HIVE CASE STUDY

By - Yesh Thakur and Sabyasachi De (DSC31)

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. This is done by tracking your clicks on their website and searching for patterns within them. This kind of data is called a clickstream data.

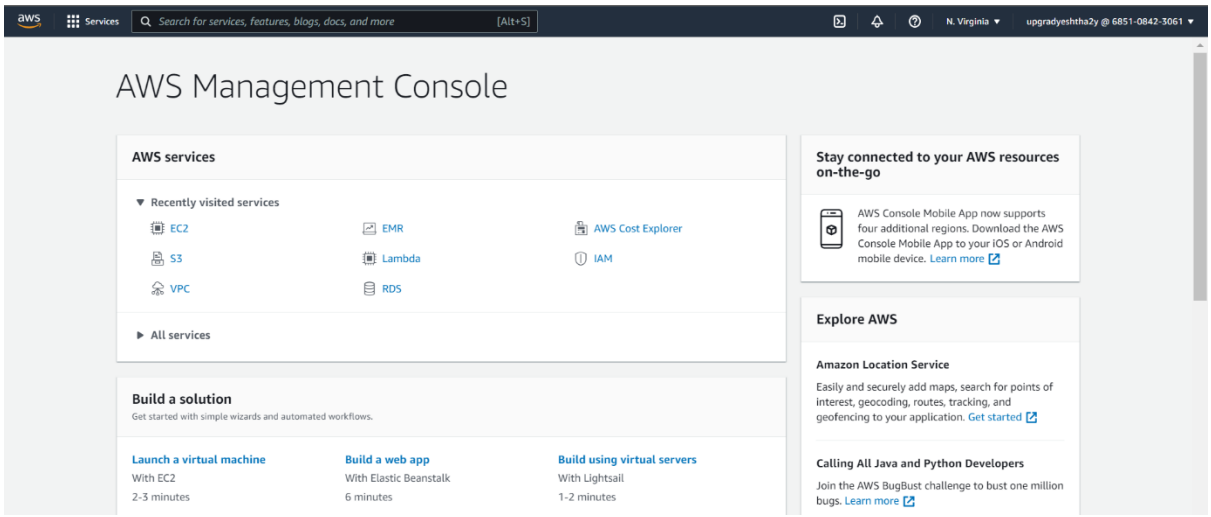
OBJECTIVE:

To extract data and gather insights from a real-life data set of an e-commerce company.

CLUSTER CREATION:

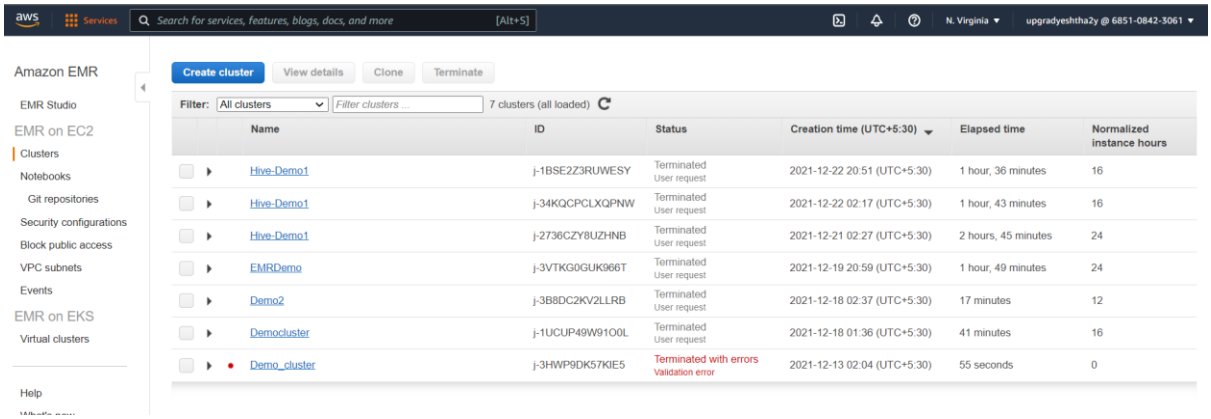
Step 1

Login to AWS and look for EMR cluster either in the Recently Viewed Services tab or search for it in the Search tab.



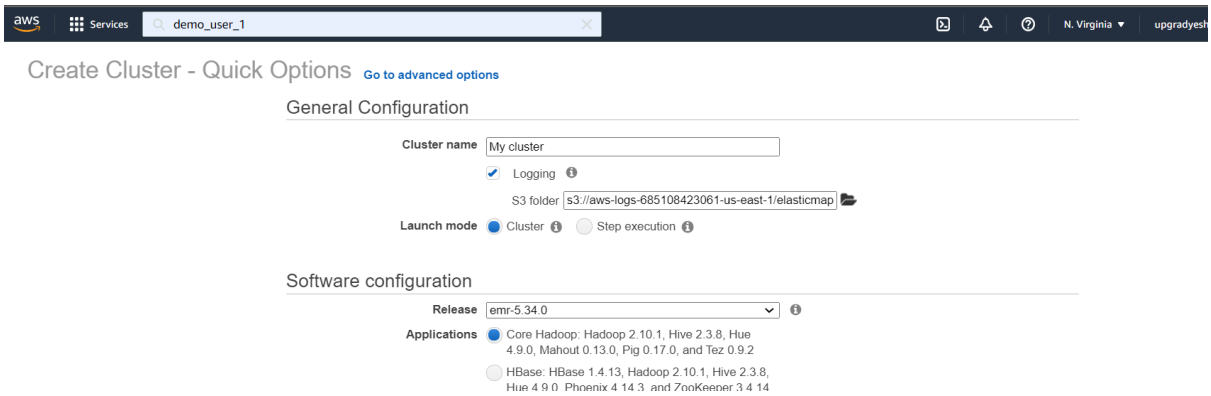
Step 2

Once you are in the EMR home page, click on ‘Create Cluster’ button.



