Hive Case Study DA track

Problem Statement : With online sales gaining popularity, tech companies are exploring ways to improve their sales by analyzing customer behavior and gaining insights about product trends. Furthermore, the websites make it easier for customers to find the products they require without much scavenging. 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.

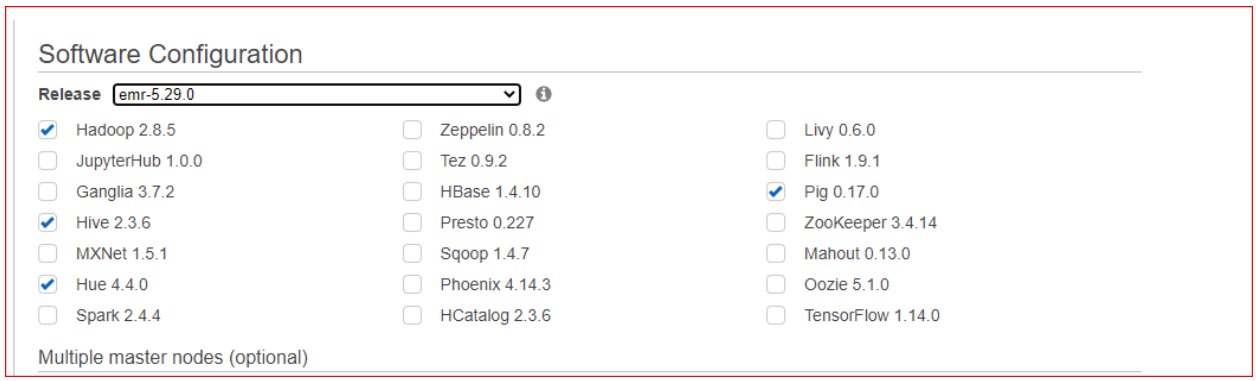
Objective: Need to analyze and gain insights about the clickstream data from a website so that we can extract insights about the customers behavior.

The steps involved in the entire process are as follows:

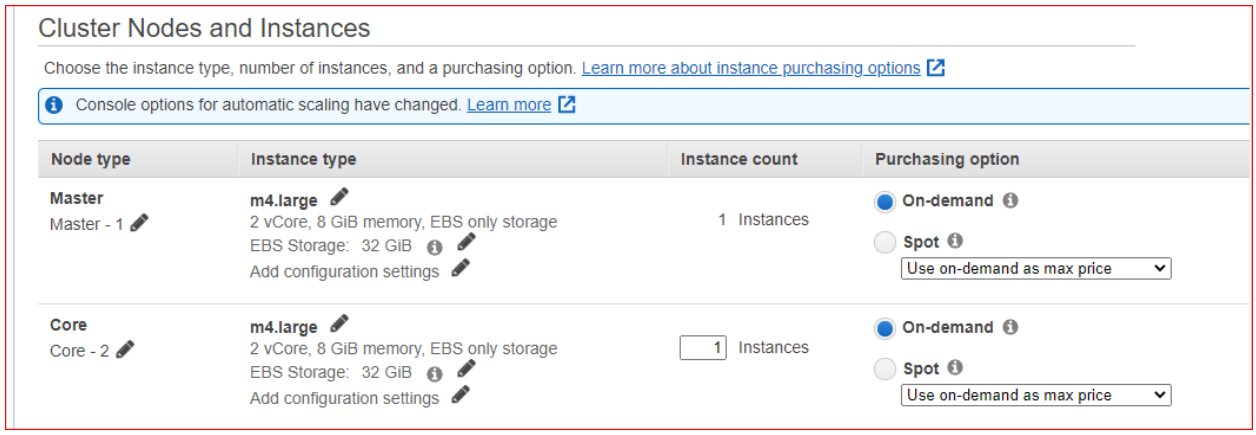
- ✓ Copying the data set into the HDFS:
  - Launch an EMR cluster that utilizes the Hive services .
  - Move the data from the S3 bucket into the HDFS .
- ✓ Creating the database and launching Hive queries on your EMR cluster:
  - Create the schema of database
  - Run Hive queries to answer the questions given below.
  - Use optimized techniques to run queries with Higher efficiency .
  - Notice improvement of the performance after using optimization on any single query.
  - Terminate the Cluster .

Launching EMR cluster

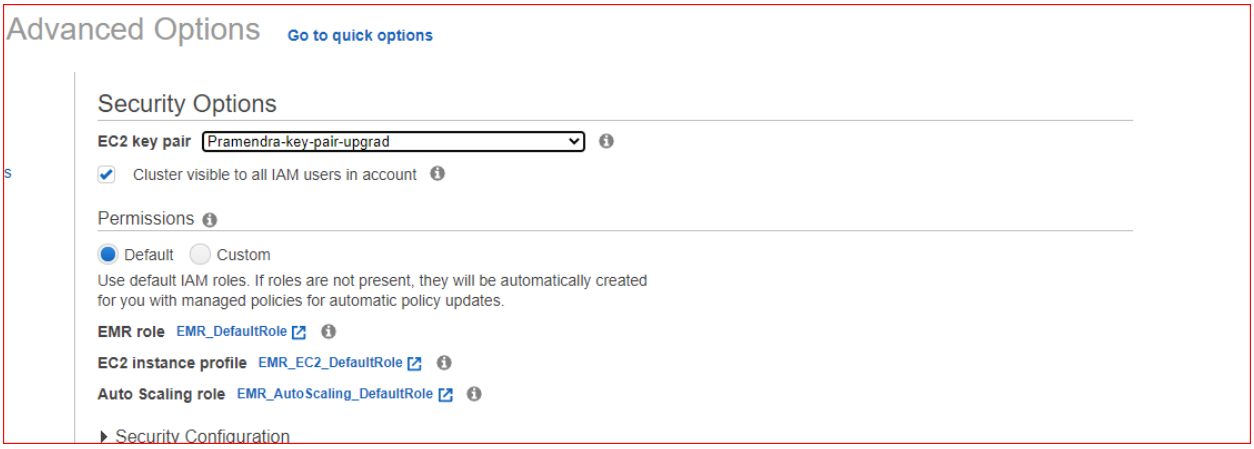
- ✓ As suggested in Assignment details , we are using below while creating clusters .
  - ✓ 2-node EMR cluster with both the master and core nodes as M4.large.
  - ✓ emr-5.29.0 release for this case study



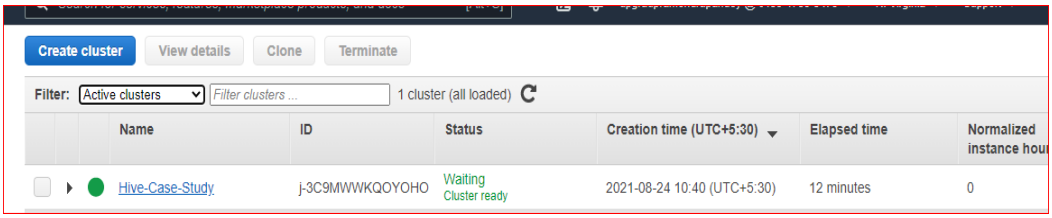
- ✓ Selecting m4.large instance and 2 node .



