Total 4 Partitioning Done using th Above command.

Performance comparison:

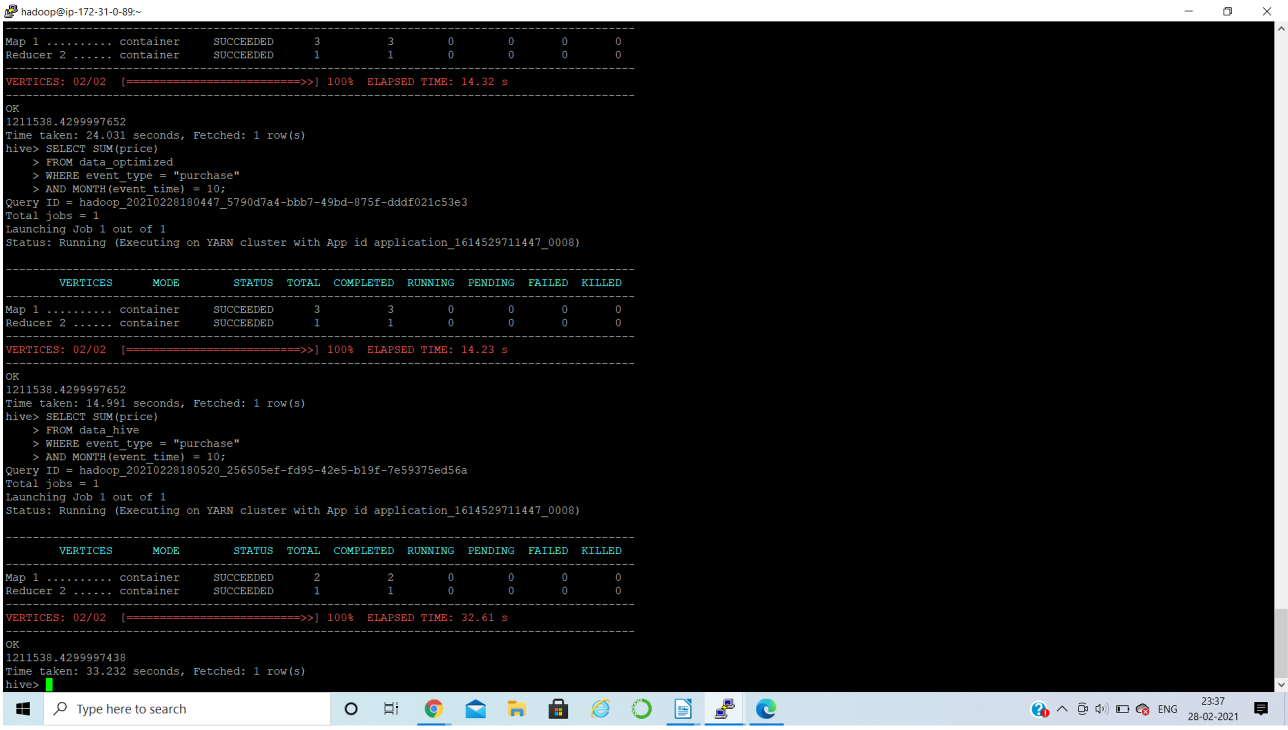
For Instance running the following command using data_hive Table and data_optimized Table for comparison:

SELECT SUM(price)
FROM data_optimized
WHERE event_type = "purchase"
AND MONTH(event_time) = 10;



```
    >      SELECT sum(case when date_format(event_time,'MM')=10 then price else 0 end) AS October,
    >             sum(case when date_format(event_time,'MM')=11 then price else 0 end) AS November
    >      FROM data_optimized WHERE date_format(event_time,'MM')in (10,11) AND event_type='purchase'
    >     )s;
Query ID = hadoop_20210228175238_a821aa71-586a-4ecc-b2fc-4eafb669be56
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1614529711447_0007)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      3          3        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 22.41 s
----------------------------------------------------------------------------------------------
OK
1211538.4299997652      1531016.8999999906      319478.47000022535
Time taken: 26.238 seconds, Fetched: 1 row(s)
hive> Select brand, October, November, November - October Difference
    > from
    >     (
    >     SELECT sum(case when date_format(event_time,'MM')=10 then price else 0 end) AS October,
    >            sum(case when date_format(event_time,'MM')=11 then price else 0 end) AS November
    >       FROM data_optimized WHERE date_format(event_time,'MM')in (10,11) AND event_type='purchase'
    >   )s;
FAILED: SemanticException [Error 10004]: Line 1:7 Invalid table alias or column reference 'brand': (possible column names are: october, november)
hive> SELECT SUM(price)
    > FROM data_optimized
    > WHERE event_type = "purchase"
    > AND MONTH(event_time) = 10;
Query ID = hadoop_20210228180408_155d7b7d-e6c4-4003-99f6-9a458933cf8f
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_1614529711447_0008)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      3          3        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 14.32 s
----------------------------------------------------------------------------------------------
OK
1211538.4299997652
Time taken: 24.031 seconds, Fetched: 1 row(s)
```

SELECT SUM(price)
FROM data_hive
WHERE event_type = "purchase"
AND MONTH(event_time) = 10;



