
E-Commerce Case Study 2022



(Image courtesy: <https://www.forbes.com>)

MAY 4

upGrad & IIITB(DS_B17_C3)

Authored by: Vijaykumar Rangvani
Sonam Gupta



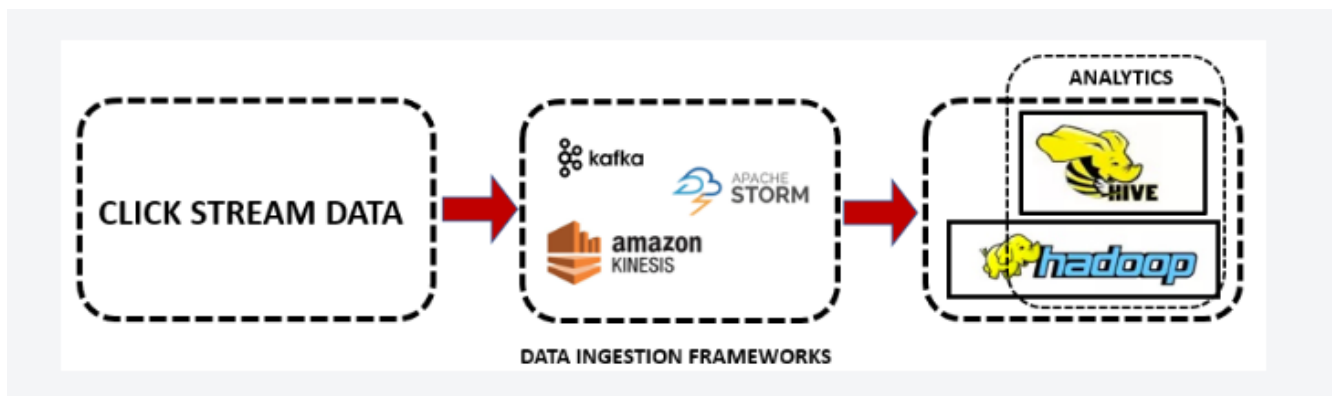
(Image courtesy: <https://en.wikipedia.org/>)

Introduction

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.

Problem Statement

One of the most popular use cases of Big Data is in ecommerce companies such as Amazon or Flipkart. So before we get into the details of the dataset, let us understand how ecommerce companies make use of these concepts to give customers product recommendations. This is done by tracking your clicks on their website and searching for patterns within them. This kind of data is called clickstream data.



The clickstream data contains all the logs as to how you navigated through the website. It also contains other details such as time spent on every page, etc. From this, they make use of data ingesting frameworks such as Apache Kafka or AWS Kinesis in order to store it in frameworks such as Hadoop.

Business Objective

Therefore, as part of this case study, as a big data analyst, we will be extracting data and will be gathering insights from a real-life data set of an e-commerce company.

Solution Steps:

Copying the data set into the HDFS:

- Launch an EMR cluster that utilizes the Hive services, and
- Move the data from the S3 bucket into the HDFS

Creating the database and launching Hive queries on your EMR cluster:

- Create the structure of your database,
- Using optimized techniques to run your queries as efficiently as possible
- Run Hive queries to answer the questions.

Cleaning up

- Drop database, and
- Terminate cluster

```
graph LR; A[Copying the data set into the HDFS] --> B[Creating the database and launching Hive queries on EMR cluster]; B --> C[Cleaning up];
```

