Problem Statement:

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

As part of this assignment, as a big data analyst, the requirement is to extract data and gather insights from a real-life data set of an e-commerce company. For this assignment, you will be working with a public clickstream dataset of a cosmetics store.

Using the clickstream dataset, your job is to extract valuable insights which generally data engineers come up within an e-retail company.

Data Description:

Dataset link:

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

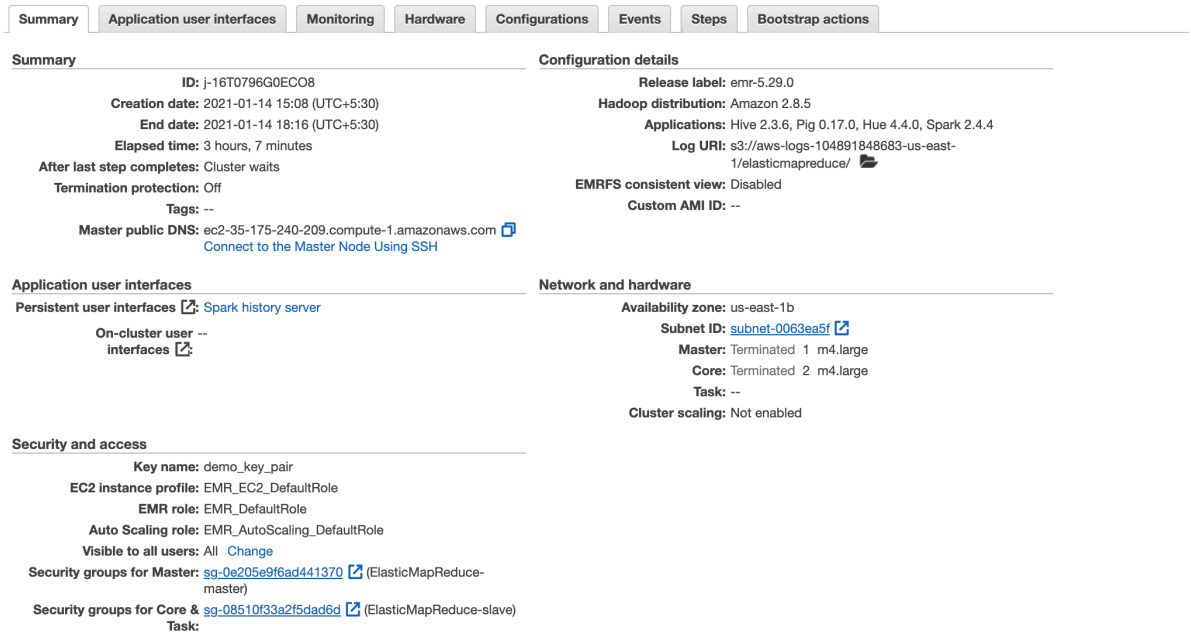
Attribute_Name	Data_type	Description
event_time	timestamp	Time at which the event took place
event_type	string	Event type may be 'view', 'cart', 'remove_from_cart', 'purchase'
product_id	string	Unique identification of the product
category_id	string	Unique identification of the product category. Each product category contains several products
category_code	string	Name (if present) of the product category
brand	string	Name of the brand
price	float	Price of the product
user_id	bigint	Permanent user id
user_session	string	Identification for the user's session. Remains same for each user's session. It changes everytime the user returns back the website after a long pause

Case Study:

1. Opening EMR Cluster

2-node EMR cluster with both the master and core nodes as M4.large

Screenshot #1



2. Launching terminal with Hadoop

create a directory in HDFS to collect the input data:
==========
hadoop fs -mkdir /user/hive/online_sales