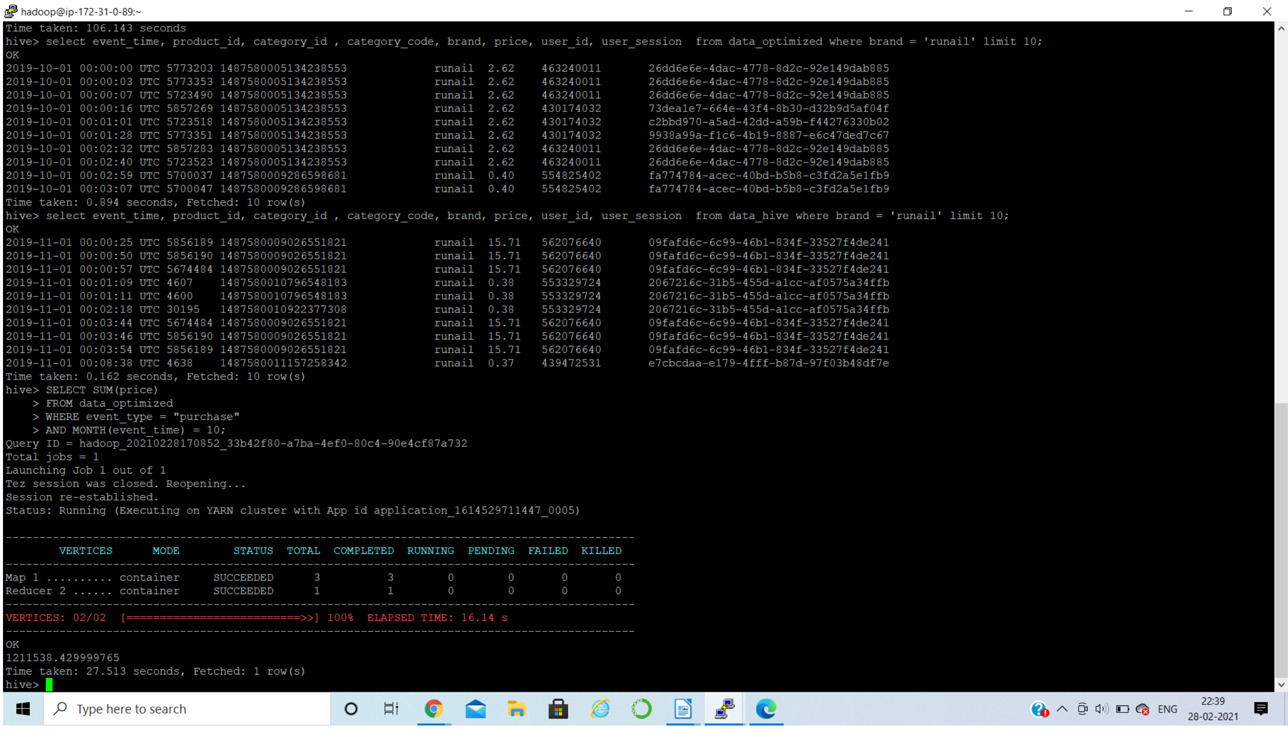
```
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      3          3        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 14.32 s
----------------------------------------------------------------------------------------------
OK
1211538.4299997652
Time taken: 24.031 seconds, Fetched: 1 row(s)
hive> SELECT SUM(price)
    > FROM data_optimized
    > WHERE event_type = "purchase"
    > AND MONTH(event_time) = 10;
Query ID = hadoop_20210228180447_5790d7a4-bbb7-49bd-875f-dddf021c53e3
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1614529711447_0008)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      3          3        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 14.23 s
----------------------------------------------------------------------------------------------
OK
1211538.4299997652
Time taken: 14.991 seconds, Fetched: 1 row(s)
hive> SELECT SUM(price)
    > FROM data_hive
    > WHERE event_type = "purchase"
    > AND MONTH(event_time) = 10;
Query ID = hadoop_20210228180520_256505ef-fd95-42e5-b19f-7e59375ed56a
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1614529711447_0008)

----------------------------------------------------------------------------------------------
        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: 32.61 s
----------------------------------------------------------------------------------------------
OK
1211538.4299997438
Time taken: 33.232 seconds, Fetched: 1 row(s)
hive>
```

Observation: Clearly the data_hive table (33.232.secs)took more time to execute the query than the data_optimized table(24.031 secs).

Case Study Questions

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

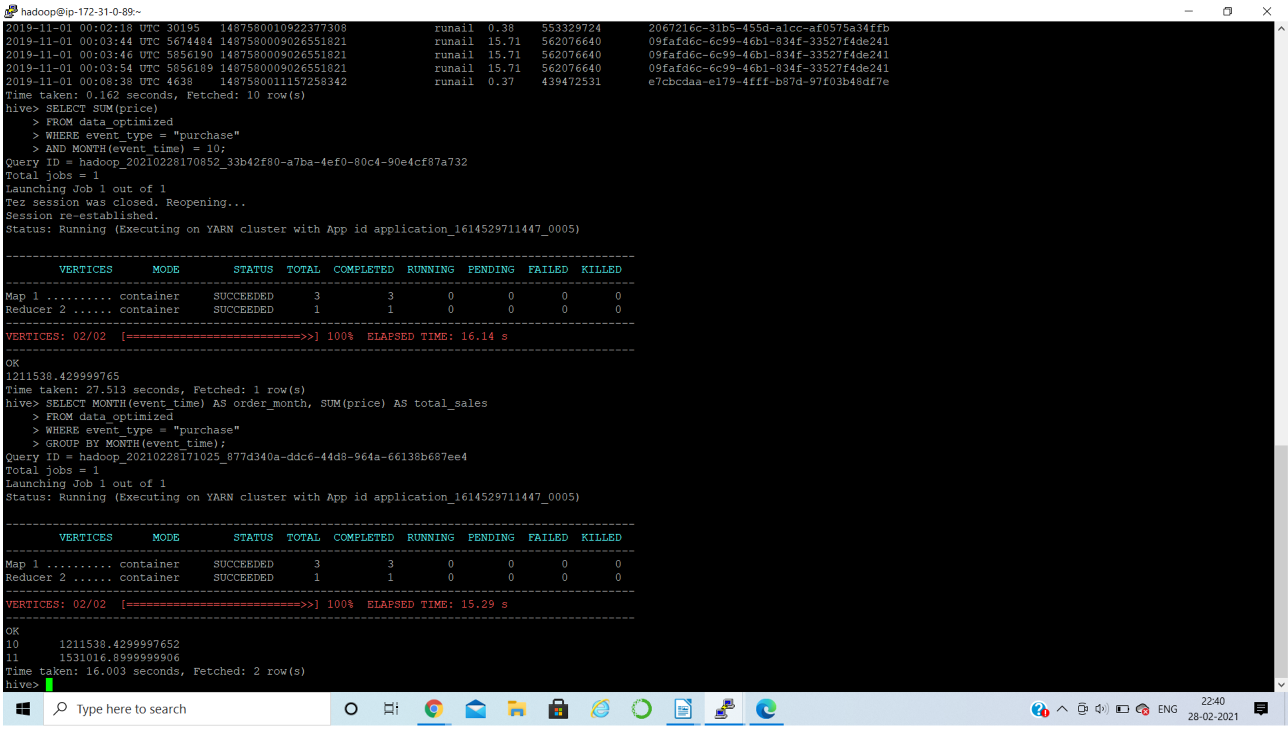
SELECT SUM(price)
FROM data_optimized
WHERE event_type = "purchase"
AND MONTH(event_time) = 10;



```
Time taken: 106.143 seconds
hive> select event_time, product_id, category_id , category_code, brand, price, user_id, user_session  from data_optimized where brand = 'runail' limit 10;
OK
2019-10-01 00:00:00 UTC 5773203 1487580005134238553             runail  2.62    463240011       26dd6e6e-4dac-4778-8d2c-92e149dab885
2019-10-01 00:00:03 UTC 5773353 1487580005134238553             runail  2.62    463240011       26dd6e6e-4dac-4778-8d2c-92e149dab885
2019-10-01 00:00:07 UTC 5723490 1487580005134238553             runail  2.62    463240011       26dd6e6e-4dac-4778-8d2c-92e149dab885
2019-10-01 00:00:16 UTC 5857269 1487580005134238553             runail  2.62    430174032       73dea1e7-664e-43f4-8b30-d32b9d5af04f
2019-10-01 00:01:01 UTC 5723518 1487580005134238553             runail  2.62    430174032       c2bbd970-a5ad-42dd-a59b-f44276330b02
2019-10-01 00:01:28 UTC 5773351 1487580005134238553             runail  2.62    430174032       9938a99a-f1c6-4b19-8887-e6c47ded7c67
2019-10-01 00:02:32 UTC 5857283 1487580005134238553             runail  2.62    463240011       26dd6e6e-4dac-4778-8d2c-92e149dab885
2019-10-01 00:02:40 UTC 5723523 1487580005134238553             runail  2.62    463240011       26dd6e6e-4dac-4778-8d2c-92e149dab885
2019-10-01 00:02:59 UTC 5700037 1487580009286598681             runail  0.40    554825402       fa774784-acec-40bd-b5b8-c3fd2a5e1fb9
2019-10-01 00:03:07 UTC 5700047 1487580009286598681             runail  0.40    554825402       fa774784-acec-40bd-b5b8-c3fd2a5e1fb9
Time taken: 0.894 seconds, Fetched: 10 row(s)
hive> select event_time, product_id, category_id , category_code, brand, price, user_id, user_session  from data_hive where brand = 'runail' limit 10;
OK
2019-11-01 00:00:25 UTC 5856189 1487580009026551821             runail  15.71   562076640       09fafd6c-6c99-46b1-834f-33527f4de241
2019-11-01 00:00:50 UTC 5856190 1487580009026551821             runail  15.71   562076640       09fafd6c-6c99-46b1-834f-33527f4de241
2019-11-01 00:00:57 UTC 5674484 1487580009026551821             runail  15.71   562076640       09fafd6c-6c99-46b1-834f-33527f4de241
2019-11-01 00:01:09 UTC 4607    1487580010796548183             runail  0.38    553329724       2067216c-31b5-455d-a1cc-af0575a34ffb
2019-11-01 00:01:11 UTC 4600    1487580010796548183             runail  0.38    553329724       2067216c-31b5-455d-a1cc-af0575a34ffb
2019-11-01 00:02:18 UTC 30195   1487580010922377308             runail  0.38    553329724       2067216c-31b5-455d-a1cc-af0575a34ffb
2019-11-01 00:03:44 UTC 5674484 1487580009026551821             runail  15.71   562076640       09fafd6c-6c99-46b1-834f-33527f4de241
2019-11-01 00:03:46 UTC 5856190 1487580009026551821             runail  15.71   562076640       09fafd6c-6c99-46b1-834f-33527f4de241
2019-11-01 00:03:54 UTC 5856189 1487580009026551821             runail  15.71   562076640       09fafd6c-6c99-46b1-834f-33527f4de241
2019-11-01 00:08:38 UTC 4638    1487580011157258342             runail  0.37    439472531       e7cbcdaa-e179-4fff-b87d-97f03b48df7e
Time taken: 0.162 seconds, Fetched: 10 row(s)
hive> SELECT SUM(price)
    > FROM data_optimized
    > WHERE event_type = "purchase"
    > AND MONTH(event_time) = 10;
Query ID = hadoop_20210228170852_33b42f80-a7ba-4ef0-80c4-90e4cf87a732
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_1614529711447_0005)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      3          3        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 16.14 s
----------------------------------------------------------------------------------------------
OK
1211538.429999765
Time taken: 27.513 seconds, Fetched: 1 row(s)
hive>
```