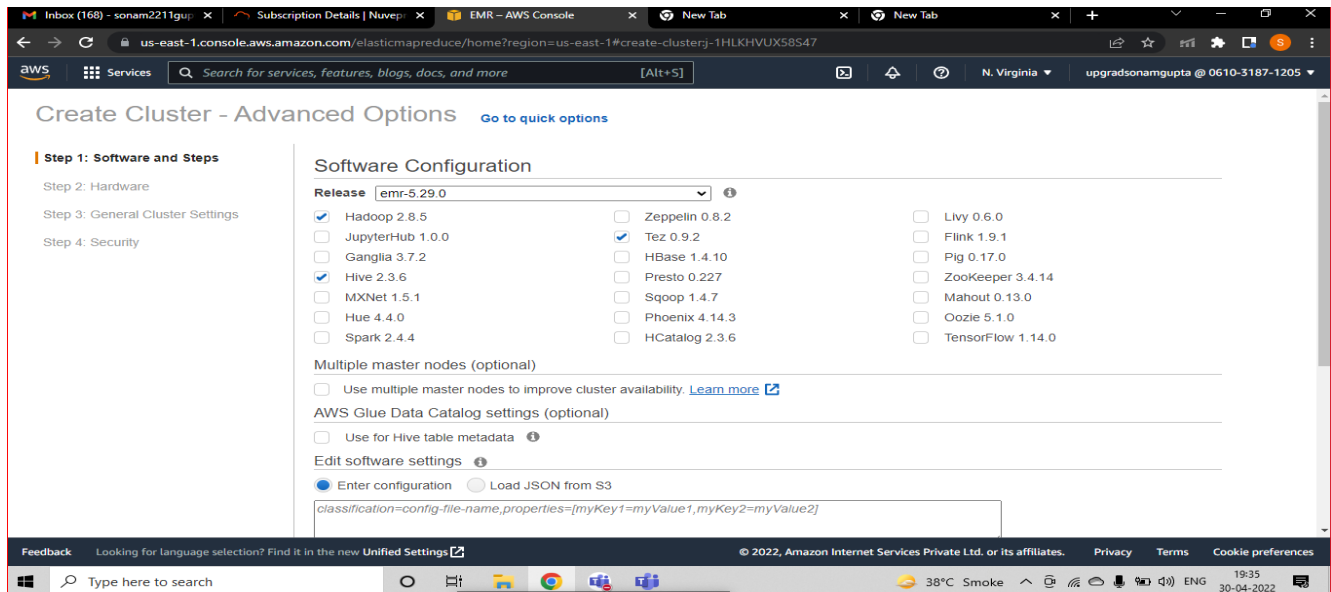
Copying the
data set into
the HDFS

Creating the
database and
launching Hive
queries on EMR
cluster

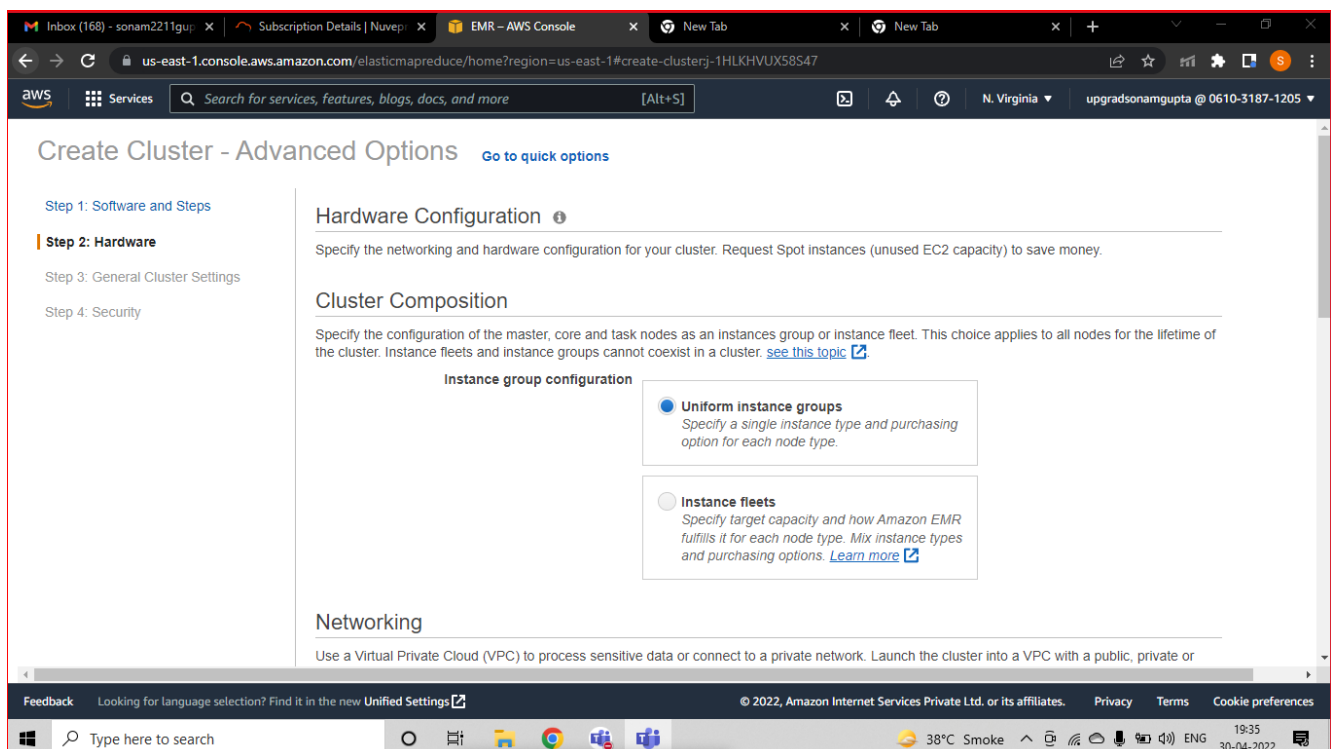
Cleaning up

Launching an EMR Cluster:

- We have created EMR Cluster having **emr-5.29.0** release along with Hive service installed.



- Next, as a part of Hardware configuration, created 2-Node EMR cluster with Master and Core node as **M4.large**.



- Verifying that EMR Cluster is up and running now and able to login into it using Putty.

[illegible]

- Uploaded files into S3 bucket under bucket named **casestudyhive**.

The screenshot displays the Amazon S3 console interface. The main content area shows the 'casestudyhive' bucket with the 'Objects' tab active. Below the tab, there is a description of objects and a search bar. A table lists two objects: '2019-Oct.csv' and '2019-Nov.csv'. The table has columns for Name, Type, Last modified, Size, and Storage class. The left sidebar contains the 'Amazon S3' navigation menu, and the top navigation bar shows the AWS logo, search bar, and user profile.

	Name	Type	Last modified	Size	Storage class
<input type="checkbox"/>	2019-Oct.csv	csv	April 30, 2022, 01:57:38 (UTC+05:30)	460.2 MB	Standard
<input type="checkbox"/>	2019-Nov.csv	csv	April 30, 2022, 01:57:38 (UTC+05:30)	520.6 MB	Standard

Importing data from S3 bucket into HDFS

- Creating the Directory named “casestudy”.

hadoop fs -mkdir /casestudy

```
[hadoop@ip-172-31-69-12 ~]$ hadoop fs -ls /
Found 4 items
drwxr-xr-x - hdfs hadoop 0 2022-04-30 14:12 /apps
drwxr-xr-x - hdfs hadoop 0 2022-04-30 14:13 /tmp
drwxr-xr-x - hdfs hadoop 0 2022-04-30 14:12 /user
drwxr-xr-x - hdfs hadoop 0 2022-04-30 14:12 /var
[hadoop@ip-172-31-69-12 ~]$ hadoop fs -mkdir /casestudy
[hadoop@ip-172-31-69-12 ~]$ aws s3 ls s3://casestudyhive/
2022-04-29 20:27:38 545839412 2019-Nov.csv
2022-04-29 20:27:38 482542278 2019-Oct.csv
[hadoop@ip-172-31-69-12 ~]$
```

- Copying files from S3 bucket to HDFS.

hadoop distcp 's3://casestudyhive/*' '/casestudy/'

```
[hadoop@ip-172-31-69-12 ~]$ hadoop distcp 's3://casestudyhive/*' '/casestudy/'
22/04/30 14:22:01 INFO tools.DistCp: Input Options: DistCpOptions{atomicCommit=false, syncFolder=false, deleteMissing=false, ignoreFailures=false, overwrite=false, skipCRC=false, blocking=true, numListStatusThreads=0, maxMaps=20, mapBandwidth=100, sslConfigurationFile='null', copyStrategy='uniformsize', preserveStatus=[], preserveRawXAttrs=false, atomicWorkPath=null, logPath=null, sourceFileListing=null, sourcePaths=[s3://casestudyhive/*], targetPath=/casestudy, targetPathExists=true, filtersFile='null'}
22/04/30 14:22:02 INFO client.RMPProxy: Connecting to ResourceManager at ip-172-31-69-12.ec2.internal/172.31.69.12:8032
22/04/30 14:22:09 INFO tools.SimpleCopyListing: Paths (files+dirs) cnt = 2; dirCnt = 0
22/04/30 14:22:09 INFO tools.SimpleCopyListing: Build file listing completed.
22/04/30 14:22:09 INFO Configuration.deprecation: io.sort.mb is deprecated. Instead, use mapreduce.task.io.sort.mb
22/04/30 14:22:09 INFO Configuration.deprecation: io.sort.factor is deprecated. Instead, use mapreduce.task.io.sort.factor
22/04/30 14:22:09 INFO tools.DistCp: Number of paths in the copy list: 2
22/04/30 14:22:09 INFO tools.DistCp: Number of paths in the copy list: 2
22/04/30 14:22:09 INFO client.RMPProxy: Connecting to ResourceManager at ip-172-31-69-12.ec2.internal/172.31.69.12:8032
22/04/30 14:22:11 INFO mapreduce.JobSubmitter: number of splits:2
22/04/30 14:22:11 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1651327969842_0001
22/04/30 14:22:12 INFO impl.YarnClientImpl: Submitted application application_1651327969842_0001
22/04/30 14:22:12 INFO mapreduce.Job: The url to track the job: http://ip-172-31-69-12.ec2.internal:20888/proxy/application_1651327969842_0001/
22/04/30 14:22:12 INFO tools.DistCp: DistCp job-id: job_1651327969842_0001
22/04/30 14:22:12 INFO mapreduce.Job: Running job: job_1651327969842_0001
```

```
[hadoop@ip-172-31-69-12 ~]$ hadoop fs -ls /casestudy
Found 2 items
-rw-r--r-- 1 hadoop hadoop 545839412 2022-04-30 14:22 /casestudy/2019-Nov.csv
-rw-r--r-- 1 hadoop hadoop 482542278 2022-04-30 14:22 /casestudy/2019-Oct.csv
[hadoop@ip-172-31-69-12 ~]$
```