Screenshot #2

```
https://aws.amazon.com/amazon-linux-ami/2018.03-release-notes/
43 package(s) needed for security, out of 74 available
Run "sudo yum update" to apply all updates.
-bash: warning: setlocale: LC_CTYPE: cannot change locale (UTF-8): No such file or d
irectory

EEEEEEEEEEEEEEEEEEEE MMMMMMMM           MMMMMMMM RRRRRRRRRRRRRRR
E::::::::::::::::::E M:::::::M         M:::::::M R::::::::::::::R
EE:::::EEEEEEEEE:::E M::::::::M       M::::::::M R:::::RRRRRR:::::R
  E::::E       EEEEE M:::::::::M     M:::::::::M RR::::R      R::::R
  E::::E             M::::::M:::M   M:::M::::::M   R:::R      R::::R
  E:::::EEEEEEEEEE   M:::::M M:::M M:::M M:::::M   R:::RRRRRR:::::R
  E::::::::::::::E   M:::::M  M:::M:::M  M:::::M   R::::::::::::RR
  E:::::EEEEEEEEEE   M:::::M   M:::::M   M:::::M   R:::RRRRRR:::::R
  E::::E             M:::::M    M:::M    M:::::M   R:::R      R::::R
  E::::E       EEEEE M:::::M     MMM     M:::::M   R:::R      R::::R
EE:::::EEEEEEEE::::E M:::::M             M:::::M   R:::R      R::::R
E::::::::::::::::::E M:::::M             M:::::M RR::::R      R::::R
EEEEEEEEEEEEEEEEEEEE MMMMMMM             MMMMMMM RRRRRRR      RRRRRR

[hadoop@ip-172-31-38-197 ~]$ hadoop fs -mkdir /user/hive/online_sales
mkdir: `/user/hive/online_sales': File exists
[hadoop@ip-172-31-38-197 ~]$ hadoop fs -ls /user/hive
Found 2 items
drwxr-xr-x   - hadoop hadoop          0 2021-01-10 04:44 /user/hive/online_sales
drwxrwxrwt   - hdfs   hadoop          0 2021-01-10 04:22 /user/hive/warehouse
[hadoop@ip-172-31-38-197 ~]$
```

3. Reading the files from S3

Import File from S3
==========
aws s3 ls e-commerce-events-ml

Check Directory
==========
hadoop fs -ls /user/hive

Screenshot #3

```
ankursugandhi — hadoop@ip-172-31-38-197:~ — ssh -i/Users/ankursugandhi/Do...

  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
EEEEEEEEEEEEEEEEEEEE MMMMMMM             MMMMMMM RRRRRRR      RRRRRR

[hadoop@ip-172-31-38-197 ~]$ hadoop fs -mkdir /user/hive/online_sales
[hadoop@ip-172-31-38-197 ~]$ aws s3 ls e-commerce-events-ml
2020-03-17 11:47:09  545839412 2019-Nov.csv
2020-03-17 11:37:31  482542278 2019-Oct.csv
[hadoop@ip-172-31-38-197 ~]$
```

4. Copying the dataset from S3 to HDFS

To move datasets from S3 to HDFS:
==========
hadoop distcp 's3://e-commerce-events-ml/*' '/user/hive/online_sales/'

Screenshot #4

```
ankursugandhi — hadoop@ip-172-31-38-197:~ — ssh -i/Users/ankursugandhi/Downloads/demo_key_pair.pem hadoop@ec2-18-232-106-66.compute-1.amazonaws....
drwxrwxrwt   - hdfs   hadoop          0 2021-01-10 04:22 /user/hive/warehouse
[hadoop@ip-172-31-38-197 ~]$ aws s3 ls e-commerce-events-ml
2020-03-17 11:47:09  545839412 2019-Nov.csv
2020-03-17 11:37:31  482542278 2019-Oct.csv
[hadoop@ip-172-31-38-197 ~]$ hadoop distcp 's3://e-commerce-events-ml/*' '/user/hive/online_sales'
21/01/10 06:08:49 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=[], pr
eserveRawXattrs=false, atomicWorkPath=null, logPath=null, sourceFileListing=null, sourcePaths=[s3://e-commerce-events-ml/*], targetPath=/user/hive/online_sales, ta
rgetPathExists=true, filtersFile='null'}
21/01/10 06:08:49 INFO client.RMProxy: Connecting to ResourceManager at ip-172-31-38-197.ec2.internal/172.31.38.197:8032
21/01/10 06:08:53 INFO tools.SimpleCopyListing: Paths (files+dirs) cnt = 2; dirCnt = 0
21/01/10 06:08:53 INFO tools.SimpleCopyListing: Build file listing completed.
21/01/10 06:08:53 INFO Configuration.deprecation: io.sort.mb is deprecated. Instead, use mapreduce.task.io.sort.mb
21/01/10 06:08:53 INFO Configuration.deprecation: io.sort.factor is deprecated. Instead, use mapreduce.task.io.sort.factor
21/01/10 06:08:54 INFO tools.DistCp: Number of paths in the copy list: 2
21/01/10 06:08:54 INFO tools.DistCp: Number of paths in the copy list: 2
21/01/10 06:08:54 INFO client.RMProxy: Connecting to ResourceManager at ip-172-31-38-197.ec2.internal/172.31.38.197:8032
21/01/10 06:08:54 INFO mapreduce.JobSubmitter: number of splits:2
21/01/10 06:08:54 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1610252607876_0001
21/01/10 06:08:55 INFO impl.YarnClientImpl: Submitted application application_1610252607876_0001
21/01/10 06:08:55 INFO mapreduce.Job: The url to track the job: http://ip-172-31-38-197.ec2.internal:20888/proxy/application_1610252607876_0001/
21/01/10 06:08:55 INFO tools.DistCp: DistCp job-id: job_1610252607876_0001
21/01/10 06:08:55 INFO mapreduce.Job: Running job: job_1610252607876_0001
21/01/10 06:09:05 INFO mapreduce.Job: Job job_1610252607876_0001 running in uber mode : false
21/01/10 06:09:05 INFO mapreduce.Job:  map 0% reduce 0%
21/01/10 06:09:25 INFO mapreduce.Job:  map 100% reduce 0%
21/01/10 06:09:39 INFO mapreduce.Job: Job job_1610252607876_0001 completed successfully
21/01/10 06:09:40 INFO mapreduce.Job: Counters: 38
        File System Counters
                FILE: Number of bytes read=0
                FILE: Number of bytes written=345678
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=896
                HDFS: Number of bytes written=1028381690
                HDFS: Number of read operations=26
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=8
                S3: Number of bytes read=1028381690
                S3: Number of bytes written=0
                S3: Number of read operations=0
                S3: Number of large read operations=0
                S3: Number of write operations=0
        Job Counters
                Launched map tasks=2
                Other local map tasks=2
                Total time spent by all maps in occupied slots (ms)=1977152
                Total time spent by all reduces in occupied slots (ms)=0
                Total time spent by all map tasks (ms)=61786
                Total vcore-milliseconds taken by all map tasks=61786
                Total megabyte-milliseconds taken by all map tasks=63268864
        Map-Reduce Framework
                Map input records=2
```