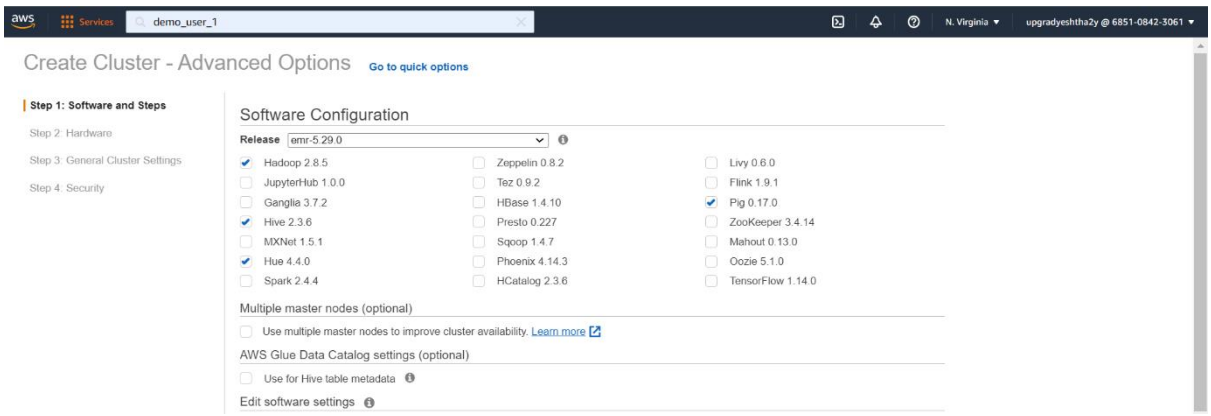
Step 3

Click on ‘Go to Advanced Options’.



Step 4

From the Software Configuration, select the ones which are required along with the release of the EMR and then click next. For this case study, we have chosen emr-5.29.0 as per the instructions.

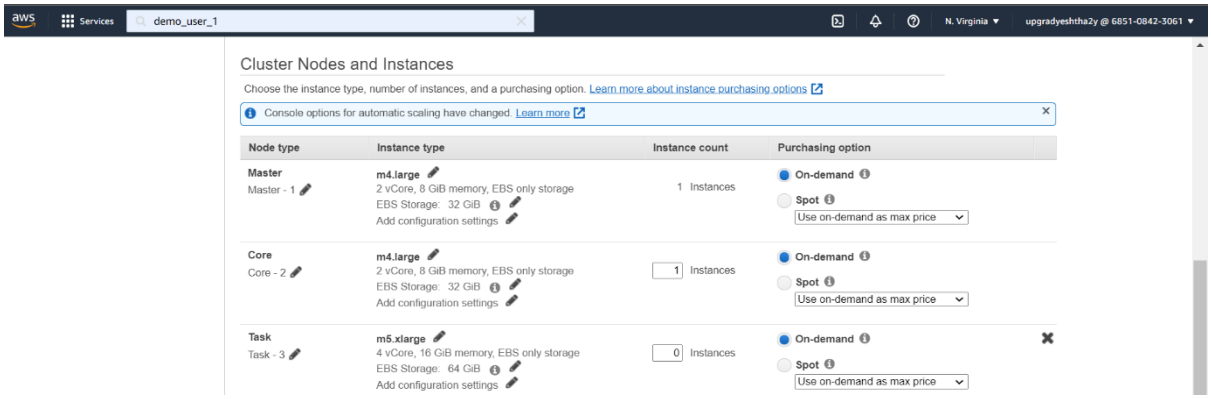


Step 5

On the Hardware Configuration page, we have to specify the required configuration i.e., instance type and instance count for the master and core nodes.

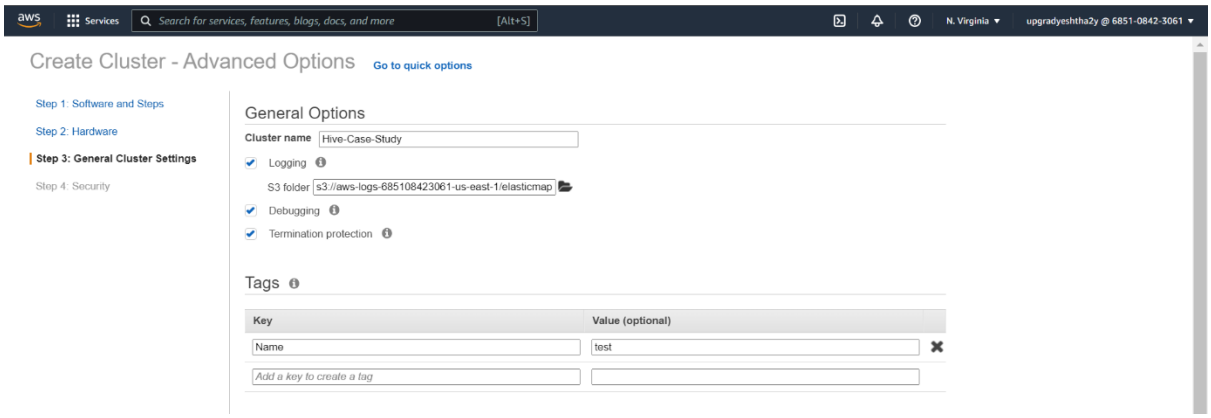
As per the instructions, we have selected m4.large for both the master and core node as it is a 2 node cluster.

Click next.



