IIT- M Advanced Certificate Program in Machine Learning and Cloud- upGrad Capstone Project

User Demographics Prediction using Telecom dataset

Data Preparation & Modelling Commands

Authors :

Mukul Pahawa

Mitesh

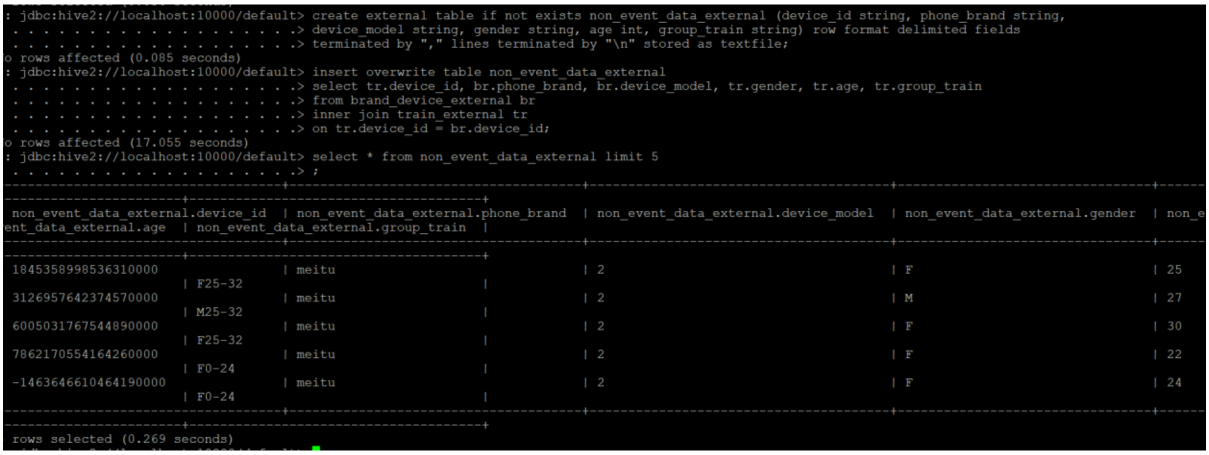
DATA PREPARATION FOR MODELLING

==

Creation of external tables and loading data

==

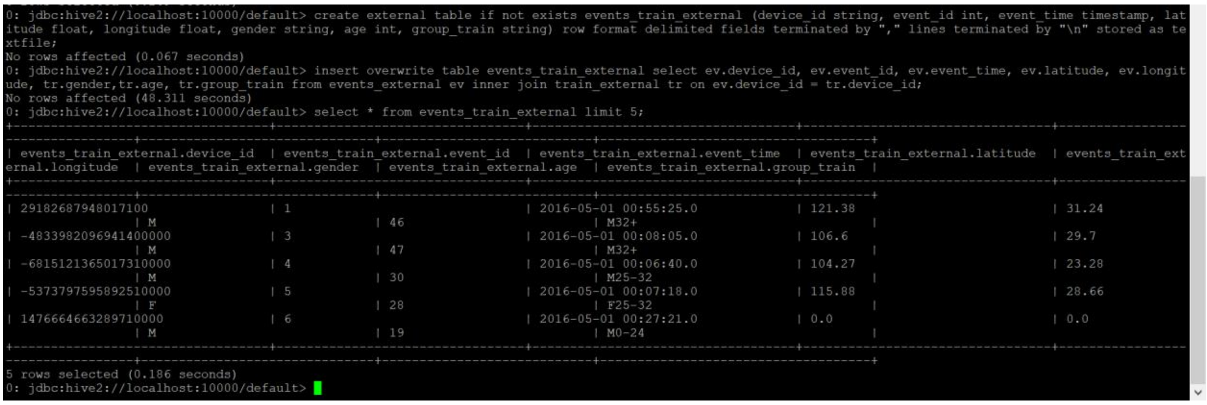
```
create external table if not exists non_event_data_external (device_id string, phone_brand string,
device_model string, gender string, age int, group_train string) row format delimited fields
terminated by "," lines terminated by "\n" stored as textfile;
insert overwrite table non_event_data_external
select tr.device_id, br.phone_brand, br.device_model, tr.gender, tr.age, tr.group_train
from brand_device_external br
inner join train_external tr
on tr.device_id = br.device_id;
```



```
: jdbc:hive2://localhost:10000/default> create external table if not exists non_event_data_external (device_id string, phone_brand string,
. . . . . . . . . . . . . . . . . . . .> device_model string, gender string, age int, group_train string) row format delimited fields
. . . . . . . . . . . . . . . . . . . .> terminated by "," lines terminated by "\n" stored as textfile;
o rows affected (0.085 seconds)
: jdbc:hive2://localhost:10000/default> insert overwrite table non_event_data_external
. . . . . . . . . . . . . . . . . . . .> select tr.device_id, br.phone_brand, br.device_model, tr.gender, tr.age, tr.group_train
. . . . . . . . . . . . . . . . . . . .> from brand_device_external br
. . . . . . . . . . . . . . . . . . . .> inner join train_external tr
. . . . . . . . . . . . . . . . . . . .> on tr.device_id = br.device_id;
o rows affected (17.055 seconds)
: jdbc:hive2://localhost:10000/default> select * from non_event_data_external limit 5
. . . . . . . . . . . . . . . . . . . .> ;
-------------------------------------+---------------------------------------+---------------------------------------+---------------------------------+------
-------------------------+---------------------------------------+
 non_event_data_external.device_id  | non_event_data_external.phone_brand  | non_event_data_external.device_model  | non_event_data_external.gender  | non_e
ent_data_external.age  | non_event_data_external.group_train  |
-------------------------------------+---------------------------------------+---------------------------------------+---------------------------------+------
-------------------------+---------------------------------------+
 1845358998536310000                | meitu                                 | 2                                     | F                               | 25
                        | F25-32                                |
 3126957642374570000                | meitu                                 | 2                                     | M                               | 27
                        | M25-32                                |
 6005031767544890000                | meitu                                 | 2                                     | F                               | 30
                        | F25-32                                |
 7862170554164260000                | meitu                                 | 2                                     | F                               | 22
                        | F0-24                                 |
 -1463646610464190000               | meitu                                 | 2                                     | F                               | 24
                        | F0-24                                 |
-------------------------------------+---------------------------------------+---------------------------------------+---------------------------------+------
-------------------------+---------------------------------------+
 rows selected (0.269 seconds)
```