5. Reading the dataset

To Check the file
==========
hadoop fs -cat /user/hive/online_sales/2019-Nov.csv | head

hadoop fs -cat /user/hive/online_sales/2019-Oct.csv | head

Screenshot #5

```
ankursugandhi — hadoop@ip-172-31-38-197:~ — ssh -i/Users/ankursugandhi/Downloads/demo_key_pair.pem hadoop@ec2-1...
[hadoop@ip-172-31-38-197 ~]$ hadoop fs -ls /user/hive/online_sales
Found 2 items
-rw-r--r--   1 hadoop hadoop  545839412 2021-01-10 06:09 /user/hive/online_sales/2019-Nov.csv
-rw-r--r--   1 hadoop hadoop  482542278 2021-01-10 06:09 /user/hive/online_sales/2019-Oct.csv
[hadoop@ip-172-31-38-197 ~]$ hadoop fs -cat /user/hive/online_sales/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
2019-11-01 00:00:32 UTC,view,5837835,1933472286753424063,,,3.49,514649199,432a4e95-375c-4b40-bd36-0fc039e77580
2019-11-01 00:00:34 UTC,remove_from_cart,5870838,1487580007675986893,,milv,0.79,429913900,2f0bff3c-252f-4fe6-afcd-5d8a6a92839a
cat: Unable to write to output stream.
[hadoop@ip-172-31-38-197 ~]$ hadoop fs -cat /user/hive/online_sales/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,2.62,430174032,73dea1e7-664e-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-84276f8ef486
2019-10-01 00:00:25 UTC,cart,5698989,1487580006317032337,,,1.27,385985999,d30965e8-1101-44ab-b45d-cc1bb9fae694
cat: Unable to write to output stream.
[hadoop@ip-172-31-38-197 ~]$
```

6. Loading Hive & creating initial tables

Use Hive command
==========
hive

Create & Use database
==========
create database online_sales;

use online_sales;

Create base table to read the input data:
==========
CREATE EXTERNAL TABLE IF NOT EXISTS sales_input(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/online_sales/'
tblproperties("skip.header.line.count"="1");

Screenshot #6

```
ankursugandhi — hadoop@ip-172-31-38-197:~ — ssh -i/Users/ankursugandhi/Downloads/dem...
[hadoop@ip-172-31-38-197 ~]$ hive

Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j2.properties Async:
alse
hive> create database online_sales;
OK
Time taken: 0.93 seconds
hive> use online_sales;
OK
Time taken: 0.098 seconds
```