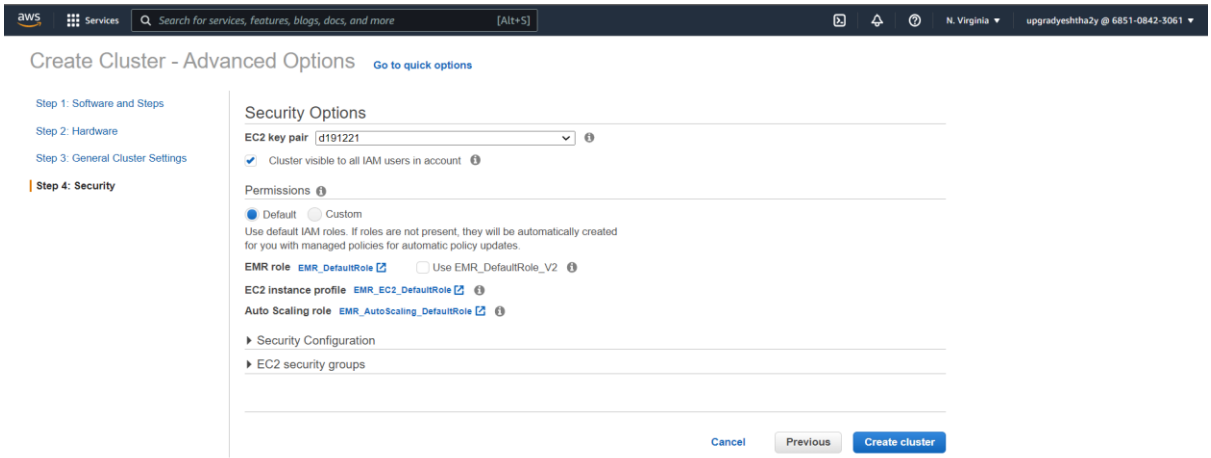
Step 6

Here, give a suitable name to your cluster.



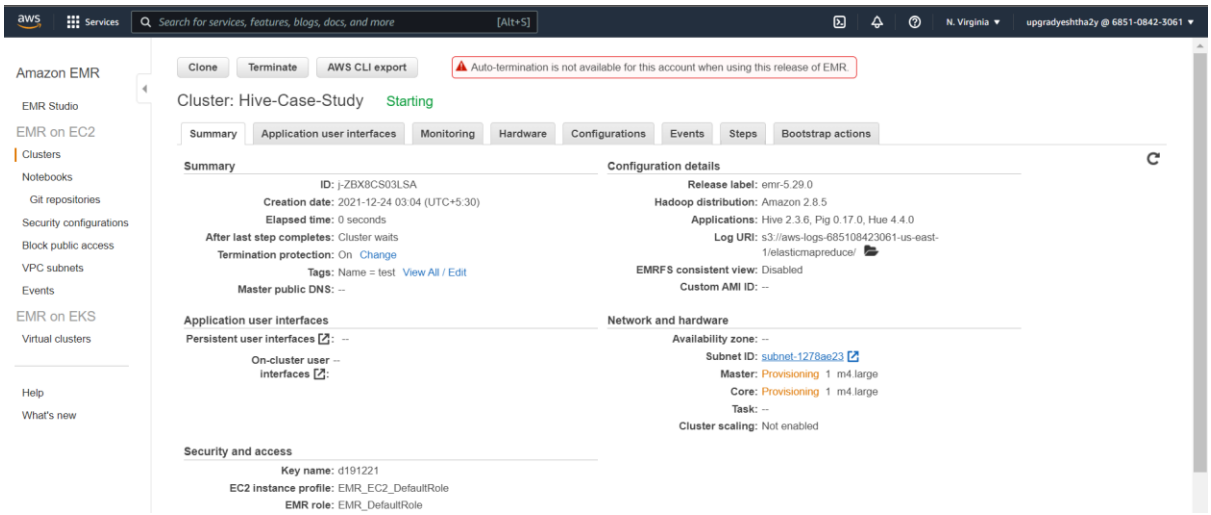
Step 7

Select an EC2 key pair (created before the cluster creation) and click on ‘Create Cluster’.



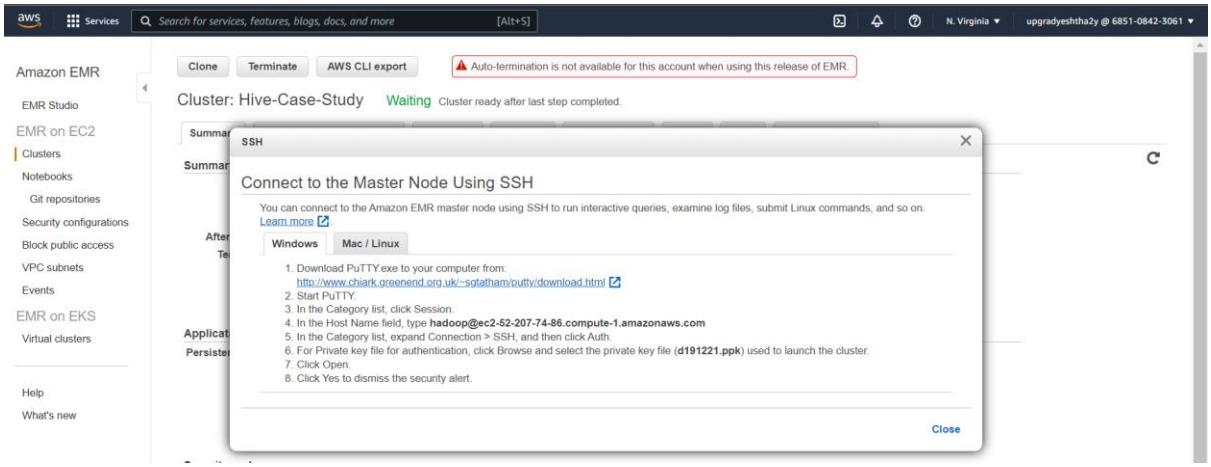
Step 8

Wait for your cluster to start and then we can go to putty.



