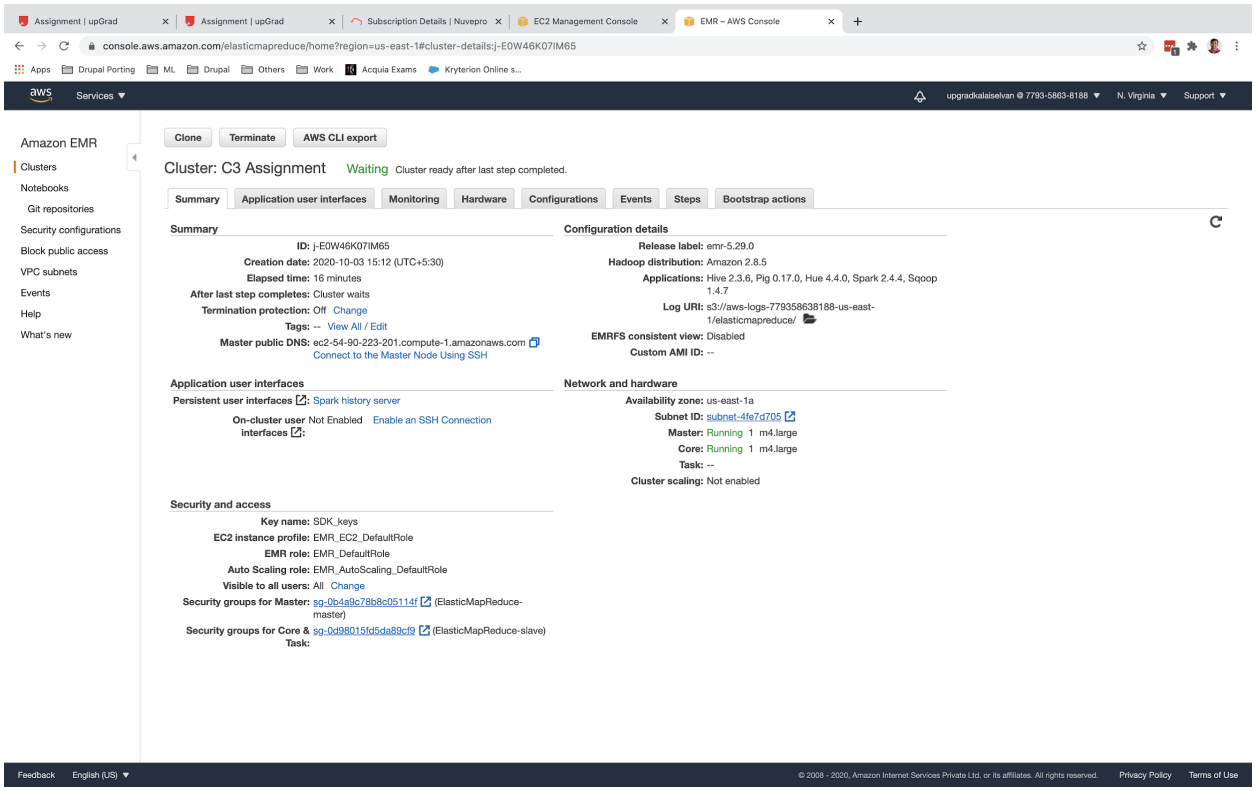
Step1: Copying the data set into the HDFS

A EMR cluster has been created with the same configuration as done for previous case study. PFB screenshot of the cluster.




```
Last login: Sat Oct  3 14:39:09 on console
kalaiselvanswamy@KalaiselvansMBP ~ % ssh -i ~/.ssh/SDK_keys.pem hadoop@ec2-54-90-223-201.compute-1.amazonaws.com
ssh: connect to host ec2-54-90-223-201.compute-1.amazonaws.com port 22: Operation timed out
kalaiselvanswamy@KalaiselvansMBP ~ % ssh -i ~/.ssh/SDK_keys.pem hadoop@ec2-54-90-223-201.compute-1.amazonaws.com
The authenticity of host 'ec2-54-90-223-201.compute-1.amazonaws.com (54.90.223.201)' can't be established.
ECDSA key fingerprint is SHA256:J9WL0v3QMINA6TwRtCMDotjeNkgiOavum51eg/z1PQQ.
Are you sure you want to continue connecting (yes/no/[fingerprint])? yes
Warning: Permanently added 'ec2-54-90-223-201.compute-1.amazonaws.com,54.90.223.201' (ECDSA) to the list of known hosts.
Last login: Sat Oct  3 09:54:20 2020

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

https://aws.amazon.com/amazon-linux-ami/2018.03-release-notes/
44 package(s) needed for security, out of 73 available
Run "sudo yum update" to apply all updates.
-bash: warning: setlocale: LC_CTYPE: cannot change locale (UTF-8): No such file or directory

EEEEEEEEEEEEEEEEEEE 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
EEEEEEEEEEEEEEEEEEE MMMMMMM             MMMMMMM RRRRRRR      RRRRRR

[hadoop@ip-172-31-18-181 ~]$
```

Now let us move the data from S3 to the EMR using hadoop distributed file copy. The command for the same is given below followed by the screenshots of the copy for both files.

```
hadoop  distcp   s3n://e-commerce-events-ml/2019-Oct.csv
/user/hadoop/e-commerce-events/2019-Oct.csv
hadoop  distcp   s3n://e-commerce-events-ml/2019-Nov.csv
/user/hadoop/e-commerce-events/2019-Nov.csv
```

```
   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
EEEEEEEEEEEEEEEEEEE MMMMMMM             MMMMMMM RRRRRRR     RRRRRR

[hadoop@ip-172-31-18-181 ~]$ hadoop distcp s3n://e-commerce-events-ml/2019-Oct.csv /user/hadoop/e-commerce-events/2019-Oct.csv
20/10/03 10:02:23 INFO tools.DistCp: Input Options: DistCpOptions{atomicCommit=false, syncFolder=false, deleteMissing=false, ignoreFailures=false, overwrite=false, skipCRC=false, blocking=true, numListstatusThreads=0, maxMaps=20, mapBandwidth=100, sslCon
figurationFile='null', copyStrategy='uniformsize', preserveStatus=[], preserveRawXattrs=false, atomicWorkPath=null, logPath=null, sourceFileListing=null, sourcePaths=[s3n://e-commerce-events-ml/2019-Oct.csv], targetPath=/user/hadoop/e-commerce-events/201
9-Oct.csv, targetPathExists=false, filtersFile='null'}
20/10/03 10:02:23 INFO client.RMProxy: Connecting to ResourceManager at ip-172-31-18-181.ec2.internal/172.31.18.181:8032
20/10/03 10:02:27 INFO tools.SimpleCopyListing: Paths (files+dirs) cnt = 1; dirCnt = 0
20/10/03 10:02:27 INFO tools.SimpleCopyListing: Build file listing completed.
20/10/03 10:02:27 INFO Configuration.deprecation: io.sort.mb is deprecated. Instead, use mapreduce.task.io.sort.mb
20/10/03 10:02:27 INFO Configuration.deprecation: io.sort.factor is deprecated. Instead, use mapreduce.task.io.sort.factor
20/10/03 10:02:27 INFO tools.DistCp: Number of paths in the copy list: 1
20/10/03 10:02:27 INFO tools.DistCp: Number of paths in the copy list: 1
20/10/03 10:02:27 INFO client.RMProxy: Connecting to ResourceManager at ip-172-31-18-181.ec2.internal/172.31.18.181:8032
20/10/03 10:02:27 INFO mapreduce.JobSubmitter: number of splits:1
20/10/03 10:02:28 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1601718673301_0001
20/10/03 10:02:28 INFO impl.YarnClientImpl: Submitted application application_1601718673301_0001
20/10/03 10:02:28 INFO mapreduce.Job: The url to track the job: http://ip-172-31-18-181.ec2.internal:20888/proxy/application_1601718673301_0001/
20/10/03 10:02:28 INFO tools.DistCp: DistCp job-id: job_1601718673301_0001
20/10/03 10:02:28 INFO mapreduce.Job: Running job: job_1601718673301_0001
20/10/03 10:02:38 INFO mapreduce.Job: Job job_1601718673301_0001 running in uber mode : false
20/10/03 10:02:38 INFO mapreduce.Job:  map 0% reduce 0%
20/10/03 10:02:56 INFO mapreduce.Job:  map 100% reduce 0%
20/10/03 10:02:57 INFO mapreduce.Job: Job job_1601718673301_0001 completed successfully
20/10/03 10:02:57 INFO mapreduce.Job: Counters: 38
        File System Counters
                FILE: Number of bytes read=0
                FILE: Number of bytes written=172886
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=360
                HDFS: Number of bytes written=482542278
                HDFS: Number of read operations=12
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=4
                S3N: Number of bytes read=482542278
                S3N: Number of bytes written=0
                S3N: Number of read operations=0
                S3N: Number of large read operations=0
                S3N: Number of write operations=0
        Job Counters
                Launched map tasks=1
                Other local map tasks=1
                Total time spent by all maps in occupied slots (ms)=510112
                Total time spent by all reduces in occupied slots (ms)=0
                Total time spent by all map tasks (ms)=15941
                Total vcore-milliseconds taken by all map tasks=15941
                Total megabyte-milliseconds taken by all map tasks=16323584
        Map-Reduce Framework
                Map input records=1
                Map output records=0
                Input split bytes=136
                Spilled Records=0
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=278
                CPU time spent (ms)=17300
                Physical memory (bytes) snapshot=456630272
                Virtual memory (bytes) snapshot=3285282816
                Total committed heap usage (bytes)=366477312
        File Input Format Counters
                Bytes Read=224
        File Output Format Counters
                Bytes Written=0
        DistCp Counters
                Bytes Copied=482542278
                Bytes Expected=482542278
                Files Copied=1
[hadoop@ip-172-31-18-181 ~]$
```

```
        File Output Format Counters
                Bytes Written=0
        DistCp Counters
                Bytes Copied=482542278
                Bytes Expected=482542278
                Files Copied=1
[hadoop@ip-172-31-18-181 ~]$ hadoop distcp s3n://e-commerce-events-ml/2019-Nov.csv  /user/hadoop/e-commerce-events/2019-Nov.csv
20/10/03 10:06:18 INFO tools.DistCp: Input Options: DistCpOptions{atomicCommit=false, syncFolder=false, deleteMissing=false, ignoreFailures=false, overwrite=false, skipCRC=false, blocking=true, numListstatusThreads=0, maxMaps=20, mapBandwidth=100, sslConfigurationFile='null', copyStrategy='uniformsize', preserveStatus=[], preserveRawXattrs=false, atomicWorkPath=null, logPath=null, sourceFileListing=null, sourcePaths=[s3n://e-commerce-events-ml/2019-Nov.csv], targetPath=/user/hadoop/e-commerce-events/2019-Nov.csv, targetPathExists=false, filtersFile='null'}
20/10/03 10:06:18 INFO client.RMProxy: Connecting to ResourceManager at ip-172-31-18-181.ec2.internal/172.31.18.181:8032
20/10/03 10:06:24 INFO tools.SimpleCopyListing: Paths (files+dirs) cnt = 1; dirCnt = 0
20/10/03 10:06:24 INFO tools.SimpleCopyListing: Build file listing completed.
20/10/03 10:06:24 INFO Configuration.deprecation: io.sort.mb is deprecated. Instead, use mapreduce.task.io.sort.mb
20/10/03 10:06:24 INFO Configuration.deprecation: io.sort.factor is deprecated. Instead, use mapreduce.task.io.sort.factor
20/10/03 10:06:24 INFO tools.DistCp: Number of paths in the copy list: 1
20/10/03 10:06:24 INFO tools.DistCp: Number of paths in the copy list: 1
20/10/03 10:06:24 INFO client.RMProxy: Connecting to ResourceManager at ip-172-31-18-181.ec2.internal/172.31.18.181:8032
20/10/03 10:06:24 INFO mapreduce.JobSubmitter: number of splits:1
20/10/03 10:06:24 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1601718673301_0002
20/10/03 10:06:25 INFO impl.YarnClientImpl: Submitted application application_1601718673301_0002
20/10/03 10:06:25 INFO mapreduce.Job: The url to track the job: http://ip-172-31-18-181.ec2.internal:20888/proxy/application_1601718673301_0002/
20/10/03 10:06:25 INFO tools.DistCp: DistCp job-id: job_1601718673301_0002
20/10/03 10:06:25 INFO mapreduce.Job: Running job: job_1601718673301_0002
20/10/03 10:06:34 INFO mapreduce.Job: Job job_1601718673301_0002 running in uber mode : false
20/10/03 10:06:34 INFO mapreduce.Job:  map 0% reduce 0%
20/10/03 10:06:51 INFO mapreduce.Job:  map 100% reduce 0%
20/10/03 10:06:53 INFO mapreduce.Job: Job job_1601718673301_0002 completed successfully
20/10/03 10:06:53 INFO mapreduce.Job: Counters: 38
        File System Counters
                FILE: Number of bytes read=0
                FILE: Number of bytes written=172886
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=360
                HDFS: Number of bytes written=545839412
                HDFS: Number of read operations=12
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=4
                S3N: Number of bytes read=545839412
                S3N: Number of bytes written=0
                S3N: Number of read operations=0
                S3N: Number of large read operations=0
                S3N: Number of write operations=0
        Job Counters
                Launched map tasks=1
                Other local map tasks=1
                Total time spent by all maps in occupied slots (ms)=553440
                Total time spent by all reduces in occupied slots (ms)=0
                Total time spent by all map tasks (ms)=17295
                Total vcore-milliseconds taken by all map tasks=17295
                Total megabyte-milliseconds taken by all map tasks=17710080
        Map-Reduce Framework
                Map input records=1
                Map output records=0
                Input split bytes=136
                Spilled Records=0
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=277
                CPU time spent (ms)=19770
                Physical memory (bytes) snapshot=54082764B
                Virtual memory (bytes) snapshot=3294130176
                Total committed heap usage (bytes)=486014976
        File Input Format Counters
                Bytes Read=224
        File Output Format Counters
                Bytes Written=0
        DistCp Counters
                Bytes Copied=545839412
                Bytes Expected=545839412
                Files Copied=1
[hadoop@ip-172-31-18-181 ~]$
```

```
[hadoop@ip-172-31-18-181 ~]$ hadoop fs -ls e-commerce-events
Found 2 items
-rw-r--r--   1 hadoop hadoop  545839412 2020-10-03 10:06 e-commerce-events/2019-Nov.csv
-rw-r--r--   1 hadoop hadoop  482542278 2020-10-03 10:02 e-commerce-events/2019-Oct.csv
[hadoop@ip-172-31-18-181 ~]$ hadoop fs -cat e-commerce-events/2019-Nov.csv | head
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
2019-11-01 00:00:24 UTC,remove_from_cart,5826182,1487580007483048900,,,3.33,553329724,2067216c-31b5-455d-a1cc-af0575a34ffb
2019-11-01 00:00:24 UTC,remove_from_cart,5826182,1487580007483048900,,,3.33,553329724,2067216c-31b5-455d-a1cc-af0575a34ffb
2019-11-01 00:00:25 UTC,view,5856189,1487580009026551821,,runail,15.71,562076640,09fafd6c-6c99-46b1-834f-33527f4de241
2019-11-01 00:00:32 UTC,view,5837835,1933472286753424063,,,3.49,514649199,432a4e95-375c-4b40-bd36-0fc039e77580
2019-11-01 00:00:34 UTC,remove_from_cart,5870838,1487580007675986893,,milv,0.79,429913900,2f0bff3c-252f-4fe6-afcd-5d8a6a92839a
cat: Unable to write to output stream.
[hadoop@ip-172-31-18-181 ~]$ hadoop fs -cat e-commerce-events/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.
```