Creating Hive Tables:

- Creating Database named “clickstream”.

CREATE DATABASE IF NOT EXISTS clickstream COMMENT "Database to store clickstream Data";

```
hive> CREATE DATABASE IF NOT EXISTS clickstream COMMENT "Database to store clickstream Data";
OK
Time taken: 2.46 seconds
hive>
```

- ---Setting up hive paramter to display the header
set hive.cli.print.header=True;
- ---Using clickstream DB
USE clickstream;

BASE TABLE (ecom_tab)

- Creating table named “ecom_table”.

```
hive> ---Creating table to store data for Oct and Nov months
hive> CREATE EXTERNAL TABLE IF NOT EXISTS ecom_table(
> 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'
> WITH SERDEPROPERTIES
> ('separatorChar' = ",", "quoteChar" = "\"", "escapeChar" = "\\")
> STORED AS TEXTFILE
> LOCATION '/casestudy/'
> TBLPROPERTIES (
> 'serialization.null.format'='',
> 'skip.header.line.count'='1');
OK
Time taken: 0.448 seconds
```

- Verifying data under table “ecom_table”.

SELECT * FROM ecom_table LIMIT 5;

```
hive> ---Verifying the data
hive> SELECT * FROM ecom_table LIMIT 5;
OK
ecom_table.event_time  ecom_table.event_type  ecom_table.product_id  ecom_table.category_id  ecom_table.category_code  ecom_table.brand  ecom_table.price
ecom_table.user_id    ecom_table.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-alcc-af0575a34ffb
2019-11-01 00:00:10 UTC view  5837166 1783999064103190764  pnb  22.22  556138645  57ed222e-a54a-4907-5944-5a875c2d7f4f
2019-11-01 00:00:11 UTC cart  5876812 1487580010100293687  jessmail  3.16  564506666  106c1951-8052-4b37-adcc-dd9644b1d5f7
2019-11-01 00:00:24 UTC remove_from_cart  5826182 1487580007483048900  3.33  553329724  2067216c-31b5-455d-alcc-af0575a34ffb
Time taken: 2.472 seconds, Fetched: 5 row(s)
```


- Setting up parameters for creating tables with bucketing and partitioning.

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

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

```
set hive.enforce.bucketing=true;
```

```
hive> ---Setting up parameters for partitioning and Bucketing
hive> set hive.exec.dynamic.partition=true;
hive> set hive.exec.dynamic.partition.mode=nonstrict;
hive> set hive.enforce.bucketing=true;
hive>
```

PARTITIONED TABLE (ecom_table_part)

- Partitioning the event type column and creating table for the same named “ecom_table_part”.

```
hive> ---Partitioning the event type column and creating table for the same
hive> CREATE EXTERNAL TABLE IF NOT EXISTS ecom_table_part(
> 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)
> STORED AS TEXTFILE
> TBLPROPERTIES (
> 'serialization.null.format'='');
OK
Time taken: 0.128 seconds
hive>
```

- Inserting Data into partitioned table

```
hive> ---Inserting Data into partitioned table
hive> INSERT INTO TABLE ecom_table_part
> PARTITION (event_type)
> SELECT event_time,
> product_id,
> category_id,
> category_code,
> brand,
> price,
> user_id,
> user_session,
> event_type
> FROM ecom_table;
Query ID = hadoop_20220430144703_4d7d4786-48cc-4cb3-8c9a-42df4a64f251
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_1651327969842_0003)
```

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: 134.54 s
Loading data to table clickstream.ecom_table_part partition (event_type=null)
Loaded : 4/4 partitions.
Time taken to load dynamic partitions: 0.813 seconds
Time taken for adding to write entity : 0.005 seconds
OK
_col0 _col1 _col2 _col3 _col4 _col5 _col6 _col7 _col8
Time taken: 147.761 seconds
hive>
```


- Verifying data under table “ecom_table_part”.

SELECT * FROM ecom_table_part LIMIT 5;

```
hive> ---Checking Data of Partitioned table
hive> SELECT * FROM ecom_table_part LIMIT 5;
OK
ecom_table_part.event_time    ecom_table_part.product_id    ecom_table_part.category_id    ecom_table_part.category_code    ecom_table_part.brand    ecom_table_part.
price    ecom_table_part.user_id    ecom_table_part.user_session    ecom_table_part.event_type
2019-10-09 15:28:14 UTC 5773158 1487580012969197740 NULL irisk 2.79 558203129 b1e0b3d3-d60b-bd68-d7e5-500a7af8920f cart
2019-10-07 20:53:09 UTC 5663062 1487580009622143014 NULL NULL 1.43 251478914 a99a5589-0f7a-40a5-9748-b19961fc4d30 cart
2019-10-07 20:53:11 UTC 5796751 1487580009416134735 NULL NULL 4.29 533966373 a834973c-30a5-4268-9fbc-597ea0865ff4 cart
2019-10-07 20:53:11 UTC 5835333 1926797403503985079 NULL NULL 4.76 552795963 24632cad-25f2-d02c-cfe7-6a59a096565a cart
2019-10-07 20:53:14 UTC 5854897 148758000944582239 NULL irisk 0.32 540100212 b445a30f-leaf-4275-9d0e-911e7bbdfbc2 cart
Time taken: 0.298 seconds, Fetched: 5 row(s)
hive>
```

BUCKETING TABLE (ecom_table_optimize)

- Creating table partitioned by event_type and bucketing the values of price.