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

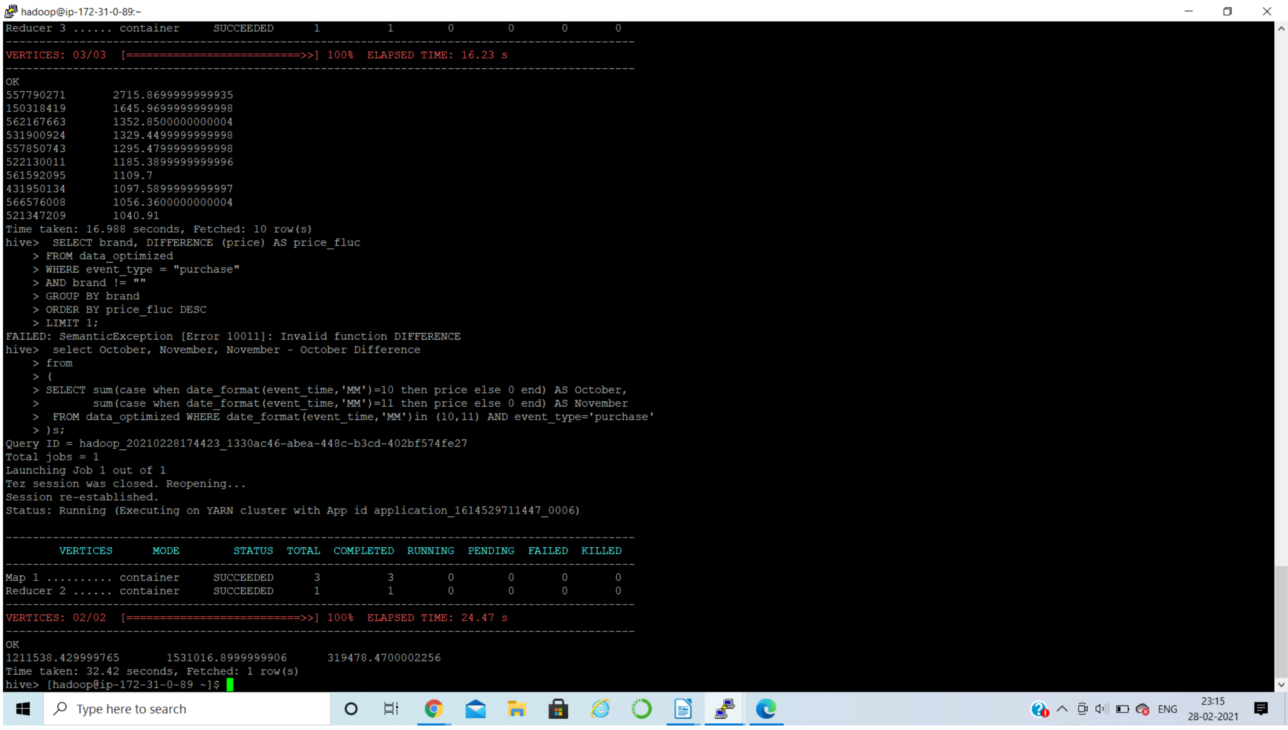
```
SELECT MONTH(event_time) AS order_month, SUM(price) AS total_sales
FROM data_optimized
WHERE event_type = "purchase"
GROUP BY MONTH(event_time);
```



```
2019-11-01 00:02:18 UTC 30195   1487580010922377308             runail  0.38    553329724       2067216c-31b5-455d-a1cc-af0575a34ffb
2019-11-01 00:03:44 UTC 5674484 1487580009026551821             runail  15.71   562076640       09fafd6c-6c99-46b1-834f-33527f4de241
2019-11-01 00:03:46 UTC 5856190 1487580009026551821             runail  15.71   562076640       09fafd6c-6c99-46b1-834f-33527f4de241
2019-11-01 00:03:54 UTC 5856189 1487580009026551821             runail  15.71   562076640       09fafd6c-6c99-46b1-834f-33527f4de241
2019-11-01 00:08:38 UTC 4638    1487580011157258342             runail  0.37    439472531       e7cbcdaa-e179-4fff-b87d-97f03b48df7e
Time taken: 0.162 seconds, Fetched: 10 row(s)
hive> SELECT SUM(price)
    > FROM data_optimized
    > WHERE event_type = "purchase"
    > AND MONTH(event_time) = 10;
Query ID = hadoop_20210228170852_33b42f80-a7ba-4ef0-80c4-90e4cf87a732
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_1614529711447_0005)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      3          3        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 16.14 s
----------------------------------------------------------------------------------------------
OK
1211538.429999765
Time taken: 27.513 seconds, Fetched: 1 row(s)
hive> SELECT MONTH(event_time) AS order_month, SUM(price) AS total_sales
    > FROM data_optimized
    > WHERE event_type = "purchase"
    > GROUP BY MONTH(event_time);
Query ID = hadoop_20210228171025_877d340a-ddc6-44d8-964a-66138b687ee4
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1614529711447_0005)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      3          3        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 15.29 s
----------------------------------------------------------------------------------------------
OK
10      1211538.4299997652
11      1531016.8999999906
Time taken: 16.003 seconds, Fetched: 2 row(s)
hive>
```