Screenshot #7

```
ankursugandhi — hadoop@ip-172-31-38-197:~ — ssh -i/Users/ankursugandhi/Downloads/demo_key_pair.pem hadoop@ec2-18-232-106-66.compute-1.ama...
OK
2019-11-01 00:00:02 UTC view    5802432 1487580009286598681                     0.32    562076640       09fafd6c-6c99-46b1-834f-33527f4de241
2019-11-01 00:00:09 UTC cart    5844397 1487580006317032337                     2.38    553329724       2067216c-31b5-455d-a1cc-af0575a34ffb
2019-11-01 00:00:10 UTC view    5837166 1783999064103190764             pnb     22.22   556138645       57ed222e-a54a-4907-9944-5a875c2d7f4f
2019-11-01 00:00:11 UTC cart    5876812 1487580010100293687             jessnai3.16     564506666       186c1951-8052-4b37-adce-dd9644b1d5f7
2019-11-01 00:00:24 UTC remove_from_cart        5826182 1487580007483048900     3.33    553329724       2067216c-31b5-455d-a1cc-af0575a34ffb
Time taken: 2.184 seconds, Fetched: 5 row(s)
hive> CREATE EXTERNAL TABLE IF NOT EXISTS sales_input(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/online_sales/'
    > tblproperties("skip.header.line.count"="1");
OK
Time taken: 0.036 seconds
hive> select * from sales_input limit 5;
OK
2019-11-01 00:00:02 UTC view    5802432 1487580009286598681                     0.32    562076640       09fafd6c-6c99-46b1-834f-33527f4de241
2019-11-01 00:00:09 UTC cart    5844397 1487580006317032337                     2.38    553329724       2067216c-31b5-455d-a1cc-af0575a34ffb
2019-11-01 00:00:10 UTC view    5837166 1783999064103190764             pnb     22.22   556138645       57ed222e-a54a-4907-9944-5a875c2d7f4f
2019-11-01 00:00:11 UTC cart    5876812 1487580010100293687             jessnail        3.16    564506666       186c1951-8052-4b37-adce-dd9644b1d5f7
2019-11-01 00:00:24 UTC remove_from_cart        5826182 1487580007483048900             3.33    553329724       2067216c-31b5-455d-a1cc-af0575a34ff
b
Time taken: 0.166 seconds, Fetched: 5 row(s)
```

7. Enable Partitioning & Bucketing

To enable partitioning and bucketing:
==========
set hive.exec.dynamic.partition.mode=nonstrict;
set hive.exec.dynamic.partition=true;
set hive.enforce.bucketing=true;

CREATE TABLE IF NOT EXISTS sales_bucket(event_time timestamp, product_id string, category_id string, category_code string, brand string, price float, user_id bigint, user_session string)
PARTITIONED BY (event_type string) CLUSTERED BY (price) into 10 buckets
ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.OpenCSVSerde'
STORED AS TEXTFILE;

Screenshot #8

```
Time taken: 0.308 seconds
hive> set hive.exec.dynamic.partition.mode=nonstrict;
hive> set hive.exec.dynamic.partition=true;
hive> set hive.enforce.bucketing=true;
hive> CREATE TABLE IF NOT EXISTS sales_bucket(event_time timestamp, event_type string, product_id string, category_id st
ring, category_code string, brand string, price float, user_id bigint, user_session string)
    > PARTITIONED BY (event_type string) CLUSTERED BY (price) into 10 buckets
    > ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.OpenCSVSerde'
    > STORED AS TEXTFILE;
FAILED: SemanticException [Error 10035]: Column repeated in partitioning columns
hive> CREATE TABLE IF NOT EXISTS sales_bucket(event_time timestamp, product_id string, category_id string, category_code
 string, brand string, price float, user_id bigint, user_session string)
    > PARTITIONED BY (event_type string) CLUSTERED BY (price) into 10 buckets
    > ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.OpenCSVSerde'
    > STORED AS TEXTFILE;
OK
Time taken: 0.134 seconds
```

8. Check improvement of the performance after using optimization

To load the optimized hive table:
==========
insert into table sales_bucket partition(event_type) select event_time, event_type, product_id, category_id, category_code, brand, price, user_id, user_session from sales_input;

Performance comparison:
==========
select * from sales_input where event_type = "cart" and price > 300;

Screenshot #9

```
hive> select * from sales_input where event_type = "cart" and price >
300;
OK
2019-11-21 15:08:13 UTC     cart  5906221     1487580006300255120
      strong      327.78      533560354  ab82d077-c051-44f2-9eee-
f934f0ef88bc
2019-11-22 06:30:06 UTC     cart  5906221     1487580006300255120
      strong      327.78      538559947  0afa2172-e3b5-4b97-b7a8-
95854deb878b
2019-11-22 06:30:07 UTC     cart  5906221     1487580006300255120
      strong      327.78      538559947  0afa2172-e3b5-4b97-b7a8-
95854deb878b
2019-11-24 09:08:14 UTC     cart  5906221     1487580006300255120
      strong      311.38      575599928  ca25f0e2-c099-46c2-b623-
b2d109343156
2019-11-24 09:08:27 UTC     cart  5906221     1487580006300255120
      strong      311.38      575599928  ca25f0e2-c099-46c2-b623-
b2d109343156
2019-11-25 07:17:45 UTC     cart  5906221     1487580006300255120
      strong      327.78      441013094  32433428-1248-42a7-a0ca-
211b769e1d33
2019-11-25 09:21:15 UTC     cart  5906221     1487580006300255120
      strong      327.78      552639549  98e8cfd3-5126-4743-b6d1-
eb7fc8447a5f
2019-11-25 17:47:12 UTC     cart  5906221     1487580006300255120
      strong      327.78      320413035  0d29f65b-2264-4e13-82b0-
29fcc4775e6a
2019-11-25 18:53:43 UTC     cart  5906221     1487580006300255120
      strong      327.78      387151543  6177736c-cae5-425f-9b6d-
cdf6297849ff
2019-11-26 18:35:01 UTC     cart  5906221     1487580006300255120
      strong      327.78      501199648  91e228f1-45a8-4e4d-b60d-
6f278da0489d
2019-11-27 16:22:02 UTC     cart  5906221     1487580006300255120
      strong      327.78      577714472  5eaaf44b-8cd2-43ce-9ae7-
4ead2cf11229
```