```
hive> ---Creating table partitioned by event_type and bucketing the values of price
hive> CREATE EXTERNAL TABLE IF NOT EXISTS ecom_table_optimize(
> 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(
> price) INTO 10 buckets
> STORED AS TEXTFILE
> TBLPROPERTIES (
> 'serialization.null.format'='');
OK
Time taken: 0.116 seconds
hive>
```

- Inserting Data into the bucketed table

```
hive> ---Inserting Data into the bucketed table
hive> INSERT INTO TABLE ecom_table_optimize
> PARTITION (event_type)
> SELECT event_time,
> product_id,
> category_id,
> category_code,
> brand,
> price,
> user_id,
> user_session,
> event_type
> FROM ecom_table;
Query ID = hadoop_20220430145157_ff14b874-a31e-4abd-a64b-e38ce83ac2a1
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1651327969842_0003)

-----
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: 165.94 s
-----
Loading data to table clickstream.ecom_table_optimize partition (event_type=null)
Loaded : 4/4 partitions.
Time taken to load dynamic partitions: 0.428 seconds
Time taken for adding to write entity : 0.001 seconds
OK
_col10    _col11    _col12    _col13    _col14    _col15    _col16    _col17    _col18
Time taken: 167.981 seconds
hive>
```

- Verifying data under table “ecom_table_optimize”.

SELECT * FROM ecom_table_optimize LIMIT 5;

```
hive> ---Checking Data of Bucketed table
hive> SELECT * FROM ecom_table_optimize LIMIT 5;
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
ecom_table_optimize.event_time  ecom_table_optimize.product_id  ecom_table_optimize.category_id  ecom_table_optimize.category_code  ecom_table_optimize.brand  ecom_table_optimize.event_type
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
019-10-05 15:18:54 UTC 5569510 1487580012594363565 NULL NULL 5.24 418229350 63a3f2de-345d-432a-94ea-ee9c5fa3494e cart
019-10-05 13:52:48 UTC 5867181 1487580007910867929 NULL staleks 9.84 558467769 09b14b87-1920-4e2e-8977-0f955fa3f12a cart
019-10-05 02:59:02 UTC 5770055 1487580013841613016 NULL kapous 2.81 488284680 1f154d01-dfd7-407d-8bdf-91e2366a95cf cart
019-10-11 09:39:58 UTC 5758946 1487580005067129686 NULL NULL 2.38 551690620 f54e0c7d-3dca-46e9-ba63-c75a1ead4a71 cart
019-10-09 11:46:19 UTC 5807858 1487580009471148064 NULL irisk 1.27 452168577 b3342942-feb1-4dfc-9749-4fff404347b6 cart
Time taken: 0.23 seconds, Fetched: 5 row(s)
hive> exit;
```

- Checking created Buckets in HDFS.

hadoop fs -ls /user/hive/warehouse/ClickStream.db/ecom_table_optimize

hadoop fs ls

/user/hive/warehouse/ClickStream.db/ecom_table_optimize/event_type=purchase

```
hadoop@ip-172-31-69-12~$ hadoop fs -ls /
Found 5 items
drwxr-xr-x - hadoop hadoop 0 2022-04-30 14:12 /apps
drwxr-xr-x - hadoop hadoop 0 2022-04-30 14:22 /casestudy
drwxr-xrwt - hdfs hadoop 0 2022-04-30 14:13 /tmp
drwxr-xr-x - hdfs hadoop 0 2022-04-30 14:12 /user
drwxr-xr-x - hdfs hadoop 0 2022-04-30 14:12 /var
hadoop@ip-172-31-69-12~$ hadoop fs -ls /casestudy/
Found 2 items
-rw-r--r-- 1 hadoop hadoop 545839412 2022-04-30 14:22 /casestudy/2019-Nov.csv
-rw-r--r-- 1 hadoop hadoop 482542278 2022-04-30 14:22 /casestudy/2019-Oct.csv
hadoop@ip-172-31-69-12~$ hadoop fs -ls /warehouse/
ls: '/warehouse/': No such file or directory
hadoop@ip-172-31-69-12~$ hadoop fs -ls /user/
Found 4 items
drwxrwxrwx - hadoop hadoop 0 2022-04-30 14:24 /user/hadoop
drwxr-xr-x - mapred mapred 0 2022-04-30 14:12 /user/history
drwxrwxrwx - hdfs hadoop 0 2022-04-30 14:12 /user/hive
drwxrwxrwx - root hadoop 0 2022-04-30 14:12 /user/root
hadoop@ip-172-31-69-12~$ hadoop fs -ls /user/hive/
Found 1 items
drwxrwxrwt - hdfs hadoop 0 2022-04-30 14:25 /user/hive/warehouse
hadoop@ip-172-31-69-12~$ hadoop fs -ls /user/hive/warehouse/
Found 1 items
drwxrwxrwt - hadoop hadoop 0 2022-04-30 14:51 /user/hive/warehouse/clickstream.db
hadoop@ip-172-31-69-12~$ hadoop fs -ls /user/hive/warehouse/clickstream.db/ecom_table_optimize
Found 4 items
drwxrwxrwt - hadoop hadoop 0 2022-04-30 14:54 /user/hive/warehouse/clickstream.db/ecom_table_optimize/event_type=cart
drwxrwxrwt - hadoop hadoop 0 2022-04-30 14:54 /user/hive/warehouse/clickstream.db/ecom_table_optimize/event_type=purchase
drwxrwxrwt - hadoop hadoop 0 2022-04-30 14:54 /user/hive/warehouse/clickstream.db/ecom_table_optimize/event_type=remove_from_cart
drwxrwxrwt - hadoop hadoop 0 2022-04-30 14:54 /user/hive/warehouse/clickstream.db/ecom_table_optimize/event_type=view
hadoop@ip-172-31-69-12~$ hadoop fs -ls /user/hive/warehouse/clickstream.db/ecom_table_optimize/event_type=purchase
Found 10 items
-rwxrwxrwt 1 hadoop hadoop 8848048 2022-04-30 14:53 /user/hive/warehouse/clickstream.db/ecom_table_optimize/event_type=purchase/000000_0
-rwxrwxrwt 1 hadoop hadoop 6082854 2022-04-30 14:54 /user/hive/warehouse/clickstream.db/ecom_table_optimize/event_type=purchase/000001_0
-rwxrwxrwt 1 hadoop hadoop 5771952 2022-04-30 14:54 /user/hive/warehouse/clickstream.db/ecom_table_optimize/event_type=purchase/000002_0
-rwxrwxrwt 1 hadoop hadoop 5810623 2022-04-30 14:53 /user/hive/warehouse/clickstream.db/ecom_table_optimize/event_type=purchase/000003_0
-rwxrwxrwt 1 hadoop hadoop 3127215 2022-04-30 14:54 /user/hive/warehouse/clickstream.db/ecom_table_optimize/event_type=purchase/000004_0
-rwxrwxrwt 1 hadoop hadoop 6592136 2022-04-30 14:53 /user/hive/warehouse/clickstream.db/ecom_table_optimize/event_type=purchase/000005_0
-rwxrwxrwt 1 hadoop hadoop 6760201 2022-04-30 14:54 /user/hive/warehouse/clickstream.db/ecom_table_optimize/event_type=purchase/000006_0
-rwxrwxrwt 1 hadoop hadoop 5320537 2022-04-30 14:54 /user/hive/warehouse/clickstream.db/ecom_table_optimize/event_type=purchase/000007_0
-rwxrwxrwt 1 hadoop hadoop 6034152 2022-04-30 14:53 /user/hive/warehouse/clickstream.db/ecom_table_optimize/event_type=purchase/000008_0
-rwxrwxrwt 1 hadoop hadoop 8035002 2022-04-30 14:54 /user/hive/warehouse/clickstream.db/ecom_table_optimize/event_type=purchase/000009_0
hadoop@ip-172-31-69-12~$
```

Optimization Techniques:

Comparing the performance:

Query: Finding top 10 customers who are making most of purchases.

1. ecom_table: table without partitioning and bucketing

