

# HIVE CASE STUDY

(DSC30)- Sahil Thakare

## 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. In this case study, we are working with clickstream data by getting insights and making decisions upon how the E-commerce websites can improve their sales.

## OBJECTIVE:

The aim is to extract the data and gather insights from a real-life data set of an e-commerce company.

## DATA:

The data used in this assignment is a public clickstream dataset of a cosmetics store. The clickstream data contains all the logs as to how one navigated through the e-commerce website. It also contains other details such as customer time spent on every page, number of clicks made, adding items to the cart, customer id, etc.



You will find the data in the link given below.

<https://e-commerce-events-ml.s3.amazonaws.com/2019-Oct.csv>

<https://e-commerce-events-ml.s3.amazonaws.com/2019-Nov.csv>

## **The implementation phase:**

- 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 structure of your database,
  - Use optimized techniques to run your queries as efficiently as possible
  - Show the improvement of the performance after using optimization on any single query.
  - Run Hive queries to answer the questions given below.
- Cleaning up
  - Drop your database, and
  - Terminate your cluster

# ❖ EMR CLUSTER CREATION

EMR Cluster Landing page > Create Cluster > Advanced Options > Selecting the release emr-5.29 and the required services.

Create Cluster - Advanced Options [Go to quick options](#)

Step 1: Software and Steps

Step 2: Hardware

Step 3: General Cluster Settings

Step 4: Security

Software Configuration

Release emr-5.29.0

☒ Hadoop 2.8.5

☐ JupyterHub 1.0.0

☐ Ganglia 3.7.2

☒ Hive 2.3.6

☐ MXNet 1.5.1

☒ Hue 4.4.0

☐ Spark 2.4.4

☐ Zeppelin 0.8.2

☐ Tez 0.9.2

☐ HBase 1.4.10

☐ Presto 0.227

☐ Sqoop 1.4.7

☐ Phoenix 4.14.3

☐ HCatalog 2.3.6

☐ Livy 0.6.0

☐ Flink 1.9.1

☒ Pig 0.17.0

☐ ZooKeeper 3.4.14

☐ Mahout 0.13.0

☐ Oozie 5.1.0

☐ TensorFlow 1.14.0

Multiple master nodes (optional)

☐ Use multiple master nodes to improve cluster availability. [Learn more](#)

AWS Glue Data Catalog settings (optional)

Hardware Configuration Page > To define the cluster & nodes: Instance type for both master & core nodes are M4. Large

## Cluster Nodes and Instances

Choose the instance type, number of instances, and a purchasing option. [Learn more about instance purchasing options](#)

Console options for automatic scaling have changed. [Learn more](#)

Node type	Instance type	Instance count	Purchasing option
Master Master - 1	m4.large 2 vCore, 8 GiB memory, EBS only storage EBS Storage: 32 GiB Add configuration settings	1 Instances	<div><div>On-demand</div><div>Spot</div><div>Use on-demand as max price</div></div>
Core Core - 2	m4.large 2 vCore, 8 GiB memory, EBS only storage EBS Storage: 32 GiB Add configuration settings	1 Instances	<div><div>On-demand</div><div>Spot</div><div>Use on-demand as max price</div></div>

Naming the Cluster:

### General Options

**Cluster name**

☒ Logging ⓘ  
S3 folder

☒ Debugging ⓘ

☒ Termination protection ⓘ

**Tags** ⓘ

Key	Value (optional)
-----	------------------

Selecting the Key-pair (Created before creating the cluster):

### Security Options

**EC2 key pair**  ▼ ⓘ

☒ Cluster visible to all IAM users in account ⓘ

**Permissions** ⓘ

☒ Default ☐ Custom  
Use default IAM roles. If roles are not present, they will be automatically created for you with managed policies for automatic policy updates.

**EMR role** [EMR\\_DefaultRole](#) ☐ Use EMR\_DefaultRole\_V2 ⓘ

**EC2 instance profile** [EMR\\_EC2\\_DefaultRole](#) ⓘ

**Auto Scaling role** [EMR\\_AutoScaling\\_DefaultRole](#) ⓘ

Cluster “Hive\_Case\_Study\_01” is successfully created and launched.

Using the “Hive\_Case\_Study” key-pair we enter the terminal.