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

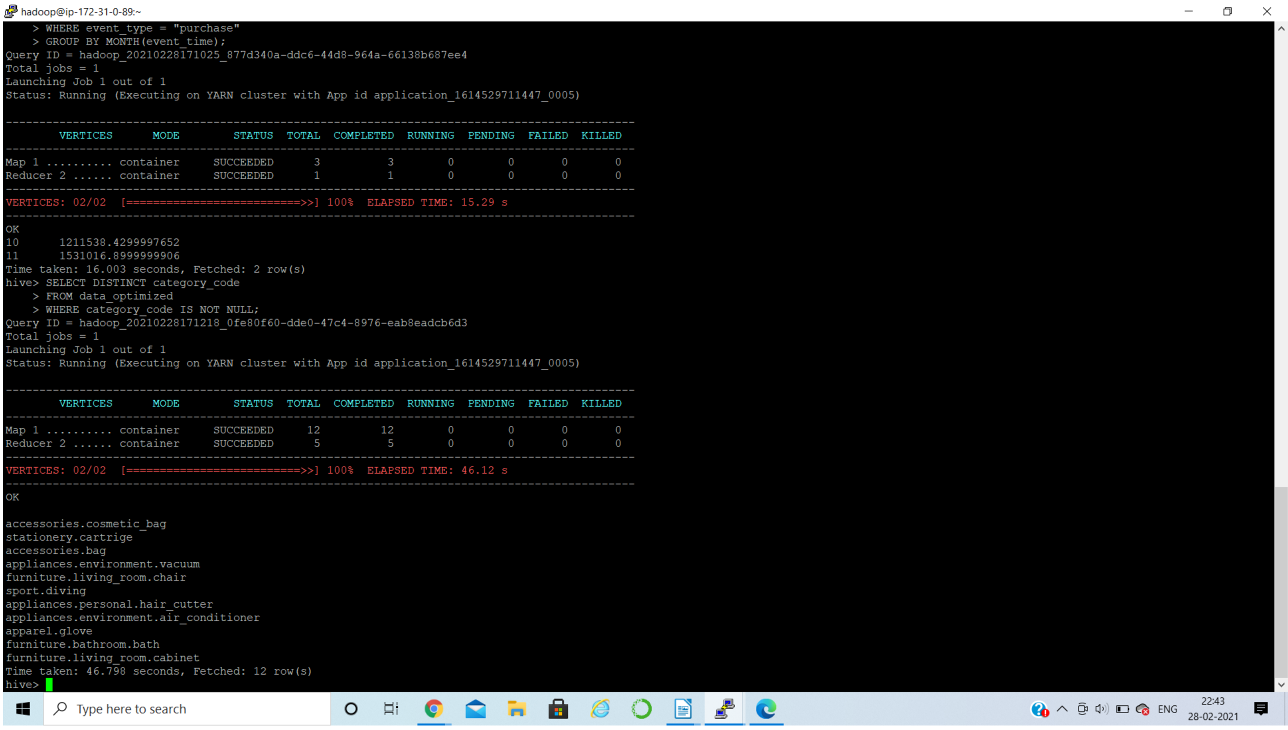
```
select October, November, November - October Difference
    from
    (
     SELECT sum(case when date_format(event_time,'MM')=10 then price else 0 end) AS October,
            sum(case when date_format(event_time,'MM')=11 then price else 0 end) AS November
     FROM data_optimized WHERE date_format(event_time,'MM')in (10,11) AND
event_type='purchase'
    )s;
```



```
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 16.23 s
----------------------------------------------------------------------------------------------
OK
557790271       2715.8699999999935
150318419       1645.9699999999998
562167663       1352.8500000000004
531900924       1329.4499999999998
557850743       1295.4799999999998
522130011       1185.3899999999996
561592095       1109.7
431950134       1097.5899999999997
566576008       1056.3600000000004
521347209       1040.91
Time taken: 16.988 seconds, Fetched: 10 row(s)
hive>  SELECT brand, DIFFERENCE (price) AS price_fluc
    > FROM data_optimized
    > WHERE event_type = "purchase"
    > AND brand != ""
    > GROUP BY brand
    > ORDER BY price_fluc DESC
    > LIMIT 1;
FAILED: SemanticException [Error 10011]: Invalid function DIFFERENCE
hive>  select October, November, November - October Difference
    > from
    > (
    > SELECT sum(case when date_format(event_time,'MM')=10 then price else 0 end) AS October,
    >        sum(case when date_format(event_time,'MM')=11 then price else 0 end) AS November
    >  FROM data_optimized WHERE date_format(event_time,'MM')in (10,11) AND event_type='purchase'
    > )s;
Query ID = hadoop_20210228174423_1330ac46-abea-448c-b3cd-402bf574fe27
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_1614529711447_0006)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      3          3        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 24.47 s
----------------------------------------------------------------------------------------------
OK
1211538.429999765       1531016.8999999906      319478.4700002256
Time taken: 32.42 seconds, Fetched: 1 row(s)
hive> [hadoop@ip-172-31-0-89 ~]$
```