Screenshot #10

```
2019-11-28 07:25:32 UTC     cart  5906221     1487580006300255120
      strong      311.38      536341436  d9df4eae-728d-41df-9b48-
c10995b2ec3d
2019-11-28 07:25:37 UTC     cart  5906221     1487580006300255120
      strong      311.38      536341436  d9df4eae-728d-41df-9b48-
c10995b2ec3d
2019-11-28 16:29:37 UTC     cart  5906221     1487580006300255120
      strong      311.38      544372170  b9ac3e41-78ec-49a8-b75b-
e1921d3b59ce
2019-11-28 19:35:39 UTC     cart  5906221     1487580006300255120
      strong      311.38      475738962  5ae37834-5e8b-4df2-8469-
ebc4f5de0229
2019-11-28 19:35:42 UTC     cart  5906221     1487580006300255120
      strong      311.38      475738962  5ae37834-5e8b-4df2-8469-
ebc4f5de0229
2019-11-30 00:37:34 UTC     cart  5906221     1487580006300255120
      strong      311.38      579340056  200c2f4e-e924-4474-8092-
5d8d8ac11647
Time taken: 0.908 seconds, Fetched: 17 row(s)
```

select * from sales_bucket where event_type = "cart" and price > 300;

Screenshot #11

```
hive> select * from sales_bucket where event_type = "cart" and price >
300;
OK
2019-11-28 07:25:32 UTC     5906221     1487580006300255120
      strong      311.38      536341436   d9df4eae-728d-41df-9b48-
c10995b2ec3d      cart
2019-11-28 07:25:37 UTC     5906221     1487580006300255120
      strong      311.38      536341436   d9df4eae-728d-41df-9b48-
c10995b2ec3d      cart
2019-11-28 16:29:37 UTC     5906221     1487580006300255120
      strong      311.38      544372170   b9ac3e41-78ec-49a8-b75b-
e1921d3b59ce      cart
2019-11-28 19:35:39 UTC     5906221     1487580006300255120
      strong      311.38      475738962   5ae37834-5e8b-4df2-8469-
ebc4f5de0229      cart
2019-11-28 19:35:42 UTC     5906221     1487580006300255120
      strong      311.38      475738962   5ae37834-5e8b-4df2-8469-
ebc4f5de0229      cart
2019-11-30 00:37:34 UTC     5906221     1487580006300255120
      strong      311.38      579340056   200c2f4e-e924-4474-8092-
5d8d8ac11647      cart
2019-11-24 09:08:14 UTC     5906221     1487580006300255120
      strong      311.38      575599928   ca25f0e2-c099-46c2-b623-
b2d109343156      cart
2019-11-24 09:08:27 UTC     5906221     1487580006300255120
      strong      311.38      575599928   ca25f0e2-c099-46c2-b623-
b2d109343156      cart
2019-11-21 15:08:13 UTC     5906221     1487580006300255120
      strong      327.78      533560354   ab82d077-c051-44f2-9eee-
f934f0ef88bc      cart
2019-11-22 06:30:06 UTC     5906221     1487580006300255120
      strong      327.78      538559947   0afa2172-e3b5-4b97-b7a8-
95854deb878b      cart
2019-11-22 06:30:07 UTC     5906221     1487580006300255120
      strong      327.78      538559947   0afa2172-e3b5-4b97-b7a8-
95854deb878b      cart
```