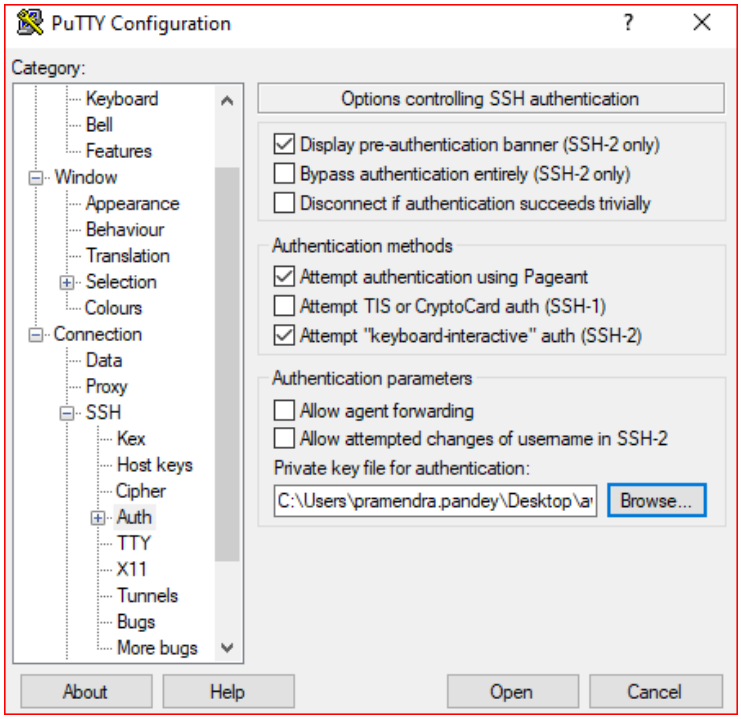
✓ Selecting key-pair



✓ Cluster is in wait status -Ready to connect , See below



✓ Now the cluster is created and in wait status and connect to cluster via putty .



```
login as: hadoop
Authenticating with public key "imported-openssh-key"
Last login: Tue Aug 24 05:26:09 2021 from 223.177.106.137

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

https://aws.amazon.com/amazon-linux-ami/2018.03-release-notes/
68 package(s) needed for security, out of 107 available
Run "sudo yum update" to apply all updates.

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-172-31-81-162 ~]$
```

✓ Command to check directories present already on HDFS

Hadoop fs -ls

```
[hadoop@ip-172-31-85-157 ~]$ hadoop fs -ls /
Found 4 items
drwxr-xr-x   - hdfs hadoop          0 2021-08-25 02:37 /apps
drwxrwxrwt   - hdfs hadoop          0 2021-08-25 02:39 /tmp
drwxr-xr-x   - hdfs hadoop          0 2021-08-25 02:37 /user
drwxr-xr-x   - hdfs hadoop          0 2021-08-25 02:37 /var
[hadoop@ip-172-31-85-157 ~]$
```

✓ Creating new temporary directory i.e., Hive_assignment' to store data file in the directory (Permanent) i.e., 'user' & further check if directory-‘ Hive_assignment’ is created .

hadoop fs -mkdir /user/Hive_assignment/

```
[hadoop@ip-172-31-85-157 ~]$ hadoop fs -mkdir /user/Hive_assignment/
[hadoop@ip-172-31-85-157 ~]$ hadoop fs -ls /user/
Found 7 items
drwxr-xr-x   - hadoop hadoop          0 2021-08-25 03:02 /user/Hive_assignment
drwxrwxrwx   - hadoop hadoop          0 2021-08-25 02:37 /user/hadoop
drwxr-xr-x   - mapred mapred          0 2021-08-25 02:37 /user/history
drwxrwxrwx   - hdfs   hadoop          0 2021-08-25 02:37 /user/hive
drwxrwxrwx   - hue    hue             0 2021-08-25 02:37 /user/hue
drwxrwxrwx   - oozie  oozie           0 2021-08-25 02:37 /user/oozie
drwxrwxrwx   - root   hadoop          0 2021-08-25 02:37 /user/root
[hadoop@ip-172-31-85-157 ~]$
```

✓ loading data file '2019-Oct.csv' from S3 storage into HDFS storage as 'October.csv' .

hadoop distcp s3://ml-pramendra-dataset/2019-Oct.csv
/user/Hive_assignment/October.csv

```
[hadoop@ip-172-31-85-157 ~]$ hadoop distcp s3://ml-pramendra-dataset/2019-Oct.csv /user/Hive_assignment/October.csv
21/08/25 03:45:19 INFO tools.DistCp: Input Options: DistCpOptions{atomicCommit=false, syncFolder=false, deleteMissing=false, ignoreFailures=false, overwrite=false, skip
CRC=false, blocking=true, numListstatusThreads=0, maxMaps=20, mapBandwidth=100, sslConfigurationFile='null', copyStrategy='uniformsize', preserveStatus=[], preserveRawX
attrs=false, atomicWorkPath=null, logPath=null, sourceFileListing=null, sourcePaths=[s3://ml-pramendra-dataset/2019-Oct.csv], targetPath=/user/Hive_assignment/October.c
sv, targetPathExists=false, filtersFile='null'}
21/08/25 03:45:19 INFO client.RMProxy: Connecting to ResourceManager at ip-172-31-85-157.ec2.internal/172.31.85.157:8032
21/08/25 03:45:23 INFO tools.SimpleCopyListing: Paths (files+dirs) cnt = 1; dirCnt = 0
21/08/25 03:45:23 INFO tools.SimpleCopyListing: Build file listing completed.
21/08/25 03:45:23 INFO Configuration.deprecation: io.sort.mb is deprecated. Instead, use mapreduce.task.io.sort.mb
21/08/25 03:45:23 INFO Configuration.deprecation: io.sort.factor is deprecated. Instead, use mapreduce.task.io.sort.factor
21/08/25 03:45:24 INFO tools.DistCp: Number of paths in the copy list: 1
21/08/25 03:45:24 INFO tools.DistCp: Number of paths in the copy list: 1
21/08/25 03:45:24 INFO client.RMProxy: Connecting to ResourceManager at ip-172-31-85-157.ec2.internal/172.31.85.157:8032
21/08/25 03:45:24 INFO mapreduce.JobSubmitter: number of splits:1
21/08/25 03:45:24 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1629859092088_0004
21/08/25 03:45:24 INFO impl.YarnClientImpl: Submitted application application_1629859092088_0004
21/08/25 03:45:24 INFO mapreduce.Job: The url to track the job: http://ip-172-31-85-157.ec2.internal:20888/proxy/application_1629859092088_0004/
21/08/25 03:45:24 INFO tools.DistCp: DistCp job-id: job_1629859092088_0004
21/08/25 03:45:24 INFO mapreduce.Job: Running job: job_1629859092088_0004
21/08/25 03:45:32 INFO mapreduce.Job: Job job_1629859092088_0004 running in uber mode : false
21/08/25 03:45:32 INFO mapreduce.Job:  map 0% reduce 0%
21/08/25 03:45:50 INFO mapreduce.Job:  map 100% reduce 0%
21/08/25 03:45:52 INFO mapreduce.Job: Job job_1629859092088_0004 completed successfully
21/08/25 03:45:52 INFO mapreduce.Job: Counters: 38
        File System Counters
                FILE: Number of bytes read=0
                FILE: Number of bytes written=172499
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=359
                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
                S3: Number of bytes read=482542278
                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=1
                Launched map tasks=1
                Other local map tasks=1
                Total time spent by all maps in occupied slots (ms)=544672
                Total time spent by all reduces in occupied slots (ms)=0
                Total time spent by all map tasks (ms)=17021
                Total vcore-milliseconds taken by all map tasks=17021
                Total megabyte-milliseconds taken by all map tasks=17429504
        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)=287
                CPU time spent (ms)=20210
                Physical memory (bytes) snapshot=574955520
                Virtual memory (bytes) snapshot=3302424576
                Total committed heap usage (bytes)=490733568
        File Input Format Counters
                Bytes Read=223
        File Output Format Counters
                Bytes Written=0
        DistCp Counters
                Bytes Copied=482542278
                Bytes Expected=482542278
                Files Copied=1
```