```
hive> use clickstream;
OK
Time taken: 0.209 seconds
hive> WITH customer_rank AS
> (
> SELECT user_id AS Customer,
> ROUND(SUM(price),2) AS Expenditure,
> RANK() OVER (ORDER BY ROUND(SUM(price),2) DESC) AS RANK
> FROM ecom_table
> WHERE event_type='purchase'
> GROUP BY user_id)
> SELECT Customer,
> Expenditure,
> RANK
> FROM customer_rank
> WHERE RANK<=10;
Query ID = hadoop_20220430150759_660c25a1-f2a9-48d4-bc9b-9d5318272b58
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1651327969842_0004)
```

	VERTICES	MODE	STATUS	TOTAL	COMPLETED	RUNNING	PENDING	FAILED	KILLED
Map 1	container	SUCCEEDED	2	2	0	0	0	0	0
Reducer 2	container	SUCCEEDED	3	3	0	0	0	0	0
Reducer 3	container	SUCCEEDED	2	2	0	0	0	0	0

```
VERTICES: 03/03 [=====] 100% ELAPSED TIME: 62.82 s
OK
557790271      2715.87 1
150318419      1645.97 2
562167663      1352.85 3
531900924      1329.45 4
557850743      1295.48 5
522130011      1185.39 6
561592095      1109.7  7
431950134      1097.59 8
566576008      1056.36 9
521347209      1040.91 10
Time taken: 68.661 seconds, Fetched: 10 row(s)
hive>
```

2. ecom_table_part: table with partitioning

```
hive> WITH customer_rank AS
> (
> SELECT user_id AS Customer,
> ROUND(SUM(price),2) AS Expenditure,
> RANK() OVER (ORDER BY ROUND(SUM(price),2) DESC) AS RANK
> FROM ecom_table_part
> WHERE event_type='purchase'
> GROUP BY user_id)
> SELECT Customer,
> Expenditure,
> RANK
> FROM customer_rank
> WHERE RANK<=10;
Query ID = hadoop_20220430151012_0ce48c21-e187-4e51-a92b-3840867cf810
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1651327969842_0004)
```

	VERTICES	MODE	STATUS	TOTAL	COMPLETED	RUNNING	PENDING	FAILED	KILLED
Map 1	container	SUCCEEDED	2	2	0	0	0	0	0
Reducer 2	container	SUCCEEDED	1	1	0	0	0	0	0
Reducer 3	container	SUCCEEDED	1	1	0	0	0	0	0

```
VERTICES: 03/03 [=====] 100% ELAPSED TIME: 18.35 s
OK
557790271      2715.87 1
150318419      1645.97 2
562167663      1352.85 3
531900924      1329.45 4
557850743      1295.48 5
522130011      1185.39 6
561592095      1109.7  7
431950134      1097.59 8
566576008      1056.36 9
521347209      1040.91 10
Time taken: 19.565 seconds, Fetched: 10 row(s)
hive>
```

3. ecom_table_optimize – table with partitioning and bucketing

```
hive> WITH customer_rank AS
> (
> SELECT user_id AS Customer,
> ROUND(SUM(price),2) AS Expenditure,
> RANK() OVER (ORDER BY ROUND(SUM(price),2) DESC) AS RANK
> FROM ecom_table_optimize
> WHERE event_type='purchase'
> GROUP BY user_id)
> SELECT Customer,
> Expenditure,
> RANK
> FROM customer_rank
> WHERE RANK<=10;
Query ID = hadoop_20220430161127_296caecb-6a84-43a1-a3f4-690507100381
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1651327969842_0004)

-----
VERTICES      MODE           STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
-----
Map 1 ..... container    SUCCEEDED    3         3          0         0         0         0
Reducer 2 ..... container    SUCCEEDED    1         1          0         0         0         0
Reducer 3 ..... container    SUCCEEDED    1         1          0         0         0         0
-----
VERTICES: 03/03 [=====>] 100% ELAPSED TIME: 24.17 s
-----
CK
557790271      2715.87 1
150318419      1645.97 2
562167663      1352.85 3
531900924      1329.45 4
557850743      1295.48 5
522130011      1185.39 6
561592095      1109.7 7
431950134      1097.59 8
566576008      1056.36 9
521347209      1040.91 10
Time taken: 25.058 seconds, Fetched: 10 row(s)
hive>
```

Performance Comparison:

Below table shows the time taken to execute the query:

Without Partition and Bucketing (ecom_table)	With Partitioning (ecom_table_part)	With Partitioning and Bucketing (ecom_table_optimize)
68.66 seconds	19.56 seconds	25.05 seconds

Conclusion:

Hence, it can be clearly observed that partitioned tables and Bucketed tables definitely has better performance in terms of query execution time.

Analysis using Hive Queries:

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