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

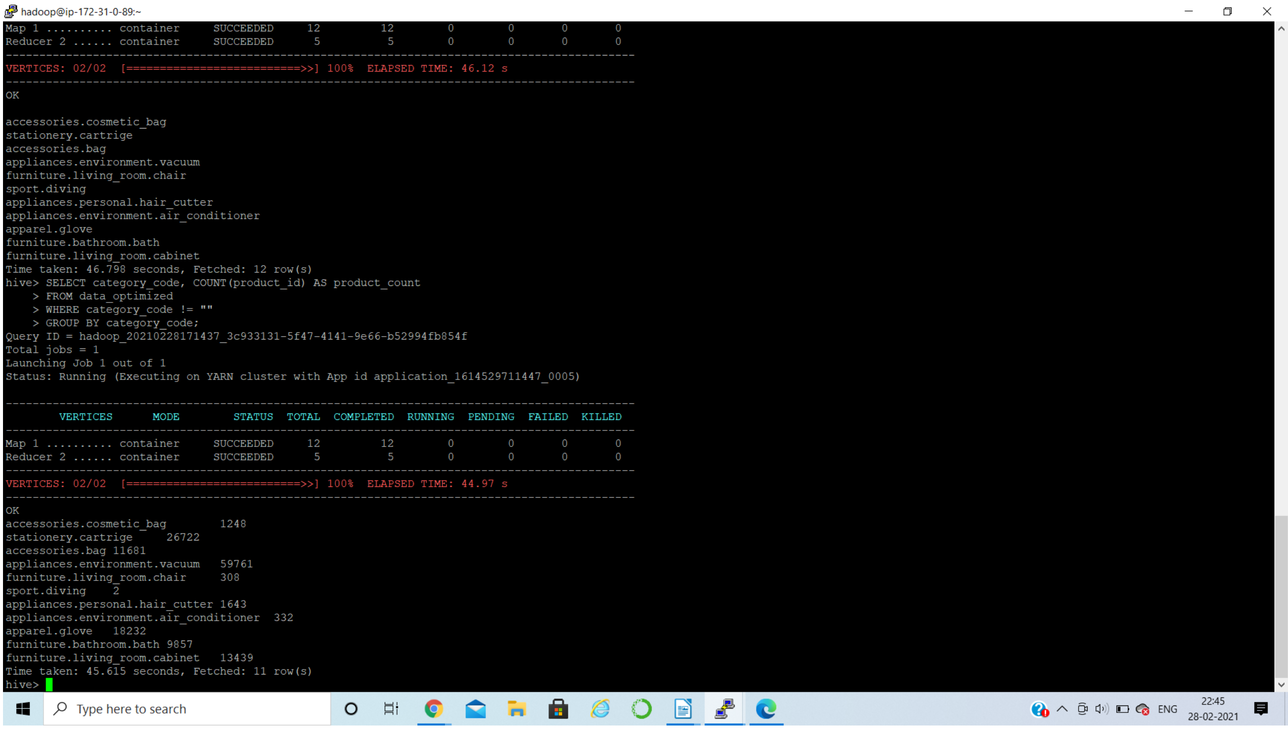
SELECT DISTINCT category_code
FROM data_optimized
WHERE category_code IS NOT NULL;



```
    > WHERE event_type = "purchase"
    > GROUP BY MONTH(event_time);
Query ID = hadoop_20210228171025_877d340a-ddc6-44d8-964a-66138b687ee4
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1614529711447_0005)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      3          3        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 15.29 s
----------------------------------------------------------------------------------------------
OK
10      1211538.4299997652
11      1531016.8999999906
Time taken: 16.003 seconds, Fetched: 2 row(s)
hive> SELECT DISTINCT category_code
    > FROM data_optimized
    > WHERE category_code IS NOT NULL;
Query ID = hadoop_20210228171218_0fe80f60-dde0-47c4-8976-eab8eadcb6d3
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1614529711447_0005)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED     12         12        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      5          5        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 46.12 s
----------------------------------------------------------------------------------------------
OK

accessories.cosmetic_bag
stationery.cartrige
accessories.bag
appliances.environment.vacuum
furniture.living_room.chair
sport.diving
appliances.personal.hair_cutter
appliances.environment.air_conditioner
apparel.glove
furniture.bathroom.bath
furniture.living_room.cabinet
Time taken: 46.798 seconds, Fetched: 12 row(s)
hive>
```

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

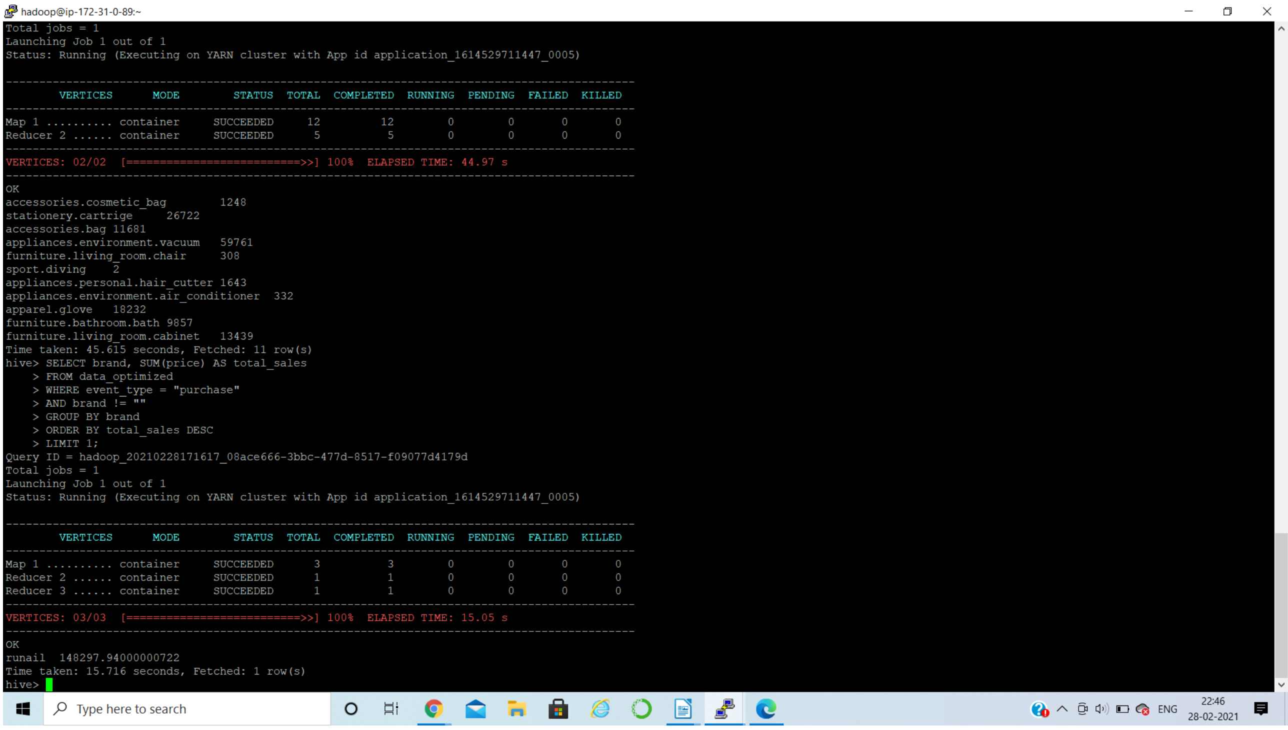
SELECT category_code, COUNT(product_id) AS product_count
FROM data_optimized
WHERE category_code != ""
GROUP BY category_code;



```
Map 1 .......... container     SUCCEEDED     12          12         0         0        0        0
Reducer 2 ...... container     SUCCEEDED      5           5         0         0        0        0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 46.12 s
----------------------------------------------------------------------------------------------
OK

accessories.cosmetic_bag
stationery.cartrige
accessories.bag
appliances.environment.vacuum
furniture.living_room.chair
sport.diving
appliances.personal.hair_cutter
appliances.environment.air_conditioner
apparel.glove
furniture.bathroom.bath
furniture.living_room.cabinet
Time taken: 46.798 seconds, Fetched: 12 row(s)
hive> SELECT category_code, COUNT(product_id) AS product_count
    > FROM data_optimized
    > WHERE category_code != ""
    > GROUP BY category_code;
Query ID = hadoop_20210228171437_3c933131-5f47-4141-9e66-b52994fb854f
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1614529711447_0005)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED     12          12         0         0        0        0
Reducer 2 ...... container     SUCCEEDED      5           5         0         0        0        0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 44.97 s
----------------------------------------------------------------------------------------------
OK
accessories.cosmetic_bag        1248
stationery.cartrige     26722
accessories.bag 11681
appliances.environment.vacuum   59761
furniture.living_room.chair     308
sport.diving    2
appliances.personal.hair_cutter 1643
appliances.environment.air_conditioner  332
apparel.glove   18232
furniture.bathroom.bath 9857
furniture.living_room.cabinet   13439
Time taken: 45.615 seconds, Fetched: 11 row(s)
hive> 
```