✓ loading data file '2019-Nov.csv' from S3 storage into HDFS storage as 'November.csv'.

hadoop distcp s3://ml-pramendra-dataset/2019-Nov.csv /user/Hive_assignment/November.csv

```
[hadoop@ip-172-31-85-157 ~]$ hadoop distcp s3://ml-pramendra-dataset/2019-Nov.csv /user/Hive_assignment/November.csv
21/08/25 03:56:54 INFO tools.DistCp: Input Options: DistCpOptions{atomicCommit=false, syncFolder=false, deleteMissing=false, ignoreFailures=false, overwrite=false, skip
CRC=false, blocking=true, numListstatusThreads=0, maxMaps=20, mapBandwidth=100, sslConfigurationFile='null', copyStrategy='uniformsize', preserveStatus=[], preserveRawX
attrs=false, atomicWorkPath=null, logPath=null, sourceFileListing=null, sourcePaths=[s3://ml-pramendra-dataset/2019-Nov.csv], targetPath=/user/Hive_assignment/November.
csv, targetPathExists=false, filtersFile='null'}
21/08/25 03:56:54 INFO client.RMProxy: Connecting to ResourceManager at ip-172-31-85-157.ec2.internal/172.31.85.157:8032
21/08/25 03:56:59 INFO tools.SimpleCopyListing: Paths (files+dirs) cnt = 1; dirCnt = 0
21/08/25 03:56:59 INFO tools.SimpleCopyListing: Build file listing completed.
21/08/25 03:56:59 INFO Configuration.deprecation: io.sort.mb is deprecated. Instead, use mapreduce.task.io.sort.mb
21/08/25 03:56:59 INFO Configuration.deprecation: io.sort.factor is deprecated. Instead, use mapreduce.task.io.sort.factor
21/08/25 03:56:59 INFO tools.DistCp: Number of paths in the copy list: 1
21/08/25 03:56:59 INFO tools.DistCp: Number of paths in the copy list: 1
21/08/25 03:56:59 INFO client.RMProxy: Connecting to ResourceManager at ip-172-31-85-157.ec2.internal/172.31.85.157:8032
21/08/25 03:56:59 INFO mapreduce.JobSubmitter: number of splits:1
21/08/25 03:56:59 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1629859092088_0005
21/08/25 03:57:00 INFO impl.YarnClientImpl: Submitted application application_1629859092088_0005
21/08/25 03:57:00 INFO mapreduce.Job: The url to track the job: http://ip-172-31-85-157.ec2.internal:20888/proxy/application_1629859092088_0005/
21/08/25 03:57:00 INFO tools.DistCp: DistCp job-id: job_1629859092088_0005
21/08/25 03:57:00 INFO mapreduce.Job: Running job: job_1629859092088_0005
21/08/25 03:57:08 INFO mapreduce.Job: Job job_1629859092088_0005 running in uber mode : false
21/08/25 03:57:08 INFO mapreduce.Job:  map 0% reduce 0%
21/08/25 03:57:25 INFO mapreduce.Job:  map 100% reduce 0%
21/08/25 03:57:29 INFO mapreduce.Job: Job job_1629859092088_0005 completed successfully
21/08/25 03:57:29 INFO mapreduce.Job: Counters: 38
        File System Counters
                FILE: Number of bytes read=0
                FILE: Number of bytes written=172501
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=359
                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
                S3: Number of bytes read=545839412
                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=1
                Other local map tasks=1
```

```
                Total time spent by all maps in occupied slots (ms)=589376
                Total time spent by all reduces in occupied slots (ms)=0
                Total time spent by all map tasks (ms)=18418
                Total vcore-milliseconds taken by all map tasks=18418
                Total megabyte-milliseconds taken by all map tasks=18860032
        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)=338
                CPU time spent (ms)=20050
                Physical memory (bytes) snapshot=614354944
                Virtual memory (bytes) snapshot=3304546304
                Total committed heap usage (bytes)=513802240
        File Input Format Counters
                Bytes Read=223
        File Output Format Counters
                Bytes Written=0
        DistCp Counters
                Bytes Copied=545839412
                Bytes Expected=545839412
                Files Copied=1
```

- ✓ Validate if data is loaded successful inside directory user//Hive_assignment

  hadoop fs -ls /user//Hive_assignment

```
[hadoop@ip-172-31-85-157 ~]$ hadoop fs -ls  /user//Hive_assignment
Found 2 items
-rw-r--r--   1 hadoop hadoop  545839412 2021-08-25 03:57 /user/Hive_assignment/November.csv
-rw-r--r--   1 hadoop hadoop  482542278 2021-08-25 03:45 /user/Hive_assignment/October.csv
[hadoop@ip-172-31-85-157 ~]$
```

- ✓ Starting Hive

```
[hadoop@ip-172-31-85-157 ~]$ hive

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

- ✓ Create External Table **ecom** which will hold the data for both the data files stored in temporary directory of HDFS.

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