```
hive> ---The total revenue generated due to purchases made in October
hive> SELECT ROUND (SUM(price),2) AS Total_Oct_Revenue
> FROM ecom_table_optimize
> WHERE event_type = 'purchase'
> GROUP BY month(event_time)
> HAVING month(event_time) = 10;
Query ID = hadoop_20220430151301_63592e57-0e36-454a-9a97-e61bfd105b41
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1651327969842_0004)
```

VERTICES	MODE	STATUS	TOTAL	COMPLETED	RUNNING	PENDING	FAILED	KILLED
Map 1	container	SUCCEEDED	3	3	0	0	0	0
Reducer 2	container	SUCCEEDED	1	1	0	0	0	0

VERTICES: 02/02 [=====] 100% ELAPSED TIME: 20.83 s

OK
1211538.43
Time taken: 21.67 seconds, Fetched: 1 row(s)

The total revenue generated due to purchases made in October is 1211538.43.

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

```
hive> ---Total sum of purchases per month in single output
hive> SELECT CASE WHEN (month(event_time) == 10) THEN 'OCT'
> ELSE 'NOV'
> END AS Month,
> ROUND (SUM(price),2) AS purchase
> FROM ecom_table_optimize
> WHERE event_type = 'purchase'
> GROUP BY month(event_time);
Query ID = hadoop_20220430151506_70d1213f-1ef5-4ab4-bf06-31dalbd19d41
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1651327969842_0004)
```

VERTICES	MODE	STATUS	TOTAL	COMPLETED	RUNNING	PENDING	FAILED	KILLED
Map 1	container	INITED	3	0	0	3	0	0
Reducer 2	container	INITED	1	0	0	1	0	0

VERTICES: 00/02 [>>-----] 0% ELAPSED TIME: 3.85 s

```
hive> END AS Month,
> ROUND (SUM(price),2) AS purchase
> FROM ecom_table_optimize
> WHERE event_type = 'purchase'
> GROUP BY month(event_time);
Query ID = hadoop_20220430151506_70d1213f-1ef5-4ab4-bf06-31dalbd19d41
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1651327969842_0004)
```

VERTICES	MODE	STATUS	TOTAL	COMPLETED	RUNNING	PENDING	FAILED	KILLED
Map 1	container	SUCCEEDED	3	3	0	0	0	0
Reducer 2	container	SUCCEEDED	1	1	0	0	0	0

VERTICES: 02/02 [=====] 100% ELAPSED TIME: 22.86 s

OK
OCT 1211538.43
NOV 1531016.9
Time taken: 23.842 seconds, Fetched: 2 row(s)

Sum of purchases in October is 1211538.43 and November is 1531016.9

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

```
hive> ---Finding Change in Revenue
hive> SELECT (Nov_Rev-Oct_Rev) AS Change_In_Revenue
> FROM (
> SELECT ROUND(SUM(CASE WHEN month(event_time)=10 THEN price ELSE 0 END),2) AS Oct_rev,
> ROUND(SUM(CASE WHEN month(event_time)=11 THEN price ELSE 0 END),2) AS Nov_rev
> FROM ecom_table_optimize
> WHERE event_type = 'purchase'
> ) AS rev_cal;
Query ID = hadoop_20220430151547_9b7f6916-b224-42de-8f92-5da9d87bee5c
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1651327969842_0004)

-----
VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
-----
Map 1 ..... container    SUCCEEDED    3         3         0         0         0         0
Reducer 2 ..... container    SUCCEEDED    1         1         0         0         0         0
-----
VERTICES: 02/02 [=====>>>] 100% ELAPSED TIME: 22.67 s
-----
OK
319478.47
Time taken: 23.506 seconds, Fetched: 1 row(s)
```

Change in revenue from October to November is 319478.47.

This indicates that there is strong positive increase in business.

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

```
hive> ---Finding Distinct categories of products
hive> SELECT DISTINCT category_code
> FROM ecom_table_optimize
> WHERE category_code IS NOT NULL;
Query ID = hadoop_20220430151650_0fcdad5e-e95e-4389-9667-53020a6148dd
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1651327969842_0004)
```

VERTICES	MODE	STATUS	TOTAL	COMPLETED	RUNNING	PENDING	FAILED	KILLED
Map 1	container	RUNNING	6	0	3	3	0	0
Reducer 2	container	INITED	4	0	0	4	0	0

VERTICES: 00/02 [>>-----] 0% ELAPSED TIME: 8.71 s

Windows taskbar showing search bar, task view, and system tray with date 30-04-2022 and time 20:47.

hadoop@ip-172-31-69-12~

VERTICES	MODE	STATUS	TOTAL	COMPLETED	RUNNING	PENDING	FAILED	KILLED
Map 1	container	SUCCEEDED	6	6	0	0	0	0
Reducer 2	container	SUCCEEDED	4	4	0	0	0	0

VERTICES: 02/02 [=====>>>] 100% ELAPSED TIME: 28.51 s

```
OK
accessories.bag
appliances.environment.vacuum
appliances.personal.hair_cutter
sport.diving
apparel.glove
furniture.bathroom.bath
furniture.living_room.cabinet
stationery.cartridge
accessories.cosmetic_bag
appliances.environment.air_conditioner
furniture.living_room.chair
Time taken: 29.349 seconds, Fetched: 11 row(s)
```

There are total 11 distinct categories of products.

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

```
hive> ---Finding total No. of products under each category
hive> SELECT category_code,
> COUNT(product_id) AS number_of_products
> FROM ecom_table_optimize
> WHERE category_code IS NOT NULL
> GROUP BY category_code
> ORDER BY number_of_products DESC;
Query ID = hadoop_20220430151744_71baa324-56c6-4b2f-95ba-dfb3637588bc
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1651327969842_0004)
```

VERTICES	MODE	STATUS	TOTAL	COMPLETED	RUNNING	PENDING	FAILED	KILLED
Map 1	container	RUNNING	6	0	3	3	0	0
Reducer 2	container	INITED	4	0	0	4	0	0
Reducer 3	container	INITED	1	0	0	1	0	0

VERTICES: 00/03 [>>-----] 0% ELAPSED TIME: 19.09 s

