Capstone Project – Mid submission

(Data Preparation)

1. Creation of External Tables

The following Hive statements is used to create the external tables.

a. non_event_train_brand

```
create external table non_event_train_brand
(
 device_id  string,
 gender  string,
 age  int,
 group_train string,
 phone_brand  string,
 device_model  string
)
row format delimited fields terminated by ',' lines terminated by '\n'
STORED AS TEXTFILE
LOCATION '/user/hadoop/mlctest/non_event_train_brand/';
```

Loading Data :

INSERT OVERWRITE TABLE non_event_train_brand select a., b.phone_brand, b.device_model from train a left join brand_device b on a.device_id = b.device_id ;*

b. Events Train:

```
create external table events_train
(
 event_id  int,
 device_id  string,
 event_timestamp  timestamp ,
```

longitude string,
latitude string
)
row format delimited fields terminated by ',' lines terminated by '\n'
STORED AS TEXTFILE
LOCATION '/user/hadoop/mlctest/events_train/';

```
hive> create external table events_train
    > (
    >  event_id  int,
    >  device_id  string,
    >  event_timestamp  timestamp ,
    >  longitude string,
    >  latitude string
    > )
    > row format delimited fields terminated by ',' lines terminated by '\n'
    > STORED AS TEXTFILE
    > LOCATION '/user/hadoop/mlctest/events_train/';
OK
Time taken: 0.051 seconds
hive> INSERT OVERWRITE TABLE events_train select b.* from train a left join events b on a.device_id = b.device_id ;
Query ID = hadoop_20210111163829_c7dcdc2c-9f54-4090-83c9-fb14c567f74a
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1610340234319_0025)
```

Loading Data :
INSERT OVERWRITE TABLE events_train select b. from train a left join events b on a.device_id = b.device_id ;*

c. **App data :**
create external table app_data
(
event_id int,
is_installed int,
is_active int,
category string
)
row format delimited fields terminated by ',' lines terminated by '\n'
STORED AS TEXTFILE
LOCATION '/user/hadoop/mlctest/app_data/';

```
hive> create external table app_data
    > (
    >  event_id  int,
    >  is_installed  int,
    >  is_active  int,
    >  category  string
    > )
    > row format delimited fields terminated by ',' lines terminated by '\n'
    > STORED AS TEXTFILE
    > LOCATION '/user/hadoop/mlctest/app_data/';
OK
Time taken: 0.071 seconds
hive> INSERT OVERWRITE TABLE app_data select a.event_id, a.is_installed, a.is_active, c.category from app_events a inner join app_labels b  on a.app_id = b.a
pp_id inner join label_categories c on b.label_id = c.label_id   ;
No Stats for default@app_events, Columns: is_installed, event_id, is_active, app_id
No Stats for default@app_labels, Columns: app_id, label_id
No Stats for default@label_categories, Columns: category, label_id
Query ID = hadoop_20210111164041_1c706c41-8bfe-4df5-8427-bc65806b1552
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1610340234319_0025)
```

Table Shape Details:

Table Name	No. Of Columns	No. Of Rows
non_event_train_brand	6	74645
events_train	5	1266933
app_data	4	209355710

```
hive> select count(*) from non_event_train_brand;
OK
74645
Time taken: 0.108 seconds, Fetched: 1 row(s)
hive>
```

```
hive> select count(*)from events_train;
OK
1266933
Time taken: 0.185 seconds, Fetched: 1 row(s)
hive>
```

```
hive> select count(*) from app_data;
OK
209355710
Time taken: 0.152 seconds, Fetched: 1 row(s)
hive>
```

The below mentioned commands were used to move the tables from Hadoop to S3 for the 3 external tables created above:

hadoop distcp -Dfs.s3a.access.key=AKIA3WANO6POFSV52Z5H -Dfs.s3a.secret.key="" hdfs://ec2-3-87-222-170.compute-1.amazonaws.com:8020/user/hadoop/mlctest/events_train/ s3a://capstone-project-cp/table_dump/events_train

hadoop distcp -Dfs.s3a.access.key=AKIA3WANO6POFSV52Z5H -Dfs.s3a.secret.key="" hdfs://ec2-3-87-222-170.compute-1.amazonaws.com:8020/user/hadoop/mlctest/non_event_train_brand/ s3a://capstone-project-cp/table_dump/non_event_train_brand

hadoop distcp -Dfs.s3a.access.key=AKIA3WANO6POFSV52Z5H -Dfs.s3a.secret.key="" hdfs://ec2-3-87-222-170.compute-

1.amazonaws.com:8020/user/hadoop/mlctest/app_data/ s3a://capstone-project-cp/table_dump/app_data

Screenshot from AWS S3 :