## ❖ HADOOP & HIVE QUERIES:

Terminal > Connecting to EMR Cluster(Putty).

```

[ec2] Using username "hadoop".
[ec2] Authenticating with public key "imported-openssh-key"
Last login: Tue Nov 30 12:50:56 2021

    _ _ | _ _ | _ )
    _ | ( _ _ /
    _ | \ _ _ | _ _ |

Amazon Linux AMI

https://aws.amazon.com/amazon-linux-ami/2018.03-release-notes/
68 package(s) needed for security, out of 106 available
Run "sudo yum update" to apply all updates.

EEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE MMMMMMMMMM MMMMMMMMMM RRRRRRRRRRRRRRRRR
E::::::::::::::::::::E M::::::::M M::::::::M R::::::::::::R
EE::::::::EEEEEEEEEE::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 M:::M R:::R R:::R
E:::EEEEEEEEEE M:::M M:::M M:::M R:::RRRRRR:::R
E::::::::::::E 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
EEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE MMMMMMMMMM RRRRRRR RRRRRR

[hadoop@ip-10-0-0-13 ~]$

```

## Creating a directory “casestudy”

```
hadoop fs -mkdir /casestudy
```

```
hadoop fs -ls /
```

```
[hadoop@ip-10-0-0-61 ~]$ hadoop fs -mkdir /casestudy
[hadoop@ip-10-0-0-61 ~]$ hadoop fs -ls /
Found 5 items
drwxr-xr-x  - hdfs  hadoop          0 2021-12-07 11:27 /apps
drwxr-xr-x  - hadoop hadoop          0 2021-12-07 11:36 /casestudy
drwxrwxrwt  - hdfs  hadoop          0 2021-12-07 11:29 /tmp
drwxr-xr-x  - hdfs  hadoop          0 2021-12-07 11:27 /user
drwxr-xr-x  - hdfs  hadoop          0 2021-12-07 11:27 /var
```

## Loading the datasets into HDFS from S3:

`hadoopdistcp 's3://e-commerce-events-ml/2019-Oct.csv' /hive02/2019_Oct.csv`

```
hadoop@ip-10-0-0-61 ~]$ hadoop distcp 's3://e-commerce-events-ml/2019-Oct.csv' /hive02/2019-Oct.csv
21/12/07 11:41:07 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://e-commerce-events-ml/2019-Oct.csv], targetPath=/hive02/2019-Oct.csv, targetPathExists=false, filtersFile='null'}
21/12/07 11:41:07 INFO client.RMPProxy: Connecting to ResourceManager at ip-10-0-0-61.ec2.internal/10.0.0.61:8032
21/12/07 11:41:12 INFO tools.SimpleCopyListing: Paths (files+dirs) cnt = 1; dirCnt = 0
21/12/07 11:41:12 INFO tools.SimpleCopyListing: Build file listing completed.
```

```
21/12/07 11:41:45 INFO mapreduce.Job: map 100% reduce 0%
21/12/07 11:41:45 INFO mapreduce.Job: Job job_1638876533904_0001 completed successfully
21/12/07 11:41:45 INFO mapreduce.Job: Counters: 38
  File System Counters
    FILE: Number of bytes read=0
    FILE: Number of bytes written=172396
    FILE: Number of read operations=0
    FILE: Number of large read operations=0
    FILE: Number of write operations=0
    HDFS: Number of bytes read=360
    HDFS: Number of bytes written=482542278
    HDFS: Number of read operations=12
    HDFS: Number of large read operations=0
    HDFS: Number of write operations=4
```

`hadoopdistcp 's3://e-commerce-events-ml/2019-Oct.csv' /hive02/2019_Nov.csv`

```
hadoop@ip-10-0-0-61 ~]$ hadoop distcp 's3://e-commerce-events-ml/2019-Nov.csv' /hive02/2019-Nov.csv
21/12/07 11:42:44 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://e-commerce-events-ml/2019-Nov.csv], targetPath=/hive02/2019-Nov.csv, targetPathExists=false, filtersFile='null'}
21/12/07 11:42:44 INFO client.RMPProxy: Connecting to ResourceManager at ip-10-0-0-61.ec2.internal/10.0.0.61:8032
21/12/07 11:42:48 INFO tools.SimpleCopyListing: Paths (files+dirs) cnt = 1; dirCnt = 0
21/12/07 11:42:48 INFO tools.SimpleCopyListing: Build file listing completed.
21/12/07 11:42:48 INFO Configuration.deprecation: io.sort.mb is deprecated. Instead, use mapreduce.task.io.sort.mb
21/12/07 11:42:48 INFO Configuration.deprecation: io.sort.factor is deprecated. Instead, use mapreduce.task.io.sort.factor
21/12/07 11:42:48 INFO tools.DistCp: Number of paths in the copy list: 1
21/12/07 11:42:48 INFO tools.DistCp: Number of paths in the copy list: 1
21/12/07 11:42:48 INFO client.RMPProxy: Connecting to ResourceManager at ip-10-0-0-61.ec2.internal/10.0.0.61:8032
```