```
Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j2.properties Async: false
hive> CREATE EXTERNAL TABLE IF NOT EXISTS ecom (event_time timestamp, event_type string, product_id string, category_id string, category_code string, brand string, pric
e float, user_id bigint, user_session string) ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.OpenCSVSerde' STORED AS TEXTfILE LOCATION '/user/Hive_assignment/' tblprop
erties("skip.header.line.count"="1");
OK
Time taken: 0.973 seconds
```

- ✓ Reading the file ecom first five rows

  set hive.cli.print.header=True ;

  Select * from ecom limit 5;

```
hive> select * from ecom limit 5;
OK
ecom.event_time ecom.event_type ecom.product_id ecom.category_id        ecom.category_code      ecom.brand      ecom.price      ecom.user_id    ecom.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
Time taken: 2.522 seconds, Fetched: 5 row(s)
```

Hive Questions/Answers

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

Solution :-

```
SELECT
SUM(price) AS Total_Revenue_October
FROM ecom
WHERE date_format(event_time, 'MM') = 10
AND event_type = 'purchase';
```

```
Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j2.properties Async: false
hive> SELECT SUM(price) AS Total_Revenue_October FROM ecom WHERE date_format(event_time, 'MM')=10 AND event_type='purchase';
Query ID = hadoop_20210825041449_8850ac3d-221b-491a-a047-6457f6d74b84
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1629859092088_0008)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 ..... container     SUCCEEDED      2          2        0        0       0       0
Reducer 2 ..... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 120.51 s
--------------------------------------------------------------------------------
OK
1211538.4299997438
Time taken: 125.797 seconds, Fetched: 1 row(s)
```

Insights :- The total revenue generated based on Purchase in the month of October of 2019 was **1,211,538/-** .

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

Solution :

```
SELECT
date_format(event_time, 'MM') AS Months,
COUNT(event_type) AS Sum_of_Purchases
FROM ecom
WHERE event_type = 'purchase'
GROUP BY date_format(event_time, 'MM');
```

```
hive> SELECT date_format(event_time, 'MM') AS Months, COUNT(event_type) AS Sum_of_Purchases FROM ecom WHERE event_type='purchase' GROUP BY date_format(event_time, 'MM')
;
Query ID = hadoop_20210825042558_5eb4864d-409a-4648-906d-d8d112d47778
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1629859092088_0010)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 ..... container     SUCCEEDED      2          2        0        0       0       0
Reducer 2 ..... container     SUCCEEDED      3          3        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 57.64 s
--------------------------------------------------------------------------------
OK
10      245624
11      322417
Time taken: 63.046 seconds, Fetched: 2 row(s)
```

Insight – In November , Purchases were higher than October Month .

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

Solution :-

```
WITH Month_Revenue
AS (SELECT
  SUM(CASE
    WHEN date_format(event_time, 'MM') = 10 THEN price
    ELSE 0
  END) AS Oct_Revenue,
  SUM(CASE
    WHEN date_format(event_time, 'MM') = 11 THEN price
    ELSE 0
  END) AS Nov_Revenue
FROM ecom
WHERE event_type = 'purchase'
AND date_format(event_time, 'MM') IN ('10', '11'))
SELECT
  Nov_Revenue,
  Oct_Revenue,
  (Nov_Revenue - Oct_Revenue) AS Revenue_Difference
FROM Month_Revenue;
```

```
hive>
    > WITH Month_Revenue AS ( SELECT SUM(CASE WHEN date_format(event_time, 'MM')=10 THEN price ELSE 0 END) AS Oct_Revenue, SUM(CASE WHEN date_format(event_time, 'MM')=1
1 THEN price ELSE 0 END) AS Nov_Revenue FROM ecom WHERE event_type= 'purchase' AND date_format(event_time, 'MM') in ('10', '11') ) SELECT Nov_Revenue, Oct_Revenue, (Nov
_Revenue - Oct_Revenue) AS Revenue_Difference FROM Month_Revenue ;
Query ID = hadoop_20210825160804_1f14b015-896e-4794-a2c0-67c33646a87b
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1629905929129_0009)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 ..........   container     SUCCEEDED      2          2        0        0       0       0
Reducer 2 ......   container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 66.49 s
----------------------------------------------------------------------------------------------
OK
nov_revenue     oct_revenue     revenue_difference
1531016.900000122       1211538.4299997438      319478.4700003781
Time taken: 73.473 seconds, Fetched: 1 row(s)
```

Insight :

- Company has better sale in November than October .
- Revenue generated in November of 2019 was more than the revenue generated in the month of October. In other words, November was more profitable for the company than October

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

Solution –

select

```
distinct(split(category_code,"\\."))[0] as categoty_code
from
ecom
where split(category_code,"\\.")[0] <> '';
```

```
hive> select distinct(split(category_code,"\\."))[0] as categoty_code from ecom where split(category_code,"\\.")[0] <> '';
Query ID = hadoop_20210825164656_7ee79610-98c8-47ee-808a-e2319b198c27
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1629905929129_0014)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 ..... container     SUCCEEDED      2          2        0        0       0       0
Reducer 2 ..... container     SUCCEEDED      5          5        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 59.52 s
--------------------------------------------------------------------------------
OK
categoty_code
furniture
appliances
accessories
apparel
sport
stationery
Time taken: 60.268 seconds, Fetched: 6 row(s)
hive>
```

Insights : There is total 6 different categories under which company sells their different products.

Category -furniture ,appliances, accessories, apparel, sport & stationery

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

Solution :

```
SELECT SPLIT(category_code,'\\.')[0] AS Category, COUNT(product_id) AS No_of_products
FROM ecom
WHERE SPLIT(category_code,'\\.')[0] <> ''
GROUP BY
SPLIT(category_code,'\\.')[0]
ORDER BY
No_of_products DESC;
```

```
hive> SELECT SPLIT(category_code,'\\.')[0] AS Category, COUNT(product_id) AS No_of_products
    > FROM ecom
    > WHERE SPLIT(category_code,'\\.')[0] <> ''
    > GROUP BY SPLIT(category_code,'\\.')[0]
    > ORDER BY No_of_products DESC;
Query ID = hadoop_20210825165323_922786e9-824a-42b1-adf6-f95a529ad34c
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1629905929129_0016)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      2          2        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      5          5        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 60.01 s
--------------------------------------------------------------------------------
OK
category        no_of_products
appliances      61736
stationery      26722
furniture       23604
apparel 18232
accessories     12929
sport   2
Time taken: 66.211 seconds, Fetched: 6 row(s)
hive>
```

Insights:

- Company has highest products registered under Appliances category i.e., **61736** products than any other categories.
- Then it is followed by stationery as second with 26,722 products, furniture as third with 23,604 products, apparel as fourth with 18232 products registered, accessories as fifth with 12929 products.
- Sports category has only 2 products registered -lowest .

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

Solutions :

