HIVE CASE STUDY

By Sanjay Gupta and Jyoti Sodhi

Problem Statement

With online sales gaining popularity, tech companies are exploring ways to improve their sales by analysing customer behaviour and gaining insights about product trends.

Furthermore, the websites make it easier for customers to find the products they require without much scavenging. Needless to say, the role of big data analysts is among the most sought-after job profiles of this decade. Therefore, as part of this assignment, we will be challenging you, as a big data analyst, to extract data and gather insights from a real-life data set of an e-commerce company.

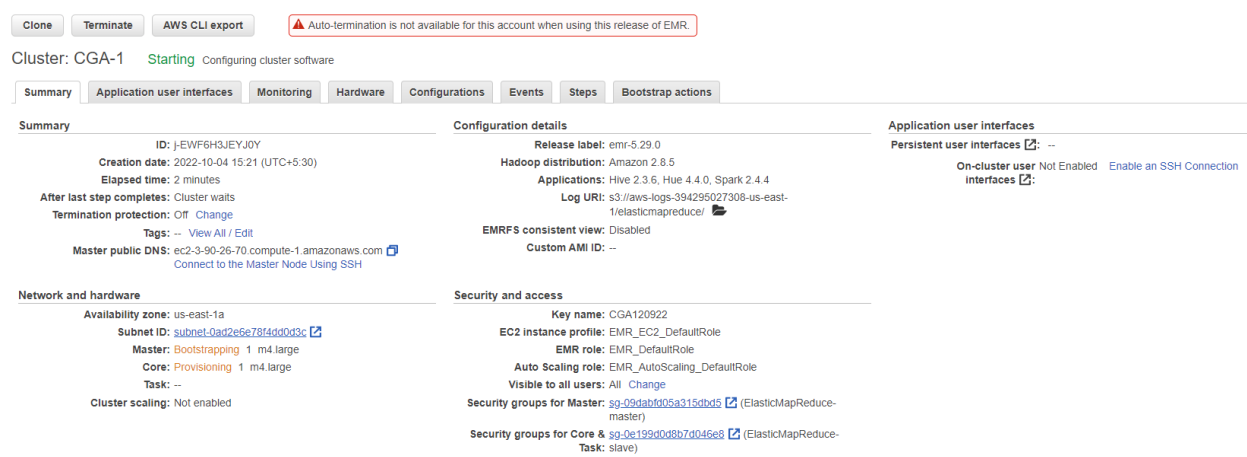
Data for the case study is in the link given below.

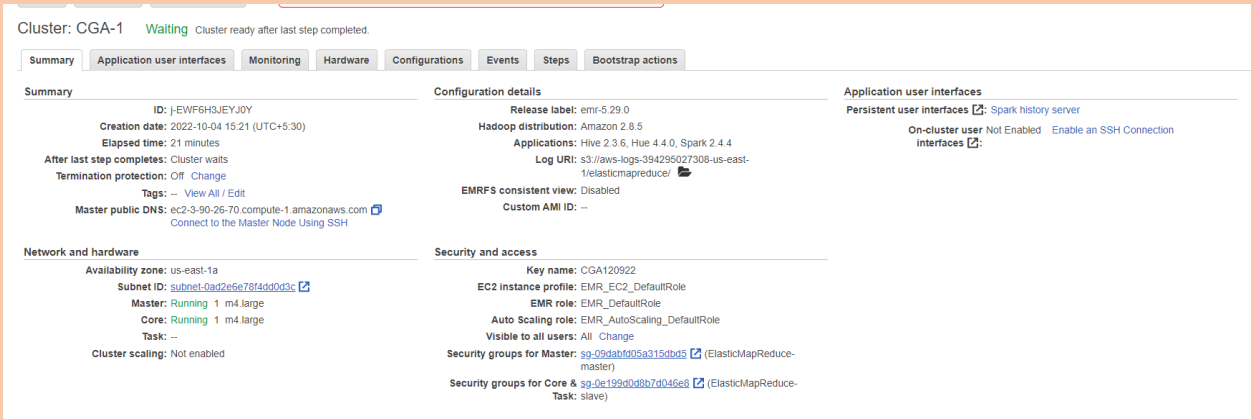
https://e-commerce-events-ml.

We login to Nuvepro dashboard, go to the console and then to EMR home page → Click on