## Viewing the data

hadoop fs -cat /hive02/2019-Oct.csv | head

```
[hadoop@ip-10-0-0-61 ~]$ hadoop fs -cat /hive02/2019-Oct.csv | head
event_time,event_type,product_id,category_id,category_code,brand,price,user_id,user_session
2019-10-01 00:00:00 UTC,cart,5773203,1487580005134238553,,runail,2.62,463240011,26dd6e6e-4dac-4778-8d2c-92e149dab885
2019-10-01 00:00:03 UTC,cart,5773353,1487580005134238553,,runail,2.62,463240011,26dd6e6e-4dac-4778-8d2c-92e149dab885
2019-10-01 00:00:07 UTC,cart,5881589,2151191071051219817,,lovely,13.48,429681830,49e8d843-adf3-428b-a2c3-fe8bc6a307c9
2019-10-01 00:00:07 UTC,cart,5723490,1487580005134238553,,runail,2.62,463240011,26dd6e6e-4dac-4778-8d2c-92e149dab885
2019-10-01 00:00:15 UTC,cart,5881449,1487580013522845895,,lovely,0.56,429681830,49e8d843-adf3-428b-a2c3-fe8bc6a307c9
2019-10-01 00:00:16 UTC,cart,5857269,1487580005134238553,,runail,2.62,430174032,73deale7-664e-43f4-8b30-d32b9d5af04f
2019-10-01 00:00:19 UTC,cart,5739055,1487580008246412266,,kapous,4.75,377667011,81326ac6-daa4-4f0a-b488-fd0956a78733
2019-10-01 00:00:24 UTC,cart,5825598,1487580009445982239,,,0.56,467916806,2f5b5546-b8cb-9ee7-7ecd-8a2f569e-5466
```

hadoop fs -cat /hive02/2019-Nov.csv | head

```
[hadoop@ip-10-0-0-61 ~]$ hadoop fs -cat /hive02/2019-Nov.csv | head
event_time,event_type,product_id,category_id,category_code,brand,price,user_id,user_session
2019-11-01 00:00:02 UTC,view,5802432,1487580009286598681,,,0.32,562076640,09fafd6c-6c99-46b1-834f-33527f4de241
2019-11-01 00:00:09 UTC,cart,5844397,1487580006317032337,,,2.38,553329724,2067216c-31b5-455d-a1cc-af0575a34ffb
2019-11-01 00:00:10 UTC,view,5837166,1783999064103190764,,pnb,22.22,556138645,57ed222e-a54a-4907-9944-5a875c2d7f4f
2019-11-01 00:00:11 UTC,cart,5876812,1487580010100293687,,jessnail,3.16,564506666,186c1951-8052-4b37-adce-dd9644b1d5f7
2019-11-01 00:00:24 UTC,remove_from_cart,5826182,1487580007483048900,,,3.33,553329724,2067216c-31b5-455d-a1cc-af0575a34ffb
2019-11-01 00:00:24 UTC,remove_from_cart,5826182,1487580007483048900,,,3.33,553329724,2067216c-31b5-455d-a1cc-af0575a34ffb
2019-11-01 00:00:25 UTC,view,5856189,1487580009026551821,,runail,15.71,562076640,09fafd6c-6c99-46b1-834f-33527f4de241
```

Data are successfully loaded.

## Launch Hive

### Creating new database

```
hive> CREATE DATABASE IF NOT EXISTS hive_assignment ;
```

```
hive> SHOW DATABASES ;
```

```
hive> DESCRIBE DATABASE hive_assignment ;
```

```
hive> CREATE DATABASE IF NOT EXISTS hive_assignment ;
OK
Time taken: 0.319 seconds
hive> SHOW DATABASES ;
OK
default
hive_assignment
Time taken: 0.021 seconds, Fetched: 2 row(s)
hive> DESCRIBE DATABASE hive_assignment ;
OK
```

### Creating retail

```
hive >CREATE EXTERNAL TABLE IF NOT EXISTS retail (event_time timestamp,
event_type string,product_id string, category_id string, category_code string, brand string,
price decimal(10,3),user_session string) ROW FORMAT SERDE
'org.apache.hadoop.hive.serde2.OpenCSVSerde' WITHSERDEPROPERTIES
("separatorChar" = "," , "quoteChar" = "\"", "escapeChar" = "\\") stored as tLOCATION
'/casestudy' TBLPROPERTIES ("skip.header.line.count"="1") ;
```

```
hive > DESCRIBE retail ;
```

### Loading data into table “retail”;

```
hive> LOAD DATA INPATH '/hive02/2019-Oct.csv' INTO TABLE retail ;
```

```
hive> LOAD DATA INPATH '/hive02/2019-Nov.csv' INTO TABLE retail ;
```

```
hive> LOAD DATA INPATH '/hive02/2019-Oct.csv' INTO TABLE retail ;
Loading data to table default.retail
OK
Time taken: 1.116 seconds
hive> LOAD DATA INPATH '/hive02/2019-Nov.csv' INTO TABLE retail ;
Loading data to table default.retail
OK
```



Performing data check:

```
hive> SELECT * FROM retail WHERE MONTH(event_time)=11 limit 5 ;
```

```
hive> SELECT * FROM retail WHERE MONTH(event_time)=10 limit 5 ;
```

```
hive> SELECT * FROM retail WHERE MONTH(event_time)=11 limit 5 ;
OK
2019-11-01 00:00:02 UTC view      5802432 1487580009286598681      0.32      562076640 0
9fafd6c-6c99-46b1-834f-33527f4de241
2019-11-01 00:00:09 UTC cart      5844397 1487580006317032337      2.38      553329724 2
067216c-31b5-455d-a1cc-af0575a34ffb
2019-11-01 00:00:10 UTC view      5837166 1783999064103190764      pnb      22.22      556138645 5
7ed222e-a54a-4907-9944-5a875c2d7f4f
2019-11-01 00:00:11 UTC cart      5876812 1487580010100293687      jessnail      3.16      564
506666 186c1951-8052-4b37-adce-dd9644b1d5f7
2019-11-01 00:00:24 UTC remove_from_cart      5826182 1487580007483048900      3.3
3      553329724      2067216c-31b5-455d-a1cc-af0575a34ffb
Time taken: 2.466 seconds, Fetched: 5 row(s)
hive> SELECT * FROM retail WHERE MONTH(event_time)=10 limit 5 ;
OK
2019-10-01 00:00:00 UTC cart      5773203 1487580005134238553      runail      2.62      463240011 2
6dd6e6e-4dac-4778-8d2c-92e149dab885
2019-10-01 00:00:03 UTC cart      5773353 1487580005134238553      runail      2.62      463240011 2
6dd6e6e-4dac-4778-8d2c-92e149dab885
2019-10-01 00:00:07 UTC cart      5881589 2151191071051219817      lovely      13.48      429681830 4
9e8d843-adf3-428b-a2c3-fe8bc6a307c9
2019-10-01 00:00:07 UTC cart      5723490 1487580005134238553      runail      2.62      463240011 2
6dd6e6e-4dac-4778-8d2c-92e149dab885
```

## DYNAMIC PARTITIONING

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

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

## PARTITION TABLE 1: retail\_part\_1

```
hive> CREATE EXTERNAL TABLE IF NOT EXISTS retail_part_1 (event_time timestamp,
product_id string,category_id string, category_code string, brand string, price decimal(10,3),
user_idbigint, user_sessionstring) PARTITIONED BY(event_type string) CLUSTERED BY
(user_id) INTO 5 buckets ROW FORMATSERDE
'org.apache.hadoop.hive.serde2.OpenCSVSerde' STORED AS textfile ;
```

```
hive> set hive.exec.dynamic.partition=true;
hive> set hive.exec.dynamic.partition.mode=nonstrict;
hive> CREATE EXTERNAL TABLE IF NOT EXISTS retail_part_1 (event_time timestamp, product_id string,
> category_id string, category_code string, brand string, price decimal(10,3), user_id bigint,
user_session
> string) PARTITIONED BY(event_type string) CLUSTERED BY (user_id) INTO 5 buckets ROW FORMAT
> SERDE 'org.apache.hadoop.hive.serde2.OpenCSVSerde' STORED AS textfile ;
OK
Time taken: 0.117 seconds
hive> DESCRIBE retail_part_1 ;
```

hive> DESCRIBE retail\_part\_1 ;

```
hive> DESCRIBE retail_part_1 ;
OK
event_time          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
event_type          string
```

hive> INSERT INTO TABLE retail\_part\_1 PARTITION (event\_type) SELECT event\_time, product\_id, category\_id, category\_code, brand, price, user\_id, user\_session, event\_type FROM retail ;

```
hive> INSERT INTO TABLE retail_part_1 PARTITION (event_type) SELECT event_time, product_id,
> category_id, category_code, brand, price, user_id, user_session, event_type FROM retail ;
Query ID = hadoop_20211207121117_940d49ad-7005-46f1-9060-f77dc4e141d0
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1638876533904_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: 164.62 s
-----
Loading data to table default.retail_part_1 partition (event_type=null)

Loaded : 4/4 partitions.
    Time taken to load dynamic partitions: 0.579 seconds
    Time taken for adding to write entity : 0.002 seconds
OK
Time taken: 166.933 seconds
```

```
hive> SELECT SUM(price) FROM retail_part_1 WHERE MONTH(event_time)=10 AND event_type='purchase' ;
```

Time Taken to execute the below query is 25.038 sec.

## PARTITION TABLE 2: retail\_part\_3

Partition on : month

```
hive> CREATE EXTERNAL TABLE IF NOT EXISTS retail_part_3 (event_time timestamp, event_type string, product_id string, category_id string, category_code string, brand string, price decimal(10,3), user_id bigint, user_session string) PARTITIONED BY(month int) CLUSTERED BY (brand) INTO 5 buckets ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.OpenCSVSerde' STORED AS textfile ;
hive> DESCRIBE retail_part_3
```

```
hive > DESCRIBE retail_part_3 ;
```

```
hive> SELECT SUM(price) FROM retail_part_1 WHERE MONTH(event_time)=10 AND
> event_type='purchase' ;
Query ID = hadoop_20211207121646_afff7385-d1f1-474c-b112-20587aade2c5
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1638876533904_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: 24.00 s
-----
OK
1211538.4299998982
Time taken: 25.038 seconds, Fetched: 1 row(s)
hive> CREATE EXTERNAL TABLE IF NOT EXISTS retail_part_3 (event_time timestamp, event_type string,
> product_id string, category_id string, category_code string, brand string, price decimal(10,3)
), user_id bigint,
> user_session string) PARTITIONED BY(month int) CLUSTERED BY (brand) INTO 5 buckets ROW FORMAT
> SERDE 'org.apache.hadoop.hive.serde2.OpenCSVSerde' STORED AS textfile ;
OK
Time taken: 0.098 seconds
hive> DESCRIBE retail_part_3 ;
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
month           int
```

```
hive> INSERT INTO TABLE retail_part_3 PARTITION (month) SELECT event_time, event_type, product_id, category_id, category_code, brand, price, user_id, user_session, MONTH(CAST(REPLACE(event_time,'UTC','') AS timestamp)) FROM retail ;
```

```
hive> INSERT INTO TABLE retail_part_3 PARTITION (month) SELECT event_time, event_type, product_id,
category_id, category_code, brand, price, user_id, user_session, MONTH(CAST(REPLACE(event_time,'UTC
','')) AS timestamp)) FROM retail ;
Query ID = hadoop_20211207122626_f63b8503-fd65-49f9-8675-0e986fdfe5aa
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_1638876533904_0005)
```

	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: 188.50 s
Loading data to table default.retail_part_3 partition (month=null)
Loaded : 2/2 partitions.
Time taken to load dynamic partitions: 0.177 seconds
Time taken for adding to write entity : 0.001 seconds
OK
Time taken: 198.216 seconds
```

Executing the same query with the new table “retail\_part\_3” to check the time.

```
hive> SELECT SUM(price) FROM retail_part_3 WHERE MONTH(event_time)=10 AND
event_type='purchase' ;
```

```
hive> SELECT SUM(price) FROM retail_part_3 WHERE MONTH(event_time)=10 AND
> event_type='purchase' ;
Query ID = hadoop_20211207123028_9bde9468-53bd-4def-84fc-a0e9552cac32
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1638876533904_0005)
```

	VERTICES	MODE	STATUS	TOTAL	COMPLETED	RUNNING	PENDING	FAILED	KILLED
Map 1	.....	container	SUCCEEDED	8	8	0	0	0	0
Reducer 2	.....	container	SUCCEEDED	1	1	0	0	0	0

```
VERTICES: 02/02 [=====>>>] 100% ELAPSED TIME: 80.43 s
OK
1211538.4299999713
Time taken: 81.213 seconds, Fetched: 1 row(s)
```

Time Taken to execute the above query is 81.213 sec.

## QUESTIONS

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

```
hive> SELECT SUM(price) FROM retail_part_1 WHERE MONTH(event_time)=10 AND
AND event_type='purchase' ;
```

```
hive> SELECT SUM(price) FROM retail_part_1 WHERE MONTH(event_time)=10 AND
> event_type='purchase' ;
Query ID = hadoop_20211207123301_aea51df1-27f5-4d88-8161-cf6efb8d6245
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1638876533904_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
-----
VERTICES: 02/02  [=====>>>] 100%  ELAPSED TIME: 26.67 s
-----
OK
1211538.4299998982
Time taken: 27.358 seconds, Fetched: 1 row(s)
```

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

```
hive> SELECT MONTH(event_time), SUM(price) as sum_purchase,
COUNT(event_type) as cnt FROM retail_part_1 WHERE event_type='purchase' GROUP
BY MONTH(event_time) ;
```

```
hive> SELECT MONTH(event_time), SUM(price) as sum_purchase, COUNT(event_type) as cnt FROM retail_pa
rt_1 WHERE event_type='purchase' GROUP BY MONTH(event_time) ;
Query ID = hadoop_20211207123359_28681afd-eldf-40de-be0b-1b574ad5be2e
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1638876533904_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
-----
VERTICES: 02/02  [=====>>>] 100%  ELAPSED TIME: 24.00 s
-----
OK
10      1211538.4299998982      245624
11      1531016.8999999384      322417
Time taken: 24.765 seconds, Fetched: 2 row(s)
```

In October month, 245624 purchases generated revenue of 1211538.4299 Similarly in November month, 322417 purchases generated revenue of 1531016.8999.

Time Taken to execute the above query is 24.765 sec.

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

hive>WITH diff AS ( SELECT SUM(CASE WHEN date\_format(event\_time,'MM')=10 THEN price ELSE 0 END) AS October, SUM(CASE WHEN date\_format(event\_time,'MM')=11 THEN price ELSE 0 END) AS November FROM retail\_part\_1 WHERE date\_format(event\_time,'MM') IN (10,11) AND event\_type='purchase') SELECT October, November, (November - October) as Difference FROM diff ;

```
hive> WITH diff AS ( SELECT SUM(CASE WHEN date_format(event_time,'MM')=10 THEN price ELSE 0 END) AS
October, SUM(CASE WHEN date_format(event_time,'MM')=11 THEN price ELSE 0 END) AS
> November FROM retail_part_1 WHERE date_format(event_time,'MM') IN (10,11) AND event_type='pur
chase') SELECT October, November, (November - October) as Difference FROM diff ;
Query ID = hadoop_20211207123613_60012341-821f-4b43-a6d8-f200377b349a
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1638876533904_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

```
VERTICES: 02/02 [=====>>>] 100% ELAPSED TIME: 44.32 s
OK
1211538.429999898      1531016.8999999384      319478.4700000405
Time taken: 44.93 seconds, Fetched: 1 row(s)
```

The change in revenue generated from October to November is 319478.4700

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

hive>SELECT DISTINCT split(category\_code,'\\.')[0] AS category FROM retail\_part\_1 WHERE split(category\_code,'\\.')[0]<>" ;

```
hive> SELECT DISTINCT split(category_code,'\\.')[0] AS category FROM retail_part_1 WHERE split(cate
gory_code,'\\.')[0]<>' ' ;
Query ID = hadoop_20211207123803_1626cc69-166f-4a8a-bb55-7bf0ba2fc6fb
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1638876533904_0005)
```

	VERTICES	MODE	STATUS	TOTAL	COMPLETED	RUNNING	PENDING	FAILED	KILLED
Map 1 .....	container	SUCCEEDED	6	6	0	0	0	0	0
Reducer 2 .....	container	SUCCEEDED	5	5	0	0	0	0	0

```
VERTICES: 02/02 [=====>>>] 100% ELAPSED TIME: 71.52 s
OK
furniture
appliances
accessories
apparel
sport
stationery
Time taken: 72.253 seconds, Fetched: 6 row(s)
```

There are 6 distinct categories of products. They are Furniture, appliances, accessories, apparel,sport, and stationary.

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

```
hive>SELECT split(category_code,'\\\.')[0] AS category, COUNT(product_id) AS prd
FROM retail_part_1 GROUP BY split(category_code,'\\\.')[0] ORDER BY prd DESC ;
```

```
hive> SELECT split(category_code,'\\\.')[0] AS category, COUNT(product_id) AS prd FROM retail_part_1
> GROUP BY split(category_code,'\\\.')[0] ORDER BY prd DESC ;
Query ID = hadoop_20211207124025_93dccb7-eac7-4694-87d4-0c7f6b21d6d8
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1638876533904_0005)

-----
VERTICES      MODE      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
-----
Map 1 ..... container  SUCCEEDED    6         6         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: 71.55 s
-----
OK
      8594895
appliances      61736
stationery      26722
furniture       23604
apparel 18232
accessories     12929
sport           2
Time taken: 72.233 seconds, Fetched: 7 row(s)
```

The “Sport” category has the least number of products, whereas “appliances” have 61736 products.

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

```
SELECT brand, SUM(price) AS Sales FROM retail_part_1 WHERE brand <>" AND
event_type='purchase' GROUP BY brand ORDER BY Sales DESC LIMIT 1 ;
```

```
hive> SELECT brand, SUM(price) AS Sales FROM retail_part_1 WHERE brand <>' ' AND event_type='purchase'
> GROUP BY brand ORDER BY Sales DESC LIMIT 1 ;
Query ID = hadoop_20211207124231_4b60156d-8dee-4bd4-8ea5-a8790cdedeb4
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1638876533904_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: 22.78 s
-----
OK
runail 148297.93999999898
Time taken: 23.509 seconds, Fetched: 1 row(s)
```

Brand “runail” has the maximum sales for both months combined.

Time Taken to execute the above query is 23.509 sec.



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

```
hive>WITH monthly_diff AS ( SELECT brand, SUM(CASE WHEN date_format(event_time,'MM')=10 THEN price ELSE 0 END) AS October, SUM(CASE WHEN date_format(event_time,'MM')=11 THEN price ELSE 0 END) AS November FROM retail_part_1 WHERE event_type='purchase' GROUP BY brand) SELECT brand, October, November, (November-October) as Sales_diff FROM monthly_diff WHERE (November-October) >0 ORDER BY Sales_diff ;
```

```
hive> WITH monthly_diff AS ( SELECT brand, SUM(CASE WHEN date_format(event_time,'MM')=10 THEN price ELSE 0 END) AS October, SUM(CASE WHEN date_format(event_time,'MM')=11 THEN price ELSE 0 END) AS November FROM retail_part_1 WHERE event_type='purchase' GROUP BY brand) SELECT brand, October, November, (November-October) as Sales_diff FROM monthly_diff WHERE (November-October) >0 ORDER BY Sales_diff ;
Query ID = hadoop_20211207125054_6c5dd93c-9708-4a24-9124-2e97b14ab794
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_1638876533904_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
Reducer 3 ..... container  SUCCEEDED    1         1         0         0         0         0
-----
VERTICES: 03/03 [=====>>>] 100% ELAPSED TIME: 39.45 s
-----
OK
```

```
OK
ovale 2.54 3.1 0.56
cosima 20.229999999999997 20.93 0.7000000000000028
grace 100.91999999999999 102.61000000000001 1.6900000000000261
helloganic 0.0 3.1 3.1
skinity 8.88 12.440000000000001 3.5600000000000005
bodyton 1376.3400000000006 1380.6400000000003 4.299999999999727
moyou 5.71 10.280000000000001 4.570000000000001
neoleor 43.41 51.7 8.290000000000006
soleo 204.19999999999998 212.52999999999999 8.3300000000000098
jaguar 1102.1100000000004 1110.65 8.539999999999736
tertio 236.16 245.79999999999995 9.639999999999958
fly 17.14 27.17 10.030000000000001
rasyan 18.799999999999997 28.939999999999998 10.14
deoproce 316.84 329.17 12.330000000000041
barbie 0.0 12.39 12.39
supertan 50.369999999999999 66.509999999999999 16.14
treaclemoon 163.37 181.49000000000004 18.120000000000033
kamill 63.009999999999999 81.490000000000002 18.480000000000032
juno 0.0 21.08 21.08
veraclara 50.11 71.210000000000001 21.100000000000001
glysolid 69.729999999999999 91.59 21.860000000000014
godefroy 401.22 425.11999999999995 23.899999999999992
binacil 0.0 24.259999999999998 24.259999999999998
olixz 38.95 63.400000000000006 24.450000000000003
profepil 93.36 118.02 24.659999999999997
```



```
runail 71539.279999999 76758.65999999984 5219.380000000849
bolarus 6013.71999999999 11371.9300000000004 5358.2100000000005
cosmoprofi 8322.809999999994 14536.9900000000042 6214.1800000000048
jessnail 26287.840000000127 33345.230000000014 7057.3900000000014
strong 29196.6300000000005 38671.270000000002 9474.6400000000014
ngarden 23161.3899999999883 33566.2100000000225 10404.820000000342
lianail 5892.8399999999865 16394.239999999996 10501.399999999976
ano 35302.0300000000006 51039.750000000007 15737.7200000000067
grattol 35445.53999999993 71472.710000000341 36027.170000000348
474679.0600000175 619509.2400000119 144830.17999999435
Time taken: 48.174 seconds, Fetched: 161 row(s)
```

Total of 161 brands have increased their sales from October to November.

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

hive>SELECT user\_id, SUM(price) AS expense FROM retail\_part\_1 WHERE event\_type='purchase' GROUP BY user\_id ORDER BY expense DESC LIMIT 10 ;

```
hive> SELECT user_id, SUM(price) AS expense FROM retail_part_1 WHERE event_type='purchase'
> GROUP BY user_id ORDER BY expense DESC LIMIT 10 ;
Query ID = hadoop_20211207125317_9e360148-c302-4522-9b79-62f8336b8ae1
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1638876533904_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
Reducer 3	.....	container	SUCCEEDED	1	1	0	0	0	0

```
VERTICES: 03/03 [=====>>] 100% ELAPSED TIME: 23.77 s
OK
557790271 2715.869999999991
150318419 1645.9700000000005
562167663 1352.85
531900924 1329.45
557850743 1295.4800000000005
522130011 1185.3899999999999
561592095 1109.7000000000005
431950134 1097.5899999999997
566576008 1056.3600000000004
521347209 1040.9099999999999
Time taken: 24.447 seconds, Fetched: 10 row(s)
hive>
```

Above is the list of the top 10 users.

## ❖ CLEANING UP

Once the analysis is completed, delete the database & terminate the cluster, and stop EC2 instance.

```
hive> show DATABASES ;
OK
default
hive_assignment
Time taken: 0.013 seconds, Fetched: 2 row(s)
hive> DROP DATABASE hive_assignment ;
OK
Time taken: 0.187 seconds
hive> show DATABASES ;
OK
default
Time taken: 0.009 seconds, Fetched: 1 row(s)
```

CloneTerminateAWS CLI export

Cluster: Hive\_Case\_Study\_01   Terminated   Terminated by user request

Summary

Application user interfaces

Monitoring

Hardware

Configurations

Events


Steps

Bootstrap actions

---

**Summary**

**Configuration details**

<b>ID:</b> j-1K0RDJ40X3OT	<b>Release label:</b> emr-5.29.0
<b>Creation date:</b> 2021-12-07 16:50 (UTC+5:30)	<b>Hadoop distribution:</b> Amazon 2.8.5
<b>End date:</b> 2021-12-07 18:37 (UTC+5:30)	<b>Applications:</b> Hive 2.3.6, Pig 0.17.0, Hue 4.4.0
<b>Elapsed time:</b> 1 hour, 47 minutes	<b>Log URI:</b> s3://aws-logs-237101383683-us-east-1/elasticmapreduce/ 
<b>After last step completes:</b> Cluster waits	<b>EMRFS consistent view:</b> Disabled
<b>Termination protection:</b> Off	<b>Custom AMI ID:</b> --
<b>Tags:</b> --	
<b>Master public DNS:</b> ec2-100.25.117.147 compute-1.amazonaws.com 