Screenshot #12

```
2019-11-25 18:53:43 UTC     5906221     1487580006300255120
      strong      327.78      387151543   6177736c-cae5-425f-9b6d-
cdf6297849ff      cart
2019-11-25 09:21:15 UTC     5906221     1487580006300255120
      strong      327.78      552639549   98e8cfd3-5126-4743-b6d1-
eb7fc8447a5f      cart
2019-11-25 17:47:12 UTC     5906221     1487580006300255120
      strong      327.78      320413035   0d29f65b-2264-4e13-82b0-
29fcc4775e6a      cart
2019-11-26 18:35:01 UTC     5906221     1487580006300255120
      strong      327.78      501199648   91e228f1-45a8-4e4d-b60d-
6f278da0489d      cart
2019-11-27 16:22:02 UTC     5906221     1487580006300255120
      strong      327.78      577714472   5eaaf44b-8cd2-43ce-9ae7-
4ead2cf11229      cart
2019-11-25 07:17:45 UTC     5906221     1487580006300255120
      strong      327.78      441013094   32433428-1248-42a7-a0ca-
211b769e1d33      cart
Time taken: 0.603 seconds, Fetched: 17 row(s)
```

9. Solutions to Questions asked

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

select sum(price) from sales_bucket where event_type = 'purchase' and month(event_time) = 10;

Output:
1211538.430000433

Screenshot #13

```
hive> select sum(price) from sales_bucket where event_type =
'purchase' and  month(event_time) = 10;
Query ID = hadoop_20210114111114_9a6a5a88-e861-408f-ba31-c5daa4ad8c93
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id
application_1610617611610_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:
16.33 s
--------------------------------------------------------------------------------
OK
1211538.430000433
Time taken: 16.993 seconds, Fetched: 1 row(s)
```

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

select sum(price) from sales_bucket where event_type = 'purchase' group by month(event_time);

Output:
1211538.430000433
1531016.9000001205

Screenshot #14

```
hive> select sum(price) from sales_bucket where event_type =
'purchase' group by month(event_time);
Query ID = hadoop_20210114111454_be1057dd-40b5-4b45-a9dc-840524ccbe2c
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id
application_1610617611610_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:
16.92 s
--------------------------------------------------------------------------------
OK
1211538.430000433
1531016.9000001205
Time taken: 17.669 seconds, Fetched: 2 row(s)
```

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

select Oct, Nov, Nov - Oct Difference

```
from
(
SELECT sum(case when date_format(event_time,'MM')=10 then price else 0 end) AS Oct,
     sum(case when date_format(event_time,'MM')=11 then price else 0 end) AS Nov
 FROM sales_bucket WHERE date_format(event_time,'MM')in (10,11) AND event_type='purchase'
)s;
```

Output:
1211538.430000433 1531016.9000001205 319478.4699996875

Screenshot #15

```
hive> select October, November, November - October Difference
    > from
    > (
    > 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 sales_bucket WHERE date_format(event_time,'MM')in (10,11)
AND event_type='purchase'
    > )s;
Query ID = hadoop_20210114112021_102ea4a8-4f3b-419b-a56e-9faf40cf1dd3
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_1610617611610_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:
37.96 s
--------------------------------------------------------------------------------
------------------------
OK
1211538.430000433     1531016.9000001205    319478.4699996875
Time taken: 46.334 seconds, Fetched: 1 row(s)
```

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

```
select distinct category_code from sales_bucket;
```

Output:

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

Screenshot #15

```
hive> select distinct category_code from sales_bucket;
Query ID = hadoop_20210114113334_eb6359c7-b4c4-4b49-bd66-bef6997d2f2d
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_1610617611610_0006)

--------------------------------------------------------------------
--------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING
PENDING  FAILED  KILLED
--------------------------------------------------------------------
--------------------------
Map 1 .......... container     SUCCEEDED     14         14        0
0       0       0
Reducer 2 ...... container     SUCCEEDED      5          5        0
0       0       0
--------------------------------------------------------------------
--------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME:
47.83 s
--------------------------------------------------------------------
--------------------------
OK

accessories.cosmetic_bag
stationery.cartrige
accessories.bag
appliances.environment.vacuum
furniture.living_room.chair
sport.diving
appliances.personal.hair_cutter
appliances.environment.air_conditioner
apparel.glove
furniture.bathroom.bath
furniture.living_room.cabinet
Time taken: 56.08 seconds, Fetched: 12 row(s)
```

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

select category_id, category_code, count(product_id) from sales_bucket group by category_code;

Screenshot #16

```
hive> select category_id, category_code, count(product_id) from
sales_bucket group by category_code;
FAILED: SemanticException [Error 10025]: Line 1:7 Expression not in
GROUP BY key 'category_id'
hive> select category_id, count(product_id) from sales_bucket group
by category_id;
Query ID = hadoop_20210114120759_924b77d8-5637-40cc-b414-1740f2d6b66a
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id
application_1610617611610_0007)

--------------------------------------------------------------------
--------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING
PENDING  FAILED  KILLED
--------------------------------------------------------------------
--------------------------
Map 1 .......... container     SUCCEEDED     14         14        0
0       0       0
Reducer 2 ...... container     SUCCEEDED      5          5        0
0       0       0
--------------------------------------------------------------------
--------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME:
50.40 s
--------------------------------------------------------------------
--------------------------
OK
1487580004966466385   16
1487580005050352469   83278
1487580005176181595   127
1487580005369119587   3
1487580005570446188   24
1487580005671109489   300570
1487580005687886706   14
1487580005922767741   640
1487580006056985476   1794
1487580006073762693   7556
1487580006174425994   466
1487580006216369036   3
```