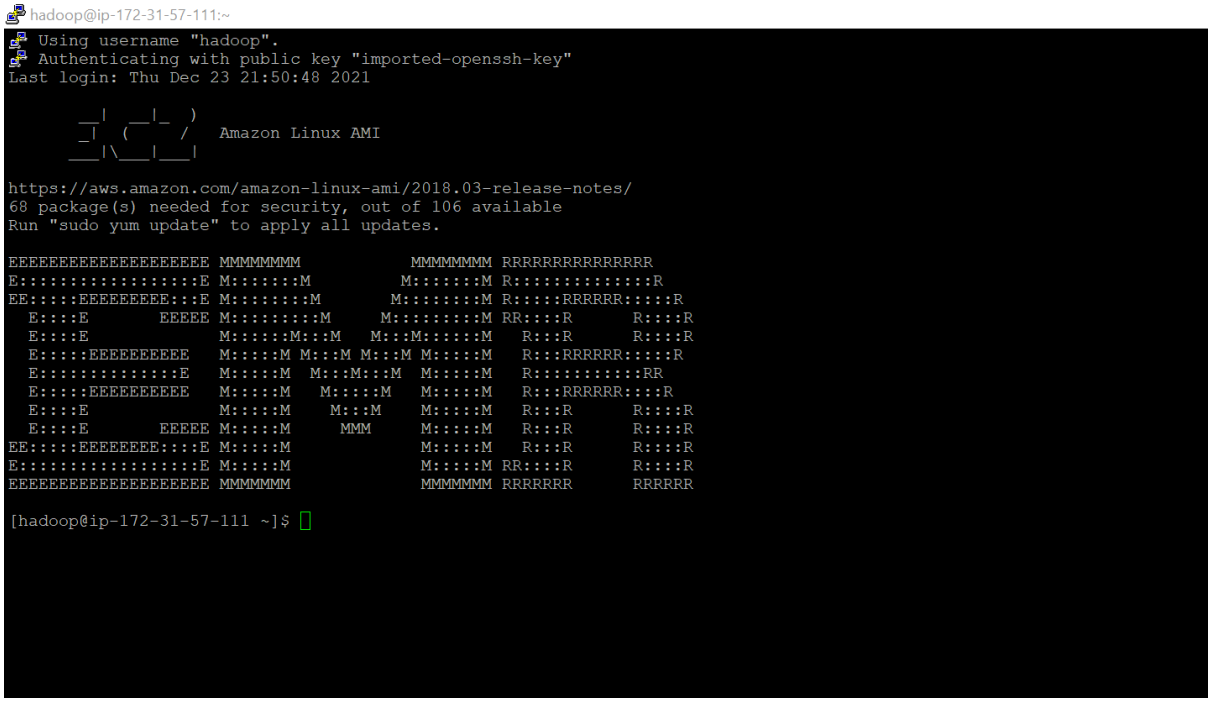
Step 9

When the cluster is in the Waiting state which means it's running, click on Master public DNS.



Step 10

- Open Putty and paste the Host Name.
- Click and expand SSH and select AUTH.
- Browse Key pair file and click Open.



Step 11

Copying the data into HDFS.

1. Creating a directory ‘casestudy’ in HDFS.
   Command: hadoop fs -mkdir /user/hive/casestudy

2. Let’s check this directory.
   Command: hadoop fs -ls /user/hive

```
[hadoop@ip-172-31-48-166 ~]$ hadoop fs -ls /user/hive
Found 2 items
drwxr-xr-x   - hadoop hadoop          0 2021-12-24 20:28 /user/hive/casestudy
drwxrwxrwt   - hdfs   hadoop          0 2021-12-24 19:34 /user/hive/warehouse
[hadoop@ip-172-31-48-166 ~]$
```

3. Loading the data from S3 buckeer to HDFS .
   Command:
   hadoop distcp s3://casestudyhive1/2019-Oct.csv /user/hive/casestudy/2019-Oct.csv

```
[hadoop@ip-172-31-58-247 ~]$ hadoop distcp s3://casestudyhive1/2019-Oct.csv /user/hive/casestudy/2019-Oct.csv
21/12/29 20:44:30 INFO tools.DistCp: Input Options: DistCpOptions{atomicCommit=false, syncFolder=false, deleteMissing=false, ignoreFailures=false, overwrite=
false, skipCRC=false, blocking=true, numListstatusThreads=0, maxMaps=20, mapBandwidth=100, sslConfigurationFile='null', copyStrategy='uniformsize', preserveS
tatus=[], preserveRawXattrs=false, atomicWorkPath=null, logPath=null, sourceFileListing=null, sourcePaths=[s3://casestudyhive1/2019-Oct.csv], targetPath=/use
r/hive/casestudy/2019-Oct.csv, targetPathExists=false, filtersFile='null'}
21/12/29 20:44:31 INFO client.RMProxy: Connecting to ResourceManager at ip-172-31-58-247.ec2.internal/172.31.58.247:8032
21/12/29 20:44:37 INFO tools.SimpleCopyListing: Paths (files+dirs) cnt = 1; dirCnt = 0
21/12/29 20:44:37 INFO tools.SimpleCopyListing: Build file listing completed.
21/12/29 20:44:37 INFO Configuration.deprecation: io.sort.mb is deprecated. Instead, use mapreduce.task.io.sort.mb
21/12/29 20:44:37 INFO Configuration.deprecation: io.sort.factor is deprecated. Instead, use mapreduce.task.io.sort.factor
21/12/29 20:44:37 INFO tools.DistCp: Number of paths in the copy list: 1
21/12/29 20:44:37 INFO tools.DistCp: Number of paths in the copy list: 1
21/12/29 20:44:37 INFO client.RMProxy: Connecting to ResourceManager at ip-172-31-58-247.ec2.internal/172.31.58.247:8032
21/12/29 20:44:37 INFO mapreduce.JobSubmitter: number of splits:1
21/12/29 20:44:38 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1640810275048_0001
```

```
                GC time elapsed (ms)=360
                CPU time spent (ms)=18770
                Physical memory (bytes) snapshot=584773632
                Virtual memory (bytes) snapshot=3310522368
                Total committed heap usage (bytes)=503316480
        File Input Format Counters
                Bytes Read=217
        File Output Format Counters
                Bytes Written=0
        DistCp Counters
                Bytes Copied=482542278
                Bytes Expected=482542278
                Files Copied=1
```