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

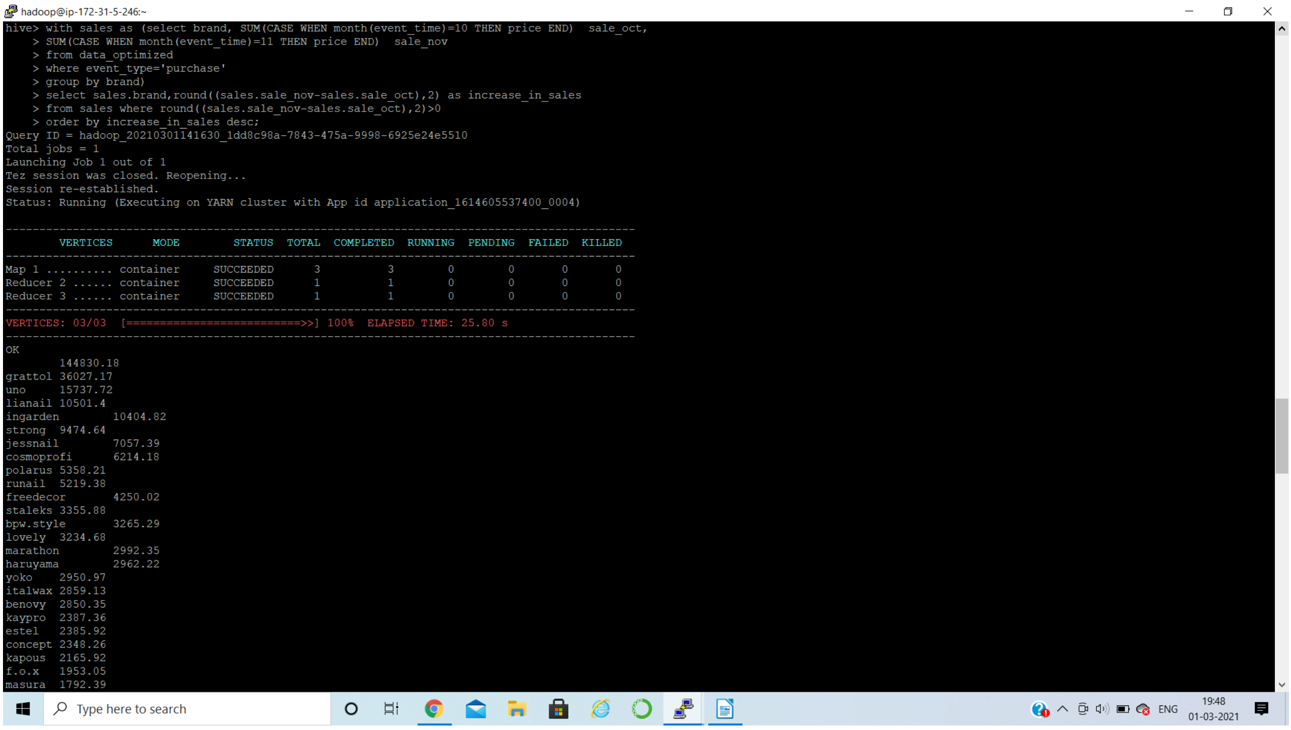
SELECT brand, SUM(price) AS total_sales
FROM data_optimized
WHERE event_type = "purchase"
AND brand != ""
GROUP BY brand
ORDER BY total_sales DESC
LIMIT 1;



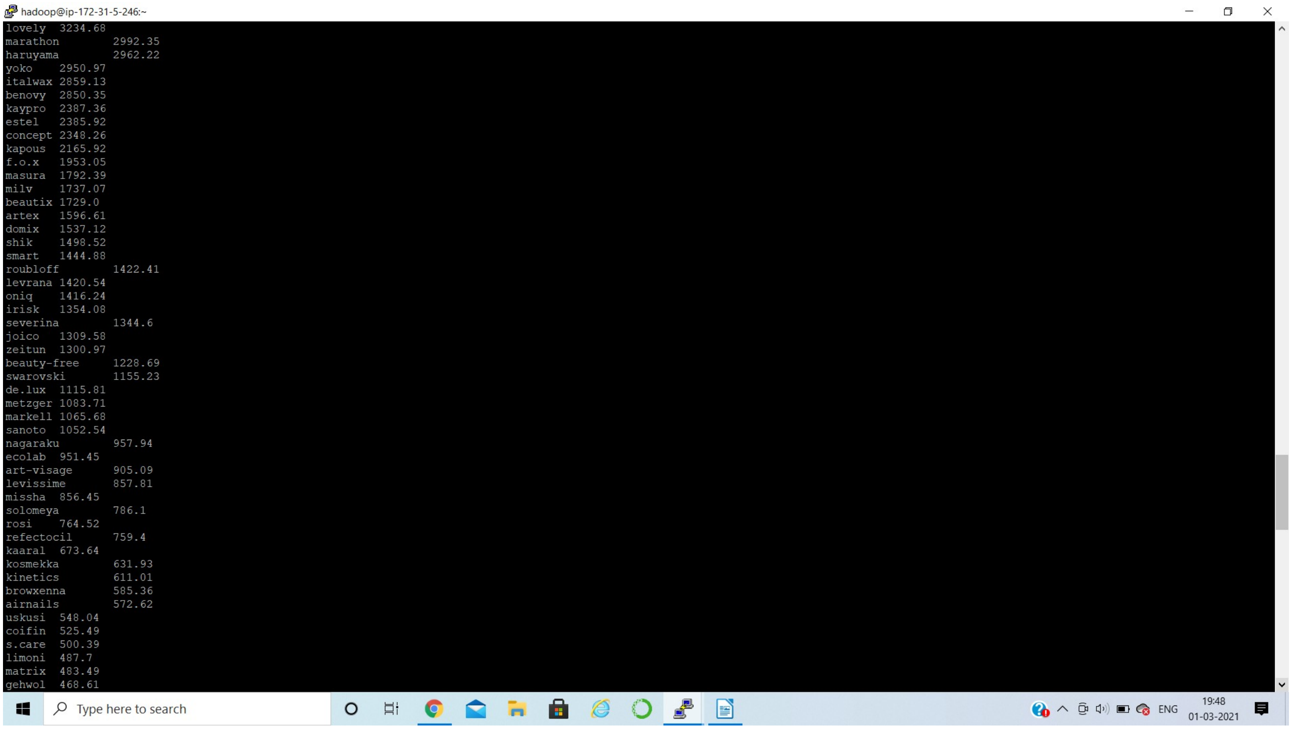
```
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1614529711447_0005)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED     12         12        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      5          5        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 44.97 s
----------------------------------------------------------------------------------------------
OK
accessories.cosmetic_bag        1248
stationery.cartrige     26722
accessories.bag 11681
appliances.environment.vacuum   59761
furniture.living_room.chair     308
sport.diving    2
appliances.personal.hair_cutter 1643
appliances.environment.air_conditioner  332
apparel.glove   18232
furniture.bathroom.bath 9857
furniture.living_room.cabinet   13439
Time taken: 45.615 seconds, Fetched: 11 row(s)
hive> SELECT brand, SUM(price) AS total_sales
    > FROM data_optimized
    > WHERE event_type = "purchase"
    > AND brand != ""
    > GROUP BY brand
    > ORDER BY total_sales DESC
    > LIMIT 1;
Query ID = hadoop_20210228171617_08ace666-3bbc-477d-8517-f09077d4179d
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1614529711447_0005)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      3          3        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 15.05 s
----------------------------------------------------------------------------------------------
OK
runail  148297.94000000722
Time taken: 15.716 seconds, Fetched: 1 row(s)
hive>
```

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

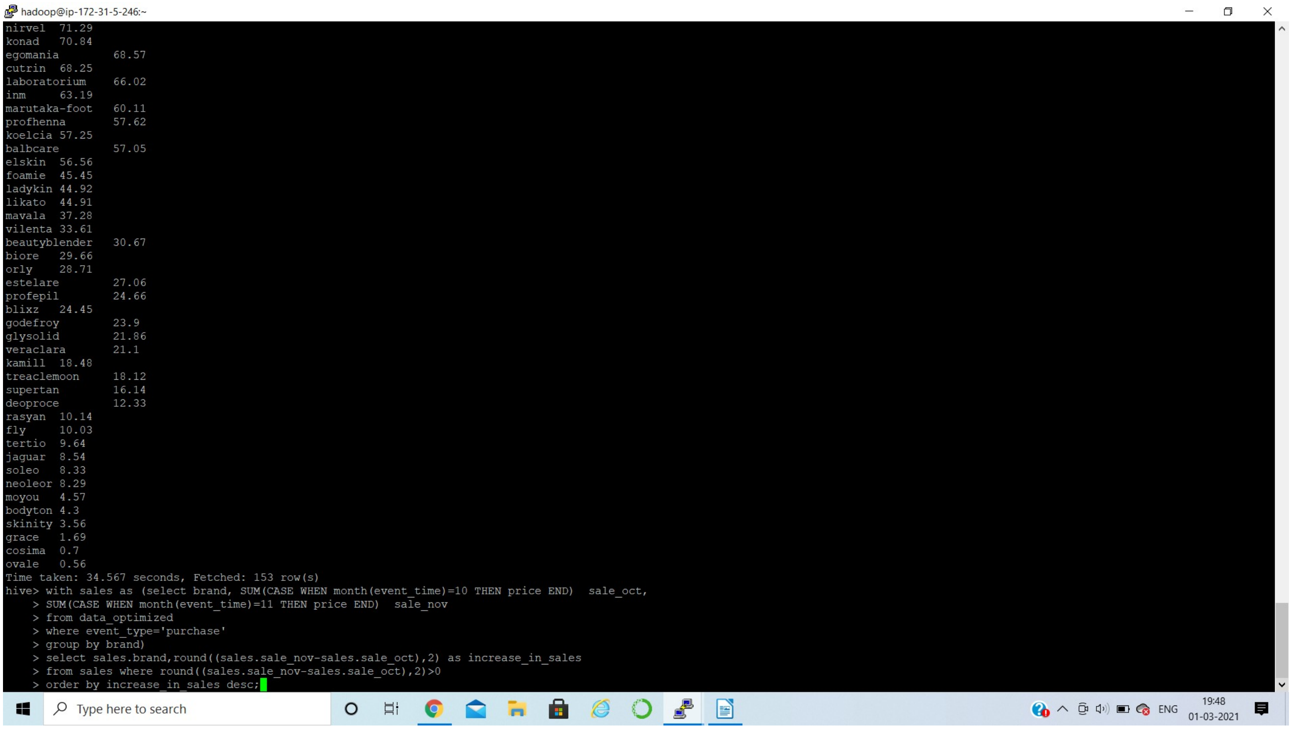
with sales as (select brand, SUM(CASE WHEN month(event_time)=10 THEN price END)
sale_oct,
SUM(CASE WHEN month(event_time)=11 THEN price END) sale_nov
from data_optimized
where event_type='purchase'
group by brand)
select sales.brand,round((sales.sale_nov-sales.sale_oct),2) as increase_in_sales
from sales where round((sales.sale_nov-sales.sale_oct),2)>0
order by increase_in_sales desc;



```
hive> with sales as (select brand, SUM(CASE WHEN month(event_time)=10 THEN price END)  sale_oct,
    > SUM(CASE WHEN month(event_time)=11 THEN price END)  sale_nov
    > from data_optimized
    > where event_type='purchase'
    > group by brand)
    > select sales.brand,round((sales.sale_nov-sales.sale_oct),2) as increase_in_sales
    > from sales where round((sales.sale_nov-sales.sale_oct),2)>0
    > order by increase_in_sales desc;
Query ID = hadoop_20210301141630_1dd8c98a-7843-475a-9998-6925e24e5510
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_1614605537400_0004)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      3          3        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 25.80 s
----------------------------------------------------------------------------------------------
OK
        144830.18
grattol 36027.17
uno     15737.72
lianail 10501.4
ingarden        10404.82
strong  9474.64
jessnail        7057.39
cosmoprofi      6214.18
polarus 5358.21
runail  5219.38
freedecor       4250.02
staleks 3355.88
bpw.style       3265.29
lovely  3234.68
marathon        2992.35
haruyama        2962.22
yoko    2950.97
italwax 2859.13
benovy  2850.35
kaypro  2387.36
estel   2385.92
concept 2348.26
kapous  2165.92
f.o.x   1953.05
masura  1792.39
```



