HIVE CASE STUDY – DA TRACK

Submitted By:

Shyam Dalsaniya

Iranna Chatti

PROBLEM STATEMENT:

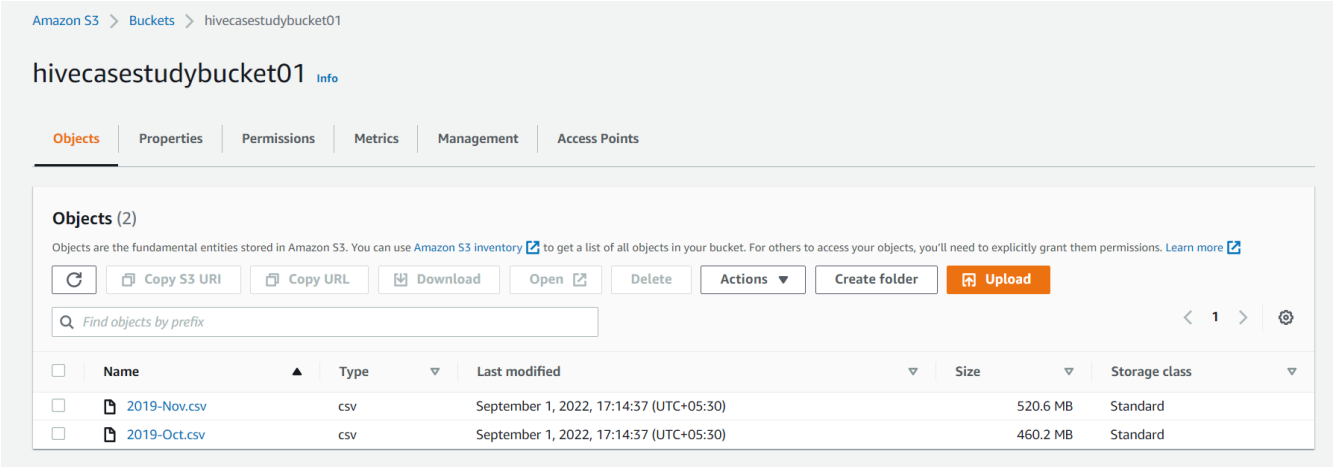
With online sales gaining popularity, tech companies are exploring ways to improve their sales by analysing customer behaviour and gaining insights about product trends. Furthermore, the websites make it easier for customers to find the products they require without much scavenging. Needless to say, the role of big data analysts is among the most sought-after job profiles of this decade. Therefore, as part of this assignment, we will be challenging you, as a big data analyst, to extract data and gather insights from a real-life data set of an e-commerce company.

The implementation phase can be divided into the following parts:

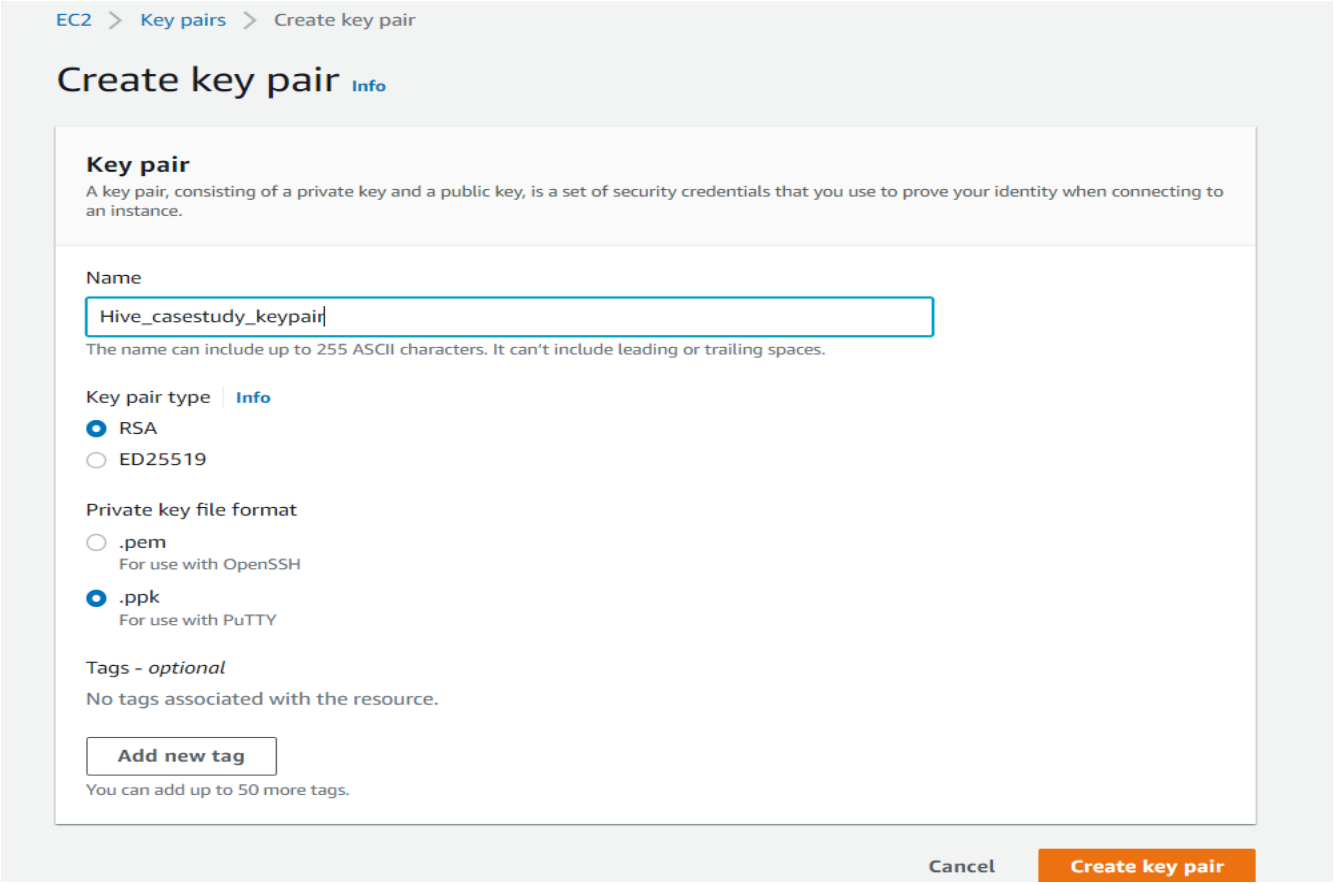
- Copying the data set into the HDFS:
  - Launch an EMR cluster that utilizes the Hive services, and
  - Move the data from the S3 bucket into the HDFS
- Creating the database and launching Hive queries on your EMR cluster:
  - Create the structure of your database,
  - Use optimized techniques to run 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

Copying the data set into the HDFS:

Uploaded both the Datasets to a S3 Bucket:



Creating a Key Pair for the EMR Cluster:



Launching an EMR cluster that utilizes the Hive services:

Amazon EMR

- EMR Studio
- EMR Serverless New

EMR on EC2

- Clusters
- Notebooks
  - Git repositories
- Security configurations
- Block public access
- VPC subnets
- Events

EMR on EKS

- Virtual clusters

- Help
- What's new

Clone | Terminate | AWS CLI export

Auto-termination is not available for this account when using this release of EMR.

Cluster: Hive_casestudy_cluster Waiting Cluster ready after last step completed.

Summary | Application user interfaces | Monitoring | Hardware | Configurations | Events | Steps | Bootstrap actions

Summary

ID: j-ERBVL6WS368M
Creation date: 2022-09-02 10:29 (UTC+5:30)
Elapsed time: 11 minutes
After last step completes: Cluster waits
Termination protection: On Change
Tags: -- View All / Edit
Master public DNS: ec2-50-17-119-221.compute-1.amazonaws.com
Connect to the Master Node Using SSH

Configuration details

Release label: emr-5.29.0
Hadoop distribution: Amazon 2.8.5
Applications: Hive 2.3.6, Pig 0.17.0, Hue 4.4.0
Log URI: s3://aws-logs-484340426713-us-east-1/elasticmapreduce/
EMRFS consistent view: Disabled
Custom AMI ID: --

Application user interfaces

Persistent user interfaces: --
On-cluster user interfaces: Not Enabled Enable an SSH Connection

Network and hardware

Availability zone: us-east-1d
Subnet ID: subnet-08fc83c6cb79f4c8d
Master: Running 1 m4.large
Core: Running 1 m4.large
Task: --
Cluster scaling: Not enabled

Security and access

Key name: Hive_casestudy_keypair
EC2 instance profile: EMR_EC2_DefaultRole
EMR role: EMR_DefaultRole
Auto Scaling role: EMR_AutoScaling_DefaultRole

Launched a EMR cluster and successfully connected to Putty:

```
login as: hadoop
Authenticating with public key "Hive_casestudy_keypair"
Last login: Fri Sep  2 05:09:11 2022

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

https://aws.amazon.com/amazon-linux-ami/2018.03-release-notes/
69 package(s) needed for security, out of 97 available
Run "sudo yum update" to apply all updates.

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-172-31-46-4 ~]$
```

Making a new directory in HDFS:

```
[hadoop@ip-172-31-46-4 ~]$ hadoop fs -mkdir /hive_casestudy-folder
[hadoop@ip-172-31-46-4 ~]$
```

Moving the data from the S3 bucket into the HDFS:

```
[hadoop@ip-172-31-46-4 ~]$ hadoop distcp s3://hivecasestudybucket01/2019-Oct.csv /hive_casestudy-folder/2019-Oct.csv
```

```
[hadoop@ip-172-31-46-4 ~]$ hadoop distcp s3://hivecasestudybucket01/2019-Nov.csv /hive_casestudy-folder/2019-Nov.csv
```

Checking for both the datasets in the directory:

```
[hadoop@ip-172-31-46-4 ~]$ hadoop fs -ls /hive_casestudy-folder/
Found 2 items
-rw-r--r--   1 hadoop hadoop  545839412 2022-09-02 05:32 /hive_casestudy-folder/2019-Nov.csv
-rw-r--r--   1 hadoop hadoop  482542278 2022-09-02 05:27 /hive_casestudy-folder/2019-Oct.csv
[hadoop@ip-172-31-46-4 ~]$
```

Logging into Hive CLI:

```
[hadoop@ip-172-31-46-4 ~]$ hive

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

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

Creating a Database and using it:

```
hive> CREATE DATABASE IF NOT EXISTS Retail_Ecom_DB ;
OK
Time taken: 1.323 seconds
hive> show databases ;
OK
default
retail_ecom_db
Time taken: 0.232 seconds, Fetched: 2 row(s)
hive> USE Retail_Ecom_DB ;
OK
Time taken: 0.062 seconds
hive>
```

Creating an External table:

```
hive> CREATE EXTERNAL TABLE IF NOT EXISTS retail_info ( 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 TBLPROPERTIES("skip.header.line.count" = "1") ;
OK
Time taken: 0.124 seconds
```

Query used:

CREATE EXTERNAL TABLE IF NOT EXISTS retail_info (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 TBLPROPERTIES("skip.header.line.count" = "1");

Describing the table 'retail_info':

```
hive> describe retail_info ;
OK
event_time              string                  from deserializer
event_type              string                  from deserializer
product_id              string                  from deserializer
category_id             string                  from deserializer
category_code           string                  from deserializer
brand                   string                  from deserializer
price                   string                  from deserializer
user_id                 string                  from deserializer
user_session            string                  from deserializer
Time taken: 0.131 seconds, Fetched: 9 row(s)
hive>
```

Loading datasets into the 'retail_info' table:

```
hive> LOAD DATA INPATH '/hive_casestudy-folder/2019-Oct.csv' into table retail_info ;
Loading data to table retail_ecom_db.retail_info
OK
Time taken: 2.043 seconds
hive> LOAD DATA INPATH '/hive_casestudy-folder/2019-Nov.csv' into table retail_info ;
Loading data to table retail_ecom_db.retail_info
OK
Time taken: 1.031 seconds
hive>
```

Setting Headers to ‘true’ and checking the data:

```
hive> set hive.cli.print.header=true ;
hive> select * from retail_info limit 3 ;
OK
retail_info.event_time  retail_info.event_type  retail_info.product_id  retail_info.category_id retail_info.category_code      retail_info.brand      retail_info.price      retail_info.u
ser_id  retail_info.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
Time taken: 2.486 seconds, Fetched: 3 row(s)
hive>
```

To set Dynamic partitioning:

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

Creating an optimized table using Partition by and Bucketing:

```
hive> CREATE EXTERNAL TABLE IF NOT EXISTS part_retail_info (event_time timestamp, product_id string, category_id string, category_code string, brand string, price float, user_id bigint, use
r_session string) partitioned by (event_type string) clustered by (user_id) into 10 buckets row format SERDE 'org.apache.hadoop.hive.serde2.OpenCSVSerde' stored as textfile ;
OK
Time taken: 0.077 seconds
hive>
```

Query used:

CREATE EXTERNAL TABLE IF NOT EXISTS part_retail_info (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 (user_id) into 10 buckets row format SERDE 'org.apache.hadoop.hive.serde2.OpenCSVSerde' stored as textfile;

Describing the table ‘part_retail_info’:

```
hive> describe part_retail_info ;
OK
col_name        data_type       comment
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

# Partition Information
# col_name              data_type               comment

event_type              string
Time taken: 0.08 seconds, Fetched: 14 row(s)
hive>
```

Loading data into the optimized table ‘part_retail_info’:

```
hive> INSERT INTO part_retail_info partition (event_type) SELECT event_time, product_id, category_id, category_code, brand, price, user_id, user_session, event_type FROM retail_info ;
Query ID = hadoop_20220902093333_5150c7b9-5f73-435b-994c-fb07baf505dd
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_1662095176557_0004)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 ......... container     SUCCEEDED      2          2        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      5          5        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 169.17 s
--------------------------------------------------------------------------------
Loading data to table retail_ecom_db.part_retail_info partition (event_type=null)

Loaded : 4/4 partitions.
        Time taken to load dynamic partitions: 0.466 seconds
        Time taken for adding to write entity : 0.003 seconds
OK
event_time      product_id      category_id     category_code   brand   price   user_id user_session    event_type
Time taken: 180.291 seconds
hive>
```

Query Used:

INSERT INTO part_retail_info partition (event_type) SELECT event_time, product_id, category_id, category_code, brand, price, user_id, user_session, event_type FROM retail_info;

Checking the data in the optimized table ‘part_retail_info’:

```
hive> SELECT * from part_retail_info limit 3 ;
OK
part_retail_info.event_time     part_retail_info.product_id     part_retail_info.category_id    part_retail_info.category_code  part_retail_info.brand  part_retail_info.price  part_retail_i
nfo.user_id     part_retail_info.user_session   part_retail_info.event_type
2019-10-09 07:51:50 UTC 5760338 1487580009311764506             zinger  0.44    515249149       b3955b69-ae1d-41d1-b0a7-ebd047c7867c    cart
2019-10-09 14:22:51 UTC 5792800 1487580005268456287                     10.32   455356130       37862e24-7ecf-4fba-b267-2de27dd23f46    cart
2019-10-09 07:51:50 UTC 5760338 1487580009311764506             zinger  0.44    515249149       b3955b69-ae1d-41d1-b0a7-ebd047c7867c    cart
Time taken: 0.178 seconds, Fetched: 3 row(s)
hive>
```

Checking the partitions that were created on ‘event_type’ column in the table ‘part_retail_info’:

```
hive> show partitions part_retail_info ;
OK
partition
event_type=cart
event_type=purchase
event_type=remove_from_cart
event_type=view
Time taken: 0.086 seconds, Fetched: 4 row(s)
hive>
```

Running Hive queries to answer the questions:

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

- ❖ Comparing Query execution efficiency between ‘retail_info’ table and ‘part_retail_info’ table:
- Unoptimized table ‘retail_info’:

```
hive> SELECT SUM(price) as Oct_Revenue from retail_info where event_type='purchase' and month(event_time)=10 ;
Query ID = hadoop_20220902095458_4d165290-f50d-442d-853b-9e169027def7
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_1662095176557_0005)

--------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      2          2        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 56.15 s
--------------------------------------------------------------------------------------------
OK
oct_revenue
1211538.4299997438
Time taken: 66.265 seconds, Fetched: 1 row(s)
hive>
```

Time taken to execute = **66.26 seconds.**

- Optimized table ‘part_retail_info’:

```
hive> SELECT SUM(price) as Oct_Revenue from part_retail_info where event_type='purchase' and month(event_time)=10 ;
Query ID = hadoop_20220902095914_1235a898-163f-45a7-a602-047abeeaa0a3
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1662095176557_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: 24.65 s
--------------------------------------------------------------------------------------------------
OK
oct_revenue
1211538.4299998833
Time taken: 25.66 seconds, Fetched: 1 row(s)
hive>
```

Time taken to execute = **25.66 seconds.**

Query used:

SELECT SUM(price) as Oct_Revenue from part_retail_info where event_type='purchase' and month(event_time) =10;

ANSWER:

The total revenue generated due to purchases made in October is 1211538.4299.

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

```
hive> SELECT ROUND(SUM(price),2) as Total_Revenue, month(event_time) as Month
    > FROM part_retail_info
    > WHERE event_type = 'purchase'
    > GROUP BY month(event_time) ;
Query ID = hadoop_20220902102430_66ed41a3-dd50-4858-9266-253ae64fe85b
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_1662095176557_0006)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      3          3        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 26.51 s
--------------------------------------------------------------------------------
OK
total_revenue   month
1211538.43      10
1531016.9       11
Time taken: 35.636 seconds, Fetched: 2 row(s)
hive>
```

Query used:

SELECT ROUND(SUM(price),2) as Total_Revenue, month(event_time) as Month FROM part_retail_info WHERE event_type = 'purchase' GROUP BY month(event_time);

ANSWER:

The total sum of purchases made for month of October is 1211538.43 and for month of November is 1531016.9.

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

```
hive> WITH month_wise_sales as (
    > SELECT ROUND(SUM(CASE WHEN month(event_time)=10 THEN price ELSE 0 END),2) as oct_revenue,
    > ROUND(SUM(CASE WHEN month(event_time)=11 THEN price ELSE 0 END),2) as nov_revenue
    > FROM part_retail_info WHERE event_type = 'purchase')
    > SELECT nov_revenue - oct_revenue as change_in_revenue
    > FROM month_wise_sales ;
Query ID = hadoop_20220902111001_d3169cd0-f98f-466a-97d4-e4b63bc640db
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_1662095176557_0007)

--------------------------------------------------------------------------------
        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: 26.73 s
--------------------------------------------------------------------------------
OK
change_in_revenue
319478.47
Time taken: 35.72 seconds, Fetched: 1 row(s)
hive>
```

Query used:

WITH month_wise_sales as (

SELECT ROUND(SUM(CASE WHEN month(event_time)=10 THEN price ELSE 0 END),2) as oct_revenue,

ROUND(SUM(CASE WHEN month(event_time)=11 THEN price ELSE 0 END),2) as nov_revenue

FROM part_retail_info WHERE event_type = 'purchase')

SELECT nov_revenue - oct_revenue as change_in_revenue

FROM month_wise_sales;

ANSWER:

The change in revenue generated due to purchases from October to November is 319478.47.

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

```
hive> SELECT DISTINCT(category_code) as product_category
    > FROM part_retail_info
    > WHERE category_code IS NOT NULL AND category_code != "" ;
Query ID = hadoop_20220902114258_c63f0e3c-45f3-4134-bb47-12f3016a5193
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_1662095176557_0009)

--------------------------------------------------------------------------------------------
        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
--------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 67.13 s
--------------------------------------------------------------------------------------------
OK
product_category
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: 75.979 seconds, Fetched: 11 row(s)
hive>
```

Query used:

SELECT DISTINCT(category_code) as product_category

FROM part_retail_info

WHERE category_code IS NOT NULL AND category_code != "";

ANSWER:

There are 11 distinct categories of products.

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

ANSWER:

```
hive> SELECT COUNT(*) AS total_products, category_code
    > FROM part_retail_info
    > GROUP BY category_code ;
Query ID = hadoop_20220902130412_3e7d1af5-22bb-4684-93f3-fbbcb59ebfdb
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_1662095176557_0011)

--------------------------------------------------------------------------------
        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
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 67.26 s
--------------------------------------------------------------------------------
OK
total_products  category_code
8594895
1248    accessories.cosmetic_bag
26722   stationery.cartrige
11681   accessories.bag
59761   appliances.environment.vacuum
308     furniture.living_room.chair
2       sport.diving
1643    appliances.personal.hair_cutter
332     appliances.environment.air_conditioner
18232   apparel.glove
9857    furniture.bathroom.bath
13439   furniture.living_room.cabinet
Time taken: 76.613 seconds, Fetched: 12 row(s)
hive>
```

NOTE: There are a total of 8594895 null values in category code.

Query used:

SELECT COUNT(*) AS total_products, category_code

FROM part_retail_info

GROUP BY category_code;

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

```
hive> SELECT brand, ROUND(SUM(price),2) as Sales
    > FROM part_retail_info
    > WHERE brand != "" AND event_type='purchase'
    > GROUP BY brand
    > ORDER BY Sales DESC LIMIT 1 ;
Query ID = hadoop_20220902132636_8521e5cc-a225-4fb0-ab02-c32f4a28fb70
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_1662095176557_0012)

--------------------------------------------------------------------------------------------------
        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: 24.36 s
--------------------------------------------------------------------------------------------------
OK
brand   sales
runail  148297.94
Time taken: 33.913 seconds, Fetched: 1 row(s)
hive>
```

Query used:

SELECT brand, ROUND(SUM(price),2) as Sales

FROM part_retail_info

WHERE brand != "" AND event_type='purchase'

GROUP BY brand

ORDER BY Sales DESC LIMIT 1;

ANSWER:

The brand '**Runail**' had the maximum sales in October and November combined.

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

Query used:

WITH month_wise_sales as (SELECT brand, ROUND(SUM(CASE WHEN month(event_time)=10 THEN price ELSE 0 END),2) as oct_sales,

ROUND(SUM(CASE WHEN month(event_time)=11 THEN price ELSE 0 END),2) as nov_sales FROM part_retail_info

WHERE event_type='purchase' GROUP BY brand)

SELECT brand FROM month_wise_sales

WHERE (nov_sales - oct_sales)>0;

ANSWER:

```
hive> WITH month_wise_sales as (
    > SELECT brand, ROUND(SUM(CASE WHEN month(event_time)=10 THEN price ELSE 0 END),2) as oct_sales,
    > ROUND(SUM(CASE WHEN month(event_time)=11 THEN price ELSE 0 END),2) as nov_sales
    > FROM part_retail_info
    > WHERE event_type='purchase'
    > GROUP BY brand)
    > SELECT brand
    > FROM month_wise_sales
    > WHERE (nov_sales - oct_sales)>0 ;
Query ID = hadoop_20220902135154_b724ca93-21b9-4e63-a7c6-9465ebc75979
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_1662095176557_0013)

--------------------------------------------------------------------------------
        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: 27.03 s
--------------------------------------------------------------------------------
OK
brand

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

```
chi
coifin
concept
cosima
cosmoprofi
cristalinas
cutrin
de.lux
deoproce
depilflax
dewal
dizao
domix
ecocraft
ecolab
egomania
elizavecca
ellips
elskin
enjoy
entity
eos
estel
estelare
f.o.x
farmavita
farmona
fedua
finish
fly
foamie
freedecor
freshbubble
gehwol
glysolid
godefroy
grace
grattol
greymy
happyfons
haruyama
helloganic
igrobeauty
ingarden
inm
insight
irisk
italwax
jaguar
jas
jessnail
joico
juno
kaaral
kamill
kapous
kares
kaypro
keen
kerasys
kims
kinetics
kiss
kocostar
koelcia
koelf
konad
kosmekka
laboratorium
lador
ladykin
latinoil
levissime
levrana
lianail
likato
limoni
lovely
lowence
mane
marathon
markell
marutaka-foot
masura
matreshka
matrix
mavala
metzger
milv
miskin
missha
moyou
nagaraku
naomi
nefertiti
neoleor
nirvel
nitrile
oniq
orly
osmo
ovale
plazan
polarus
profepil
profhenna
protokeratin
provoc
rasyan
refectocil
rosi
roubloff
runail
s.care
sanoto
severina
shary
shik
skinity
skinlite
smart
soleo
solomeya
sophin
staleks
strong
supertan
swarovski
tertio
treaclemoon
trind
uno
uskusi
veraclara
vilenta
yoko
yu-r
zeitun
Time taken: 35.991 seconds, Fetched: 161 row(s)
hive>
```

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.

ANSWER:

```
hive> SELECT user_id, SUM(price) as total_spent FROM part_retail_info WHERE event_type='purchase'
    > GROUP BY user_id
    > ORDER BY total_spent DESC
    > LIMIT 10 ;
Query ID = hadoop_20220902143240_cd25de40-7ca1-4f4b-b383-b03adea4afaf
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1662095176557_0015)

--------------------------------------------------------------------------------------------
        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: 29.21 s
--------------------------------------------------------------------------------------------
OK
user_id total_spent
557790271       2715.8699999999976
150318419       1645.9700000000003
562167663       1352.8500000000001
531900924       1329.449999999999
557850743       1295.4800000000002
522130011       1185.39
561592095       1109.7000000000007
431950134       1097.589999999999
566576008       1056.3600000000004
521347209       1040.91
Time taken: 29.963 seconds, Fetched: 10 row(s)
hive>
```

Query used:

SELECT user_id, SUM(price) as total_spent

FROM part_retail_info

WHERE event_type='purchase'

GROUP BY user_id

ORDER BY total_spent DESC

LIMIT 10;

Cleaning up:

Dropping tables:

```
hive> drop table retail_info ;
OK
Time taken: 0.112 seconds
hive> drop table part_retail_info ;
OK
Time taken: 0.156 seconds
hive>
```

Dropping Database:

```
hive> drop database retail_ecom_db ;
OK
Time taken: 0.248 seconds
hive>
```

Terminating the EMR Cluster:

Cluster: Hive_casestudy_cluster Terminated Terminated by user request

Summary | Application user interfaces | Monitoring | Hardware | Configurations | Events | Steps | Bootstrap actions

Summary

ID:	j-ERBVL6WS368M
Creation date:	2022-09-02 10:29 (UTC+5:30)
End date:	2022-09-02 20:16 (UTC+5:30)
Elapsed time:	9 hours, 46 minutes
After last step completes:	Cluster waits
Termination protection:	Off
Tags:	--
Master public DNS:	ec2-50-17-119-221.compute-1.amazonaws.com Connect to the Master Node Using SSH

Configuration details

Release label:	emr-5.29.0
Hadoop distribution:	Amazon 2.8.5
Applications:	Hive 2.3.6, Pig 0.17.0, Hue 4.4.0
Log URI:	s3://aws-logs-484340426713-us-east-1/elasticmapreduce/
EMRFS consistent view:	Disabled
Custom AMI ID:	--

Application user interfaces

Persistent user interfaces:	--
On-cluster user interfaces:	--

Network and hardware

Availability zone:	us-east-1d
Subnet ID:	subnet-08fc83c6cb79f4c8d
Master:	Terminated 1 m4.large
Core:	Terminated 1 m4.large
Task:	--
Cluster scaling:	Not enabled

THANK YOU!!

Submitted By:

Shyam Dalsaniya

Iranna Chatti