hadoop distcp s3://casestudyhive1/2019-Nov.csv /user/hive/casestudy/2019-Nov.csv

```
[hadoop@ip-172-31-58-247 ~]$ hadoop distcp s3://casestudyhive1/2019-Nov.csv /user/hive/casestudy/2019-Nov.csv
21/12/29 20:48:23 INFO tools.DistCp: Input Options: DistCpOptions{atomicCommit=false, syncFolder=false, deleteMissing=false, ignoreFailures=false, overwrite=
false, skipCRC=false, blocking=true, numListstatusThreads=0, maxMaps=20, mapBandwidth=100, sslConfigurationFile='null', copyStrategy='uniformsize', preserveS
tatus=[], preserveRawXattrs=false, atomicWorkPath=null, logPath=null, sourceFileListing=null, sourcePaths=[s3://casestudyhive1/2019-Nov.csv], targetPath=/use
r/hive/casestudy/2019-Nov.csv, targetPathExists=false, filtersFile='null'}
21/12/29 20:48:23 INFO client.RMProxy: Connecting to ResourceManager at ip-172-31-58-247.ec2.internal/172.31.58.247:8032
21/12/29 20:48:29 INFO tools.SimpleCopyListing: Paths (files+dirs) cnt = 1; dirCnt = 0
21/12/29 20:48:29 INFO tools.SimpleCopyListing: Build file listing completed.
21/12/29 20:48:29 INFO Configuration.deprecation: io.sort.mb is deprecated. Instead, use mapreduce.task.io.sort.mb
21/12/29 20:48:29 INFO Configuration.deprecation: io.sort.factor is deprecated. Instead, use mapreduce.task.io.sort.factor
21/12/29 20:48:29 INFO tools.DistCp: Number of paths in the copy list: 1
21/12/29 20:48:29 INFO tools.DistCp: Number of paths in the copy list: 1
21/12/29 20:48:29 INFO client.RMProxy: Connecting to ResourceManager at ip-172-31-58-247.ec2.internal/172.31.58.247:8032
21/12/29 20:48:30 INFO mapreduce.JobSubmitter: number of splits:1
```

```
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=299
                CPU time spent (ms)=20170
                Physical memory (bytes) snapshot=594878464
                Virtual memory (bytes) snapshot=3308191744
                Total committed heap usage (bytes)=506462208
        File Input Format Counters
                Bytes Read=217
        File Output Format Counters
                Bytes Written=0
        DistCp Counters
                Bytes Copied=545839412
                Bytes Expected=545839412
                Files Copied=1
```

4. Now, checking the files in the directory.
   Command: hadoop fs -ls /user/hive/casestudy/

```
[hadoop@ip-172-31-48-166 ~]$ hadoop fs -ls /user/hive/casestudy/
Found 2 items
-rw-r--r--   1 hadoop hadoop  482542278 2021-12-24 20:34 /user/hive/casestudy/2019_Nov.csv
-rw-r--r--   1 hadoop hadoop  482542278 2021-12-24 20:33 /user/hive/casestudy/2019_Oct.csv
[hadoop@ip-172-31-48-166 ~]$
```

Step 12

Data sets are loaded now let's launch Hive.

Creating a new database 'hive_casestudy'

Command:

create database if not exists hive_casestudy ;

show databases;

use hive_casestudy;

```
[hadoop@ip-172-31-58-247 ~]$ hive

Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j2.p
hive> create database if not exists hive_casestudy ;
OK
Time taken: 1.095 seconds
hive> show databases;
OK
default
hive_casestudy
Time taken: 0.214 seconds, Fetched: 2 row(s)
hive> use hive_casestudy;
OK
Time taken: 0.122 seconds
```

Creating a table.

create external table if not exists clickstream (event_time timestamp, event_type string, product_id string, category_id string, category_code string, brand string, price float, user_id bigint, user_session string) row format serde 'org.apache.hadoop.hive.serde2.OpenCSVSerde' stored as textfile location '/user/hive/casestudy/' tblproperties ("skip.header.line.count"="1");

```
hive> create external table if not exists clickstream (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 '/u
ser/hive/casestudy/' tblproperties ("skip.header.line.count"="1");
OK
Time taken: 0.367 seconds
```

Now, we need to optimize the table ‘clickstream’ through partitioning and bucketing for faster query results.

set hive.exec.dynamic.partition=true;

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

set hive.enforce.bucketing = true;

```
Time taken: 0.381 seconds
hive> set hive.exec.dynamic.partition=true;
hive> set hive.exec.dynamic.partition.mode=nonstrict;
hive> set hive.enforce.bucketing = true;
hive>
```

Creating a new table with dynamic partitions and buckets and then inserting the data.

create table if not exists dyn_part_buck_clickstream (event_time string, product_id string, category_id string, category_code string, brand string, price float, user_id bigint, user_session string) partitioned by (event_type string) clustered by (category_code) into 13 buckets row format delimited fields terminated by ',' lines terminated by '\n' stored as textfile;

```
hive> create table if not exists dyn_part_buck_clickstream (event_time string, product_id string, category_id string, category_code string, brand string, pri
ce float, user_id bigint, user_session string) partitioned by (event_type string) clustered by (category_code) into 13 buckets row format delimited fields te
rminated by ',' lines terminated by '\n' stored as textfile;
OK
Time taken: 0.115 seconds
```

insert into table dyn_part_buck_clickstream partition (event_type) select event_time, product_id, category_id, category_code, brand, price, user_id, user_session, event_type from clickstream;

```
hive> insert into table dyn_part_buck_clickstream partition (event_type) select event_time, product_id, category_id, category_code, brand, price, user_id, us
er_session, event_type from clickstream;
Query ID = hadoop_20211230195527_667535ca-bb1d-4589-96ad-1da7720aeb02
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_1640892795316_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: 169.28 s
--------------------------------------------------------------------------------
Loading data to table hive_casestudy.dyn_part_buck_clickstream partition (event_type=null)

Loaded : 4/4 partitions.
         Time taken to load dynamic partitions: 0.94 seconds
         Time taken for adding to write entity : 0.003 seconds
OK
Time taken: 186.607 seconds
```