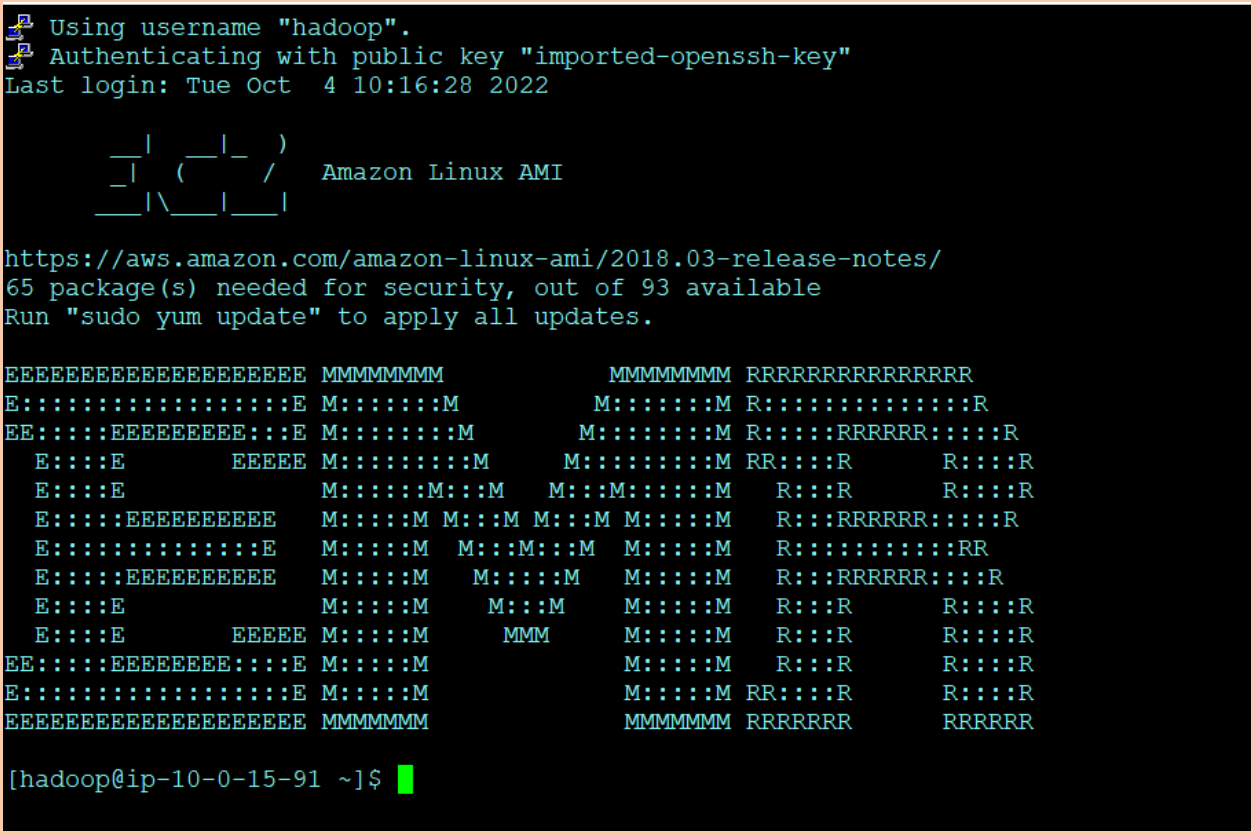
Create Cluster → select release EMR 5.29.0 and select required service for the case study.

1) Launching an EMR cluster that utilizes Hive services





2) Launch an EMR Cluster and connect to master node through SSH



3) Creating a folder “hivecasestudy ”on the Hadoop file system

```
[hadoop@ip-10-0-15-91 ~]$ hadoop fs -ls /user
Found 8 items
drwxrwxrwx   - hadoop hadoop          0 2022-10-04 10:59 /user/hadoop
drwxr-xr-x   - mapred mapred          0 2022-10-04 09:58 /user/history
drwxrwxrwx   - hdfs   hadoop          0 2022-10-04 11:01 /user/hive
drwxrwxrwx   - hue    hue             0 2022-10-04 09:58 /user/hue
drwxrwxrwx   - livy   livy            0 2022-10-04 09:58 /user/livy
drwxrwxrwx   - oozie  oozie           0 2022-10-04 09:58 /user/oozie
drwxrwxrwx   - root   hadoop          0 2022-10-04 09:58 /user/root
drwxrwxrwx   - spark  spark           0 2022-10-04 09:58 /user/spark
[hadoop@ip-10-0-15-91 ~]$ hadoop fs -ls /user/hive
Found 2 items
drwxr-xr-x   - hadoop hadoop          0 2022-10-04 11:01 /user/hive/hivecasestudy
drwxrwxrwt   - hdfs   hadoop          0 2022-10-04 09:58 /user/hive/warehouse
[hadoop@ip-10-0-15-91 ~]$ hadoop fs -mkdir /user/hive/hivecasestudy
mkdir: `/user/hive/hivecasestudy': File exists
```