Screenshot #17

```
1982860263572898112   6507
1998040849203594085   5995
2007399943458784057   18070
2035665444290953519   7792
2069171133327868014   2028
2093602042093240877   3188
2106514244487873093   1472
2134354356349173879   257
2141560642253881670   12861
2166295400451933025   11
2193074740493550411   1749
2193074740552270669   13772
2195085258339123402   25
Time taken: 51.039 seconds, Fetched: 500 row(s)
```

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

select brand, sum(price) as pr from sales_bucket where event_type = 'purchase' group by brand order by pr desc limit 2;

Output:

1094188.3000002217
148297.94000001193

Screenshot #18

```
hive> select brand, sum(price) as pr from sales_bucket group by brand
order by pr desc;
Query ID = hadoop_20210114122352_90746d61-ed2f-4237-85de-86e05f578b6f
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_1610617611610_0008)

--------------------------------------------------------------------------
-------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING
PENDING  FAILED  KILLED
--------------------------------------------------------------------------
-------------------------
Map 1 .......... container     SUCCEEDED     14         14        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:
46.30 s
--------------------------------------------------------------------------
-------------------------
OK
     2.6194508600006673E7
strong     4927445.599999605
jessnail   3905094.1099998252
runail     3838847.3299999256
irisk 2660064.559999657
```

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

select brand, sum(price) from sales_bucket where event_type = 'purchase' and (month(event_time) = 10) < (month(event_time) = 11) group by brand;

Screenshot #18

```
hive> select brand, sum(price) from sales_bucket where event_type =
'purchase' and  (month(event_time) = 10) < (month(event_time) = 11)
group by brand;
Query ID = hadoop_20210114124027_48cca2be-a43e-472c-a5ea-5775535e6bbd
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_1610617611610_0009)

--------------------------------------------------------------------
-------------------------
        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:
22.76 s
--------------------------------------------------------------------
-------------------------
OK
      619509.2399999945
airnails   5691.5200000000095
almea 973.8700000000003
ardell     843.6500000000003
art-visage 2997.8000000000056
artex 4327.249999999993
aura  177.50999999999996
balbcare   212.3799999999997
barbie     12.39
batiste    874.1699999999998
beautix    12222.949999999979
beauty-free      1782.8600000000104
beautyblender    109.40999999999998
```

Output continued in the next page

Screenshot #19

```
sawa  45.5
severina   6120.479999999983
shary 1176.4899999999993
shik  4839.720000000007
siberina   337.6499999999999
skinity    12.440000000000001
skinlite   890.4500000000008
skipofit   8.49
smart 5902.140000000007
soleo 212.52999999999966
solomeya   2685.8000000000025
sophin     1515.5200000000013
staleks    11875.61000000004
strong     38671.26999999998
sun   65.9
sunuv 8042.150000000003
supertan   66.50999999999999
swarovski  3043.1600000000053
tannymaxx  171.28
tazol 7.18
tertio     245.79999999999998
thuya 2604.9400000000005
tosowoong  27.3
treaclemoon      181.49
trind 542.9600000000002
uno   51039.749999999956
uskusi     5690.31000000004
veraclara  71.21
vilenta    231.20999999999998
vosev 316.7
weaver     6.48
yoko  11707.879999999994
ypsed 436.32
yu-r  673.7099999999999
zeitun     2009.6299999999994
zinger     6684.86000000003
Time taken: 32.449 seconds, Fetched: 214 row(s)
```