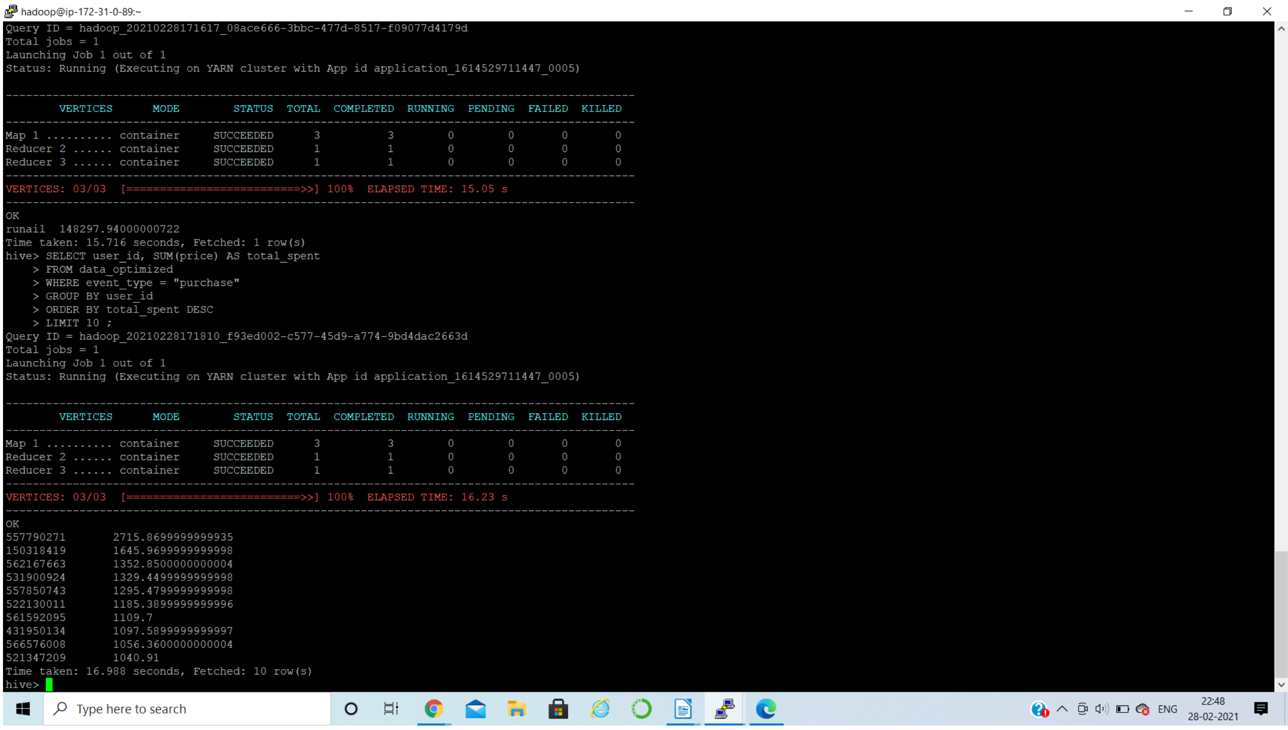
```
lovely  3234.68
marathon        2992.35
haruyama        2962.22
yoko    2950.97
italwax 2859.13
benovy  2850.35
kaypro  2387.36
estel   2385.92
concept 2348.26
kapous  2165.92
f.o.x   1953.05
masura  1792.39
milv    1737.07
beautix 1729.0
artex   1596.61
domix   1537.12
shik    1498.52
smart   1444.88
roubloff        1422.41
levrana 1420.54
oniq    1416.24
irisk   1354.08
severina        1344.6
joico   1309.58
zeitun  1300.97
beauty-free     1228.69
swarovski       1155.23
de.lux  1115.81
metzger 1083.71
markell 1065.68
sanoto  1052.54
nagaraku        957.94
ecolab  951.45
art-visage      905.09
levissime       857.81
missha  856.45
solomeya        786.1
rosi    764.52
refectocil      759.4
kaaral  673.64
kosmekka        631.93
kinetics        611.01
browxenna       585.36
airnails        572.62
uskusi  548.04
coifin  525.49
s.care  500.39
limoni  487.7
matrix  483.49
gehwol  468.61
```



```
nirvel  71.29
konad   70.84
egomania        68.57
cutrin  68.25
laboratorium    66.02
inm     63.19
marutaka-foot   60.11
profhenna       57.62
koelcia 57.25
balbcare        57.05
elskin  56.56
foamie  45.45
ladykin 44.92
likato  44.91
mavala  37.28
vilenta 33.61
beautyblender   30.67
biore   29.66
orly    28.71
estelare        27.06
profepil        24.66
blixz   24.45
godefroy        23.9
glysolid        21.86
veraclara       21.1
kamill  18.48
treaclemoon     18.12
supertan        16.14
deoproce        12.33
rasyan  10.14
fly     10.03
tertio  9.64
jaguar  8.54
soleo   8.33
neoleor 8.29
moyou   4.57
bodyton 4.3
skinity 3.56
grace   1.69
cosima  0.7
ovale   0.56
Time taken: 34.567 seconds, Fetched: 153 row(s)
hive> with sales as (select brand, SUM(CASE WHEN month(event_time)=10 THEN price END)  sale_oct,
    > SUM(CASE WHEN month(event_time)=11 THEN price END)  sale_nov
    > from data_optimized
    > where event_type='purchase'
    > group by brand)
    > select sales.brand,round((sales.sale_nov-sales.sale_oct),2) as increase_in_sales
    > from sales where round((sales.sale_nov-sales.sale_oct),2)>0
    > order by increase_in_sales desc;
```

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

```
SELECT user_id, SUM(price) AS total_spent
FROM data_optimized
WHERE event_type = "purchase"
GROUP BY user_id
ORDER BY total_spent DESC
LIMIT 10 ;
```



```
Query ID = hadoop_20210228171617_08ace666-3bbc-477d-8517-f09077d4179d
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1614529711447_0005)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      3          3        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 15.05 s
----------------------------------------------------------------------------------------------
OK
runail  148297.94000000722
Time taken: 15.716 seconds, Fetched: 1 row(s)
hive> SELECT user_id, SUM(price) AS total_spent
    > FROM data_optimized
    > WHERE event_type = "purchase"
    > GROUP BY user_id
    > ORDER BY total_spent DESC
    > LIMIT 10 ;
Query ID = hadoop_20210228171810_f93ed002-c577-45d9-a774-9bd4dac2663d
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1614529711447_0005)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      3          3        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 16.23 s
----------------------------------------------------------------------------------------------
OK
557790271       2715.8699999999935
150318419       1645.9699999999998
562167663       1352.8500000000004
531900924       1329.4499999999998
557850743       1295.4799999999998
522130011       1185.3899999999996
561592095       1109.7
431950134       1097.5899999999997
566576008       1056.3600000000004
521347209       1040.91
Time taken: 16.988 seconds, Fetched: 10 row(s)
hive> 
```