4) Moving the data from S3 bucket into the HDFS

```
hadoop distcp 's3://e-commerce-events-ml/*' '/user/hive/hivecasestudy/'
```

```
mkdir: `/user/hive/hivecasestudy': File exists
[hadoop@ip-10-0-15-91 ~]$ hadoop distcp 's3://e-commerce-events-ml/*' '/user/hive/hivecasestudy/'
22/10/04 11:19:12 INFO tools.DistCp: Input Options: DistCpOptions{atomicCommit=false, syncFolder=false, deleteMissing=false, ignoreFailures=false, ove
apBandwidth=100, sslConfigurationFile='null', copyStrategy='uniformsize', preserveStatus=[], preserveRawXattrs=false, atomicWorkPath=null, logPath=nul
/user/hive/hivecasestudy, targetPathExists=true, filtersFile='null'}
22/10/04 11:19:12 INFO client.RMProxy: Connecting to ResourceManager at ip-10-0-15-91.ec2.internal/10.0.15.91:8032
22/10/04 11:19:19 INFO tools.SimpleCopyListing: Paths (files+dirs) cnt = 2; dirCnt = 0
22/10/04 11:19:19 INFO tools.SimpleCopyListing: Build file listing completed.
22/10/04 11:19:19 INFO Configuration.deprecation: io.sort.mb is deprecated. Instead, use mapreduce.task.io.sort.mb
22/10/04 11:19:19 INFO Configuration.deprecation: io.sort.factor is deprecated. Instead, use mapreduce.task.io.sort.factor
22/10/04 11:19:19 INFO tools.DistCp: Number of paths in the copy list: 2
22/10/04 11:19:20 INFO tools.DistCp: Number of paths in the copy list: 2
22/10/04 11:19:20 INFO client.RMProxy: Connecting to ResourceManager at ip-10-0-15-91.ec2.internal/10.0.15.91:8032
22/10/04 11:19:20 INFO mapreduce.JobSubmitter: number of splits:2
22/10/04 11:19:20 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1664877578682_0002
22/10/04 11:19:21 INFO impl.YarnClientImpl: Submitted application application_1664877578682_0002
22/10/04 11:19:21 INFO mapreduce.Job: The url to track the job: http://ip-10-0-15-91.ec2.internal:20888/proxy/application_1664877578682_0002/
22/10/04 11:19:21 INFO tools.DistCp: DistCp job-id: job_1664877578682_0002
22/10/04 11:19:21 INFO mapreduce.Job: Running job: job_1664877578682_0002
22/10/04 11:19:29 INFO mapreduce.Job: Job job_1664877578682_0002 running in uber mode : false
22/10/04 11:19:29 INFO mapreduce.Job:  map 0% reduce 0%
22/10/04 11:19:50 INFO mapreduce.Job:  map 50% reduce 0%
22/10/04 11:19:51 INFO mapreduce.Job:  map 100% reduce 0%
22/10/04 11:20:06 INFO mapreduce.Job: Job job_1664877578682_0002 completed successfully
22/10/04 11:20:06 INFO mapreduce.Job: Counters: 38
        File System Counters
                FILE: Number of bytes read=0
                FILE: Number of bytes written=345574
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=900
                HDFS: Number of bytes written=1028381690
                HDFS: Number of read operations=26
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=8
                S3: Number of bytes read=1028381690
                S3: Number of bytes written=0
                S3: Number of read operations=0
                S3: Number of large read operations=0
                S3: Number of write operations=0
        Job Counters
                Launched map tasks=2
                Other local map tasks=2
                Total time spent by all maps in occupied slots (ms)=2130720
                Total time spent by all reduces in occupied slots (ms)=0
                Total time spent by all map tasks (ms)=66585
                Total vcore-milliseconds taken by all map tasks=66585
                Total megabyte-milliseconds taken by all map tasks=68183040
        Map-Reduce Framework
                Map input records=2
                Map output records=0
                Input split bytes=274
                Spilled Records=0
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=1365
                CPU time spent (ms)=43910
```

```
    File System Counters
            FILE: Number of bytes read=0
            FILE: Number of bytes written=345574
            FILE: Number of read operations=0
            FILE: Number of large read operations=0
            FILE: Number of write operations=0
            HDFS: Number of bytes read=900
            HDFS: Number of bytes written=1028381690
            HDFS: Number of read operations=26
            HDFS: Number of large read operations=0
            HDFS: Number of write operations=8
            S3: Number of bytes read=1028381690
            S3: Number of bytes written=0
            S3: Number of read operations=0
            S3: Number of large read operations=0
            S3: Number of write operations=0
    Job Counters
            Launched map tasks=2
            Other local map tasks=2
            Total time spent by all maps in occupied slots (ms)=2130720
            Total time spent by all reduces in occupied slots (ms)=0
            Total time spent by all map tasks (ms)=66585
            Total vcore-milliseconds taken by all map tasks=66585
            Total megabyte-milliseconds taken by all map tasks=68183040
    Map-Reduce Framework
            Map input records=2
            Map output records=0
            Input split bytes=274
            Spilled Records=0
            Failed Shuffles=0
            Merged Map outputs=0
            GC time elapsed (ms)=1365
            CPU time spent (ms)=43910
            Physical memory (bytes) snapshot=1095065600
            Virtual memory (bytes) snapshot=6592835584
            Total committed heap usage (bytes)=954204160
    File Input Format Counters
            Bytes Read=626
    File Output Format Counters
            Bytes Written=0
    DistCp Counters
            Bytes Copied=1028381690
            Bytes Expected=1028381690
            Files Copied=2
```