```
hive> SELECT category_code,
> COUNT(product_id) AS number_of_products
> FROM ecom_table_optimize
> WHERE category_code IS NOT NULL
> GROUP BY category_code
> ORDER BY number_of_products DESC;
Query ID = hadoop_20220430151744_71baa324-56c6-4b2f-95ba-dfb3637588bc
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1651327969842_0004)
```

VERTICES	MODE	STATUS	TOTAL	COMPLETED	RUNNING	PENDING	FAILED	KILLED
Map 1	container	SUCCEEDED	6	6	0	0	0	0
Reducer 2	container	SUCCEEDED	4	4	0	0	0	0
Reducer 3	container	SUCCEEDED	1	1	0	0	0	0

VERTICES: 03/03 [=====] 100% ELAPSED TIME: 28.54 s

OK

category_code	number_of_products
appliances.environment.vacuum	59761
stationery.cartridge	26722
apparel.glove	18232
furniture.living_room.cabinet	13439
accessories.bag	11681
furniture.bathroom.bath	9857
appliances.personal.hair_cutter	1643
accessories.cosmetic_bag	1248
appliances.environment.air_conditioner	332
furniture.living_room.chair	308
sport.diving	2

Time taken: 29.155 seconds, Fetched: 11 row(s)

As we already know that there are 11 distinct categories of products, the above query result shows the total number of available under each category.

Category having highest products available is appliances environment vacuum.

Category having lowest products available is sport diving.

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

```
hive> ---Finding Brand having max sales in Oct & Nov combined
hive> SELECT brand,
> ROUND(SUM(price),2) AS Sales
> FROM ecom_table_optimize
> WHERE brand IS NOT NULL AND
> event_type = 'purchase'
> GROUP BY brand
> ORDER BY Sales DESC LIMIT 1;
Query ID = hadoop_20220430151859_c262c14f-5ce0-4b5d-b43a-853636233a87
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1651327969842_0004)
```

VERTICES	MODE	STATUS	TOTAL	COMPLETED	RUNNING	PENDING	FAILED	KILLED
Map 1	container	INITED	3	0	0	3	0	0
Reducer 2	container	INITED	1	0	0	1	0	0
Reducer 3	container	INITED	1	0	0	1	0	0

VERTICES: 00/03 [>>-----] 0% ELAPSED TIME: 3.85 s

runail 148297.94

Runail is the brand that made the maximum sales in October and November combined with sales amount of 148297.94

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

```
hive> SELECT oct.brand AS Brand,
> ROUND((nov.Nov_sale - oct.Oct_sale),2) AS increase_In_Sale
> FROM(
> SELECT brand,SUM(price) AS Oct_Sale
> FROM ecom_table_optimize
> WHERE event_type = 'purchase'
> AND brand IS NOT NULL
> AND month(event_time) = 10
> GROUP BY brand
> ) oct
> JOIN (SELECT brand,SUM(price) AS Nov_sale
> FROM ecom_table_optimize
> WHERE event_type = 'purchase'
> AND brand IS NOT NULL
> AND month(event_time) = 11
> GROUP BY brand
> ) nov
> ON
> oct.brand = nov.brand
> WHERE
> nov.Nov_sale > oct.Oct_sale
> ORDER BY Increase_In_Sale DESC;
Query ID = hadoop_20220430152805_9083b2b3-5379-43b4-bdbe-702be1cfd7a5
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_1651327969842_0005)
```

	VERTICES	MODE	STATUS	TOTAL	COMPLETED	RUNNING	PENDING	FAILED	KILLED
Map 1	container	INITED	3	0	0	3	0	0	0
Map 5	container	INITED	3	0	0	3	0	0	0
Reducer 2	container	INITED	1	0	0	1	0	0	0
Reducer 3	container	INITED	1	0	0	1	0	0	0
Reducer 4	container	INITED	1	0	0	1	0	0	0
Reducer 6	container	INITED	1	0	0	1	0	0	0

VERTICES: 00/06 [>-----] 0% ELAPSED TIME: 1.09 s

```
hadoop@ip-172-31-09-12:~$
$ cat /tmp/hadoop-hive/hive_warehouse/warehouse/brand_sales.txt
brand    oct_sale    nov_sale
Kattol  36027.17
Uno      15737.72
Lianail  10501.4
Ingarden 10404.82
Kronog  9474.64
Kessnail 7057.39
Cosmoprofi 6214.18
Polarus  5358.21
Runail   5219.38
Trendecor 4250.02
Kaleks  3355.88
Upw.style 3265.29
Lovely  3234.68
Sarathon 2992.35
Saruyama 2962.22
Joko     2950.97
Ktalwax  2859.13
Benovy   2850.35
Saypro   2387.36
Estel    2385.92
Concept  2348.26
Mapous   2165.92
C.o.x    1953.05
Sasura   1792.39
Daily    1737.07
Kautix   1729.0
Krtex    1596.41
Kmix     1537.12
Shik     1498.52
Smart    1444.88
Dubloff  1422.41
Levrana  1420.54
Uniq     1416.24
Risk     1354.08
Severina 1344.6
Cico     1309.58
Keltun   1300.97
Beauty-free 1228.69
Kwazowski 1155.23
De.lux   1115.81
Ktger    1083.71
Markell  1065.68
Kamoto    1052.54
Kagaraku 957.94
```

```
hadoop@ip-172-31-69-12~  
yu-r 402.3  
kiss 395.78  
lador 387.92  
ellips 360.19  
gas 338.47  
lowence 324.91  
nitriple 315.4  
shary 304.53  
kims 302.0  
happyfons 289.67  
kocostar 284.08  
insight 278.26  
candy 264.42  
bluesky 258.29  
beauugreen 256.84  
protokeratin 255.54  
trind 244.89  
entity 239.55  
skinlite 238.51  
provoc 235.83  
fedua 211.43  
ecocraft 200.79  
keen 199.27  
mane 193.47  
freshbubble 183.64  
chi 179.67  
cristalinas 157.32  
farmona 150.97  
latinoil 135.07  
miskin 135.03  
elizavecca 133.77  
nefertiti 133.12  
finish 132.0  
igrobeauty 131.41  
dizao 126.38  
osmo 116.73  
batiste 101.77  
carmex 98.28  
eos 98.27  
depilflax 96.71  
enjoy 95.22  
kerasys 94.29  
aura 93.56  
plazan 92.64
```

```
hadoop@ip-172-31-69-12~  
koelf 84.56  
nirvel 71.29  
konad 70.84  
egomania 68.57  
cutrin 68.25  
laboratorium 66.02  
inm 63.19  
marutaka-foot 60.11  
profhenna 57.62  
koelcia 57.25  
palbcare 57.05  
elskin 56.56  
foamie 45.45  
ladykin 44.92  
likato 44.91  
mavala 37.28  
vilenta 33.61  
beautyblender 30.67  
piore 29.66  
orly 28.71  
avoclare 27.06  
profepil 24.66  
elixx 24.45  
codefroy 23.9  
plysolid 21.86  
veraclara 21.1  
kamill 18.48  
treaclemoon 18.12  
supertan 16.14  
heoproce 12.33  
rasyan 10.14  
fly 10.03  
tertio 9.64  
jaguar 8.54  
oleo 8.33  
neoleor 8.29  
moyou 4.57  
bodyton 4.3  
skinity 3.56  
grace 1.69  
cosima 0.7  
ovale 0.56  
Time taken: 38.966 seconds, Fetched: 152 row(s)  
hive>
```

Total 152 brands have increased their sales value from October to November.

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.