Screenshots #20 to 23

```
bergamo     144.3
bespecial   70.5
binacil     24.259999999999998
bioaqua     1398.1200000000001
biofollica 257.93999999999994
biore 90.30999999999999
blixz 63.4
bluesky     10565.529999999775
bodyton     1380.6399999999999
bpw.style   14837.440000000077
browxenna   14916.730000000072
candy 799.3799999999997
carmex      243.35999999999999
chi   538.61
cnd   29166.589999999946
coifin      1428.4899999999998
concept     13380.399999999938
consly      153.85000000000002
coocla      133.0
cosima      20.929999999999996
cosmoprofi 14536.990000000053
coxir 185.05
cristalinas      584.9499999999999
cruset      145.28
cutrin      367.62
de.lux      2775.509999999994
deoproce    329.16999999999996
depilflax   2803.779999999999
dermal      257.08
dewal 61.29
dizao 945.5099999999994
domix 12009.169999999984
dr.gloderm 11.07
ecocraft    241.95000000000005
ecolab      1214.3000000000002
egomania    146.04000000000002
elizavecca 204.3
ellips      606.04
elskin      307.6500000000001
emil  4098.82
enas  14.1
```

```
enas  14.1
enigma      224.85000000000002
enjoy 136.57
entity      719.2599999999996
eos   152.61
estel 24142.66999999991
estelare    471.87
eunyul      234.15000000000026
f.o.x 8577.279999999999
fancy 50.620000000000005
farmavita   1291.9699999999996
farmona     1843.4299999999996
farmstay    1074.0599999999997
fedua 263.81
finish      230.38
fly   27.17
foamie      80.49
freedecor   7671.799999999991
freshbubble      502.3399999999998
frozen      12.18
gehwol      1557.6800000000003
glysolid    91.58999999999999
godefroy    425.12
grace 102.60999999999999
grattol     71472.7100000016
greymy      489.48999999999984
happyfons   1091.5900000000006
haruyama    12352.910000000102
helloganic 3.1
i-laq 366.71999999999986
igrobeauty 645.0700000000004
ingarden    33566.210000000065
inm   351.2099999999999
inoface     70.47
insight     1721.9600000000005
irisk 46946.03999999892
italwax     24799.369999999755
jaguar      1110.65
jas   3657.430000000002
jessnail    33345.230000000425
joico 2015.1
```

```
joico 2015.1
juno  21.08
kaaral      5086.070000000007
kamill      81.49000000000001
kapous      14093.080000000027
kares 59.45
kaypro      3268.699999999999
keen  435.62
kerasys     525.2
keune 375.1
kims  632.0400000000001
kinetics    6945.260000000038
kiss  817.3300000000002
kocostar    594.9300000000001
koelcia     112.75
koelf 507.28999999999985
konad 810.6699999999987
koreatida   46.11
kosmekka    1813.37
labay 41.78
laboratorium      312.52
lador 2471.5300000000007
ladykin     170.57
lakme 602.19
lamixx      672.2600000000001
latinoil    384.59000000000003
lebelage    218.27000000000004
levissime   3085.3100000000027
levrana     3664.100000000002
lianail     16394.240000000096
likato      340.97
limoni      1796.6000000000004
litaline    135.74
lovely      11939.059999999932
lowence     567.75
lsanic      959.3399999999999
mane  260.26
marathon    10273.099999999999
markell     2834.430000000003
marutaka-foot    109.33000000000001
masura      33058.469999999696
matreshka   182.67000000000002
```

```
matrix      3726.7400000000002
mavala      446.32000000000005
max   8664.230000000001
meisterwerk      340.01000000000005
metzger     6457.160000000007
milv  5642.009999999955
miskin      293.06999999999994
missha      2150.2799999999997
moyou 10.280000000000001
nagaraku    5327.68
naomi 389.0
naturmed    67.31
nefertiti   366.64
neoleor     51.7
nirvel      234.32999999999998
nitrile     1162.679999999999
nitrimax    1809.6400000000035
oniq  9841.650000000023
orly  931.09
osmo  762.31
ovale 3.1
parachute   307.30999999999995
petitfee    864.7999999999997
philips     6.86
plazan      194.01
pnb   6372.489999999998
polarus     11371.930000000008
pole  5527.239999999981
profepil    118.02000000000001
profhenna   736.8499999999996
protokeratin      456.7900000000001
provoc      1063.8200000000013
rasyan      28.939999999999998
refectocil 3475.5800000000063
riche 202.41000000000003
rocknailstar      1.9
rosi  3841.560000000005
roubloff    4913.770000000014
runail      76758.65999999491
s.care      913.0699999999999
sanoto      1209.6799999999998
sawa  45.5
```

viii. 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

select user_id, sum(price) as pr from sales_bucket where event_type = 'purchase'
group by user_id order by pr desc limit 10;

Screenshot #24

```
hive> select user_id, sum(price) as pr from sales_bucket where
event_type = 'purchase' group by user_id order by pr desc limit 10;
Query ID = hadoop_20210114123135_fa96e2d3-815d-47f2-9f2c-72d566090a4e
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id
application_1610617611610_0008)

--------------------------------------------------------------------
--------------------------
        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:
17.54 s
--------------------------------------------------------------------
--------------------------
OK
557790271  2715.87
150318419  1645.9699999999998
562167663  1352.85
531900924  1329.4499999999998
557850743  1295.48
522130011  1185.3899999999996
561592095  1109.7000000000003
431950134  1097.5899999999995
566576008  1056.3600000000006
521347209  1040.9099999999996
Time taken: 18.264 seconds, Fetched: 10 row(s)
```

== End-of-document==