create external table if not exists events_train_external (device_id string, event_id int, event_time timestamp, latitude float, longitude float, gender string, age int, group_train string) row format delimited fields terminated by "," lines terminated by "\n" stored as textfile;

insert overwrite table events_train_external select ev.device_id, ev.event_id, ev.event_time, ev.latitude, ev.longitude, tr.gender,tr.age, tr.group_train from events_external ev inner join train_external tr on ev.device_id = tr.device_id;



```
0: jdbc:hive2://localhost:10000/default> create external table if not exists events_train_external (device_id string, event_id int, event_time timestamp, lat
itude float, longitude float, gender string, age int, group_train string) row format delimited fields terminated by "," lines terminated by "\n" stored as te
xtfile;
No rows affected (0.067 seconds)
0: jdbc:hive2://localhost:10000/default> insert overwrite table events_train_external select ev.device_id, ev.event_id, ev.event_time, ev.latitude, ev.longit
ude, tr.gender,tr.age, tr.group_train from events_external ev inner join train_external tr on ev.device_id = tr.device_id;
No rows affected (48.311 seconds)
0: jdbc:hive2://localhost:10000/default> select * from events_train_external limit 5;
+---------------------------------+-------------------------------+---------------------------------+----------------------------------+----------------
----------------+-------------------------------+---------------------------+------------------------------------+
| events_train_external.device_id  | events_train_external.event_id  | events_train_external.event_time  | events_train_external.latitude  | events_train_ext
ernal.longitude  | events_train_external.gender  | events_train_external.age  | events_train_external.group_train  |
+---------------------------------+-------------------------------+---------------------------------+----------------------------------+----------------
----------------+-------------------------------+---------------------------+------------------------------------+
| 29182687948017100               | 1                              | 2016-05-01 00:55:25.0            | 121.38                           | 31.24
                 | M                             | 46                         | M32+                               |
| -4833982096941400000            | 3                              | 2016-05-01 00:08:05.0            | 106.6                            | 29.7
                 | M                             | 47                         | M32+                               |
| -6815121365017310000            | 4                              | 2016-05-01 00:06:40.0            | 104.27                           | 23.28
                 | M                             | 30                         | M25-32                             |
| -5373797595892510000            | 5                              | 2016-05-01 00:07:18.0            | 115.88                           | 28.66
                 | F                             | 28                         | F25-32                             |
| 1476664663289710000             | 6                              | 2016-05-01 00:27:21.0            | 0.0                              | 0.0
                 | M                             | 19                         | M0-24                              |
+---------------------------------+-------------------------------+---------------------------------+----------------------------------+----------------
----------------+-------------------------------+---------------------------+------------------------------------+
5 rows selected (0.186 seconds)
0: jdbc:hive2://localhost:10000/default>
```

create external table if not exists app_data_external (event_id int, app_id string, is_installed int, is_active int, label_id int, category string) row format delimited fields terminated by "," lines terminated by "\n" stored as textfile;