```
SELECT
brand  , sum(price) as Total_Sale from ecom  where ( event_type = 'purchase' and brand <> '' )
group by
brand
order by
Total_Sale desc
limit 1 ;
```

```
hive> select brand  , sum(price) as Total_Sale from ecom  where ( event_type = 'purchase' and brand <> '' )
    > group by brand
    > order by Total_Sale desc
    > limit 1 ;
Query ID = hadoop_20210825173709_d36d9e7e-8145-4efa-8748-8458be200566
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1629905929129_0022)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      2          2        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: 57.58 s
--------------------------------------------------------------------------------
OK
runail  148297.9400000003
Time taken: 72.328 seconds, Fetched: 1 row(s)
hive>
```

Insights :

- Runail is the brand that has highest / maximum sales in the month of October and November of 2019 combined.

Question 7: Which brands increased their sales from October to November?
Solution :-

```
WITH Monthly_Revenue AS (
SELECT brand,
SUM(CASE WHEN date_format(event_time, 'MM')=10 THEN price ELSE 0 END) AS Oct_Revenue,
SUM(CASE WHEN date_format(event_time, 'MM')=11 THEN price ELSE 0 END) AS Nov_Revenue
FROM ecom
WHERE event_type='purchase'
AND
date_format(event_time, 'MM') IN ('10', '11')
GROUP BY brand
)
SELECT brand, Oct_Revenue, Nov_Revenue, Nov_Revenue-Oct_Revenue AS Sales_Difference
FROM Monthly_Revenue
WHERE (Nov_Revenue - Oct_Revenue)>0
ORDER BY Sales_Difference;
```

```
hive>
    > WITH Monthly_Revenue AS (
    > SELECT brand,
    > SUM(CASE WHEN date_format(event_time, 'MM')=10 THEN price ELSE 0 END) AS Oct_Revenue,
    > SUM(CASE WHEN date_format(event_time, 'MM')=11 THEN price ELSE 0 END) AS Nov_Revenue
    > FROM ecom
    > WHERE event_type='purchase'
    > AND
    > date_format(event_time, 'MM') IN ('10', '11')
    > GROUP BY brand
    > )
    > SELECT brand, Oct_Revenue, Nov_Revenue, Nov_Revenue-Oct_Revenue AS Sales_Difference
    > FROM Monthly_Revenue
    > WHERE (Nov_Revenue - Oct_Revenue)>0
    > ORDER BY Sales_Difference;
Query ID = hadoop_20210825174404_a1f941d7-2d87-4536-99a1-6aa21a9c59e3
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1629905929129_0023)

----------------------------------------------------------------------------------------------
        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
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 68.45 s
----------------------------------------------------------------------------------------------
OK
brand   oct_revenue     nov_revenue     sales_difference
ovale   2.54    3.1     0.56
cosima  20.23   20.929999999999993      0.6999999999999922
grace   100.92000000000002      102.61000000000001      1.6899999999999977
helloganic      0.0     3.1     3.1
skinity 8.88    12.440000000000001      3.5600000000000005
bodyton 1376.3399999999974      1380.6399999999992      4.3000000000017735
moyou   5.71    10.280000000000001      4.570000000000001
neoleor 43.41   51.7    8.290000000000006
soleo   204.2000000000003       212.5299999999998       8.329999999999501
```

```
de.lux  1659.699999999967       2775.509999999968       1115.8100000000009
swarovski       1887.9299999999873      3043.160000000003       1155.2300000000157
beauty-free     554.1700000000006       1782.8600000000163      1228.6900000000155
zeitun  708.6600000000004       2009.63 1300.9699999999998
joico   705.52  2015.1000000000015      1309.5800000000015
severina        4775.88 6120.480000000023       1344.600000000023
irisk   45591.96000000588       46946.040000002184      1354.0799999963056
oniq    8425.41000000003        9841.650000000018       1416.239999999987
levrana 2243.560000000002       3664.099999999998       1420.5399999999959
roubloff        3491.360000000003       4913.769999999991       1422.4099999999885
smart   4457.260000000004       5902.140000000017       1444.8800000000128
shik    3341.2  4839.720000000007       1498.5200000000068
domix   10472.04999999994       12009.170000000022      1537.1200000000827
artex   2730.639999999998       4327.250000000017       1596.6100000000192
beautix 10493.949999999966      12222.949999999913      1728.9999999999472
milv    3904.9399999999964      5642.01000000008        1737.0700000000838
masura  31266.07999999821       33058.46999999708       1792.3899999988753
f.o.x   6624.229999999982       8577.280000000004       1953.050000000022
kapous  11927.159999999898      14093.080000000158      2165.92000000026
concept 11032.139999999925      13380.39999999993       2348.2600000000057
estel   21756.750000000342      24142.67000000022       2385.919999999878
kaypro  881.3399999999998       3268.699999999995       2387.359999999995
benovy  409.6200000000002       3259.970000000001       2850.350000000001
italwax 21940.239999999732      24799.369999999893      2859.130000000161
yoko    8756.909999999949       11707.879999999996      2950.9700000000466
haruyama        9390.689999999991       12352.91000000013       2962.2200000001394
marathon        7280.749999999997       10273.1 2992.350000000003
lovely  8704.379999999952       11939.060000000045      3234.680000000093
bpw.style       11572.150000001699      14837.440000000812      3265.289999999113
staleks 8519.730000000003       11875.61000000008       3355.8800000000774
freedecor       3421.779999999971       7671.800000000175       4250.020000000204
runail  71539.27999999933       76758.66000000098       5219.380000001649
polarus 6013.720000000003       11371.930000000018      5358.2100000000155
cosmoprofi      8322.81000000007        14536.99000000016       6214.180000000089
jessnail        26287.839999999916      33345.22999999992       7057.390000000007
strong  29196.62999999994       38671.269999999924      9474.639999999985
ingarden        23161.390000000138      33566.21000000009       10404.819999999949
lianail 5892.839999999975       16394.240000000245      10501.40000000027
uno     35302.02999999977       51039.749999998035      15737.719999998262
grattol 35445.5400000011        71472.71000000068       36027.169999999576
        474679.0599999623       619509.2399999934       144830.18000003108
Time taken: 73.121 seconds, Fetched: 161 row(s)
```

Insights:

- Here are some 161 brands with increment in the selling from October to November.
- 'Grattol' brand has the highest total increment i.e., 36,027 /- and 'Ovale' seems to have least increment of 0.56 /- from October to November.
- Among all these brands list, 'Runail' which was the best brand in terms of selling in October and November combined is also in the top 10 brands with high increment for October (71539.28 /-) to November (76758.61 /-) i.e., increment of total 5219.38 /-.
- 'Runail' is the best and popular brand among all other brands within people.

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.

Solution :-

```
SELECT user_id, SUM(price) as Total_Expense
FROM ecom
WHERE event_type='purchase'
GROUP BY user_id
ORDER BY Total_Expense DESC
LIMIT 10;
```

```
hive>
    > SELECT user_id, SUM(price) as Total_Expense
    > FROM ecom
    > WHERE event_type='purchase'
    > GROUP BY user_id
    > ORDER BY Total_Expense DESC
    > LIMIT 10;
Query ID = hadoop_20210825175111_211217f5-adf5-492f-ac53-03b93ff48256
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_1629905929129_0024)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 ..........   container     SUCCEEDED      2          2        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: 57.88 s
--------------------------------------------------------------------------------
OK
user_id total_expense
557790271       2715.869999999991
150318419       1645.97
562167663       1352.8500000000004
531900924       1329.4500000000003
557850743       1295.4800000000002
522130011       1185.3899999999994
561592095       1109.6999999999996
431950134       1097.5899999999995
566576008       1056.3600000000017
521347209       1040.9099999999999
Time taken: 67.71 seconds, Fetched: 10 row(s)
hive>
```

Insights:

- Top 10 users or buyers who have spend the most and could be rewarded with a Premium Customer plan to attract more people in the coming future.
- We are selecting this query to be executed using Optimized table to check that does optimized table reduces execution time with proper partitioning and bucketing
. • Time taken to execute this query on Base table (non-optimized table) is **67.71** seconds.

Optimized Table

To create table with Partitioning and Bucketing below commands need to be executed

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

set hive.exec.dynamic.partition=true;

set hive.enforce.bucketing=true;

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

Optimization Steps:-

- ✓ Command to create table **'Dyn_Part_Buck_Shopping'** with partition on **'event_type'** attribute and bucket(**cluster**) on 'price' attribute.

CREATE TABLE IF NOT EXISTS Dyn_Part_Buck_Shopping(event_time timestamp, product_id string, category_id string, category_code string, brand string, price float, user_id bigint, user_session string) PARTITIONED BY (event_type string) CLUSTERED BY (price) INTO 7 BUCKETS ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.OpenCSVSerde' STORED AS TEXTFILE;

```
hive>
    > CREATE TABLE IF NOT EXISTS Dyn_Part_Buck_Shopping(
    > event_time timestamp, product_id string, category_id string, category_code string, brand string, price
    > float, user_id bigint, user_session string
    > )
    > PARTITIONED BY (event_type string)
    > CLUSTERED BY (price) INTO 7 BUCKETS
    > ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.OpenCSVSerde'
    > STORED AS TEXTFILE;
OK
Time taken: 0.202 seconds
hive>
```

- ✓ Loading Data into partitioned and bucketed table we need to get it from already created table i.e., 'ecom'

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

```
hive>
    > CREATE TABLE IF NOT EXISTS Dyn_Part_Buck_Shopping(
    > event_time timestamp, product_id string, category_id string, category_code string, brand string, price
    > float, user_id bigint, user_session string
    > )
    > PARTITIONED BY (event_type string)
    > CLUSTERED BY (price) INTO 7 BUCKETS
    > ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.OpenCSVSerde'
    > STORED AS TEXTFILE;
OK
Time taken: 0.202 seconds
hive> INSERT INTO TABLE Dyn_Part_Buck_Shopping
    > PARTITION (event_type)
    > SELECT event_time, product_id, category_id, category_code, brand, price, user_id, user_session,
    > event_type
    > FROM ecom;
Query ID = hadoop_20210826041847_263c2b9f-fd31-4075-87cf-b23b114f9ab3
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_1629950559635_0004)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      2          2        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      5          5        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 166.58 s
--------------------------------------------------------------------------------
Loading data to table default.dyn_part_buck_shopping partition (event_type=null)

Loaded : 4/4 partitions.
        Time taken to load dynamic partitions: 0.959 seconds
        Time taken for adding to write entity : 0.005 seconds
OK
event_time      product_id      category_id     category_code   brand   price   user_id user_session    event_type
Time taken: 179.717 seconds
hive>
```

- ✓ check successful existence of Partitioned and Bucketed table ‘Dyn_Part_Buck_Shopping’ in hive warehouse

hadoop fs -ls /user/hive/warehouse//dyn_part_buck_shopping

```
[hadoop@ip-172-31-93-75 ~]$ hadoop fs -ls /user/hive/warehouse//dyn_part_buck_shopping
Found 4 items
drwxrwxrwt   - hadoop hadoop          0 2021-08-26 04:21 /user/hive/warehouse/dyn_part_buck_shopping/event_type=cart
drwxrwxrwt   - hadoop hadoop          0 2021-08-26 04:21 /user/hive/warehouse/dyn_part_buck_shopping/event_type=purchase
drwxrwxrwt   - hadoop hadoop          0 2021-08-26 04:21 /user/hive/warehouse/dyn_part_buck_shopping/event_type=remove_from_cart
drwxrwxrwt   - hadoop hadoop          0 2021-08-26 04:21 /user/hive/warehouse/dyn_part_buck_shopping/event_type=view
[hadoop@ip-172-31-93-75 ~]$
```

- ✓ check existence of partitions (event_type = purchase) in the table

hadoop fs -ls /user/hive/warehouse//dyn_part_buck_shopping//event_type=purchase

```
[hadoop@ip-172-31-93-75 ~]$ hadoop fs -ls /user/hive/warehouse//dyn_part_buck_shopping//event_type=purchase
Found 7 items
-rwxrwxrwt   1 hadoop hadoop   13052654 2021-08-26 04:20 /user/hive/warehouse/dyn_part_buck_shopping/event_type=purchase/000000_0
-rwxrwxrwt   1 hadoop hadoop    9399111 2021-08-26 04:21 /user/hive/warehouse/dyn_part_buck_shopping/event_type=purchase/000001_0
-rwxrwxrwt   1 hadoop hadoop   12636711 2021-08-26 04:21 /user/hive/warehouse/dyn_part_buck_shopping/event_type=purchase/000002_0
-rwxrwxrwt   1 hadoop hadoop   10650131 2021-08-26 04:20 /user/hive/warehouse/dyn_part_buck_shopping/event_type=purchase/000003_0
-rwxrwxrwt   1 hadoop hadoop    7226455 2021-08-26 04:20 /user/hive/warehouse/dyn_part_buck_shopping/event_type=purchase/000004_0
-rwxrwxrwt   1 hadoop hadoop   10737803 2021-08-26 04:20 /user/hive/warehouse/dyn_part_buck_shopping/event_type=purchase/000005_0
-rwxrwxrwt   1 hadoop hadoop    7825305 2021-08-26 04:21 /user/hive/warehouse/dyn_part_buck_shopping/event_type=purchase/000006_0
[hadoop@ip-172-31-93-75 ~]$
```

- ✓ to check existence of partitions (event_type = cart) in the table

hadoop fs -ls /user/hive/warehouse//dyn_part_buck_shopping//event_type=cart

```
[hadoop@ip-172-31-93-75 ~]$ hadoop fs -ls /user/hive/warehouse//dyn_part_buck_shopping//event_type=cart
Found 7 items
-rwxrwxrwt   1 hadoop hadoop   57724286 2021-08-26 04:20 /user/hive/warehouse/dyn_part_buck_shopping/event_type=cart/000000_0
-rwxrwxrwt   1 hadoop hadoop   43094161 2021-08-26 04:20 /user/hive/warehouse/dyn_part_buck_shopping/event_type=cart/000001_0
-rwxrwxrwt   1 hadoop hadoop   56823661 2021-08-26 04:21 /user/hive/warehouse/dyn_part_buck_shopping/event_type=cart/000002_0
-rwxrwxrwt   1 hadoop hadoop   49030059 2021-08-26 04:21 /user/hive/warehouse/dyn_part_buck_shopping/event_type=cart/000003_0
-rwxrwxrwt   1 hadoop hadoop   31050141 2021-08-26 04:20 /user/hive/warehouse/dyn_part_buck_shopping/event_type=cart/000004_0
-rwxrwxrwt   1 hadoop hadoop   48253679 2021-08-26 04:20 /user/hive/warehouse/dyn_part_buck_shopping/event_type=cart/000005_0
-rwxrwxrwt   1 hadoop hadoop   34272441 2021-08-26 04:20 /user/hive/warehouse/dyn_part_buck_shopping/event_type=cart/000006_0
[hadoop@ip-172-31-93-75 ~]$
```

- ✓ check existence of partitions (event_type = remove_from_cart) in the table

hadoop fs -ls /user/hive/warehouse//dyn_part_buck_shopping//event_type=remove_from_cart

```
[hadoop@ip-172-31-93-75 ~]$ hadoop fs -ls /user/hive/warehouse//dyn_part_buck_shopping//event_type=remove_from_cart
Found 7 items
-rwxrwxrwt   1 hadoop hadoop   39017824 2021-08-26 04:21 /user/hive/warehouse/dyn_part_buck_shopping/event_type=remove_from_cart/000000_0
-rwxrwxrwt   1 hadoop hadoop   29421828 2021-08-26 04:20 /user/hive/warehouse/dyn_part_buck_shopping/event_type=remove_from_cart/000001_0
-rwxrwxrwt   1 hadoop hadoop   38713899 2021-08-26 04:21 /user/hive/warehouse/dyn_part_buck_shopping/event_type=remove_from_cart/000002_0
-rwxrwxrwt   1 hadoop hadoop   31959876 2021-08-26 04:21 /user/hive/warehouse/dyn_part_buck_shopping/event_type=remove_from_cart/000003_0
-rwxrwxrwt   1 hadoop hadoop   19751571 2021-08-26 04:20 /user/hive/warehouse/dyn_part_buck_shopping/event_type=remove_from_cart/000004_0
-rwxrwxrwt   1 hadoop hadoop   31335021 2021-08-26 04:21 /user/hive/warehouse/dyn_part_buck_shopping/event_type=remove_from_cart/000005_0
-rwxrwxrwt   1 hadoop hadoop   22175799 2021-08-26 04:21 /user/hive/warehouse/dyn_part_buck_shopping/event_type=remove_from_cart/000006_0
[hadoop@ip-172-31-93-75 ~]$
```

- ✓ check existence of partitions (event_type = view) in the table

  hadoop fs -ls /user/hive/warehouse//dyn_part_buck_shopping//event_type=view

```
[hadoop@ip-172-31-93-75 ~]$ hadoop fs -ls /user/hive/warehouse//dyn_part_buck_shopping//event_type=view
Found 7 items
-rwxrwxrwt   1 hadoop hadoop   88831872 2021-08-26 04:21 /user/hive/warehouse/dyn_part_buck_shopping/event_type=view/000000_0
-rwxrwxrwt   1 hadoop hadoop   73953212 2021-08-26 04:21 /user/hive/warehouse/dyn_part_buck_shopping/event_type=view/000001_0
-rwxrwxrwt   1 hadoop hadoop   85620113 2021-08-26 04:21 /user/hive/warehouse/dyn_part_buck_shopping/event_type=view/000002_0
-rwxrwxrwt   1 hadoop hadoop   71874121 2021-08-26 04:21 /user/hive/warehouse/dyn_part_buck_shopping/event_type=view/000003_0
-rwxrwxrwt   1 hadoop hadoop   48335545 2021-08-26 04:21 /user/hive/warehouse/dyn_part_buck_shopping/event_type=view/000004_0
-rwxrwxrwt   1 hadoop hadoop   72515614 2021-08-26 04:21 /user/hive/warehouse/dyn_part_buck_shopping/event_type=view/000005_0
-rwxrwxrwt   1 hadoop hadoop   56694677 2021-08-26 04:21 /user/hive/warehouse/dyn_part_buck_shopping/event_type=view/000006_0
[hadoop@ip-172-31-93-75 ~]$
```

- ✓ Running the Same query on optimized table to check performance Improvement –

Running Optimized Query for Question Number -8 where company wants to reward the top 10 users of its website with a Golden Customer plan. Write a query to generate a list of top 10 users who
spend the most .

SELECT user_id, SUM(price) AS Total_Expenditure FROM Dyn_Part_Buck_Shopping WHERE event_type='purchase' GROUP BY user_id ORDER BY Total_Expenditure DESC LIMIT 10;

```
Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j2.properties Async: false
hive> SELECT user_id, SUM(price) AS Total_Expenditure
    > FROM Dyn_Part_Buck_Shopping
    > WHERE event_type='purchase'
    > GROUP BY user_id
    > ORDER BY Total_Expenditure DESC
    > LIMIT 10;
Query ID = hadoop_20210826053419_ae3de138-7ccc-4458-bea4-f57bd0758c69
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1629950559635_0005)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      3          3        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 27.34 s
--------------------------------------------------------------------------------
OK
557790271       2715.869999999996
150318419       1645.97
562167663       1352.8500000000001
531900924       1329.4500000000003
557850743       1295.4800000000005
522130011       1185.3899999999999
561592095       1109.7
431950134       1097.5900000000001
566576008       1056.3600000000006
521347209       1040.9100000000003
Time taken: 32.961 seconds, Fetched: 10 row(s)
hive>
```

Insights :

we can see there is significant drop in the execution time of the same query i.e., *previously the execution was measured as 69.71 seconds and now it is 32.96 seconds with the difference of* ***36.75*** *seconds*. Hence, with proper partitioning and bucketing on table we can reduce execution time of the query .

Running Optimized Query for Question Number -1 – where we need to Find the total revenue generated due to purchases made in October.

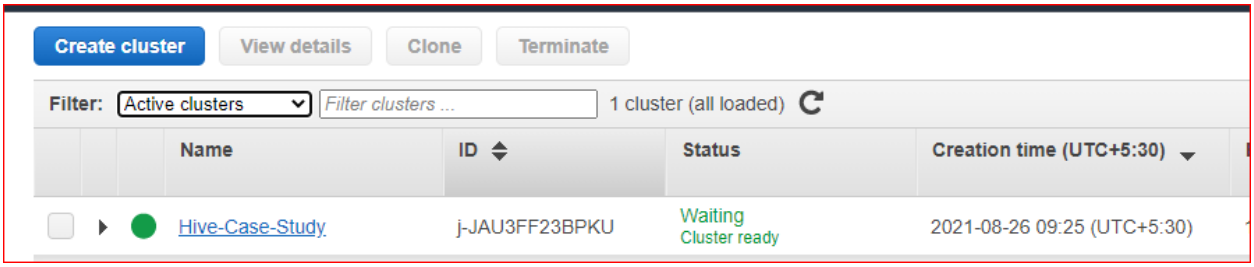
SELECT SUM(price) AS Total_Revenue_October FROM Dyn_Part_Buck_Shopping WHERE date_format(event_time, 'MM')=10 AND event_type='purchase';

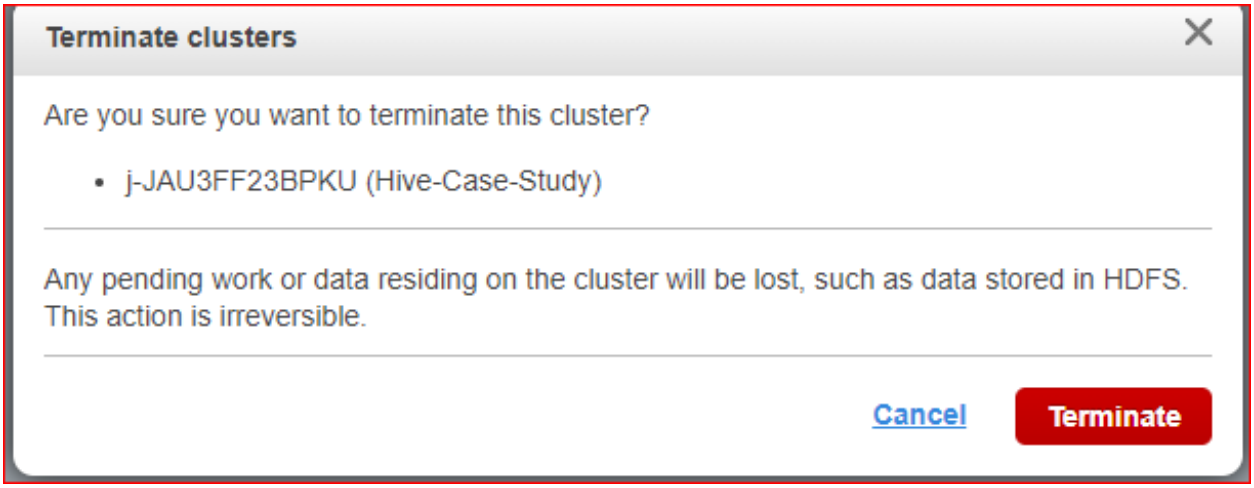
```
hive> SELECT SUM(price) AS Total_Revenue_October FROM  Dyn_Part_Buck_Shopping WHERE date_format(event_time, 'MM')=10 AND event_type='purchase';
Query ID = hadoop_20210826054119_c85b29b4-b9a2-4480-b1a8-dcf0ff8dabc9
Total jobs = 1
Launching Job 1 out of 1
Tez session was closed. Reopening...
Session re-established.
Status: Running (Executing on YARN cluster with App id application_1629950559635_0006)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      3          3        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 28.65 s
--------------------------------------------------------------------------------
OK
1211538.4299997753
Time taken: 37.555 seconds, Fetched: 1 row(s)
hive>
```

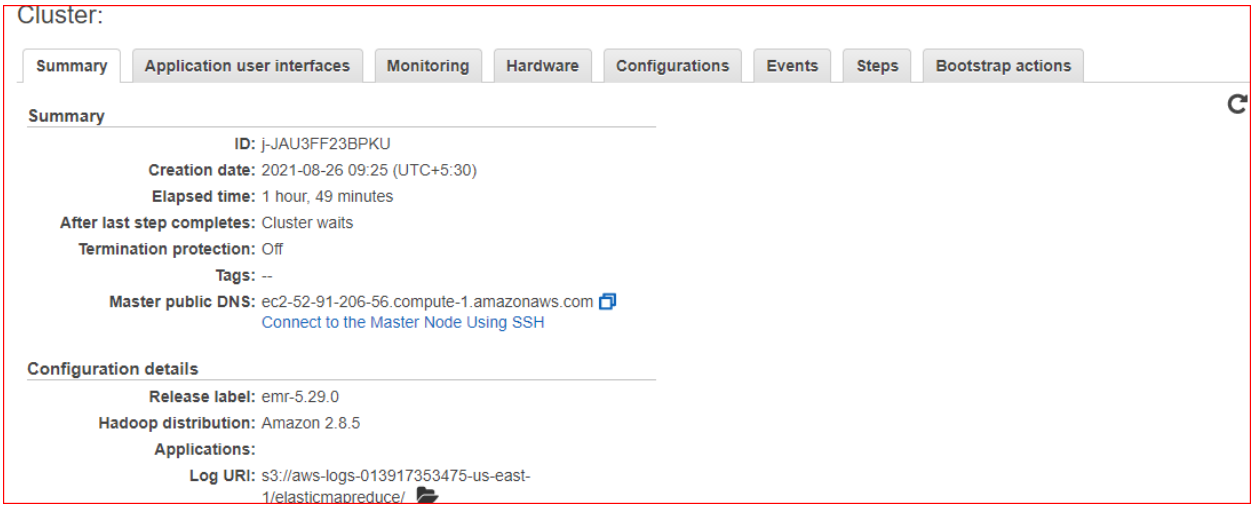
Insights :

we can see there is significant drop in the execution time of the same query i.e**.,** *previously the execution was measured as 125.797 seconds and now it is 37.555 seconds with the difference of* ***88.242*** seconds- *Huge Performance Improvement* .

Terminating the Cluster :







Thanks for Watching