5) Checking if the files are correctly imported to HDFS

```
hadoop fs -ls/user/hive/hivecasestudy/
```

```
[hadoop@ip-10-0-15-91 ~]$ hadoop fs -ls/user/hive/hivecasestudy/
-ls/user/hive/hivecasestudy/: Unknown command
Usage: hadoop fs [generic options]
        [-appendToFile <localsrc> ... <dst>]
        [-cat [-ignoreCrc] <src> ...]
        [-checksum <src> ...]
        [-chgrp [-R] GROUP PATH...]
        [-chmod [-R] <MODE[,MODE]... | OCTALMODE> PATH...]
        [-chown [-R] [OWNER][:[GROUP]] PATH...]
        [-copyFromLocal [-f] [-p] [-l] [-d] <localsrc> ... <dst>]
        [-copyToLocal [-f] [-p] [-ignoreCrc] [-crc] <src> ... <localdst>]
        [-count [-q] [-h] [-v] [-t [<storage type>]] [-u] [-x] <path> ...]
        [-cp [-f] [-p | -p[topax]] [-d] <src> ... <dst>]
        [-createSnapshot <snapshotDir> [<snapshotName>]]
        [-deleteSnapshot <snapshotDir> <snapshotName>]
        [-df [-h] [<path> ...]]
        [-du [-s] [-h] [-x] <path> ...]
        [-expunge]
        [-find <path> ... <expression> ...]
        [-get [-f] [-p] [-ignoreCrc] [-crc] <src> ... <localdst>]
        [-getfacl [-R] <path>]
        [-getfattr [-R] {-n name | -d} [-e en] <path>]
        [-getmerge [-nl] [-skip-empty-file] <src> <localdst>]
        [-help [cmd ...]]
        [-ls [-C] [-d] [-h] [-q] [-R] [-t] [-S] [-r] [-u] [<path> ...]]
        [-mkdir [-p] <path> ...]
        [-moveFromLocal <localsrc> ... <dst>]
        [-moveToLocal <src> <localdst>]
        [-mv <src> ... <dst>]
        [-put [-f] [-p] [-l] [-d] <localsrc> ... <dst>]
        [-renameSnapshot <snapshotDir> <oldName> <newName>]
        [-rm [-f] [-r|-R] [-skipTrash] [-safely] <src> ...]
        [-rmdir [--ignore-fail-on-non-empty] <dir> ...]
        [-setfacl [-R] [{-b|-k} {-m|-x <acl_spec>} <path>]|[--set <acl_spec> <path>]]
        [-setfattr {-n name [-v value] | -x name} <path>]
        [-setrep [-R] [-w] <rep> <path> ...]
        [-stat [format] <path> ...]
        [-tail [-f] <file>]
        [-test -[defsz] <path>]
        [-text [-ignoreCrc] <src> ...]
        [-touchz <path> ...]
        [-truncate [-w] <length> <path> ...]
        [-usage [cmd ...]]

Generic options supported are
-conf <configuration file>     specify an application configuration file
-D <property=value>            use value for given property
-fs <file:///|hdfs://namenode:port> specify default filesystem URL to use, overrides 'fs.defaultFS' property from configurations.
-jt <local|resourcemanager:port>    specify a ResourceManager
-files <comma separated list of files>    specify comma separated files to be copied to the map reduce cluster
-libjars <comma separated list of jars>    specify comma separated jar files to include in the classpath.
-archives <comma separated list of archives>    specify comma separated archives to be unarchived on the compute machines.

The general command line syntax is
command [genericOptions] [commandOptions]
```

We can confirm the databases were loaded successfully.

6) Changing cmd to the hive

```
[hadoop@ip-10-0-15-91 ~]$ hive

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

7) Create database if not exists casestudy;

```
hive> create databases if not exists casestudy ;
NoViableAltException(79@[846:1: ddlStatement : ( createDatabaseStatement | switchDatabaseStatement | dr
Statement | metastoreCheck | createViewStatement | createMaterializedViewStatement | dropViewStatement
ropFunctionStatement | reloadFunctionStatement | dropMacroStatement | analyzeStatement | lockStatement
ges | ( revokePrivileges )=> revokePrivileges | showGrants | showRoleGrants | showRolePrincipals | show
        at org.antlr.runtime.DFA.noViableAlt(DFA.java:158)
        at org.antlr.runtime.DFA.predict(DFA.java:116)
        at org.apache.hadoop.hive.ql.parse.HiveParser.ddlStatement(HiveParser.java:3757)
        at org.apache.hadoop.hive.ql.parse.HiveParser.execStatement(HiveParser.java:2382)
        at org.apache.hadoop.hive.ql.parse.HiveParser.statement(HiveParser.java:1333)
        at org.apache.hadoop.hive.ql.parse.ParseDriver.parse(ParseDriver.java:208)
        at org.apache.hadoop.hive.ql.parse.ParseUtils.parse(ParseUtils.java:77)
        at org.apache.hadoop.hive.ql.parse.ParseUtils.parse(ParseUtils.java:70)
        at org.apache.hadoop.hive.ql.Driver.compile(Driver.java:468)
        at org.apache.hadoop.hive.ql.Driver.compileInternal(Driver.java:1317)
        at org.apache.hadoop.hive.ql.Driver.runInternal(Driver.java:1457)
        at org.apache.hadoop.hive.ql.Driver.run(Driver.java:1237)
        at org.apache.hadoop.hive.ql.Driver.run(Driver.java:1227)
        at org.apache.hadoop.hive.cli.CliDriver.processLocalCmd(CliDriver.java:233)
        at org.apache.hadoop.hive.cli.CliDriver.processCmd(CliDriver.java:184)
        at org.apache.hadoop.hive.cli.CliDriver.processLine(CliDriver.java:403)
        at org.apache.hadoop.hive.cli.CliDriver.executeDriver(CliDriver.java:821)
        at org.apache.hadoop.hive.cli.CliDriver.run(CliDriver.java:759)
        at org.apache.hadoop.hive.cli.CliDriver.main(CliDriver.java:686)
        at sun.reflect.NativeMethodAccessorImpl.invoke0(Native Method)
        at sun.reflect.NativeMethodAccessorImpl.invoke(NativeMethodAccessorImpl.java:62)
        at sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccessorImpl.java:43)
        at java.lang.reflect.Method.invoke(Method.java:498)
        at org.apache.hadoop.util.RunJar.run(RunJar.java:239)
        at org.apache.hadoop.util.RunJar.main(RunJar.java:153)
FAILED: ParseException line 1:7 cannot recognize input near 'create' 'databases' 'if' in ddl statement
hive> create database if not exists casestudy ;
OK
Time taken: 0.608 seconds
```

8) Checking data base