Step 2: Creating the database and launching Hive queries on your EMR cluster
Let us first create a database with some comments.

```
create database if not exists sks_c3_assignment comment "Kalaiselvan C3 assignment Database";
use sks_c3_assignment;
```

```
[hadoop@ip-172-31-18-181 ~]$ hive

Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j2.properties Async: false
hive> create database if not exists sks_c3_assignment comment "Kalaiselvan C3 assignment Database";
OK
Time taken: 0.752 seconds
hive> use sks_c3_assignment;
OK
Time taken: 0.051 seconds
hive> show databases;
OK
default
sks_c3_assignment
Time taken: 0.178 seconds, Fetched: 2 row(s)
hive>
```

Set Hive CLI headers:

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

Create table for CSV delimited by ‘,’ and lines separated by newline. This is an ordinary table without any special storage format and without any partition.

NOTE: I am having event_time as string as there are string values like UTC in the file as shown above.

```
create table if not exists ecom_events_2019 (event_time string, event_type string, product_id string, category_id string, category_code string, brand string, price float, user_id bigint, user_session string) row format delimited fields terminated by ‘,’ lines terminated by '\n' tblproperties("skip.header.line.count"="1");
```

```
hive> set hive.cli.print.header=true;
hive> create table if not exists ecom_events_2019 (event_time string, event_type string, product_id string, category_id string, category_code string, brand string, price float, user_id bigint, user_
session string) row format delimited fields terminated by ',' lines terminated by '\n' tblproperties("skip.header.line.count"="1");
OK
Time taken: 0.634 seconds
hive> show tables;
OK
tab_name
ecom_events_2019
Time taken: 0.055 seconds, Fetched: 1 row(s)
hive>
```

Now let us load both the data into the same table as there are queries which require data from both files.

```
load data inpath 'e-commerce-events/2019-Oct.csv' into table
ecom_events_2019;
load data inpath 'e-commerce-events/2019-Nov.csv' into table
ecom_events_2019;
```

```
hive> load data inpath 'e-commerce-events/2019-Oct.csv' into table ecom_events_2019;
Loading data to table sks_c3_assignment.ecom_events_2019
OK
Time taken: 0.87 seconds
hive> load data inpath 'e-commerce-events/2019-Nov.csv' into table ecom_events_2019;
Loading data to table sks_c3_assignment.ecom_events_2019
OK
Time taken: 0.637 seconds
```

Now let us switch to beeline and we can see the data has been loaded into the Hive table successfully in correct order.

```
[hadoop@ip-172-31-10-181 ~]$ beeline -u jdbc:hive2://localhost:10000/default -n hadoop
Connecting to jdbc:hive2://localhost:10000/default
Connected to: Apache Hive (version 2.3.6-amzn-1)
Driver: Hive JDBC (version 2.3.6-amzn-1)
Transaction isolation: TRANSACTION_REPEATABLE_READ
Beeline version 2.3.6-amzn-1 by Apache Hive
0: jdbc:hive2://localhost:10000/default> show databases;
INFO  : Compiling command(queryId=hive_20201003101821_c5c3ce90-4a46-40d2-8f18-e934cfe9c387): show databases
INFO  : Semantic Analysis Completed
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:database_name, type:string, comment:from deserializer)], properties:null)
INFO  : Completed compiling command(queryId=hive_20201003101821_c5c3ce90-4a46-40d2-8f18-e934cfe9c387); Time taken: 0.922 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20201003101821_c5c3ce90-4a46-40d2-8f18-e934cfe9c387): show databases
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing command(queryId=hive_20201003101821_c5c3ce90-4a46-40d2-8f18-e934cfe9c387); Time taken: 0.042 seconds
INFO  : OK
+----------------+
|  database_name |
+----------------+
| default        |
| sks_c3_assignment  |
+----------------+
2 rows selected (1.358 seconds)
0: jdbc:hive2://localhost:10000/default> use sks_c3_assignment;
INFO  : Compiling command(queryId=hive_20201003101836_d15aab99-1bf2-4619-8385-8088aa69233d): use sks_c3_assignment
INFO  : Semantic Analysis Completed
INFO  : Returning Hive schema: Schema(fieldSchemas:null, properties:null)
INFO  : Completed compiling command(queryId=hive_20201003101836_d15aab99-1bf2-4619-8385-8088aa69233d); Time taken: 0.02 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20201003101836_d15aab99-1bf2-4619-8385-8088aa69233d): use sks_c3_assignment
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing command(queryId=hive_20201003101836_d15aab99-1bf2-4619-8385-8088aa69233d); Time taken: 0.012 seconds
INFO  : OK
No rows affected (0.072 seconds)
0: jdbc:hive2://localhost:10000/default> show tables;
INFO  : Compiling command(queryId=hive_20201003101844_4d4d33aa-7cd2-4b80-a13d-5f0ca3e6c72e): show tables
INFO  : Semantic Analysis Completed
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:tab_name, type:string, comment:from deserializer)], properties:null)
INFO  : Completed compiling command(queryId=hive_20201003101844_4d4d33aa-7cd2-4b80-a13d-5f0ca3e6c72e); Time taken: 0.009 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20201003101844_4d4d33aa-7cd2-4b80-a13d-5f0ca3e6c72e): show tables
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing command(queryId=hive_20201003101844_4d4d33aa-7cd2-4b80-a13d-5f0ca3e6c72e); Time taken: 0.048 seconds
INFO  : OK
+--------------------+
|      tab_name      |
+--------------------+
| ecom_events_2019  |
+--------------------+
1 row selected (0.098 seconds)
0: jdbc:hive2://localhost:10000/default> select * from ecom_events_2019 limit 5;
INFO  : Compiling command(queryId=hive_20201003101858_fa451c13-ebdb-41dc-a73b-45a91ead9f51): select * from ecom_events_2019 limit 5
INFO  : Semantic Analysis Completed
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:ecom_events_2019.event_time, type:string, comment:null), FieldSchema(name:ecom_events_2019.event_type, type:string, comment:null), FieldSchema(name:ecom_events_2019.product_id, type:string, comment:null), FieldSchema(name:ecom_e
vents_2019.category_id, type:string, comment:null), FieldSchema(name:ecom_events_2019.category_code, type:string, comment:null), FieldSchema(name:ecom_events_2019.brand, type:string, comment:null), FieldSchema(name:ecom_events_2019.price, type:float, comment:null), FieldSchema(name:ecom_events_20
19.user_id, type:bigint, comment:null), FieldSchema(name:ecom_events_2019.user_session, type:string, comment:null)], properties:null)
INFO  : Completed compiling command(queryId=hive_20201003101858_fa451c13-ebdb-41dc-a73b-45a91ead9f51); Time taken: 1.961 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20201003101858_fa451c13-ebdb-41dc-a73b-45a91ead9f51): select * from ecom_events_2019 limit 5
INFO  : Completed executing command(queryId=hive_20201003101858_fa451c13-ebdb-41dc-a73b-45a91ead9f51); Time taken: 0.002 seconds
INFO  : OK
+-----------------------------+-----------------------------+-------------------------------+--------------------------------+---------------------------------+-----------------------------+--------------------------+-----------------------------+--------------------------------------+
| ecom_events_2019.event_time | ecom_events_2019.event_type | ecom_events_2019.product_id | ecom_events_2019.category_id | ecom_events_2019.category_code | ecom_events_2019.brand | ecom_events_2019.price | ecom_events_2019.user_id |     ecom_events_2019.user_session    |
+-----------------------------+-----------------------------+-------------------------------+--------------------------------+---------------------------------+-----------------------------+--------------------------+-----------------------------+--------------------------------------+
| 2019-11-01 00:00:02 UTC     | view                        | 5802432                       | 1487580009286598681            |                                 |                             | 0.32                     | 562076640                   | 09fafd6c-6c99-46b1-834f-33527fade241 |
| 2019-11-01 00:00:09 UTC     | cart                        | 50w4397                       | 1487580006317032337            |                                 |                             | 2.38                     | 553329724                   | 20c7216c-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                   | 20c7216c-31b5-455d-a1cc-af0575a34ffb |
+-----------------------------+-----------------------------+-------------------------------+--------------------------------+---------------------------------+-----------------------------+--------------------------+-----------------------------+--------------------------------------+
5 rows selected (2.208 seconds)
0: jdbc:hive2://localhost:10000/default>
```

Now let us look at the questions that need to be answered.

Question 1:

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

Here we need to find the sum of the total price of all products which has an event as purchase and month is October. We can extract the month from the timestamp string and match it to 10th month.

Query:

```
select sum(price) as OctPurchase from ecom_events_2019 where
event_type = 'purchase' and month(event_time) = 10;
```

Screenshot:

```
0: jdbc:hive2://localhost:10000/default> select sum(price) as OctPurchase from ecom_events_2019 where event_type = 'purchase' and month(event_time) = 10;
INFO  : Compiling command(queryId=hive_20201003102601_8d64140a-56df-445c-a812-b441d8d9d885): select sum(price) as OctPurchase from ecom_events_2019 where event_type = 'purchase' and month(event_time) = 10
INFO  : Semantic Analysis Completed
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:octpurchase, type:double, comment:null)], properties:null)
INFO  : Completed compiling command(queryId=hive_20201003102601_8d64140a-56df-445c-a812-b441d8d9d885); Time taken: 0.898 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20201003102601_8d64140a-56df-445c-a812-b441d8d9d885): select sum(price) as OctPurchase from ecom_events_2019 where event_type = 'purchase' and month(event_time) = 10
INFO  : Query ID = hive_20201003102601_8d64140a-56df-445c-a812-b441d8d9d885
INFO  : Total jobs = 1
INFO  : Launching Job 1 out of 1
INFO  : Starting task [Stage-1:MAPRED] in serial mode
INFO  : Tez session hasn't been created yet. Opening session
INFO  : Dag name: select sum(price) as OctPurchase from e...10(Stage-1)
INFO  : Status: Running (Executing on YARN cluster with App id application_1601718673301_0004)

INFO  : Map 1: -/-      Reducer 2: 0/1
INFO  : Map 1: 0/2      Reducer 2: 0/1
INFO  : Map 1: 0/2      Reducer 2: 0/1
INFO  : Map 1: 0/2      Reducer 2: 0/1
INFO  : Map 1: 0(+1)/2  Reducer 2: 0/1
INFO  : Map 1: 0(+2)/2  Reducer 2: 0/1
INFO  : Map 1: 0(+2)/2  Reducer 2: 0/1
INFO  : Map 1: 0(+2)/2  Reducer 2: 0/1
INFO  : Map 1: 0(+2)/2  Reducer 2: 0/1
INFO  : Map 1: 0(+2)/2  Reducer 2: 0/1
INFO  : Map 1: 0(+1)/2  Reducer 2: 0/1
INFO  : Map 1: 1(+1)/2  Reducer 2: 0(+1)/1
INFO  : Map 1: 2/2      Reducer 2: 0(+1)/1
INFO  : Map 1: 2/2      Reducer 2: 1/1
INFO  : Completed executing command(queryId=hive_20201003102601_8d64140a-56df-445c-a812-b441d8d9d885); Time taken: 32.612 seconds
INFO  : OK
+---------------------+
|     octpurchase     |
+---------------------+
| 1211538.4295325726  |
+---------------------+
1 row selected (33.634 seconds)
0: jdbc:hive2://localhost:10000/default>
```

Partitioning:

As we can see most of the queries so far and also the forthcoming queries are focussed on the purchase event_type. So let us partition the data which has been purchased.

```
create table if not exists purchases_2019 (event_time string,
product_id string, category_id string, category_code string, brand
string, price float, user_id bigint, user_session string) partitioned
by (event_type string) row format delimited fields terminated by ','
lines terminated by '\n';
```

```
0: jdbc:hive2://localhost:10000/default> create table if not exists purchases_2019 (event_time string, product_id string, category_id string, category_code string, brand string, price float, user_id
 bigint, user_session string) partitioned by (event_type string) row format delimited fields terminated by ',' lines terminated by '\n';
INFO  : Compiling command(queryId=hive_20201003104012_1f7a3bf9-44ea-4546-a3c3-9d7e9c8203d0): create table if not exists purchases_2019 (event_time string, product_id string, category_id string, cate
gory_code string, brand string, price float, user_id bigint, user_session string) partitioned by (event_type string) row format delimited fields terminated by ',' lines terminated by '\n'
INFO  : Semantic Analysis Completed
INFO  : Returning Hive schema: Schema(fieldSchemas:null, properties:null)
INFO  : Completed compiling command(queryId=hive_20201003104012_1f7a3bf9-44ea-4546-a3c3-9d7e9c8203d0); Time taken: 0.018 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20201003104012_1f7a3bf9-44ea-4546-a3c3-9d7e9c8203d0): create table if not exists purchases_2019 (event_time string, product_id string, category_id string, cate
gory_code string, brand string, price float, user_id bigint, user_session string) partitioned by (event_type string) row format delimited fields terminated by ',' lines terminated by '\n'
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing command(queryId=hive_20201003104012_1f7a3bf9-44ea-4546-a3c3-9d7e9c8203d0); Time taken: 0.147 seconds
INFO  : OK
No rows affected (0.183 seconds)
```

```
insert into table purchases_2019 partition (event_type = 'purchase')
select event_time, product_id, category_id, category_code, brand,
```

```
price, user_id, user_session from ecom_events_2019 where event_type =
'purchase';
```

```
0: jdbc:hive2://localhost:10000/default> insert into table purchases_2019 partition (event_type = 'purchase') select event_time, product_id, category_id, category_code, brand, price, user_id, user_session from ecom_events_2019 where event_type = 'purchase';
INFO  : Compiling command(queryId=hive_20201003104252_447a03d3-87bb-4936-9503-4c7cb6a468af): insert into table purchases_2019 partition (event_type = 'purchase') select event_time, product_id, category_id, category_code, brand, price, user_id, user_session from ecom_events_2019 where event_type =
 'purchase'
INFO  : Semantic Analysis Completed
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:event_time, 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(n
ame: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, comment:null)], properties:null)
INFO  : Completed compiling command(queryId=hive_20201003104252_447a03d3-87bb-4936-9503-4c7cb6a468af); Time taken: 0.344 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20201003104252_447a03d3-87bb-4936-9503-4c7cb6a468af): insert into table purchases_2019 partition (event_type = 'purchase') select event_time, product_id, category_id, category_code, brand, price, user_id, user_session from ecom_events_2019 where event_type =
 'purchase'
INFO  : Query ID = hive_20201003104252_447a03d3-87bb-4936-9503-4c7cb6a468af
INFO  : Total jobs = 1
INFO  : Launching Job 1 out of 1
INFO  : Starting task [Stage-1:MAPRED] in serial mode
INFO  : Tez session hasn't been created yet. Opening session
INFO  : Dag name: insert into table purchases_201...'purchase'(Stage-1)
INFO  : Status: Running (Executing on YARN cluster with App id application_1601718673301_0006)

INFO  : Map 1: 0/2
INFO  : Map 1: 0/2
INFO  : Map 1: 0(+1)/2
INFO  : Map 1: 0(+2)/2
INFO  : Map 1: 0(+2)/2
INFO  : Map 1: 0(+2)/2
INFO  : Map 1: 0(+2)/2
INFO  : Map 1: 0(+2)/2
INFO  : Map 1: 1(+1)/2
INFO  : Map 1: 2/2
INFO  : Starting task [Stage-2:DEPENDENCY_COLLECTION] in serial mode
INFO  : Starting task [Stage-0:MOVE] in serial mode
INFO  : Loading data to table sks_c3_assignment.purchases_2019 partition (event_type=purchase) from hdfs://ip-172-31-18-181.ec2.internal:8020/user/hive/warehouse/sks_c3_assignment.db/purchases_2019/event_type=purchase/.hive-staging_hive_2020-10-03_10-42-52_050_2294195763986892544-5/-ext-10000
INFO  : Starting task [Stage-3:STATS] in serial mode
INFO  : Completed executing command(queryId=hive_20201003104252_447a03d3-87bb-4936-9503-4c7cb6a468af); Time taken: 30.454 seconds
INFO  : OK
No rows affected (30.817 seconds)
0: jdbc:hive2://localhost:10000/default> select * from purchases_2019 limit 5;
INFO  : Compiling command(queryId=hive_20201003104343_9c00d080-fbb1-4724-bd2f-5e5e62c30980): select * from purchases_2019 limit 5
INFO  : Semantic Analysis Completed
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:purchases_2019.event_time, type:string, comment:null), FieldSchema(name:purchases_2019.product_id, type:string, comment:null), FieldSchema(name:purchases_2019.category_id, type:string, comment:null), FieldSchema(name:purchases_2
019.category_code, type:string, comment:null), FieldSchema(name:purchases_2019.brand, type:string, comment:null), FieldSchema(name:purchases_2019.price, type:float, comment:null), FieldSchema(name:purchases_2019.user_id, type:bigint, comment:null), FieldSchema(name:purchases_2019.user_session, ty
pe:string, comment:null), FieldSchema(name:purchases_2019.event_type, type:string, comment:null)], properties:null)
INFO  : Completed compiling command(queryId=hive_20201003104343_9c00d080-fbb1-4724-bd2f-5e5e62c30980); Time taken: 0.254 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20201003104343_9c00d080-fbb1-4724-bd2f-5e5e62c30980): select * from purchases_2019 limit 5
INFO  : Completed executing command(queryId=hive_20201003104343_9c00d080-fbb1-4724-bd2f-5e5e62c30980); Time taken: 0.001 seconds
INFO  : OK
```

purchases_2019.event_time	purchases_2019.product_id	purchases_2019.category_id	purchases_2019.category_code	purchases_2019.brand	purchases_2019.price	purchases_2019.user_id	purchases_2019.user_session	purchases_2019.event_type
2019-11-01 00:01:57 UTC	5839412	1487580006551913373		lovely	3.16	460304619	9f777569-bdf3-47e5-a3d4-dfc26beb29cb	purchase
2019-11-01 00:01:57 UTC	5823969	1487580005268456287		uno	17.46	460304619	9f777569-bdf3-47e5-a3d4-dfc26beb29cb	purchase
2019-11-01 00:01:57 UTC	5810480	1487580011283087468			22.54	460304619	9f777569-bdf3-47e5-a3d4-dfc26beb29cb	purchase
2019-11-01 00:04:33 UTC	24380	1487580012994363565		depilflax	5.24	564451209	861ab2f1-b2e5-886f-a93b-5b067eff081f	purchase
2019-11-01 00:04:33 UTC	26765	1487580013522845895		ardell	7.16	564451209	861ab2f1-b2e5-886f-a93b-5b067eff081f	purchase

```
5 rows selected (0.32 seconds)
0: jdbc:hive2://localhost:10000/default>
```

Below is the screenshot of the partition created in HDFS.

```
[hadoop@ip-172-31-18-181 ~]$ hadoop fs -ls /user/hive/warehouse/
Found 1 items
drwxrwxrwt   - hadoop hadoop          0 2020-10-03 10:40 /user/hive/warehouse/sks_c3_assignment.db
[hadoop@ip-172-31-18-181 ~]$ hadoop fs -ls /user/hive/warehouse//sks_c3_assignment.db
Found 2 items
drwxrwxrwt   - hadoop hadoop          0 2020-10-03 10:16 /user/hive/warehouse/sks_c3_assignment.db/ecom_events_2019
drwxrwxrwt   - hadoop hadoop          0 2020-10-03 10:42 /user/hive/warehouse/sks_c3_assignment.db/purchases_2019
[hadoop@ip-172-31-18-181 ~]$ hadoop fs -ls /user/hive/warehouse/sks_c3_assignment.db/purchases_2019
Found 1 items
drwxrwxrwt   - hadoop hadoop          0 2020-10-03 10:43 /user/hive/warehouse/sks_c3_assignment.db/purchases_2019/event_type=purchase
[hadoop@ip-172-31-18-181 ~]$
```

Now let us again run the 1st question in the partitioned table to see the difference in time taken. With a non-partitioned table the time taken was 32.6 seconds.

```
0: jdbc:hive2://localhost:10000/default> select sum(price) as OctPurchase from purchases_2019 where month(event_time) = 10;
INFO  : Compiling command(queryId=hive_20201003105503_87435d76-d0c6-47ec-82c2-2f7a6f7b22bc): select sum(price) as OctPurchase from purchases_2019 where month(event_time) = 10
INFO  : Semantic Analysis Completed
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:octpurchase, type:double, comment:null)], properties:null)
INFO  : Completed compiling command(queryId=hive_20201003105503_87435d76-d0c6-47ec-82c2-2f7a6f7b22bc); Time taken: 0.368 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20201003105503_87435d76-d0c6-47ec-82c2-2f7a6f7b22bc): select sum(price) as OctPurchase from purchases_2019 where month(event_time) = 10
INFO  : Query ID = hive_20201003105503_87435d76-d0c6-47ec-82c2-2f7a6f7b22bc
INFO  : Total jobs = 1
INFO  : Launching Job 1 out of 1
INFO  : Starting task [Stage-1:MAPRED] in serial mode
INFO  : Tez session hasn't been created yet. Opening session
INFO  : Dag name: select sum(price) as OctPurchase from p...10(Stage-1)
INFO  : Status: Running (Executing on YARN cluster with App id application_1601718673301_0007)

INFO  : Map 1: 0/2      Reducer 2: 0/1
INFO  : Map 1: 0/2      Reducer 2: 0/1
INFO  : Map 1: 0(+2)/2  Reducer 2: 0/1
INFO  : Map 1: 0(+2)/2  Reducer 2: 0/1
INFO  : Map 1: 0(+2)/2  Reducer 2: 0/1
INFO  : Map 1: 1(+1)/2  Reducer 2: 0(+1)/1
INFO  : Map 1: 2/2      Reducer 2: 0(+1)/1
INFO  : Map 1: 2/2      Reducer 2: 0/1
INFO  : Map 1: 2/2      Reducer 2: 1/1
INFO  : Completed executing command(queryId=hive_20201003105503_87435d76-d0c6-47ec-82c2-2f7a6f7b22bc); Time taken: 24.322 seconds
INFO  : OK
+---------------------+
|     octpurchase     |
+---------------------+
| 1211538.4295325726  |
+---------------------+
1 row selected (24.753 seconds)
0: jdbc:hive2://localhost:10000/default>
```

As you can see the time taken with the partitioned table is only 24.3 seconds for the same query and same result.

Question 2:

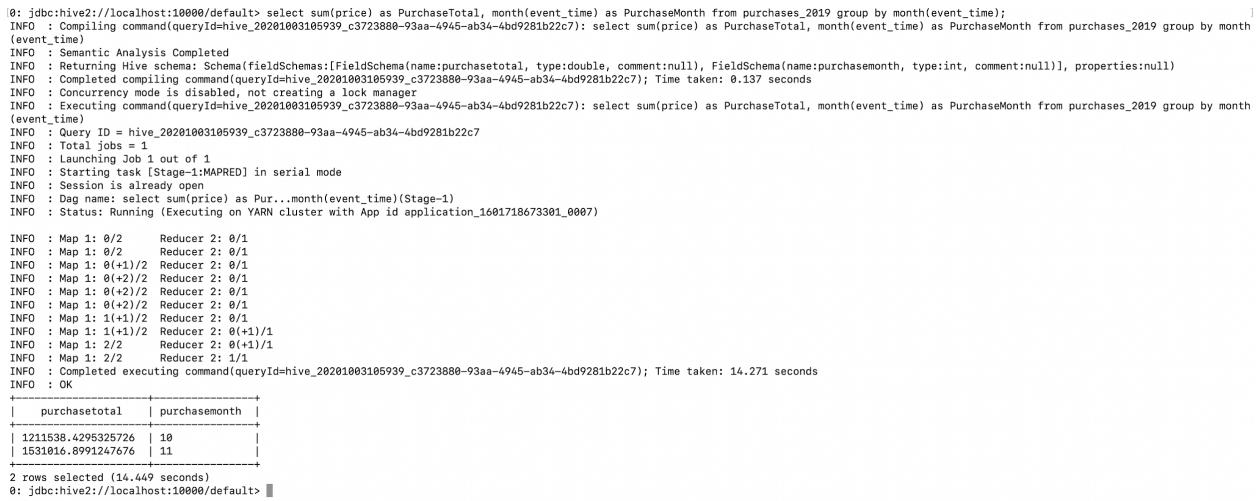
Write a query to yield the total sum of purchases per month in a single output.

Here since we need to show data/month let us group by based on month.