Checking both the tables:

describe clickstream;

describe dyn_part_buck_clickstream;

```
hive> describe clickstream;
OK
event_time              string                  from deserializer
event_type              string                  from deserializer
product_id              string                  from deserializer
category_id             string                  from deserializer
category_code           string                  from deserializer
brand                   string                  from deserializer
price                   string                  from deserializer
user_id                 string                  from deserializer
user_session            string                  from deserializer
Time taken: 0.047 seconds, Fetched: 9 row(s)
hive> describe dyn_part_buck_clickstream;
OK
event_time              string
product_id              string
category_id             string
category_code           string
brand                   string
price                   float
user_id                 bigint
user_session            string
event_type              string

# Partition Information
# col_name              data_type               comment

event_type              string
Time taken: 0.133 seconds, Fetched: 14 row(s)
```

Getting the headers back.

set hive.cli.print.header=true;

We have created two tables:

1st table which contains data for both October and November.

2nd table which has the same data but optimized with partitions and buckets.

Step 13

Query Analysis

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

Unoptimized Query:

SELECT SUM(price) as total_rev FROM clickstream WHERE MONTH(event_time) = 10 and event_type = 'purchase';

```
hive> SELECT SUM(price) as total_rev FROM clickstream WHERE MONTH(event_time) = 10 and event_type = 'purchase';
Query ID = hadoop_20211229214117_0b1ba86c-3f37-4863-be5e-0bc94e251182
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1640810275048_0006)

--------------------------------------------------------------------------------------------
        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: 65.03 s
--------------------------------------------------------------------------------------------
OK
total_rev
1211538.4299997438
Time taken: 65.675 seconds, Fetched: 1 row(s)
```

Optimized Query using dyn_part_buck_clickstream:

SELECT SUM(price) as total_rev FROM dyn_part_buck_clickstream WHERE MONTH(event_time) = 10 and event_type = 'purchase';

```
hive> SELECT SUM(price) as total_rev FROM dyn_part_buck_clickstream WHERE MONTH(event_time) = 10 and event_type = 'purchase';
Query ID = hadoop_20211230231835_29779f34-5d56-4b25-8de4-41f6ce79c843
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1640905004182_0003)

--------------------------------------------------------------------------------------------
        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: 15.13 s
--------------------------------------------------------------------------------------------
OK
total_rev
1211538.4295325726
Time taken: 15.879 seconds, Fetched: 1 row(s)
```

The total revenue generated is 1211538.4295325726. Unoptimized query took 65.675 secs while optimized query took 15.879 secs.

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

Unoptimized Query:

SELECT MONTH(event_time), COUNT(event_type) AS pur_cnt from clickstream WHERE event_type = 'purchase' GROUP BY MONTH(event_time);

```
hive> SELECT MONTH(event_time), COUNT(event_type) AS pur_cnt from clickstream WHERE event_type = 'purchase' GROUP BY MONTH(event_time);
Query ID = hadoop_20211229214456_b24ed3ca-3716-456d-ba31-818c9efba1e2
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1640810275048_0006)

--------------------------------------------------------------------------------
        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: 60.79 s
--------------------------------------------------------------------------------
OK
_c0     pur_cnt
10      245624
11      322417
Time taken: 61.451 seconds, Fetched: 2 row(s)
```

Optimized Query using dyn_part_buck_clickstream:

SELECT MONTH(event_time), COUNT(event_type) AS pur_cnt from dyn_part_buck_clickstream WHERE event_type = 'purchase' GROUP BY MONTH(event_time);

```
hive> SELECT MONTH(event_time), COUNT(event_type) AS pur_cnt from dyn_part_buck_clickstream WHERE event_type = 'purchase' GROUP BY MONTH(event_time);
Query ID = hadoop_20211231181315_749c5841-8f64-4520-82ec-ef590b8de71a
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1640973564708_0003)

--------------------------------------------------------------------------------
        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: 16.96 s
--------------------------------------------------------------------------------
OK
_c0     pur_cnt
10      245624
11      322417
Time taken: 18.047 seconds, Fetched: 2 row(s)
```

The total sum of the purchase for October is 245624 and for November it is 322417. Unoptimized query took 61.451 secs while optimized query took 18.047 secs.

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

Unoptimized Query:

WITH diff_revenue AS (SELECT SUM(CASE WHEN MONTH(event_time) = 10 THEN price ELSE 0 END) AS oct_pur, SUM(CASE WHEN MONTH(event_time) = 11 THEN price ELSE 0 END) AS nov_pur FROM clickstream WHERE event_type = 'purchase') SELECT (nov_pur - oct_pur) AS difference_revenue FROM diff_revenue;

```
hive> WITH diff_revenue AS (SELECT SUM(CASE WHEN MONTH(event_time) = 10 THEN price ELSE 0 END) AS oct_pur, SUM(CASE WHEN MONTH(event_time) = 11 THEN price EL
SE 0 END) AS nov_pur FROM clickstream WHERE event_type = 'purchase') SELECT (nov_pur - oct_pur) AS difference_revenue FROM diff_revenue;
Query ID = hadoop_20211229213710_8d21ecff-8c55-479a-b196-dde72d32ca14
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_1640810275048_0006)

--------------------------------------------------------------------------------
        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: 56.08 s
--------------------------------------------------------------------------------
OK
difference_revenue
319478.4700003781
Time taken: 65.493 seconds, Fetched: 1 row(s)
```

Optimized Query:

WITH diff_revenue AS (SELECT SUM(CASE WHEN MONTH(event_time) = 10 THEN price ELSE 0 END) AS oct_pur, SUM(CASE WHEN MONTH(event_time) = 11 THEN price ELSE 0 END) AS nov_pur FROM dyn_part_buck_clickstream WHERE event_type = 'purchase') SELECT (nov_pur - oct_pur) AS difference_revenue FROM diff_revenue;

```
hive> WITH diff_revenue AS (SELECT SUM(CASE WHEN MONTH(event_time) = 10 THEN price ELSE 0 END) AS oct_pur, SUM(CASE WHEN MONTH(event_time) = 11 THEN price EL
SE 0 END) AS nov_pur FROM dyn_part_buck_clickstream WHERE event_type = 'purchase') SELECT (nov_pur - oct_pur) AS difference_revenue FROM diff_revenue;
Query ID = hadoop_20211231181909_65b2c46c-8246-46a5-b5ce-e71eb74329f1
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1640973564708_0003)

--------------------------------------------------------------------------------
        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: 18.73 s
--------------------------------------------------------------------------------
OK
difference_revenue
319478.469592195
Time taken: 19.568 seconds, Fetched: 1 row(s)
```

The change in revenue is 319478.469592195. Unoptimized query took 65.493 secs while optimized query took 19.560 secs.

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

Unoptimized Query:

SELECT DISTINCT(category_code) FROM clickstream where category_code !='';

```
hive> SELECT DISTINCT(category_code) FROM clickstream where category_code !='';
Query ID = hadoop_20211229220013_0fb8e7f1-cfb9-43f0-95d9-83db0652deaf
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1640810275048_0007)

--------------------------------------------------------------------------------------------
        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: 58.03 s
--------------------------------------------------------------------------------------------
OK
category_code
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: 58.732 seconds, Fetched: 11 row(s)
```

Optimized Query:

SELECT DISTINCT(category_code) FROM dyn_part_buck_clickstream where category_code !='';

```
hive> SELECT DISTINCT(category_code) FROM dyn_part_buck_clickstream where category_code !='';
Query ID = hadoop_20211230201428_39b977e8-2c7d-4f51-bc92-fd15ebe79d47
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1640892795316_0005)

--------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      7          7        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      4          4        0        0       0       0
--------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 29.47 s
--------------------------------------------------------------------------------------------
OK
category_code
accessories.bag
appliances.environment.vacuum
appliances.personal.hair_cutter
sport.diving
apparel.glove
furniture.bathroom.bath
furniture.living_room.cabinet
stationery.cartrige
accessories.cosmetic_bag
appliances.environment.air_conditioner
furniture.living_room.chair
Time taken: 30.289 seconds, Fetched: 11 row(s)
```

There are 11 distinct categories in total. Unoptimized query took 58.732 secs while optimized query took 30.289 secs.

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

Unoptimized Query:

select count(product_id) as prod_id, category_code from clickstream where category_code !='' group by category_code;

```
hive> select count(product_id) as prod_id, category_code from clickstream where category_code !='' group by category_code;
Query ID = hadoop_20211230203546_7f6e9328-094e-45a6-b5f9-790eb78a9fb4
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_1640892795316_0006)

--------------------------------------------------------------------------------
        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: 60.36 s
--------------------------------------------------------------------------------
OK
prod_id category_code
1248    accessories.cosmetic_bag
26722   stationery.cartrige
11681   accessories.bag
59761   appliances.environment.vacuum
308     furniture.living_room.chair
2       sport.diving
1643    appliances.personal.hair_cutter
332     appliances.environment.air_conditioner
18232   apparel.glove
9857    furniture.bathroom.bath
13439   furniture.living_room.cabinet
Time taken: 71.505 seconds, Fetched: 11 row(s)
```

Optimized Query:

select count(product_id) as prod_id, category_code from dyn_part_buck_clickstream where category_code !='' group by category_code;

```
hive> select count(product_id) as prod_id, category_code from dyn_part_buck_clickstream where category_code !='' group by category_code;
Query ID = hadoop_20211230204322_17600065-97f4-42ab-ba24-4f98dbbffdb7
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1640892795316_0006)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      7          7        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      4          4        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 29.58 s
--------------------------------------------------------------------------------
OK
prod_id category_code
11681   accessories.bag
59761   appliances.environment.vacuum
1643    appliances.personal.hair_cutter
2       sport.diving
18232   apparel.glove
9857    furniture.bathroom.bath
13439   furniture.living_room.cabinet
26722   stationery.cartrige
1248    accessories.cosmetic_bag
332     appliances.environment.air_conditioner
308     furniture.living_room.chair
Time taken: 30.228 seconds, Fetched: 11 row(s)
```

Unoptimized query took 71.505 secs while optimized query took 30.228 secs.

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

Unoptimized Query:

select brand, sum(price) as total_sales from clickstream where event_type = 'purchase' and brand !='' group by brand order by total_sales desc limit 1;

```
hive> select brand, sum(price) as total_sales from clickstream where event_type = 'purchase' and brand !='' group by brand order by total_sales desc limit 1;

Query ID = hadoop_20211230210315_ddf68b29-051d-4a18-bf1c-8d941c79df66
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_1640892795316_0007)

--------------------------------------------------------------------------------
        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: 63.24 s
--------------------------------------------------------------------------------
OK
brand   total_sales
runail  148297.9400000003
Time taken: 73.686 seconds, Fetched: 1 row(s)
```

Optimized Query:

select brand, sum(price) as total_sales from dyn_part_buck_clickstream where event_type = 'purchase' and brand !='' group by brand order by total_sales desc limit 1;

```
hive> select brand, sum(price) as total_sales from dyn_part_buck_clickstream where event_type = 'purchase' and brand !='' group by brand order by total_sales
 desc limit 1;
Query ID = hadoop_20211230210644_5f55f170-8e2f-4381-872a-eb377f73a49e
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1640892795316_0007)

--------------------------------------------------------------------------------
        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
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 20.18 s
--------------------------------------------------------------------------------
OK
brand   total_sales
runail  148297.93996394053
Time taken: 21.105 seconds, Fetched: 1 row(s)
```

Top brand is runail with total sales as 148297.93996394053. Unoptimized query took 73.686 secs while optimized query took 21.105 secs.