```
hive> show databases ;
OK
casestudy
default
Time taken: 0.276 seconds, Fetched: 2 row(s)
hive> use casestudy ;
OK
Time taken: 0.074 seconds
```

9) Creating an External Table, comm:

hive> create table if not exists comm (event_time timestamp, event_type string , product_id string, category_id string , ctaegory_code string, brand string, price decimal (10,2),user_id bigint, user_session string) row format serde 'org.apache.hadoop.hive.serde2.OpenCSVSerde' stored as textfile location '/user/hive/hivecasestudy/' tblproperties ("skip.header.line.count" = "1");

```
hive> create table if not exists comm (event_time timestamp, event_type string , product_id string, category_id string , ctaegory_code string, brand string, price d
e 'org.apache.hadoop.hive.serde2.OpenCSVSerde' stored as textfile location '/user/hive/hivecasestudy/' tblproperties ("skip.header.line.count" = "1" );
OK
Time taken: 0.612 seconds
hive> select * from comm limit 5 ;
OK
2019-11-01 00:00:02 UTC view    5802432 1487580009286598681                     0.32    562076640       09fafd6c-6c99-46b1-834f-33527f4de241
2019-11-01 00:00:09 UTC cart    5844397 1487580006317032337                     2.38    553329724       2067216c-31b5-455d-a1cc-af0575a34ffb
2019-11-01 00:00:10 UTC view    5837166 1783999064103190764             pnb     22.22   556138645       57ed222e-a54a-4907-9944-5a875c2d7f4f
2019-11-01 00:00:11 UTC cart    5876812 1487580010100293687             jessnail        3.16    564506666       186c1951-8052-4b37-adce-dd9644b1d5f7
2019-11-01 00:00:24 UTC remove_from_cart        5826182 1487580007483048900                     3.33    553329724       2067216c-31b5-455d-a1cc-af0575a34ffb
Time taken: 3.323 seconds, Fetched: 5 row(s)
hive> select month(event_time) as month , sum(price) as total_price from comm where event_type ='purchase' and month(event_time)=10 group by month(event_time) ;
Query ID = hadoop_20221004113517_f7ea71bd-c086-4651-a49b-f6a91966645f
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1664877578682_0003)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      2          2        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      2          2        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 59.27 s
--------------------------------------------------------------------------------
OK
10      1211538.4299997438
Time taken: 62.642 seconds, Fetched: 1 row(s)
```

10) hive> select * from comm limit 5 ; query to view the dataset

```
hive> select * from comm limit 5 ;
OK
2019-11-01 00:00:02 UTC view    5802432 1487580009286598681                     0.32    562076640       09fafd6c-6c99-46b1-834f-33527f4de241
2019-11-01 00:00:09 UTC cart    5844397 1487580006317032337                     2.38    553329724       2067216c-31b5-455d-a1cc-af0575a34ffb
2019-11-01 00:00:10 UTC view    5837166 1783999064103190764             pnb     22.22   556138645       57ed222e-a54a-4907-9944-5a875c2d7f4f
2019-11-01 00:00:11 UTC cart    5876812 1487580010100293687             jessnail        3.16    564506666       186c1951-8052-4b37-adce-dd9644b1d5f7
2019-11-01 00:00:24 UTC remove_from_cart        5826182 1487580007483048900                     3.33    553329724       2067216c-31b5-455d-a1cc-af0575a34ffb
Time taken: 3.323 seconds, Fetched: 5 row(s)
```

We are required to provide answers to the questions given below:

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

hive> select month(event_time) as month , sum(price) as total_price from comm where event_type ='purchase' and month(event_time)=10 group by month(event_time) ;

```
hive> select month(event_time) as month , sum(price) as total_price from comm where event_type ='purchase' and month(event_time)=10 group by month(event_time) ;
Query ID = hadoop_20221004113517_f7ea71bd-c086-4651-a49b-f6a91966645f
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1664877578682_0003)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      2          2        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      2          2        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 59.27 s
--------------------------------------------------------------------------------
OK
10      1211538.4299997438
Time taken: 62.642 seconds, Fetched: 1 row(s)
```

Total revenue generated due to purchases made in October 1211538.429

Here the query takes 62.642 seconds which can be optimized by creating dynamic partition and then compare the execution time.

Dynamic Partitioning and Bucketing:

hive> set hive.vectorized.execution.enabled =true ;

hive> set hive.vectorized.execution.enabled ;

hive> set hive.exec.dynamic.partition.mode = nonstrict ;

hive> set hive.exec.dynamic.partition.mode = true ;

```
hive> set hive.vectorized.execution.enabled =true ;
hive> set hive.vectorized.execution.enabled ;
hive.vectorized.execution.enabled=true
hive> set hive.exec.dynamic.partition.mode = nonstrict ;
hive> set hive.exec.dynamic.partition.mode = true ;
```

Creating a table by name comm_part to store the dataset which we partitioned by using 'month int'and clustered by 'event_type'.

hive> create table if not exists comm_part (event_time timestamp, event_type string , product_id string, category_id string , category_code string, brand string, price float ,user_id bigint, user_session string)partitioned by (month int) clustered by (event_type) into 4 buckets row format serde 'org.apache.hadoop.hive.serde2.OpenCSVSerde' stored as textfile ;

```
hive> create table if not exists comm_part (event_time timestamp, event_type string , product_id string, category_id string , category_code string, brand string, price float ,user_id bigint, user_session string )partitioned by (m
th int) clustered by (event_type) into 4 buckets row format serde 'org.apache.hadoop.hive.serde2.OpenCSVSerde' stored as textfile ;
OK
```

Loading the data into the new table:

hive> insert into table comm_part partition (month) select cast (replace(event_time ,'UTC','') as timestamp) , event_type , product_id , category_id , ctaegory_code , brand , price , user_id ,user_session , month(cast(replace(event_time , 'UTC', '') as timestamp)) from comm ;

```
hive> insert into table comm_part partition (month) select cast (replace(event_time ,'UTC','') as timestamp) , event_type , product_id , cat
_time , 'UTC', '') as timestamp)) from comm ;
Query ID = hadoop_20221004120211_4236ed93-c42e-4ed4-b812-370a9427063e
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1664877578682_0005)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 ......... container     SUCCEEDED      2          2        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      4          4        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 206.24 s
--------------------------------------------------------------------------------
Loading data to table default.comm_part partition (month=null)

Loaded : 2/2 partitions.
         Time taken to load dynamic partitions: 0.691 seconds
         Time taken for adding to write entity : 0.004 seconds
OK
Time taken: 209.438 seconds
```

Checking partition database

select * from comm_part limit 5 ;

```
hive> select * from comm_part limit 5 ;
OK
2019-10-31 23:58:09     cart    5820756 1487580006317032337                     2.22    553329724       2067216c-31b5-455d-a1cc-af0575a34ffb    10
2019-10-31 23:58:09     cart    5665855 1487580013900333275             irisk   2.06    566272508       8f8a6160-24b9-47b3-881d-75798a7d45ad    10
2019-10-31 23:57:08     cart    5850569 1998040852064109417                     6.35    566272508       8f8a6160-24b9-47b3-881d-75798a7d45ad    10
2019-10-31 23:57:05     cart    5850570 1998040852064109417                     6.35    566272508       8f8a6160-24b9-47b3-881d-75798a7d45ad    10
2019-10-31 23:56:54     cart    4653    1487580011157258342             runail  0.37    562691482       9025c3a5-9c56-49c4-9d3d-95a6c15b69a3    10
Time taken: 0.237 seconds, Fetched: 5 row(s)
```

Now executing the same Q1 in the partition database:

hive> select month , sum(price) as total_price from comm_part where event_type ='purchase' and month =10 group by month ;

```
hive> select month , sum(price) as total_price from comm_part where event_type ='purchase' and month =10 group by month ;
Query ID = hadoop_20221004120713_8d5cb65e-8e2a-4c6c-b91b-286b0cfff5ff
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1664877578682_0005)

--------------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      6          6        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      2          2        0        0       0       0
--------------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 44.56 s
--------------------------------------------------------------------------------------------------
OK
10      1211538.429999938
Time taken: 46.004 seconds, Fetched: 1 row(s)
```

We can notice how the time taken reduced drastically due to partitioning and bucketing. Now it took only 46 sec.

The total sales in the month of October is 1211538.42

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

hive> select month , sum(price) as total_price from comm_part where event_type ='purchase' group by month ;

```
hive> select month , sum(price) as total_price from comm_part where event_type ='purchase' group by month ;
Query ID = hadoop_20221004120856_afce9a9c-1c20-45af-962a-77f1754844a8
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1664877578682_0005)

--------------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      8          8        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      3          3        0        0       0       0
--------------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 70.25 s
--------------------------------------------------------------------------------------------------
OK
10      1211538.4299998013
11      1531016.8999998884
Time taken: 70.948 seconds, Fetched: 2 row(s)
```

Total revenue generated due to purchases made in October is 1211538.429

Total revenue generated due to purchases made in November is 1531016.899

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

hive> select month , sum(price) as total_price, sum(price)-lag(sum(price)) over (order by month) from comm_part where event_type ='purchase' group by month ;

```
hive> select month , sum(price) as total_price, sum(price)-lag(sum(price)) over (order by month) from comm_part where event_type ='purchase' group by month ;
Query ID = hadoop_20221004121051_3c08781c-e818-4c24-b88f-ca793c7b5cd7
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1664877578682_0005)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      8          8        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      3          3        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      2          2        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 70.84 s
--------------------------------------------------------------------------------
OK
10      1211538.4299998013      NULL
11      1531016.8999998884      319478.47000008705
Time taken: 71.614 seconds, Fetched: 2 row(s)
```

We can see the difference in the revenue is 319478.47

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

hive> select distinct(category_code) from comm_part where category_code is not null ;

```
hive> select distinct(category_code) from comm_part where category_code is not null ;
Query ID = hadoop_20221004121249_9299fd23-9542-4530-a8fc-0c0e545f36d8
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1664877578682_0005)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      8          8        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      5          5        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 70.74 s
--------------------------------------------------------------------------------
OK

accessories.cosmetic_bag
stationery.cartrige
accessories.bag
appliances.environment.vacuum
furniture.living_room.chair
sport.diving
appliances.personal.hair_cutter
appliances.environment.air_conditioner
apparel.glove
furniture.bathroom.bath
furniture.living_room.cabinet
Time taken: 71.439 seconds, Fetched: 12 row(s)
```

We can see the distinct categories are Furniture, Appliances, Accessories, Apparel, Sport, Stationery

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

hive> select category_code , count(product_id) as Product_count from comm_part where category_code is not null group by category_code ;

```
hive> select category_code , count(product_id) as Product_count from comm_part where category_code is not null group by category_code ;
Query ID = hadoop_20221004121442_13421b43-bcf1-440c-b5bf-756103b6abef
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1664877578682_0005)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      8          8        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      5          5        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 71.11 s
--------------------------------------------------------------------------------
OK
        8594895
accessories.cosmetic_bag        1248
stationery.cartrige     26722
accessories.bag 11681
appliances.environment.vacuum   59761
furniture.living_room.chair     308
sport.diving    2
appliances.personal.hair_cutter 1643
appliances.environment.air_conditioner  332
apparel.glove   18232
furniture.bathroom.bath 9857
furniture.living_room.cabinet   13439
Time taken: 71.763 seconds, Fetched: 12 row(s)
```

The total number of products under each category is as follows:

Category	Sub-category	Number of Sub-Category	Number of Category
Appliances	vacuum	59761	61736
	hair_cutter	1643	
	air_conditioner	332	
Stationery		26722	26722
Furniture	chair	308	23604
	bath	9857	
	room.cabinet	13439	
Apparel		18232	18232
Accessories	bag	11681	12929
	cosmetic_bag	1248	
Sport	Sport	2	2

Appliances 61736, Stationery 26722, Furniture 23604, Apparel 18232, Accessories 12929, Sport 2

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

hive> select brand , sum(price) as total_sales from comm_part where event_type ='purchase' group by brand order by total_sales desc ;

```
hive> select brand , sum(price) as total_sales from comm_part where event_type ='purchase' group by brand order by total_sales  desc ;
Query ID = hadoop_20221004121757_1db3680a-017a-458e-a06a-14ef122296fb
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1664877578682_0005)

--------------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      8          8        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      3          3        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 63.38 s
--------------------------------------------------------------------------------------------------
OK
        1094188.3000000368
runail  148297.9399999991
grattol 106918.25000000291
irisk   92538.00000000003
uno     86341.77999999987
strong  67867.90000000002
masura  64324.54999999887
jessnail        59633.070000000334
cnd     59240.77000000009
ingarden        56727.60000000019
italwax 46739.609999999826
estel   45899.420000000086
browxenna       29248.09999999998
bpw.style       26409.590000000564
kapous  26020.24000000001
concept 24412.539999999855
cosmoprofi      22859.800000000007
beautix 22716.89999999998
domix   22481.219999999914
lianail 22287.080000000024
haruyama        21743.59999999996
max     21412.000000000004
bluesky 20872.769999999924
lovely  20643.439999999973
yoko    20464.789999999986
staleks 20395.34000000002
```

We can see that Runail is the brand with the maximum sales for oct and nov. Total sales is 148297.94

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

hive> WITH monthly_sales AS (select brand , sum(CASE WHEN month(event_time)='10' then price else 0 end) as oct_rev,sum(CASE WHEN month(event_time)='11' then price else 0 end) as nov_rev from comm_part where event_type ='purchase' group by brand) select brand,nov_rev,oct_rev,(nov_rev-oct_rev) as inc_sales from monthly_sales where (nov_rev-oct_rev)>0 order by inc_sales desc

```
hive> WITH  monthly_sales AS (select brand , sum(CASE WHEN month(event_time)='10' then price else 0 end) as oct_rev,sum(CASE WHEN month(event_time)='11'
 group by brand ) select brand,nov_rev,oct_rev,(nov_rev-oct_rev) as  inc_sales from monthly_sales where (nov_rev-oct_rev)>0 order by inc_sales desc;
Query ID = hadoop_20221004122746_77d48fcf-63e8-4f26-8362-d5e5525ecd82
Total jobs = 1
Launching Job 1 out of 1
Tez session was closed. Reopening...
Session re-established.
Status: Running (Executing on YARN cluster with App id application_1664877578682_0006)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 ......... container     SUCCEEDED      8          8        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      3          3        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 77.10 s
--------------------------------------------------------------------------------
OK
        619509.24000001 474679.06000002683      144830.17999998317
grattol 71472.71000000264       35445.54000000027       36027.17000000237
uno     51039.74999999982       35302.03000000006       15737.719999999761
lianail 16394.24000000005       5892.8399999999765      10501.400000000072
ingarden        33566.21000000015       23161.390000000043      10404.820000000109
strong  38671.270000000004      29196.630000000012      9474.639999999992
jessnail        33345.230000000156      26287.840000000018      7057.389999999978
cosmoprofi      14536.99000000002       8322.809999999989       6214.180000000031
polarus 11371.930000000004      6013.72 5358.210000000004
runail  76758.65999999973       71539.27999999939       5219.380000000339
freedecor       7671.799999999954       3421.78 4250.019999999953
staleks 11875.610000000008      8519.73000000001        3355.879999999974
bpw.style       14837.440000000282      11572.150000000285      3265.2899999999972
lovely  11939.059999999983      8704.379999999988       3234.679999999995
marathon        10273.100000000006      7280.750000000004       2992.350000000002
haruyama        12352.910000000047      9390.68999999991        2962.2200000001376
yoko    11707.87999999999       8756.909999999998       2950.969999999992
italwax 24799.369999999926      21940.239999999896      2859.13000000003
benovy  3259.9699999999993      409.61999999999995      2850.3499999999995
kaypro  3268.7  881.34  2387.3599999999997
estel   24142.670000000082      21756.750000000007      2385.9200000000747
concept 13380.399999999934      11032.139999999918      2348.2600000000166
```

```
konad   810.67  739.83000000000002      70.83999999999998
egomania        146.04000000000002      77.47   68.57000000000002
cutrin  367.62  299.37000000000006      68.24999999999994
laboratorium    312.52  246.49999999999994      66.02000000000004
inm     351.21  288.02  63.19
dewal   61.28999999999999      0.0     61.28999999999999
marutaka-foot   109.33  49.21999999999999       60.11000000000001
kares   59.45   0.0     59.45
profhenna       736.8499999999999       679.2299999999998       57.62000000000012
koelcia 112.75  55.49999999999999       57.25000000000001
balbcare        212.38000000000002      155.33000000000004      57.04999999999998
elskin  307.65000000000003      251.09000000000003      56.56
foamie  80.49   35.04   45.449999999999996
ladykin 170.57  125.64999999999999      44.92
likato  340.9699999999999       296.06  44.90999999999991
mavala  446.32000000000005      409.04  37.28000000000003
vilenta 231.20999999999995      197.59999999999994      33.610000000000014
beautyblender   109.41  78.74000000000001       30.669999999999987
biore   90.31   60.650000000000006      29.659999999999997
orly    931.0900000000003       902.3800000000001       28.71000000000015
estelare        471.87  444.8100000000003       27.059999999999718
profepil        118.02000000000002      93.36000000000001       24.66000000000001
blixz   63.39999999999999       38.94999999999999       24.450000000000003
binacil 24.259999999999998      0.0     24.259999999999998
godefroy        425.12  401.22  23.899999999999977
glysolid        91.59   69.72999999999999       21.860000000000014
veraclara       71.21000000000001       50.11   21.10000000000001
juno    21.08   0.0     21.08
kamill  81.49000000000001       63.010000000000005      18.480000000000004
treaclemoon     181.49  163.37000000000006      18.119999999999948
supertan        66.51   50.370000000000005      16.14
barbie  12.39   0.0     12.39
deoproce        329.17  316.84000000000003      12.329999999999984
rasyan  28.939999999999998      18.799999999999997      10.14
fly     27.17   17.14   10.030000000000001
tertio  245.79999999999998      236.16000000000003      9.639999999999958
jaguar  1110.65 1102.11 8.540000000000191
soleo   212.52999999999983      204.19999999999948      8.330000000000354
neoleor 51.7    43.41   8.290000000000006
moyou   10.280000000000001      5.71    4.570000000000001
bodyton 1380.6400000000003      1376.3400000000001      4.300000000000182
skinity 12.440000000000001      8.88    3.5600000000000005
helloganic      3.1     0.0     3.1
grace   102.60999999999999      100.91999999999996      1.6900000000000261
cosima  20.93   20.229999999999997      0.7000000000000028
ovale   3.1     2.54    0.56
Time taken: 88.032 seconds, Fetched: 161 row(s)
```

From the output we can see that 161 brand were able to increase their sales from the month of October to November

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

hive> select user_id , sum(price)as total_spent from comm_part where event_type='purchase' group by user_id order by total_spent desc limit 10 ;

```
hive> select user_id , sum(price)as total_spent from comm_part where event_type='purchase' group by user_id order by total_spent desc limit 10 ;
Query ID = hadoop_20221004134940_3e6f885b-f0f0-4046-ace5-acf522d43de5
Total jobs = 1
Launching Job 1 out of 1
Tez session was closed. Reopening...
Session re-established.
Status: Running (Executing on YARN cluster with App id application_1664877578682_0009)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      8          8        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      3          3        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 74.05 s
--------------------------------------------------------------------------------
OK
557790271       2715.869999999991
150318419       1645.9699999999998
562167663       1352.8499999999997
531900924       1329.4500000000003
557850743       1295.4800000000002
522130011       1185.3899999999999
561592095       1109.7000000000007
431950134       1097.5899999999997
566576008       1056.36
521347209       1040.9099999999999
Time taken: 88.185 seconds, Fetched: 10 row(s)
```

We can see the top 10 users id with total purchases in the output who can be included in the Gold Plan.

Finishing Up

Once we are done, we can drop the databases, quit the hive and then terminate the EMR cluster.