```
hive> WITH Customer_Rank AS
> (
> SELECT user_id AS Customer,
> ROUND(SUM(price),2) AS Expenditure,
> RANK () OVER (ORDER BY ROUND(SUM(price),2) DESC) AS Rank
> FROM ecom_table_optimize
> WHERE event_type = 'purchase'
> GROUP BY user_id
> )
> SELECT Customer,
> Expenditure,
> Rank
> FROM Customer_Rank
> WHERE Rank <= 10;
Query ID = hadoop_20220430153303_a2928b81-4527-41b9-9a4d-3bae36151f21
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1651327969842_0005)
```

	VERTICES	MODE	STATUS	TOTAL	COMPLETED	RUNNING	PENDING	FAILED	KILLED
Map 1	container	SUCCEEDED	3	3	0	0	0	0	0
Reducer 2	container	SUCCEEDED	1	1	0	0	0	0	0
Reducer 3	container	SUCCEEDED	1	1	0	0	0	0	0

```
VERTICES: 03/03 [=====] 100% ELAPSED TIME: 23.21 s
-----
OK
557750271      2715.87 1
150310419      1645.87 2
562167663      1352.85 3
531900924      1329.45 4
557850743      1295.48 5
522130011      1185.39 6
561592095      1109.7  7
431950134      1097.59 8
566576008      1056.36 9
521347209      1040.91 10
Time taken: 23.995 seconds, Fetched: 10 row(s)
hive>
```

Query output provides list of top 10 customers who are eligible for Golden plan.

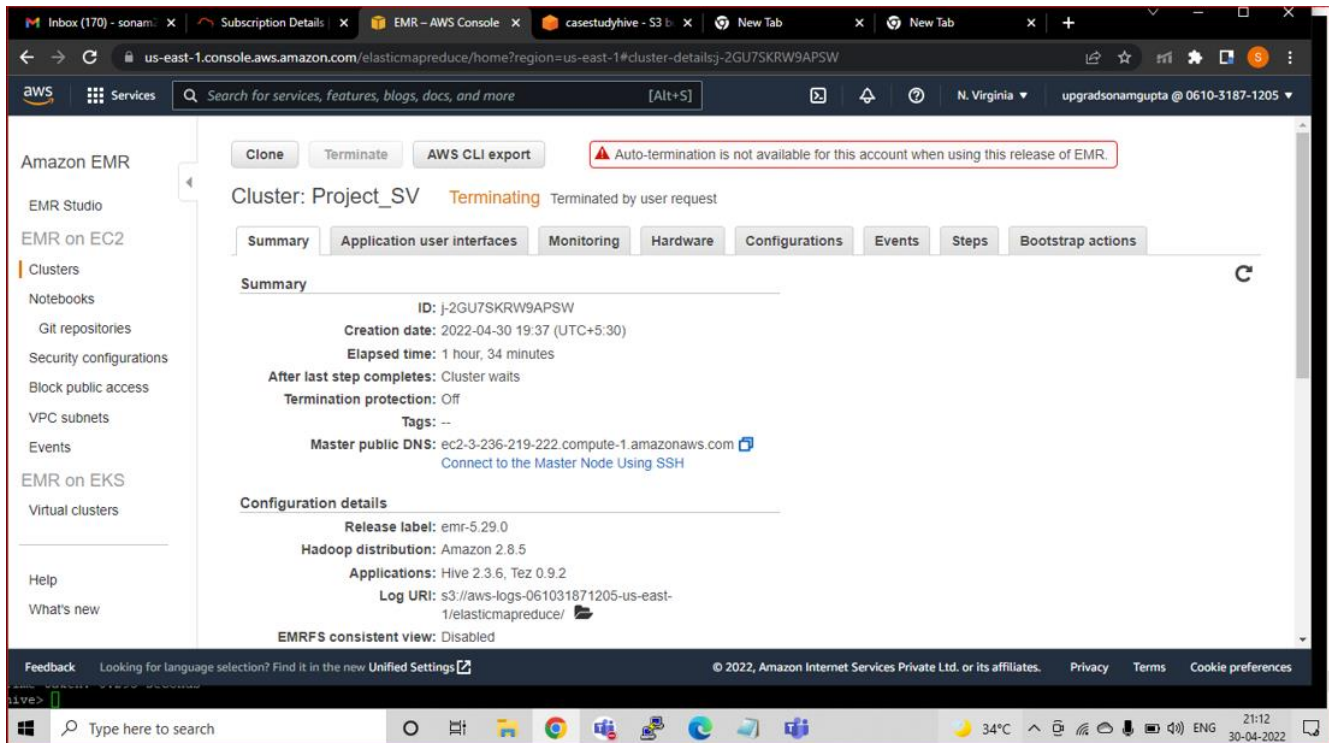
Dropping Database

- At completion, we dropped all created tables and database.

```
hive> SHOW TABLES;
OK
ecom_table
ecom_table_optimize
ecom_table_part
Time taken: 0.042 seconds, Fetched: 3 row(s)
hive> DROP TABLE ecom_table;
OK
Time taken: 0.412 seconds
hive> DROP TABLE ecom_table_optimize;
OK
Time taken: 0.145 seconds
hive> DROP TABLE ecom_table_part;
OK
Time taken: 0.187 seconds
hive> SHOW TABLES;
OK
Time taken: 0.017 seconds
hive> SHOW DATABASES;
OK
clickstream
default
Time taken: 0.015 seconds, Fetched: 2 row(s)
hive> DROP DATABASE clickstream;
FAILED: SemanticException [Error 10072]: Database does not exist: clickstream
hive> DROP DATABASE clickstream;
OK
Time taken: 0.295 seconds
hive>
```

Cluster Termination:

- At completion, we terminated the created cluster named **Project_SV**.



Insights:

- Change in revenue from October to November is 319478.47. This indicates that there is strong positive increase in business.
- Runail is the brand that made the maximum sales in October and November combined with sales amount of 148297.94
- Total 152 brands have increased their sales value from October to November.
- There are 11 distinct categories of products
- Category having highest products available is appliances environment vacuum and lowest products available is sport diving