Query:

```
select sum(price) as PurchaseTotal, month(event_time) as
PurchaseMonth from purchases_2019 group by month(event_time);
```

Screenshot:



```
0: jdbc:hive2://localhost:10000/default> select sum(price) as PurchaseTotal, month(event_time) as PurchaseMonth from purchases_2019 group by month(event_time);
INFO  : Compiling command(queryId=hive_20201003105939_c3723880-93aa-4945-ab34-4bd9281b22c7): select sum(price) as PurchaseTotal, month(event_time) as PurchaseMonth from purchases_2019 group by month
(event_time)
INFO  : Semantic Analysis Completed
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:purchasetotal, type:double, comment:null), FieldSchema(name:purchasemonth, type:int, comment:null)], properties:null)
INFO  : Completed compiling command(queryId=hive_20201003105939_c3723880-93aa-4945-ab34-4bd9281b22c7); Time taken: 0.137 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20201003105939_c3723880-93aa-4945-ab34-4bd9281b22c7): select sum(price) as PurchaseTotal, month(event_time) as PurchaseMonth from purchases_2019 group by month
(event_time)
INFO  : Query ID = hive_20201003105939_c3723880-93aa-4945-ab34-4bd9281b22c7
INFO  : Total jobs = 1
INFO  : Launching Job 1 out of 1
INFO  : Starting task [Stage-1:MAPRED] in serial mode
INFO  : Session is already open
INFO  : Dag name: select sum(price) as Pur...month(event_time)(Stage-1)
INFO  : Status: Running (Executing on YARN cluster with App id application_1601718673301_0007)

INFO  : Map 1: 0/2      Reducer 2: 0/1
INFO  : Map 1: 0/2      Reducer 2: 0/1
INFO  : Map 1: 0(+1)/2  Reducer 2: 0/1
INFO  : Map 1: 0(+2)/2  Reducer 2: 0/1
INFO  : Map 1: 0(+2)/2  Reducer 2: 0/1
INFO  : Map 1: 0(+2)/2  Reducer 2: 0/1
INFO  : Map 1: 1(+1)/2  Reducer 2: 0/1
INFO  : Map 1: 1(+1)/2  Reducer 2: 0(+1)/1
INFO  : Map 1: 2/2      Reducer 2: 0(+1)/1
INFO  : Map 1: 2/2      Reducer 2: 1/1
INFO  : Completed executing command(queryId=hive_20201003105939_c3723880-93aa-4945-ab34-4bd9281b22c7); Time taken: 14.271 seconds
INFO  : OK
+--------------------+-----------------+
|    purchasetotal   | purchasemonth   |
+--------------------+-----------------+
| 1211538.4295325726 | 10              |
| 1531016.8991247676 | 11              |
+--------------------+-----------------+
2 rows selected (14.449 seconds)
0: jdbc:hive2://localhost:10000/default>
```

Question 3:

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

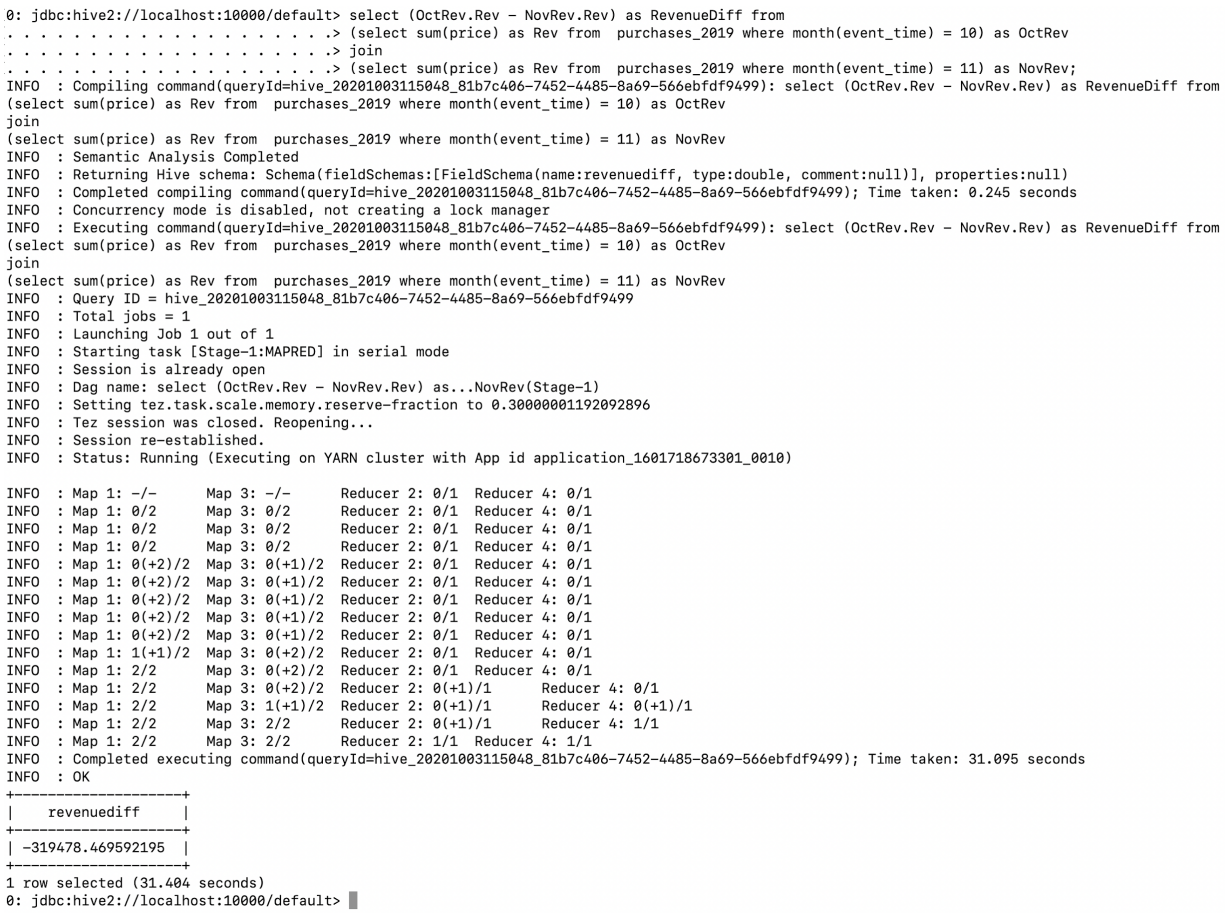
Here we need to find the difference in revenue between 2 months. We will use the below subquery with a cartesian product between 2 tables for this purpose.

NOTE: In order to run below query we need to set hive.strict.checks.cartesian.product property to false.

Query:

```
select (OctRev.Rev - NovRev.Rev) as RevenueDiff
from
(select sum(price) as Rev from  purchases_2019 where
month(event_time) = 10) as OctRev
join
(select sum(price) as Rev from  purchases_2019 where
month(event_time) = 11) as NovRev;
```

Screenshot:



```
0: jdbc:hive2://localhost:10000/default> select (OctRev.Rev - NovRev.Rev) as RevenueDiff from
. . . . . .> (select sum(price) as Rev from  purchases_2019 where month(event_time) = 10) as OctRev
. . . . . .> join
. . . . . .> (select sum(price) as Rev from  purchases_2019 where month(event_time) = 11) as NovRev;
INFO  : Compiling command(queryId=hive_20201003115048_81b7c406-7452-4485-8a69-566ebfdf9499): select (OctRev.Rev - NovRev.Rev) as RevenueDiff from
(select sum(price) as Rev from  purchases_2019 where month(event_time) = 10) as OctRev
join
(select sum(price) as Rev from  purchases_2019 where month(event_time) = 11) as NovRev
INFO  : Semantic Analysis Completed
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:revenuediff, type:double, comment:null)], properties:null)
INFO  : Completed compiling command(queryId=hive_20201003115048_81b7c406-7452-4485-8a69-566ebfdf9499); Time taken: 0.245 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20201003115048_81b7c406-7452-4485-8a69-566ebfdf9499): select (OctRev.Rev - NovRev.Rev) as RevenueDiff from
(select sum(price) as Rev from  purchases_2019 where month(event_time) = 10) as OctRev
join
(select sum(price) as Rev from  purchases_2019 where month(event_time) = 11) as NovRev
INFO  : Query ID = hive_20201003115048_81b7c406-7452-4485-8a69-566ebfdf9499
INFO  : Total jobs = 1
INFO  : Launching Job 1 out of 1
INFO  : Starting task [Stage-1:MAPRED] in serial mode
INFO  : Session is already open
INFO  : Dag name: select (OctRev.Rev - NovRev.Rev) as...NovRev(Stage-1)
INFO  : Setting tez.task.scale.memory.reserve-fraction to 0.30000001192092896
INFO  : Tez session was closed. Reopening...
INFO  : Session re-established.
INFO  : Status: Running (Executing on YARN cluster with App id application_1601718673301_0010)

INFO  : Map 1: -/-      Map 3: -/-      Reducer 2: 0/1  Reducer 4: 0/1
INFO  : Map 1: 0/2      Map 3: 0/2      Reducer 2: 0/1  Reducer 4: 0/1
INFO  : Map 1: 0/2      Map 3: 0/2      Reducer 2: 0/1  Reducer 4: 0/1
INFO  : Map 1: 0/2      Map 3: 0/2      Reducer 2: 0/1  Reducer 4: 0/1
INFO  : Map 1: 0(+2)/2  Map 3: 0(+1)/2  Reducer 2: 0/1  Reducer 4: 0/1
INFO  : Map 1: 0(+2)/2  Map 3: 0(+1)/2  Reducer 2: 0/1  Reducer 4: 0/1
INFO  : Map 1: 0(+2)/2  Map 3: 0(+1)/2  Reducer 2: 0/1  Reducer 4: 0/1
INFO  : Map 1: 0(+2)/2  Map 3: 0(+1)/2  Reducer 2: 0/1  Reducer 4: 0/1
INFO  : Map 1: 0(+2)/2  Map 3: 0(+1)/2  Reducer 2: 0/1  Reducer 4: 0/1
INFO  : Map 1: 1(+1)/2  Map 3: 0(+2)/2  Reducer 2: 0/1  Reducer 4: 0/1
INFO  : Map 1: 2/2      Map 3: 0(+2)/2  Reducer 2: 0/1  Reducer 4: 0/1
INFO  : Map 1: 2/2      Map 3: 0(+2)/2  Reducer 2: 0(+1)/1      Reducer 4: 0/1
INFO  : Map 1: 2/2      Map 3: 1(+1)/2  Reducer 2: 0(+1)/1      Reducer 4: 0(+1)/1
INFO  : Map 1: 2/2      Map 3: 2/2      Reducer 2: 0(+1)/1      Reducer 4: 1/1
INFO  : Map 1: 2/2      Map 3: 2/2      Reducer 2: 1/1  Reducer 4: 1/1
INFO  : Completed executing command(queryId=hive_20201003115048_81b7c406-7452-4485-8a69-566ebfdf9499); Time taken: 31.095 seconds
INFO  : OK
+--------------------+
|    revenuediff     |
+--------------------+
| -319478.469592195  |
+--------------------+
1 row selected (31.404 seconds)
0: jdbc:hive2://localhost:10000/default>
```

The minus value indicates that the revenue in November was more than the revenue in October and there was an 3,19,478 increase in revenue in November when compared to October

Time taken: 31 seconds

Further Optimizations:

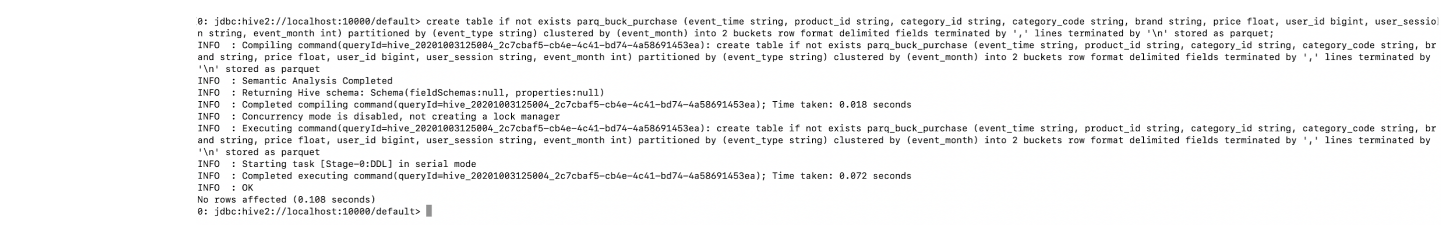
Also we can see that we need data based on a month. So let us create a new table and split the above partition into 2 buckets based on month(event_time).

For this in order to perform bucketing let us create another table ecom_events_month_2019 with month(event_time) as a separate column.

Let us use a dynamic partition this time. Also when we look at the operations required, they are more of a columnar operations. So let us store the table as parquet format.

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

```
create table if not exists parq_buck_purchase (event_time string,
product_id string, category_id string, category_code string, brand
string, price float, user_id bigint, user_session string,
event_month int) partitioned by (event_type string) clustered by
(event_month) into 2 buckets row format delimited fields terminated
by ‘,' lines terminated by '\n’ stored as parquet;
```



Now let us insert data into the final optimized table.

```
insert into parq_buck_purchase partition(event_type) select
event_time, product_id, category_id, category_code, brand, price,
user_id, user_session, event_month, event_type from
ecom_events_month_2019;
```

```
0: jdbc:hive2://localhost:10000/default> insert into parq_buck_purchase partition(event_type) select event_time, product_id, category_id, category_code, brand, price, user_id, user_session, event_month, event_type from ecom_events_month_2019;
INFO  : Compiling command(queryId=hive_20201003125138_f7a3b74b-da09-43f4-9a1d-a3f58368c2cc): insert into parq_buck_purchase partition(event_type) select event_time, product_id, category_id, category_code, brand, price, user_id, user_session, event_month, event_type from ecom_events_month_2019
INFO  : Semantic Analysis Completed
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:event_time, 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:pri
ce, type:float, comment:null), FieldSchema(name:user_id, type:bigint, comment:null), FieldSchema(name:user_session, type:string, comment:null), FieldSchema(name:event_month, type:int, comment:null), FieldSchema(name:event_type, type:string, comment:null)], properties:null)
INFO  : Completed compiling command(queryId=hive_20201003125138_f7a3b74b-da09-43f4-9a1d-a3f58368c2cc); Time taken: 0.141 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20201003125138_f7a3b74b-da09-43f4-9a1d-a3f58368c2cc): insert into parq_buck_purchase partition(event_type) select event_time, product_id, category_id, category_code, brand, price, user_id, user_session, event_month, event_type from ecom_events_month_2019
INFO  : Query ID = hive_20201003125138_f7a3b74b-da09-43f4-9a1d-a3f58368c2cc
INFO  : Total jobs = 1
INFO  : Launching Job 1 out of 1
INFO  : Starting task [Stage-1:MAPRED] in serial mode
INFO  : Session is already open
INFO  : Dag name: insert into parq_buck...om_events_month_2019(Stage-1)
INFO  : Status: Running (Executing on YARN cluster with App id application_1601718673301_0013)

INFO  : Map 1: 0/8      Reducer 2: 0/5
INFO  : Map 1: 0/8      Reducer 2: 0/5
INFO  : Map 1: 0/8      Reducer 2: 0/5
INFO  : Map 1: 0(+1)/8  Reducer 2: 0/5
INFO  : Map 1: 0(+2)/8  Reducer 2: 0/5
INFO  : Map 1: 0(+3)/8  Reducer 2: 0/5
INFO  : Map 1: 0(+3)/8  Reducer 2: 0/5
INFO  : Map 1: 0(+3)/8  Reducer 2: 0/5
INFO  : Map 1: 0(+3)/8  Reducer 2: 0/5
INFO  : Map 1: 0(+3)/8  Reducer 2: 0/5
INFO  : Map 1: 0(+3)/8  Reducer 2: 0/5
INFO  : Map 1: 0(+3)/8  Reducer 2: 0/5
INFO  : Map 1: 0(+3)/8  Reducer 2: 0/5
INFO  : Map 1: 1(+3)/8  Reducer 2: 0/5
INFO  : Map 1: 2(+2)/8  Reducer 2: 0/5
INFO  : Map 1: 3(+3)/8  Reducer 2: 0/5
INFO  : Map 1: 3(+3)/8  Reducer 2: 0/5
INFO  : Map 1: 3(+3)/8  Reducer 2: 0/5
INFO  : Map 1: 3(+3)/8  Reducer 2: 0/5
INFO  : Map 1: 3(+3)/8  Reducer 2: 0/5
INFO  : Map 1: 4(+3)/8  Reducer 2: 0/5
INFO  : Map 1: 5(+3)/8  Reducer 2: 0/5
INFO  : Map 1: 6(+2)/8  Reducer 2: 0(+1)/5
INFO  : Map 1: 6(+2)/8  Reducer 2: 0(+1)/5
INFO  : Map 1: 6(+2)/8  Reducer 2: 0(+1)/5
INFO  : Map 1: 6(+2)/8  Reducer 2: 0(+1)/5
INFO  : Map 1: 7(+1)/8  Reducer 2: 0(+1)/5
INFO  : Map 1: 7(+1)/8  Reducer 2: 0(+2)/5
INFO  : Map 1: 8/8      Reducer 2: 0(+3)/5
INFO  : Map 1: 8/8      Reducer 2: 0(+3)/5
INFO  : Map 1: 8/8      Reducer 2: 0(+3)/5
INFO  : Map 1: 8/8      Reducer 2: 0(+3)/5
INFO  : Map 1: 8/8      Reducer 2: 0(+3)/5
INFO  : Map 1: 8/8      Reducer 2: 0(+3)/5
INFO  : Map 1: 8/8      Reducer 2: 0(+3)/5
INFO  : Map 1: 8/8      Reducer 2: 0(+3)/5
INFO  : Map 1: 8/8      Reducer 2: 0(+3)/5
INFO  : Map 1: 8/8      Reducer 2: 0(+3)/5
INFO  : Map 1: 8/8      Reducer 2: 0(+3)/5
INFO  : Map 1: 8/8      Reducer 2: 0(+3)/5
INFO  : Map 1: 8/8      Reducer 2: 0(+3)/5
INFO  : Map 1: 8/8      Reducer 2: 0(+3)/5
INFO  : Map 1: 8/8      Reducer 2: 0(+3)/5
INFO  : Map 1: 8/8      Reducer 2: 0(+3)/5
INFO  : Map 1: 8/8      Reducer 2: 0(+3)/5
INFO  : Map 1: 8/8      Reducer 2: 0(+3)/5
INFO  : Map 1: 8/8      Reducer 2: 0(+3)/5
INFO  : Map 1: 8/8      Reducer 2: 0(+3)/5
INFO  : Map 1: 8/8      Reducer 2: 1(+3)/5
INFO  : Map 1: 8/8      Reducer 2: 2(+2)/5
INFO  : Map 1: 8/8      Reducer 2: 2(+3)/5
INFO  : Map 1: 8/8      Reducer 2: 3(+2)/5
INFO  : Map 1: 8/8      Reducer 2: 3(+2)/5
INFO  : Map 1: 8/8      Reducer 2: 3(+2)/5
INFO  : Map 1: 8/8      Reducer 2: 3(+2)/5
INFO  : Map 1: 8/8      Reducer 2: 4(+1)/5
INFO  : Map 1: 8/8      Reducer 2: 4(+1)/5
INFO  : Map 1: 8/8      Reducer 2: 4(+1)/5
INFO  : Map 1: 8/8      Reducer 2: 4(+1)/5
INFO  : Map 1: 8/8      Reducer 2: 5/5
INFO  : Starting task [Stage-2:DEPENDENCY_COLLECTION] in serial mode
INFO  : Starting task [Stage-0:MOVE] in serial mode
INFO  : Loading data to table sks_c3_assignment.parq_buck_purchase partition (event_type=null) from hdfs://ip-172-31-18-181.ec2.internal:8020/user/hive/warehouse/sks_c3_assignment.db/parq_buck_purchase/.hive-staging_hive_2020-10-03_12-51-38_574_3524779510099987357-7/-ext-10000
INFO  :

INFO  :          Time taken to load dynamic partitions: 0.272 seconds
INFO  :          Time taken for adding to write entity : 0.0 seconds
INFO  : Starting task [Stage-3:STATS] in serial mode
INFO  : Completed executing command(queryId=hive_20201003125138_f7a3b74b-da09-43f4-9a1d-a3f58368c2cc); Time taken: 140.61 seconds
INFO  : OK
No rows affected (140.778 seconds)
0: jdbc:hive2://localhost:10000/default>
```

```
0: jdbc:hive2://localhost:10000/default> show tables;
INFO  : Compiling command(queryId=hive_20201003125601_38a0dc71-9221-4360-9f63-b4043240c72f): show tables
INFO  : Semantic Analysis Completed
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:tab_name, type:string, comment:from deserializer)], properties:null)
INFO  : Completed compiling command(queryId=hive_20201003125601_38a0dc71-9221-4360-9f63-b4043240c72f); Time taken: 0.094 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20201003125601_38a0dc71-9221-4360-9f63-b4043240c72f): show tables
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing command(queryId=hive_20201003125601_38a0dc71-9221-4360-9f63-b4043240c72f); Time taken: 0.02 seconds
INFO  : OK
+-------------------------+
|        tab_name         |
+-------------------------+
| ecom_events_2019        |
| ecom_events_month_2019  |
| parq_buck_purchase      |
| purchases_2019          |
+-------------------------+
4 rows selected (0.126 seconds)
0: jdbc:hive2://localhost:10000/default> select * from parq_buck_purchase limit 5;
INFO  : Compiling command(queryId=hive_20201003125609_0ed968aa-1653-4f45-8c7f-e3eac20bb614): select * from parq_buck_purchase limit 5
INFO  : Semantic Analysis Completed
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:parq_buck_purchase.event_time, type:string, comment:null), FieldSchema(name:parq_buck_purchase.product_id, type:string, comment:null), FieldSchema(name:parq_buck_purchase.category_id, type:string, comment:null), FieldSchema(name:parq_buck_purchase.category_code, type:string, comment:nul
l), FieldSchema(name:parq_buck_purchase.brand, type:string, comment:null), FieldSchema(name:parq_buck_purchase.price, type:float, comment:null), FieldSchema(name:parq_buck_purchase.user_id, type:bigint, comment:null), FieldSchema(name:parq_buck_purchase.user_session, type:string, comment:null), FieldSchema(name:parq_buck_purchase.event_month, type:int, c
omment:null), FieldSchema(name:parq_buck_purchase.event_type, type:string, comment:null)], properties:null)
INFO  : Completed compiling command(queryId=hive_20201003125609_0ed968aa-1653-4f45-8c7f-e3eac20bb614); Time taken: 0.175 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20201003125609_0ed968aa-1653-4f45-8c7f-e3eac20bb614): select * from parq_buck_purchase limit 5
INFO  : Completed executing command(queryId=hive_20201003125609_0ed968aa-1653-4f45-8c7f-e3eac20bb614); Time taken: 0.001 seconds
INFO  : OK
| parq_buck_purchase.event_time | parq_buck_purchase.product_id | parq_buck_purchase.category_id | parq_buck_purchase.category_code | parq_buck_purchase.brand | parq_buck_purchase.price | parq_buck_purchase.user_id | parq_buck_purchase.user_session | parq_buck_purchase.event_month | parq_buck_purchase.event_type |
| 2019-10-26 22:15:11 UTC | 5698419 | 1487580007550157767 | | ingarden | 0.95 | 564519467 | 92f20ef7-05e7-418c-abc4-10ac89c099db | 10 | cart |
| 2019-10-28 08:45:02 UTC | 5767950 | 1487580005595612013 | | | 3.33 | 553964934 | a60501f6-febe-4c87-ac59-0a5d41b38479 | 10 | cart |
| 2019-10-28 17:04:54 UTC | 5809877 | 1602943681873052386 | | grattol | 4.76 | 565131871 | 8a4bf360-232e-4e9c-b740-3e0b52ac7f94 | 10 | cart |
| 2019-10-28 17:04:55 UTC | 5587652 | 1487580008145748965 | | roubloff | 3.49 | 516370119 | 24e68dc3-6d7a-4ce4-88b6-fd70c33876bc | 10 | cart |
| 2019-10-31 16:28:20 UTC | 5716844 | 1487580013950664926 | | | 10.32 | 491894039 | 51d0ddcf-7ffc-448c-ab39-6c60522dff0c | 10 | cart |
5 rows selected (0.891 seconds)
0: jdbc:hive2://localhost:10000/default>
```

We can see the data populated in the final optimized table parq_buck_purchase.

```
[hadoop@ip-172-31-18-181 ~]$ hadoop fs -ls /user/hive/warehouse/sks_c3_assignment.db/parq_buck_purchase
Found 4 items
drwxrwxrwt   - hadoop hadoop          0 2020-10-03 12:53 /user/hive/warehouse/sks_c3_assignment.db/parq_buck_purchase/event_type=cart
drwxrwxrwt   - hadoop hadoop          0 2020-10-03 12:53 /user/hive/warehouse/sks_c3_assignment.db/parq_buck_purchase/event_type=purchase
drwxrwxrwt   - hadoop hadoop          0 2020-10-03 12:53 /user/hive/warehouse/sks_c3_assignment.db/parq_buck_purchase/event_type=remove_from_cart
drwxrwxrwt   - hadoop hadoop          0 2020-10-03 12:53 /user/hive/warehouse/sks_c3_assignment.db/parq_buck_purchase/event_type=view
[hadoop@ip-172-31-18-181 ~]$
```

In the above table we can see 4 partitions created dynamically for each event type. Now let us check the buckets in each partition.

```
[hadoop@ip-172-31-18-181 ~]$ hadoop fs -ls /user/hive/warehouse/sks_c3_assignment.db/parq_buck_purchase/event_type=cart
Found 2 items
-rwxrwxrwt   1 hadoop hadoop   89313368 2020-10-03 12:53 /user/hive/warehouse/sks_c3_assignment.db/parq_buck_purchase/event_type=cart/000000_0
-rwxrwxrwt   1 hadoop hadoop   93813729 2020-10-03 12:53 /user/hive/warehouse/sks_c3_assignment.db/parq_buck_purchase/event_type=cart/000001_0
[hadoop@ip-172-31-18-181 ~]$ hadoop fs -ls /user/hive/warehouse/sks_c3_assignment.db/parq_buck_purchase/event_type=purchase
Found 2 items
-rwxrwxrwt   1 hadoop hadoop    6106454 2020-10-03 12:53 /user/hive/warehouse/sks_c3_assignment.db/parq_buck_purchase/event_type=purchase/000000_0
-rwxrwxrwt   1 hadoop hadoop   10210823 2020-10-03 12:52 /user/hive/warehouse/sks_c3_assignment.db/parq_buck_purchase/event_type=purchase/000001_0
[hadoop@ip-172-31-18-181 ~]$ hadoop fs -ls /user/hive/warehouse/sks_c3_assignment.db/parq_buck_purchase/event_type=remove_from_cart
Found 2 items
-rwxrwxrwt   1 hadoop hadoop   50038840 2020-10-03 12:53 /user/hive/warehouse/sks_c3_assignment.db/parq_buck_purchase/event_type=remove_from_cart/000000_0
-rwxrwxrwt   1 hadoop hadoop   63278928 2020-10-03 12:53 /user/hive/warehouse/sks_c3_assignment.db/parq_buck_purchase/event_type=remove_from_cart/000001_0
[hadoop@ip-172-31-18-181 ~]$ hadoop fs -ls /user/hive/warehouse/sks_c3_assignment.db/parq_buck_purchase/event_type=view
Found 2 items
-rwxrwxrwt   1 hadoop hadoop  146096722 2020-10-03 12:53 /user/hive/warehouse/sks_c3_assignment.db/parq_buck_purchase/event_type=view/000000_0
-rwxrwxrwt   1 hadoop hadoop  161165126 2020-10-03 12:53 /user/hive/warehouse/sks_c3_assignment.db/parq_buck_purchase/event_type=view/000001_0
[hadoop@ip-172-31-18-181 ~]$
```

As you can see above each partition is further divided into 2 buckets based on 2 months.

Now let us try running the 3rd question in the newly optimized table and check the time taken. We need to add condition here since the previous query was run on an already filtered table with event type as purchase.

```
select (OctRev.Rev - NovRev.Rev) as RevenueDiff
from
(select sum(price) as Rev from  parq_buck_purchase where
month(event_time) = 10 and event_type = 'purchase') as OctRev
join
(select sum(price) as Rev from  parq_buck_purchase where
month(event_time) = 11 and event_type = 'purchase') as NovRev;
```

```
0: jdbc:hive2://localhost:10000/default> select (OctRev.Rev - NovRev.Rev) as RevenueDiff
. . . . . .> from
. . . . . .> (select sum(price) as Rev from  parq_buck_purchase where month(event_time) = 10 and event_type = 'purchase') as OctRev
. . . . . .> join
. . . . . .> (select sum(price) as Rev from  parq_buck_purchase where month(event_time) = 11 and event_type = 'purchase') as NovRev;
INFO  : Compiling command(queryId=hive_20201003130835_cfb8a59c-7462-433f-9075-8afa5002daa4): select (OctRev.Rev - NovRev.Rev) as RevenueDiff
from
(select sum(price) as Rev from  parq_buck_purchase where month(event_time) = 10 and event_type = 'purchase') as OctRev
join
(select sum(price) as Rev from  parq_buck_purchase where month(event_time) = 11 and event_type = 'purchase') as NovRev
INFO  : Semantic Analysis Completed
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:revenuediff, type:double, comment:null)], properties:null)
INFO  : Completed compiling command(queryId=hive_20201003130835_cfb8a59c-7462-433f-9075-8afa5002daa4); Time taken: 0.684 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20201003130835_cfb8a59c-7462-433f-9075-8afa5002daa4): select (OctRev.Rev - NovRev.Rev) as RevenueDiff
from
(select sum(price) as Rev from  parq_buck_purchase where month(event_time) = 10 and event_type = 'purchase') as OctRev
join
(select sum(price) as Rev from  parq_buck_purchase where month(event_time) = 11 and event_type = 'purchase') as NovRev
INFO  : Query ID = hive_20201003130835_cfb8a59c-7462-433f-9075-8afa5002daa4
INFO  : Total jobs = 1
INFO  : Launching Job 1 out of 1
INFO  : Starting task [Stage-1:MAPRED] in serial mode
INFO  : Session is already open
INFO  : Dag name: select (OctRev.Rev - NovRev.Rev) as...NovRev(Stage-1)
INFO  : Setting tez.task.scale.memory.reserve-fraction to 0.30000001192092896
INFO  : Status: Running (Executing on YARN cluster with App id application_1601718673301_0014)

INFO  : Map 1: 0/1      Map 3: 0/1      Reducer 2: 0/1  Reducer 4: 0/1
INFO  : Map 1: 0/1      Map 3: 0/1      Reducer 2: 0/1  Reducer 4: 0/1
INFO  : Map 1: 0(+1)/1  Map 3: 0(+1)/1  Reducer 2: 0/1  Reducer 4: 0/1
INFO  : Map 1: 0(+1)/1  Map 3: 0(+1)/1  Reducer 2: 0/1  Reducer 4: 0/1
INFO  : Map 1: 0(+1)/1  Map 3: 0(+1)/1  Reducer 2: 0/1  Reducer 4: 0/1
INFO  : Map 1: 0(+1)/1  Map 3: 0(+1)/1  Reducer 2: 0/1  Reducer 4: 0/1
INFO  : Map 1: 0(+1)/1  Map 3: 1/1      Reducer 2: 0/1  Reducer 4: 0(+1)/1
INFO  : Map 1: 1/1      Map 3: 1/1      Reducer 2: 0(+1)/1      Reducer 4: 0(+1)/1
INFO  : Map 1: 1/1      Map 3: 1/1      Reducer 2: 0(+1)/1      Reducer 4: 1/1
INFO  : Map 1: 1/1      Map 3: 1/1      Reducer 2: 1/1  Reducer 4: 1/1
INFO  : Completed executing command(queryId=hive_20201003130835_cfb8a59c-7462-433f-9075-8afa5002daa4); Time taken: 17.143 seconds
INFO  : OK
+--------------------+
|    revenuediff     |
+--------------------+
| -319478.469592195  |
+--------------------+
1 row selected (17.859 seconds)
0: jdbc:hive2://localhost:10000/default>
```

Time taken : 17.86 seconds

As you can see above there is a 13 second difference between both queries even though they fetch the same data. The final optimized table with partitions and buckets stored as parquet fetches data faster than a table with just partitions.

Now let us proceed with answering rest of the questions with our most optimized table `parq_buck_purchase`

Question 4:
Find distinct categories of products.

Query:

```
select distinct(category_code) as categories from parq_buck_purchase;
```

Screenshot:

```
0: jdbc:hive2://localhost:10000/default> select distinct(category_code) as categories from parq_buck_purchase;
INFO  : Compiling command(queryId=hive_20201003131437_c879da2c-965a-472b-8939-33b7bfaf87cf): select distinct(category_code) as categories from parq_buck_purchase
INFO  : Semantic Analysis Completed
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:categories, type:string, comment:null)], properties:null)
INFO  : Completed compiling command(queryId=hive_20201003131437_c879da2c-965a-472b-8939-33b7bfaf87cf); Time taken: 0.115 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20201003131437_c879da2c-965a-472b-8939-33b7bfaf87cf): select distinct(category_code) as categories from parq_buck_purchase
INFO  : Query ID = hive_20201003131437_c879da2c-965a-472b-8939-33b7bfaf87cf
INFO  : Total jobs = 1
INFO  : Launching Job 1 out of 1
INFO  : Starting task [Stage-1:MAPRED] in serial mode
INFO  : Session is already open
INFO  : Dag name: select distinct(categor...parq_buck_purchase(Stage-1)
INFO  : Tez session was closed. Reopening...
INFO  : Session re-established.
INFO  : Status: Running (Executing on YARN cluster with App id application_1601718673301_0015)

INFO  : Map 1: -/-      Reducer 2: 0/1
INFO  : Map 1: 0/6      Reducer 2: 0/1
INFO  : Map 1: 0/6      Reducer 2: 0/1
INFO  : Map 1: 0/6      Reducer 2: 0/1
INFO  : Map 1: 0(+1)/6  Reducer 2: 0/1
INFO  : Map 1: 0(+3)/6  Reducer 2: 0/1
INFO  : Map 1: 0(+3)/6  Reducer 2: 0/1
INFO  : Map 1: 0(+3)/6  Reducer 2: 0/1
INFO  : Map 1: 0(+3)/6  Reducer 2: 0/1
INFO  : Map 1: 1(+2)/6  Reducer 2: 0/1
INFO  : Map 1: 1(+3)/6  Reducer 2: 0/1
INFO  : Map 1: 2(+2)/6  Reducer 2: 0/1
INFO  : Map 1: 3(+2)/6  Reducer 2: 0/1
INFO  : Map 1: 3(+3)/6  Reducer 2: 0/1
INFO  : Map 1: 3(+2)/6  Reducer 2: 0/1
INFO  : Map 1: 4(+2)/6  Reducer 2: 0(+1)/1
INFO  : Map 1: 6/6      Reducer 2: 0(+1)/1
INFO  : Map 1: 6/6      Reducer 2: 1/1
INFO  : Completed executing command(queryId=hive_20201003131437_c879da2c-965a-472b-8939-33b7bfaf87cf); Time taken: 31.051 seconds
INFO  : OK
+-----------------------------------------+
|               categories                |
+-----------------------------------------+
|                                         |
| 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                     |
+-----------------------------------------+
12 rows selected (31.229 seconds)
0: jdbc:hive2://localhost:10000/default>
```

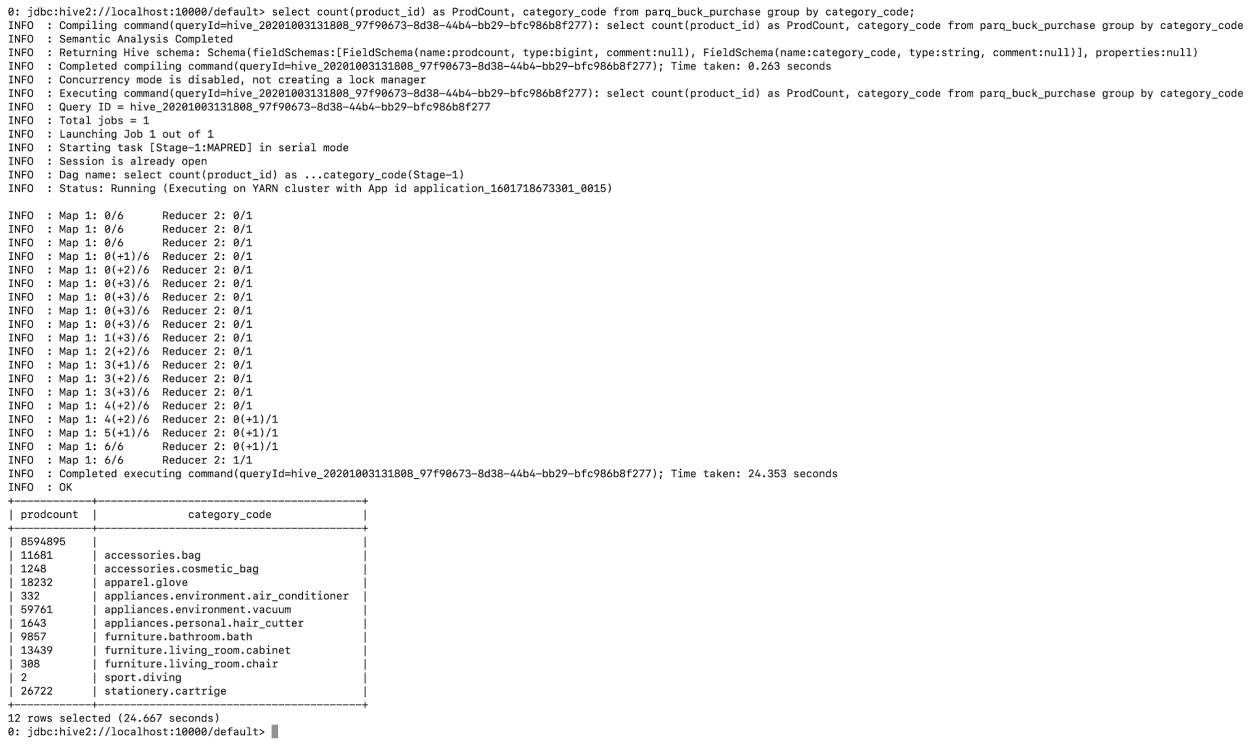
As you can see above there are 11 distinct categories and there are products with no category also.

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

Query:

```
select count(product_id) as ProdCount, category_code from
parq_buck_purchase group by category_code;
```

Screenshot:



```
0: jdbc:hive2://localhost:10000/default> select count(product_id) as ProdCount, category_code from parq_buck_purchase group by category_code;
INFO  : Compiling command(queryId=hive_20201003131808_97f90673-8d38-44b4-bb29-bfc986b8f277): select count(product_id) as ProdCount, category_code from parq_buck_purchase group by category_code
INFO  : Semantic Analysis Completed
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:prodcount, type:bigint, comment:null), FieldSchema(name:category_code, type:string, comment:null)], properties:null)
INFO  : Completed compiling command(queryId=hive_20201003131808_97f90673-8d38-44b4-bb29-bfc986b8f277); Time taken: 0.263 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20201003131808_97f90673-8d38-44b4-bb29-bfc986b8f277): select count(product_id) as ProdCount, category_code from parq_buck_purchase group by category_code
INFO  : Query ID = hive_20201003131808_97f90673-8d38-44b4-bb29-bfc986b8f277
INFO  : Total jobs = 1
INFO  : Launching Job 1 out of 1
INFO  : Starting task [Stage-1:MAPRED] in serial mode
INFO  : Session is already open
INFO  : Dag name: select count(product_id) as ...category_code(Stage-1)
INFO  : Status: Running (Executing on YARN cluster with App id application_1601718673301_0015)

INFO  : Map 1: 0/6      Reducer 2: 0/1
INFO  : Map 1: 0/6      Reducer 2: 0/1
INFO  : Map 1: 0/6      Reducer 2: 0/1
INFO  : Map 1: 0(+1)/6  Reducer 2: 0/1
INFO  : Map 1: 0(+2)/6  Reducer 2: 0/1
INFO  : Map 1: 0(+3)/6  Reducer 2: 0/1
INFO  : Map 1: 0(+3)/6  Reducer 2: 0/1
INFO  : Map 1: 0(+3)/6  Reducer 2: 0/1
INFO  : Map 1: 0(+3)/6  Reducer 2: 0/1
INFO  : Map 1: 1(+3)/6  Reducer 2: 0/1
INFO  : Map 1: 2(+2)/6  Reducer 2: 0/1
INFO  : Map 1: 3(+1)/6  Reducer 2: 0/1
INFO  : Map 1: 3(+2)/6  Reducer 2: 0/1
INFO  : Map 1: 3(+3)/6  Reducer 2: 0/1
INFO  : Map 1: 4(+2)/6  Reducer 2: 0/1
INFO  : Map 1: 4(+2)/6  Reducer 2: 0(+1)/1
INFO  : Map 1: 5(+1)/6  Reducer 2: 0(+1)/1
INFO  : Map 1: 6/6      Reducer 2: 0(+1)/1
INFO  : Map 1: 6/6      Reducer 2: 1/1
INFO  : Completed executing command(queryId=hive_20201003131808_97f90673-8d38-44b4-bb29-bfc986b8f277); Time taken: 24.353 seconds
INFO  : OK
+------------+-----------------------------------------+
| prodcount  |             category_code               |
+------------+-----------------------------------------+
| 8594895    |                                         |
| 11681      | accessories.bag                         |
| 1248       | accessories.cosmetic_bag                |
| 18232      | apparel.glove                           |
| 332        | appliances.environment.air_conditioner  |
| 59761      | appliances.environment.vacuum           |
| 1643       | appliances.personal.hair_cutter         |
| 9857       | furniture.bathroom.bath                 |
| 13439      | furniture.living_room.cabinet           |
| 308        | furniture.living_room.chair             |
| 2          | sport.diving                            |
| 26722      | stationery.cartrige                     |
+------------+-----------------------------------------+
12 rows selected (24.667 seconds)
0: jdbc:hive2://localhost:10000/default>
```

You can see above the highest number of products have no category assigned.

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

Query:

```
Select brand from (select sum(price) as Sales, brand from
purchases_2019 group by brand order by Sales desc limit 1) as
TopBrands;
```

Let us use the purchases_2019 table for the forthcoming queries since we need data only which are purchased.

Screenshot:

```
0: jdbc:hive2://localhost:10000/default> Select brand from (select sum(price) as Sales, brand from purchases_2019 group by brand order by Sales desc limit 1) as TopBrands;
INFO  : Compiling command(queryId=hive_20201003140926_3201b806-18ad-4ae3-b55b-58228660bb94): Select brand from (select sum(price) as Sales, brand from purchases_2019 group by brand order by Sales de
sc limit 1) as TopBrands
INFO  : Semantic Analysis Completed
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:brand, type:string, comment:null)], properties:null)
INFO  : Completed compiling command(queryId=hive_20201003140926_3201b806-18ad-4ae3-b55b-58228660bb94); Time taken: 0.104 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20201003140926_3201b806-18ad-4ae3-b55b-58228660bb94): Select brand from (select sum(price) as Sales, brand from purchases_2019 group by brand order by Sales de
sc limit 1) as TopBrands
INFO  : Query ID = hive_20201003140926_3201b806-18ad-4ae3-b55b-58228660bb94
INFO  : Total jobs = 1
INFO  : Launching Job 1 out of 1
INFO  : Starting task [Stage-1:MAPRED] in serial mode
INFO  : Session is already open
INFO  : Dag name: Select brand from (select sum(pr...TopBrands(Stage-1)
INFO  : Status: Running (Executing on YARN cluster with App id application_1601718673301_0019)

INFO  : Map 1: 0/2      Reducer 2: 0/1  Reducer 3: 0/1
INFO  : Map 1: 0/2      Reducer 2: 0/1  Reducer 3: 0/1
INFO  : Map 1: 0(+1)/2  Reducer 2: 0/1  Reducer 3: 0/1
INFO  : Map 1: 0(+2)/2  Reducer 2: 0/1  Reducer 3: 0/1
INFO  : Map 1: 0(+2)/2  Reducer 2: 0/1  Reducer 3: 0/1
INFO  : Map 1: 0(+2)/2  Reducer 2: 0/1  Reducer 3: 0/1
INFO  : Map 1: 1(+1)/2  Reducer 2: 0/1  Reducer 3: 0/1
INFO  : Map 1: 1(+1)/2  Reducer 2: 0(+1)/1      Reducer 3: 0/1
INFO  : Map 1: 2/2      Reducer 2: 0(+1)/1      Reducer 3: 0/1
INFO  : Map 1: 2/2      Reducer 2: 1/1  Reducer 3: 0(+1)/1
INFO  : Map 1: 2/2      Reducer 2: 1/1  Reducer 3: 1/1
INFO  : Completed executing command(queryId=hive_20201003140926_3201b806-18ad-4ae3-b55b-58228660bb94); Time taken: 13.676 seconds
INFO  : OK
+--------+
| brand  |
+--------+
|        |
+--------+
1 row selected (13.803 seconds)
0: jdbc:hive2://localhost:10000/default>
```

As you can see above the highest sales happened in products which did not contain any brand. So let us see which is the next highest (i.e.) top among available brands.

```
Select brand from (select sum(price) as Sales, brand from purchases_2019 group by brand order by Sales desc limit 2) as TopBrands;
```

```
0: jdbc:hive2://localhost:10000/default> Select brand from (select sum(price) as Sales, brand from purchases_2019 group by brand order by Sales desc limit 2) as TopBrands;
INFO  : Compiling command(queryId=hive_20201003140811_02ff6bb1-df06-4b39-a5c3-2aec7737122f): Select brand from (select sum(price) as Sales, brand from purchases_2019 group by brand order by Sales de
sc limit 2) as TopBrands
INFO  : Semantic Analysis Completed
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:brand, type:string, comment:null)], properties:null)
INFO  : Completed compiling command(queryId=hive_20201003140811_02ff6bb1-df06-4b39-a5c3-2aec7737122f); Time taken: 0.115 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20201003140811_02ff6bb1-df06-4b39-a5c3-2aec7737122f): Select brand from (select sum(price) as Sales, brand from purchases_2019 group by brand order by Sales de
sc limit 2) as TopBrands
INFO  : Query ID = hive_20201003140811_02ff6bb1-df06-4b39-a5c3-2aec7737122f
INFO  : Total jobs = 1
INFO  : Launching Job 1 out of 1
INFO  : Starting task [Stage-1:MAPRED] in serial mode
INFO  : Session is already open
INFO  : Dag name: Select brand from (select sum(pr...TopBrands(Stage-1)
INFO  : Status: Running (Executing on YARN cluster with App id application_1601718673301_0019)

INFO  : Map 1: 0/2      Reducer 2: 0/1  Reducer 3: 0/1
INFO  : Map 1: 0/2      Reducer 2: 0/1  Reducer 3: 0/1
INFO  : Map 1: 0(+2)/2  Reducer 2: 0/1  Reducer 3: 0/1
INFO  : Map 1: 0(+2)/2  Reducer 2: 0/1  Reducer 3: 0/1
INFO  : Map 1: 0(+2)/2  Reducer 2: 0/1  Reducer 3: 0/1
INFO  : Map 1: 1(+1)/2  Reducer 2: 0(+1)/1      Reducer 3: 0/1
INFO  : Map 1: 2/2      Reducer 2: 0(+1)/1      Reducer 3: 0/1
INFO  : Map 1: 2/2      Reducer 2: 1/1  Reducer 3: 0(+1)/1
INFO  : Map 1: 2/2      Reducer 2: 1/1  Reducer 3: 1/1
INFO  : Completed executing command(queryId=hive_20201003140811_02ff6bb1-df06-4b39-a5c3-2aec7737122f); Time taken: 14.046 seconds
INFO  : OK
+----------+
|  brand   |
+----------+
|          |
| runail   |
+----------+
2 rows selected (14.187 seconds)
0: jdbc:hive2://localhost:10000/default>
```

Runail is the brand with the highest sales amount of all available brands.

Question 7:

Which brands increased their sales from October to November?

We need to first get the details of each brand on their revenue for October and similarly for November, then compare which rows have November Total sales greater than October total sales and fetch those brands distinctly.

Here also we use `purchases_2019` since we need only purchase data.

Query:

```
select distinct(novdet.brand) as brands
from
(Select sum(price) as OctSales, brand from purchases_2019 group by
month(event_time), brand having month(event_time) = 10) as OctDet
Inner Join
(select sum(price) as NovSales, brand from purchases_2019 group by
month(event_time), brand having month(event_time) = 11) as NovDet
On (OctDet.brand = NovDet.brand)
Where OctDet.OctSales < NovDet.NovSales;
```

As you can see in screenshots below out of a total of 219 brands 153 brands increased their sales from October to November.

Screenshot:

```
0: jdbc:hive2://localhost:10000/default> select distinct(novdet.brand) as brands
. . . . . . . . . . . . . . . . . . . .> from
. . . . . . . . . . . . . . . . . . . .> (Select sum(price) as OctSales, brand from purchases_2019 group by
. . . . . . . . . . . . . . . . . . . .> month(event_time), brand having month(event_time) = 10) as OctDet
. . . . . . . . . . . . . . . . . . . .> inner join
. . . . . . . . . . . . . . . . . . . .> (select sum(price) as NovSales, brand from purchases_2019 group by
. . . . . . . . . . . . . . . . . . . .> month(event_time), brand having month(event_time) = 11) as NovDet
. . . . . . . . . . . . . . . . . . . .> On (OctDet.brand = NovDet.brand)
. . . . . . . . . . . . . . . . . . . .> Where OctDet.OctSales < NovDet.NovSales;
INFO  : Compiling command(queryId=hive_20201003155034_9e96a774-421c-4054-b05a-b598c9ba6ef6): select distinct(novdet.brand) as brands
from
(Select sum(price) as OctSales, brand from purchases_2019 group by
month(event_time), brand having month(event_time) = 10) as OctDet
inner join
(select sum(price) as NovSales, brand from purchases_2019 group by
month(event_time), brand having month(event_time) = 11) as NovDet
On (OctDet.brand = NovDet.brand)
Where OctDet.OctSales < NovDet.NovSales
INFO  : Semantic Analysis Completed
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:brands, type:string, comment:null)], properties:null)
INFO  : Completed compiling command(queryId=hive_20201003155034_9e96a774-421c-4054-b05a-b598c9ba6ef6); Time taken: 0.167 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20201003155034_9e96a774-421c-4054-b05a-b598c9ba6ef6): select distinct(novdet.brand) as brands
from
(Select sum(price) as OctSales, brand from purchases_2019 group by
month(event_time), brand having month(event_time) = 10) as OctDet
inner join
(select sum(price) as NovSales, brand from purchases_2019 group by
month(event_time), brand having month(event_time) = 11) as NovDet
On (OctDet.brand = NovDet.brand)
Where OctDet.OctSales < NovDet.NovSales
INFO  : Query ID = hive_20201003155034_9e96a774-421c-4054-b05a-b598c9ba6ef6
INFO  : Total jobs = 1
INFO  : Launching Job 1 out of 1
INFO  : Starting task [Stage-1:MAPRED] in serial mode
INFO  : Session is already open
INFO  : Dag name: select distinct(novdet.bra...NovDet.NovSales(Stage-1)
INFO  : Status: Running (Executing on YARN cluster with App id application_1601718673301_0023)

INFO  : Map 1: 0/2      Map 5: 0/2      Reducer 2: 0/1  Reducer 3: 0/1  Reducer 4: 0/1  Reducer 6: 0/1
INFO  : Map 1: 0/2      Map 5: 0/2      Reducer 2: 0/1  Reducer 3: 0/1  Reducer 4: 0/1  Reducer 6: 0/1
INFO  : Map 1: 0/2      Map 5: 0/2      Reducer 2: 0/1  Reducer 3: 0/1  Reducer 4: 0/1  Reducer 6: 0/1
INFO  : Map 1: 0(+1)/2  Map 5: 0/2      Reducer 2: 0/1  Reducer 3: 0/1  Reducer 4: 0/1  Reducer 6: 0/1
INFO  : Map 1: 0(+1)/2  Map 5: 0(+1)/2  Reducer 2: 0/1  Reducer 3: 0/1  Reducer 4: 0/1  Reducer 6: 0/1
INFO  : Map 1: 0(+1)/2  Map 5: 0(+2)/2  Reducer 2: 0/1  Reducer 3: 0/1  Reducer 4: 0/1  Reducer 6: 0/1
INFO  : Map 1: 0(+1)/2  Map 5: 0(+2)/2  Reducer 2: 0/1  Reducer 3: 0/1  Reducer 4: 0/1  Reducer 6: 0/1
INFO  : Map 1: 0(+1)/2  Map 5: 0(+2)/2  Reducer 2: 0/1  Reducer 3: 0/1  Reducer 4: 0/1  Reducer 6: 0/1
INFO  : Map 1: 0(+1)/2  Map 5: 0(+2)/2  Reducer 2: 0/1  Reducer 3: 0/1  Reducer 4: 0/1  Reducer 6: 0/1
INFO  : Map 1: 0(+2)/2  Map 5: 1(+1)/2  Reducer 2: 0/1  Reducer 3: 0/1  Reducer 4: 0/1  Reducer 6: 0/1
INFO  : Map 1: 1(+1)/2  Map 5: 1(+1)/2  Reducer 2: 0/1  Reducer 3: 0/1  Reducer 4: 0/1  Reducer 6: 0(+1)/1
INFO  : Map 1: 1(+1)/2  Map 5: 2/2      Reducer 2: 0(+1)/1      Reducer 3: 0/1  Reducer 4: 0/1  Reducer 6: 0(+1)/1
INFO  : Map 1: 1(+1)/2  Map 5: 2/2      Reducer 2: 0(+1)/1      Reducer 3: 0/1  Reducer 4: 0/1  Reducer 6: 1/1
INFO  : Map 1: 2/2      Map 5: 2/2      Reducer 2: 0(+1)/1      Reducer 3: 0/1  Reducer 4: 0/1  Reducer 6: 1/1
INFO  : Map 1: 2/2      Map 5: 2/2      Reducer 2: 1/1  Reducer 3: 0(+1)/1      Reducer 4: 0/1  Reducer 6: 1/1
INFO  : Map 1: 2/2      Map 5: 2/2      Reducer 2: 1/1  Reducer 3: 1/1  Reducer 4: 0/1  Reducer 6: 1/1
INFO  : Map 1: 2/2      Map 5: 2/2      Reducer 2: 1/1  Reducer 3: 1/1  Reducer 4: 0(+1)/1      Reducer 6: 1/1
INFO  : Map 1: 2/2      Map 5: 2/2      Reducer 2: 1/1  Reducer 3: 1/1  Reducer 4: 1/1  Reducer 6: 1/1
INFO  : Completed executing command(queryId=hive_20201003155034_9e96a774-421c-4054-b05a-b598c9ba6ef6); Time taken: 24.47 seconds
INFO  : OK
+-----------------+
|     brands      |
+-----------------+
|                 |
| airnails        |
| art-visage      |
| artex           |
| aura            |
```

| balbcare |
| batiste |
| beautix |
| beauty-free |
| beautyblender |
| beauugreen |
| benovy |
| bioaqua |
| biore |
| blixz |
| bluesky |
| bodyton |
| bpw.style |
| browxenna |
| candy |
| carmex |
| chi |
| coifin |
| concept |
| cosima |
| cosmoprofi |
| cristalinas |
| cutrin |
| de.lux |
| deoproce |
| depilflax |
| dizao |
| domix |
| ecocraft |
| ecolab |
| egomania |
| elizavecca |
| ellips |
| elskin |
| enjoy |
| entity |
| eos |
| estel |
| estelare |
| f.o.x |
| farmavita |
| farmona |
| fedua |
| finish |
| fly |
| foamie |
| freedecor |
| freshbubble |
| gehwol |
| glysolid |
| godefroy |
| grace |
| grattol |
| greymy |
| happyfons |
| haruyama |
| igrobeauty |
| ingarden |
| inm |
| insight |
| irisk |
| italwax |
| jaguar |
| jas |
| jessnail |
| joico |
| kaaral |
| kamill |
| kapous |
| kaypro |
| keen |
| kerasys |
| kims |
| kinetics |
| kiss |
| kocostar |
| koelcia |
| koelf |
| konad |

| kocostar |
| koelcia |
| koelf |
| konad |
| kosmekka |
| laboratorium |
| lador |
| ladykin |
| latinoil |
| levissime |
| levrana |
| lianail |
| likato |
| limoni |
| lovely |
| lowence |
| mane |
| marathon |
| markell |
| marutaka-foot |

| brands |
| masura |
| matrix |
| mavala |
| metzger |
| milv |
| miskin |
| missha |
| moyou |
| nagaraku |
| nefertiti |
| neoleor |
| nirvel |
| nitrile |
| oniq |
| orly |
| osmo |
| ovale |
| plazan |
| polarus |
| profepil |
| profhenna |
| protokeratin |
| provoc |
| rasyan |
| refectocil |
| rosi |
| roubloff |
| runail |
| s.care |
| sanoto |
| severina |
| shary |
| shik |
| skinity |
| skinlite |
| smart |
| soleo |
| solomeya |
| sophin |
| staleks |
| strong |
| supertan |
| swarovski |
| tertio |
| treaclemoon |
| trind |
| una |
| uskusi |
| veraclara |
| vilenta |
| yoko |
| yu-r |
| zeitun |

153 rows selected (24.663 seconds)
0: jdbc:hive2://localhost:10000/default>

Question 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.

Query:

```
select user_id from (select sum(price) as UserRevenue, user_id from purchases_2019 group by user_id order by UserRevenue desc limit 10) as TopUsers;
```

Screenshot:

```
0: jdbc:hive2://localhost:10000/default> select user_id from (select sum(price) as UserRevenue, user_id from purchases_2019 group by user_id order by UserRevenue desc limit 10) as TopUsers;
INFO  : Compiling command(queryId=hive_20201003140654_c3ce6e06-7a01-4851-925e-a925e6842e18): select user_id from (select sum(price) as UserRevenue, user_id from purchases_2019 group by user_id order
 by UserRevenue desc limit 10) as TopUsers
INFO  : Semantic Analysis Completed
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:user_id, type:bigint, comment:null)], properties:null)
INFO  : Completed compiling command(queryId=hive_20201003140654_c3ce6e06-7a01-4851-925e-a925e6842e18); Time taken: 0.111 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20201003140654_c3ce6e06-7a01-4851-925e-a925e6842e18): select user_id from (select sum(price) as UserRevenue, user_id from purchases_2019 group by user_id order
 by UserRevenue desc limit 10) as TopUsers
INFO  : Query ID = hive_20201003140654_c3ce6e06-7a01-4851-925e-a925e6842e18
INFO  : Total jobs = 1
INFO  : Launching Job 1 out of 1
INFO  : Starting task [Stage-1:MAPRED] in serial mode
INFO  : Session is already open
INFO  : Dag name: select user_id from (select sum(p...TopUsers(Stage-1)
INFO  : Status: Running (Executing on YARN cluster with App id application_1601718673301_0019)

INFO  : Map 1: 0/2      Reducer 2: 0/1  Reducer 3: 0/1
INFO  : Map 1: 0/2      Reducer 2: 0/1  Reducer 3: 0/1
INFO  : Map 1: 0(+1)/2  Reducer 2: 0/1  Reducer 3: 0/1
INFO  : Map 1: 0(+2)/2  Reducer 2: 0/1  Reducer 3: 0/1
INFO  : Map 1: 0(+2)/2  Reducer 2: 0/1  Reducer 3: 0/1
INFO  : Map 1: 0(+2)/2  Reducer 2: 0/1  Reducer 3: 0/1
INFO  : Map 1: 1(+1)/2  Reducer 2: 0/1  Reducer 3: 0/1
INFO  : Map 1: 2/2      Reducer 2: 0(+1)/1      Reducer 3: 0/1
INFO  : Map 1: 2/2      Reducer 2: 1/1  Reducer 3: 0(+1)/1
INFO  : Map 1: 2/2      Reducer 2: 1/1  Reducer 3: 1/1
INFO  : Completed executing command(queryId=hive_20201003140654_c3ce6e06-7a01-4851-925e-a925e6842e18); Time taken: 15.292 seconds
INFO  : OK
+------------+
|  user_id   |
+------------+
| 557790271  |
| 150318419  |
| 562167663  |
| 531900924  |
| 557850743  |
| 522130011  |
| 561592095  |
| 431950134  |
| 566576008  |
| 521347209  |
+------------+
10 rows selected (15.418 seconds)
0: jdbc:hive2://localhost:10000/default>
```

Above are the top 10 users.

Step 3: Cleaning up

We need to drop the tables first then the database.

Query:

```
drop table ecom_events_2019;
drop table ecom_events_month_2019;
drop table parq_buck_purchase;
drop table purchases_2019;
```

Screenshot:

```
0: jdbc:hive2://localhost:10000/default> show tables;
INFO  : Compiling command(queryId=hive_20201003161229_06927e83-13d2-469c-97af-7c0c96c957e5): show tables
INFO  : Semantic Analysis Completed
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:tab_name, type:string, comment:from deserializer)], properties:null)
INFO  : Completed compiling command(queryId=hive_20201003161229_06927e83-13d2-469c-97af-7c0c96c957e5); Time taken: 0.007 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20201003161229_06927e83-13d2-469c-97af-7c0c96c957e5): show tables
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing command(queryId=hive_20201003161229_06927e83-13d2-469c-97af-7c0c96c957e5); Time taken: 0.009 seconds
INFO  : OK
+------------------------+
|        tab_name        |
+------------------------+
| ecom_events_2019       |
| ecom_events_month_2019 |
| parq_buck_purchase     |
| purchases_2019         |
+------------------------+
4 rows selected (0.069 seconds)
0: jdbc:hive2://localhost:10000/default> drop table ecom_events_2019;
INFO  : Compiling command(queryId=hive_20201003161234_9867ff2f-66a3-488f-a252-bdc9099340c8): drop table ecom_events_2019
INFO  : Semantic Analysis Completed
INFO  : Returning Hive schema: Schema(fieldSchemas:null, properties:null)
INFO  : Completed compiling command(queryId=hive_20201003161234_9867ff2f-66a3-488f-a252-bdc9099340c8); Time taken: 0.015 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20201003161234_9867ff2f-66a3-488f-a252-bdc9099340c8): drop table ecom_events_2019
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing command(queryId=hive_20201003161234_9867ff2f-66a3-488f-a252-bdc9099340c8); Time taken: 0.065 seconds
INFO  : OK
No rows affected (0.094 seconds)
0: jdbc:hive2://localhost:10000/default> drop table ecom_events_month_2019;
INFO  : Compiling command(queryId=hive_20201003161241_75749c05-01ee-4815-90c8-17f39a3b3dfd): drop table ecom_events_month_2019
INFO  : Semantic Analysis Completed
INFO  : Returning Hive schema: Schema(fieldSchemas:null, properties:null)
INFO  : Completed compiling command(queryId=hive_20201003161241_75749c05-01ee-4815-90c8-17f39a3b3dfd); Time taken: 0.016 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20201003161241_75749c05-01ee-4815-90c8-17f39a3b3dfd): drop table ecom_events_month_2019
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing command(queryId=hive_20201003161241_75749c05-01ee-4815-90c8-17f39a3b3dfd); Time taken: 0.096 seconds
INFO  : OK
No rows affected (0.124 seconds)
0: jdbc:hive2://localhost:10000/default> drop table parq_buck_purchase;
INFO  : Compiling command(queryId=hive_20201003161249_74f18123-cc72-42d5-8a4d-b522979d6daf): drop table parq_buck_purchase
INFO  : Semantic Analysis Completed
INFO  : Returning Hive schema: Schema(fieldSchemas:null, properties:null)
INFO  : Completed compiling command(queryId=hive_20201003161249_74f18123-cc72-42d5-8a4d-b522979d6daf); Time taken: 0.018 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20201003161249_74f18123-cc72-42d5-8a4d-b522979d6daf): drop table parq_buck_purchase
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing command(queryId=hive_20201003161249_74f18123-cc72-42d5-8a4d-b522979d6daf); Time taken: 0.09 seconds
INFO  : OK
No rows affected (0.12 seconds)
0: jdbc:hive2://localhost:10000/default> drop table purchases_2019;
INFO  : Compiling command(queryId=hive_20201003161256_29e75e4c-ca75-4ede-83e8-48e18a23fc7c): drop table purchases_2019
INFO  : Semantic Analysis Completed
INFO  : Returning Hive schema: Schema(fieldSchemas:null, properties:null)
INFO  : Completed compiling command(queryId=hive_20201003161256_29e75e4c-ca75-4ede-83e8-48e18a23fc7c); Time taken: 0.011 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20201003161256_29e75e4c-ca75-4ede-83e8-48e18a23fc7c): drop table purchases_2019
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing command(queryId=hive_20201003161256_29e75e4c-ca75-4ede-83e8-48e18a23fc7c); Time taken: 0.056 seconds
INFO  : OK
No rows affected (0.074 seconds)
0: jdbc:hive2://localhost:10000/default> show tables;
INFO  : Compiling command(queryId=hive_20201003161300_e4ba0ce5-dd0d-4f3b-85ad-7cf0c36ebca3): show tables
INFO  : Semantic Analysis Completed
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:tab_name, type:string, comment:from deserializer)], properties:null)
INFO  : Completed compiling command(queryId=hive_20201003161300_e4ba0ce5-dd0d-4f3b-85ad-7cf0c36ebca3); Time taken: 0.013 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20201003161300_e4ba0ce5-dd0d-4f3b-85ad-7cf0c36ebca3): show tables
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing command(queryId=hive_20201003161300_e4ba0ce5-dd0d-4f3b-85ad-7cf0c36ebca3); Time taken: 0.006 seconds
INFO  : OK
+-----------+
| tab_name  |
+-----------+
+-----------+
No rows selected (0.047 seconds)
0: jdbc:hive2://localhost:10000/default> show databases;
INFO  : Compiling command(queryId=hive_20201003161330_e7b08199-fb50-4ca4-99d0-a4d85a494509): show databases
INFO  : Semantic Analysis Completed
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:database_name, type:string, comment:from deserializer)], properties:null)
INFO  : Completed compiling command(queryId=hive_20201003161330_e7b08199-fb50-4ca4-99d0-a4d85a494509); Time taken: 0.003 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20201003161330_e7b08199-fb50-4ca4-99d0-a4d85a494509): show databases
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing command(queryId=hive_20201003161330_e7b08199-fb50-4ca4-99d0-a4d85a494509); Time taken: 0.006 seconds
INFO  : OK
+--------------------+
|   database_name    |
+--------------------+
| default            |
| sks_c3_assignment  |
+--------------------+
2 rows selected (0.029 seconds)
0: jdbc:hive2://localhost:10000/default> drop database sks_c3_assignment;
INFO  : Compiling command(queryId=hive_20201003161337_59f86a85-4111-431f-9c56-e4c821b061b7): drop database sks_c3_assignment
INFO  : Semantic Analysis Completed
INFO  : Returning Hive schema: Schema(fieldSchemas:null, properties:null)
INFO  : Completed compiling command(queryId=hive_20201003161337_59f86a85-4111-431f-9c56-e4c821b061b7); Time taken: 0.009 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20201003161337_59f86a85-4111-431f-9c56-e4c821b061b7): drop database sks_c3_assignment
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing command(queryId=hive_20201003161337_59f86a85-4111-431f-9c56-e4c821b061b7); Time taken: 0.054 seconds
INFO  : OK
No rows affected (0.098 seconds)
0: jdbc:hive2://localhost:10000/default> show databases;
INFO  : Compiling command(queryId=hive_20201003161344_e2265dc2-a79d-4f6d-b3a9-258a52bc1e9a): show databases
INFO  : Semantic Analysis Completed
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:database_name, type:string, comment:from deserializer)], properties:null)
INFO  : Completed compiling command(queryId=hive_20201003161344_e2265dc2-a79d-4f6d-b3a9-258a52bc1e9a); Time taken: 0.004 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20201003161344_e2265dc2-a79d-4f6d-b3a9-258a52bc1e9a): show databases
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing command(queryId=hive_20201003161344_e2265dc2-a79d-4f6d-b3a9-258a52bc1e9a); Time taken: 0.004 seconds
INFO  : OK
+----------------+
| database_name  |
+----------------+
| default        |
+----------------+
1 row selected (0.022 seconds)
0: jdbc:hive2://localhost:10000/default>
```