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

Unoptimized Query:

WITH brand_sales AS (SELECT brand, sum(CASE WHEN month(event_time)=10 THEN price else 0 END) AS oct_sales, sum(CASE WHEN month(event_time)=11 THEN price else 0 END) AS nov_sales from clickstream where event_type = 'purchase' group by brand) select brand from brand_sales where (nov_sales - oct_sales)>0;

```
hive> WITH brand_sales AS (SELECT brand, sum(CASE WHEN month(event_time)=10 THEN price else 0 END) AS oct_sales, sum(CASE WHEN month(event_time)=11 THEN pric
e else 0 END) AS nov_sales from clickstream where event_type = 'purchase' group by brand) select brand from brand_sales where (nov_sales - oct_sales)>0;
Query ID = hadoop_20211230213144_47b9fc60-c5f3-4fea-9573-5bd6fe761b8f
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_1640892795316_0008)

--------------------------------------------------------------------------------
        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: 69.79 s
--------------------------------------------------------------------------------
OK
brand
airnails
artex
binacil
bioaqua
blixz
bluesky
bpw.style
carmex
chi
concept
cosima
cosmoprofi
deoproce
depilflax
dewal
dizao
egomania
elizavecca
ellips
finish
freshbubble
grattol
haruyama
helloganic
```

```
candy
coifin
cristalinas
cutrin
domix
ecocraft
elskin
enjoy
entity
eos
estel
estelare
farmavita
fedua
foamie
glysolid
godefroy
inm
irisk
kamill
kares
kaypro
keen
kinetics
koelcia
lianail
lowence
matreshka
mavala
missha
moyou
nagaraku
profepil
rasyan
refectocil
skinity
smart
solomeya
swarovski
trind
uno
yu-r
Time taken: 80.648 seconds, Fetched: 161 row(s)
hive>
```

Optimized Query:

WITH brand_sales AS (SELECT brand, sum(CASE WHEN month(event_time)=10 THEN price else 0 END) AS oct_sales, sum(CASE WHEN month(event_time)=11 THEN price else 0 END) AS nov_sales from dyn_part_buck_clickstream where event_type = 'purchase' group by brand) select brand from brand_sales where (nov_sales - oct_sales)>0;

```
hive> WITH brand_sales AS (SELECT brand, sum(CASE WHEN month(event_time)=10 THEN price else 0 END) AS oct_sales, sum(CASE WHEN month(event_time)=11 THEN pric
e else 0 END) AS nov_sales from dyn_part_buck_clickstream where event_type = 'purchase' group by brand) select brand from brand_sales where (nov_sales - oct_
sales)>0;
Query ID = hadoop_20211230213810_f4a13410-789b-455a-8d5c-c9f28f3e408f
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1640892795316_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: 19.02 s
--------------------------------------------------------------------------------
OK
brand

airnails
art-visage
artex
aura
balbcare
barbie
batiste
beautix
beauty-free
beautyblender
beauugreen
benovy
binacil
bioaqua
biore
blixz
bluesky
bodyton
bpw.style
browxenna
candy
carmex
chi
coifin
concept
```

```
nirvel
nitrile
oniq
orly
osmo
ovale
plazan
polarus
profepil
profhenna
protokeratin
provoc
rasyan
refectocil
rosi
roubloff
runail
s.care
sanoto
severina
shary
shik
skinity
skinlite
smart
soleo
solomeya
sophin
staleks
strong
supertan
swarovski
tertio
treaclemoon
trind
uno
uskusi
veraclara
vilenta
yoko
yu-r
zeitun
Time taken: 19.634 seconds, Fetched: 161 row(s)
hive>
```

There are a total of 161 brands that have increased sales from October to November. Unoptimized query took 80.648 secs while optimized query took 19.634 secs.

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.

Unoptimized Query:

WITH spending_summary AS (select user_id, sum(price) as total_spending from clickstream where event_type = 'purchase' group by user_id order by total_spending desc) select user_id from spending_summary limit 10;

```
hive> WITH spending_summary AS (select user_id, sum(price) as total_spending from clickstream where event_type = 'purchase' group by user_id order by total_s
pending desc) select user_id from spending_summary limit 10;
Query ID = hadoop_20211230215428_35c8adc7-978c-444c-a2c4-0f8b10cc71d2
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_1640892795316_0009)

--------------------------------------------------------------------------------
        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: 62.65 s
--------------------------------------------------------------------------------
OK
user_id
557790271
150318419
562167663
531900924
557850743
522130011
561592095
431950134
566576008
521347209
Time taken: 72.797 seconds, Fetched: 10 row(s)
```

Optimized Query:

WITH spending_summary AS (select user_id, sum(price) as total_spending from dyn_part_buck_clickstream where event_type = 'purchase' group by user_id order by total_spending desc) select user_id from spending_summary limit 10;

```
hive> WITH spending_summary AS (select user_id, sum(price) as total_spending from dyn_part_buck_clickstream where event_type = 'purchase' group by user_id or
der by total_spending desc) select user_id from spending_summary limit 10;
Query ID = hadoop_20211230220036_6406003b-2b0b-4cb4-a0d9-2cb2bfa61b72
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1640892795316_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
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 17.05 s
--------------------------------------------------------------------------------
OK
user_id
557790271
150318419
562167663
531900924
557850743
522130011
561592095
431950134
566576008
521347209
Time taken: 18.038 seconds, Fetched: 10 row(s)
hive>
```

Above are the top 10 customers who spend the most who should be awarded with Golden customer plan. Unoptimized query took 72.797 secs while optimized query took 18.038 secs.

Dropping database:

drop database hive_casestudy cascade;

```
hive> drop database hive_casestudy cascade;
OK
Time taken: 0.533 seconds
hive> show databases;
OK
database_name
default
Time taken: 0.012 seconds, Fetched: 1 row(s)
hive>
```

Terminating the cluster: