HIVE Case Study (By Avinash & Anirudh)

Copying the data set into the HDFS:

- Launched EMR cluster
- Creating a folder in HDFS:

  ```
  hadoop fs –mkdir /tmp/test-folder
  ```

- Moving the data from S3 to the created folder in HDFS:

  ```
  hadoop distcp s3://aviss/2019-Nov.csv /tmp/test-folder/

  hadoop distcp s3://aviss/2019-Oct.csv /tmp/test-folder/
  ```

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

- Creating the database:

  ```
  create database if not exists case_study;

  use case_study;
  ```

  

- Creating External table with table name “sales”:

  ```
  create external table if not exists sales(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”=“,”) stored as textfile location '/tmp/test-
  folder/' tblproperties(“skip.header.line.count”=“1”);
  ```

```
hive> create external table if not exists sales(event_time timestamp, event_type string, product_id string, categor
y_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"=",") stored as textfile locatio
n '/tmp/test-folder/' tblproperties("skip.header.line.count"="1");
OK
Time taken: 0.335 seconds
hive> select * from sales limit 5;
OK
2019-11-01 00:00:02 UTC view    5802432 1487580009286598681                     0.32    562076640       09fafd6c-6c
99-46b1-834f-33527f4de241
2019-11-01 00:00:09 UTC cart    5844397 1487580006317032337                     2.38    553329724       2067216c-31
b5-455d-a1cc-af0575a34ffb
2019-11-01 00:00:10 UTC view    5837166 1783999064103190764             pnb     22.22   556138645       57ed222e-a5
4a-4907-9944-5a875c2d7f4f
2019-11-01 00:00:11 UTC cart    5876812 1487580010100293687             jessnail        3.16    564506666       186
c1951-8052-4b37-adce-dd9644b1d5f7
2019-11-01 00:00:24 UTC remove_from_cart        5826182 1487580007483048900                     3.33    553329724 2
067216c-31b5-455d-a1cc-af0575a34ffb
Time taken: 2.621 seconds, Fetched: 5 row(s)
hive> set hive.cli.print.header=true;
hive> select * from sales limit 5;
OK
sales.event_time        sales.event_type        sales.product_id        sales.category_id       sales.category_cods
ales.brand      sales.price     sales.user_id   sales.user_session
2019-11-01 00:00:02 UTC view    5802432 1487580009286598681                     0.32    562076640       09fafd6c-6c
99-46b1-834f-33527f4de241
2019-11-01 00:00:09 UTC cart    5844397 1487580006317032337                     2.38    553329724       2067216c-31
b5-455d-a1cc-af0575a34ffb
2019-11-01 00:00:10 UTC view    5837166 1783999064103190764             pnb     22.22   556138645       57ed222e-a5
4a-4907-9944-5a875c2d7f4f
2019-11-01 00:00:11 UTC cart    5876812 1487580010100293687             jessnail        3.16    564506666       186
c1951-8052-4b37-adce-dd9644b1d5f7
2019-11-01 00:00:24 UTC remove_from_cart        5826182 1487580007483048900                     3.33    553329724 2
067216c-31b5-455d-a1cc-af0575a34ffb
Time taken: 0.213 seconds, Fetched: 5 row(s)
hive>
```

- Using dynamic partitions creating two tables(event_type & category_code) for the efficiency of the queries:

  set hive.exec.dynamic.partition=true;

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

```
hive> create table if not exists dynamic_event_type (event_time string, product_id string, category_id string, cate
gory_code string, brand string, price string, user_id string, user_session string) partitioned by(event_type string
) row format delimited fields terminated by "," lines terminated by "\n";
OK
Time taken: 0.084 seconds
hive> insert into table dynamic_event_type partition(event_type) select event_time, product_id, category_id, catego
ry_code, brand, price, user_id, user_session, event_type from sales;
Query ID = hadoop_20210427192634_7c5f7446-a0c4-44e8-b6ed-db54af325e55
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_1619547757909_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: 111.90 s
--------------------------------------------------------------------------------
Loading data to table case_study.dynamic_event_type partition (event_type=null)

Loaded : 4/4 partitions.
         Time taken to load dynamic partitions: 0.78 seconds
         Time taken for adding to write entity : 0.005 seconds
OK
event_time      product_id      category_id     category_code   brand   price   user_id user_session    event_type
Time taken: 122.314 seconds
hive>
```

```
hive> create table if not exists dynamic_category (event_time string, event_type string, product_id string, categor
y_id string, brand string, price string, user_id string, user_session string) partitioned by(category_code string)
row format delimited fields terminated by "," lines terminated by "\n";
OK
Time taken: 0.162 seconds
hive> insert into table dynamic_category partition(category_code) select event_time, event_type, product_id, catego
ry_id, brand, price, user_id, user_session, category_code from sales;
Query ID = hadoop_20210427193140_38a41820-177a-4ff2-bf93-e21a0d659848
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1619547757909_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: 112.86 s
--------------------------------------------------------------------------------
Loading data to table case_study.dynamic_category partition (category_code=null)

Loaded : 12/12 partitions.
         Time taken to load dynamic partitions: 0.846 seconds
         Time taken for adding to write entity : 0.003 seconds
OK
event_time      event_type      product_id      category_id     brand   price   user_id user_session    category_co
de
Time taken: 115.145 seconds
hive>
```

- Questions:

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

   Without Optimization/not using partition:

```
hive> select sum(price) as total_revenue_oct from sales where event_type='purchase' and month(event_time)=10;
Query ID = hadoop_20210427195322_8b651930-e1a9-4d8b-8658-7e996c058e36
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_1619547757909_0006)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      2          2        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 51.53 s
--------------------------------------------------------------------------------
OK
total_revenue_oct
1211538.4299997438
Time taken: 61.054 seconds, Fetched: 1 row(s)
```

   Using Partition:

```
hive> select sum(price) as total_revenue_oct from dynamic_event_type where event_type='purchase' and month(event_ti
me)=10;
Query ID = hadoop_20210427195451_97f6a919-98b6-4846-89ae-59ac4e0ac0fa
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1619547757909_0006)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      2          2        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 14.41 s
--------------------------------------------------------------------------------
OK
total_revenue_oct
1211538.4299996954
Time taken: 15.286 seconds, Fetched: 1 row(s)
hive>
```

Note: In the above query by using dynamic partition, we saved time.

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

   Note: Using partitioned table “dynamic_event_type” in the below query.

```
hive> select month(event_time) as month ,count(event_type) as sum_purchases from dynamic_event_type where event_typ
e='purchase' group by month(event_time);
Query ID = hadoop_20210427200442_191fc1a5-cdef-4fdb-9679-9f93fab004fa
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1619547757909_0006)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      2          2        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 13.72 s
--------------------------------------------------------------------------------
OK
month   sum_purchases
10      245624
11      322417
Time taken: 14.32 seconds, Fetched: 2 row(s)
hive>
```

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

   Note: Using partitioned table “dynamic_event_type” in the below query.

```
hive> select November - October as change_in_revenue
    > from(SELECT sum(case when date_format(event_time,'MM')=10 then price else 0 end) AS October,sum(case when dat
e_format(event_time,'MM')=11 then price else 0 end) AS November FROM dynamic_event_type WHERE date_format(event_tim
e,'MM')in (10,11) AND event_type='purchase')s;
Query ID = hadoop_20210427201517_73b74cd9-7177-4e52-84df-efe2313eb3b9
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1619547757909_0007)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      2          2        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 27.91 s
--------------------------------------------------------------------------------
OK
change_in_revenue
319478.47000008775
Time taken: 28.492 seconds, Fetched: 1 row(s)
hive>
```

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

   Note: Using partitioned table “dynamic_category” in the below query.

```
hive> select distinct(category_code) as categories from dynamic_category where category_code <>'';
Query ID = hadoop_20210427202104_03ed2f18-f293-4d4e-8d81-b3f5b7d203f8
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1619547757909_0007)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      8          8        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      4          4        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 24.26 s
--------------------------------------------------------------------------------
OK
categories
__HIVE_DEFAULT_PARTITION__
accessories.bag
appliances.environment.vacuum
appliances.personal.hair_cutter
sport.diving
apparel.glove
furniture.bathroom.bath
furniture.living_room.cabinet
stationery.cartrige
accessories.cosmetic_bag
appliances.environment.air_conditioner
furniture.living_room.chair
Time taken: 24.919 seconds, Fetched: 12 row(s)
hive>
```

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

   Note: Using partitioned table “dynamic_category” in the below query.

```
hive> select category_code as category, count(product_id) as total_products from dynamic_category group by category
_code order by total_products;
Query ID = hadoop_20210427202639_86aee51b-da82-4727-a7bb-d42b0325327e
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1619547757909_0007)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      8          8        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      4          4        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 29.32 s
--------------------------------------------------------------------------------
OK
category        total_products
sport.diving    2
furniture.living_room.chair     308
appliances.environment.air_conditioner  332
accessories.cosmetic_bag        1248
appliances.personal.hair_cutter 1643
furniture.bathroom.bath 9857
accessories.bag 11681
furniture.living_room.cabinet   13439
apparel.glove   18232
stationery.cartrige     26722
appliances.environment.vacuum   59761
__HIVE_DEFAULT_PARTITION__      8594895
Time taken: 29.94 seconds, Fetched: 12 row(s)
hive>
```

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

   Note: Using partitioned table “dynamic_event_type” in the below query.

```
hive> select brand, sum(price) as total from dynamic_event_type where event_type='purchase' and brand <>'' group by
 brand order by total desc limit 1;
Query ID = hadoop_20210427203239_1cf83405-8490-47e6-83dd-5dc6226646d7
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_1619547757909_0008)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      2          2        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 15.08 s
--------------------------------------------------------------------------------
OK
brand   total
runail  148297.94000000233
Time taken: 23.997 seconds, Fetched: 1 row(s)
hive>
```

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

   Note: Using partitioned table “dynamic_event_type” in the below query.

```
hive> with brand_sales as
    > (
    > select brand,
    > sum(case when month(event_time)=10 then price else 0 end) as oct_sales ,
    > sum(case when month(event_time)=11 then price else 0 end ) as nov_sales from
    > dynamic_event_type where event_type='purchase' group by brand
    > )
    > select brand from brand_sales where nov_sales>oct_sales;
Query ID = hadoop_20210427205054_670d4506-bae7-49e5-91b3-df6198556a49
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_1619547757909_0010)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      2          2        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 18.04 s
--------------------------------------------------------------------------------
OK
brand

airnails
art-visage
artex
aura
balbcare
barbie
batiste
beautix
beauty-free
beautyblender
```

```
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
```

```
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
```

```
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
```

```
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
```

```
swarovski
tertio
treaclemoon
trind
uno
uskusi
veraclara
vilenta
yoko
yu-r
zeitun
Time taken: 27.521 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.

   Note: Using partitioned table “dynamic_event_type” in the below query.

```
hive> select user_id, round(sum(price), 0) as money_spent from dynamic_event_type where event_type='purchase' group
 by user_id order by money_spent desc limit 10;
Query ID = hadoop_20210427210635_d1b8ff62-9d79-4dad-99ad-ef83fda1131f
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1619547757909_0011)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 ..........   container     SUCCEEDED      2          2        0        0       0       0
Reducer 2 ......   container     SUCCEEDED      1          1        0        0       0       0
Reducer 3 ......   container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 15.79 s
--------------------------------------------------------------------------------
OK
user_id money_spent
557790271       2716.0
150318419       1646.0
562167663       1353.0
531900924       1329.0
557850743       1295.0
522130011       1185.0
561592095       1110.0
431950134       1098.0
566576008       1056.0
521347209       1041.0
Time taken: 16.356 seconds, Fetched: 10 row(s)
hive>
```

- Cleaning up
  Dropping the tables and database:

```
hive> drop table dynamic_event_type;
OK
Time taken: 0.206 seconds
hive> drop table dynamic_category;
OK
Time taken: 0.253 seconds
hive> drop table sales;
OK
Time taken: 0.061 seconds
hive> drop database case_study;
OK
Time taken: 0.044 seconds
hive>
```

Deleting files and directory on HDFS:

```
[hadoop@ip-172-31-15-173 ~]$ hadoop fs -ls /tmp/test-folder
Found 2 items
-rw-r--r--   1 hadoop hadoop  545839412 2021-04-27 18:38 /tmp/test-folder/2019-Nov.csv
-rw-r--r--   1 hadoop hadoop  482542278 2021-04-27 18:40 /tmp/test-folder/2019-Oct.csv
[hadoop@ip-172-31-15-173 ~]$ hadoop fs -rmdir /tmp/test-folder
rmdir: `/tmp/test-folder': Directory is not empty
[hadoop@ip-172-31-15-173 ~]$ hadoop fs -rm /tmp/test-folder/2019-Nov.csv
Deleted /tmp/test-folder/2019-Nov.csv
[hadoop@ip-172-31-15-173 ~]$ hadoop fs -rm /tmp/test-folder/2019-Oct.csv
Deleted /tmp/test-folder/2019-Oct.csv
[hadoop@ip-172-31-15-173 ~]$ hadoop fs -rmdir /tmp/test-folder
[hadoop@ip-172-31-15-173 ~]$ hadoop fs -ls /tmp
Found 2 items
drwxrwxrwx   - mapred mapred          0 2021-04-27 18:21 /tmp/hadoop-yarn
drwx-wx-wx   - hive   hadoop          0 2021-04-27 19:04 /tmp/hive
[hadoop@ip-172-31-15-173 ~]$
```

Terminating the EMR cluster.

-END-