insert overwrite table app_data_external

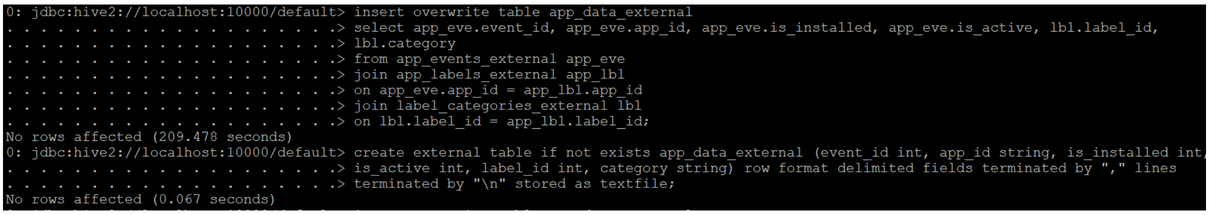
select app_eve.event_id, app_eve.app_id, app_eve.is_installed, app_eve.is_active, lbl.label_id, lbl.category from app_events_external app_eve

join app_labels_external app_lbl

on app_eve.app_id = app_lbl.app_id

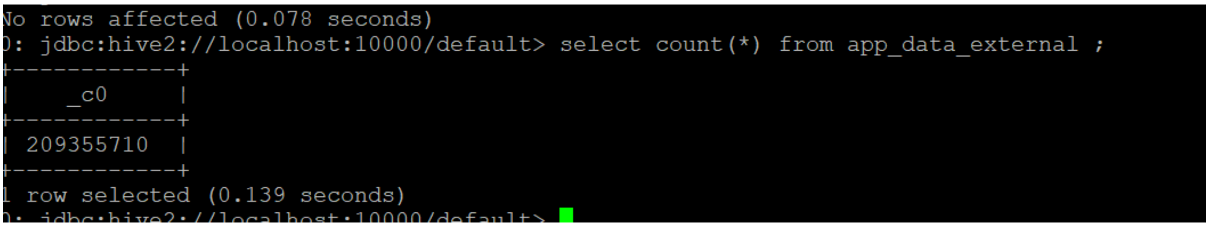
join label_categories_external lbl

on lbl.label_id = app_lbl.label_id;



```
0: jdbc:hive2://localhost:10000/default> insert overwrite table app_data_external
. . . . . . . . . . . . . . . . . . . . .> select app_eve.event_id, app_eve.app_id, app_eve.is_installed, app_eve.is_active, lbl.label_id,
. . . . . . . . . . . . . . . . . . . . .> lbl.category
. . . . . . . . . . . . . . . . . . . . .> from app_events_external app_eve
. . . . . . . . . . . . . . . . . . . . .> join app_labels_external app_lbl
. . . . . . . . . . . . . . . . . . . . .> on app_eve.app_id = app_lbl.app_id
. . . . . . . . . . . . . . . . . . . . .> join label_categories_external lbl
. . . . . . . . . . . . . . . . . . . . .> on lbl.label_id = app_lbl.label_id;
No rows affected (209.478 seconds)
0: jdbc:hive2://localhost:10000/default> create external table if not exists app_data_external (event_id int, app_id string, is_installed int,
. . . . . . . . . . . . . . . . . . . . .> is_active int, label_id int, category string) row format delimited fields terminated by "," lines
. . . . . . . . . . . . . . . . . . . . .> terminated by "\n" stored as textfile;
No rows affected (0.067 seconds)
```

select count(*) from app_data_external



```
No rows affected (0.078 seconds)
0: jdbc:hive2://localhost:10000/default> select count(*) from app_data_external ;
+------------+
|    _c0     |
+------------+
| 209355710  |
+------------+
1 row selected (0.139 seconds)
0: jdbc:hive2://localhost:10000/default>
```

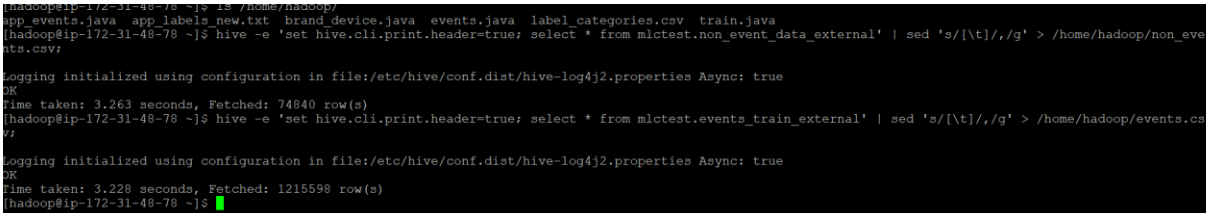
CSV File Creation from external HIVE Tables

==

hive -e 'set hive.cli.print.header=true; select * from mlctest.non_event_data_external' | sed 's/[\t]/,/g' > /home/hadoop/non_events.csv;

hive -e 'set hive.cli.print.header=true; select * from mlctest.events_train_external' | sed 's/[\t]/,/g' > /home/hadoop/events.csv;

hive -e 'set hive.cli.print.header=true; select * from mlctest.app_data_external' | sed 's/[\t]/,/g' > /home/hadoop/appdata.csv;



```
[hadoop@ip-172-31-48-78 ~]$ ls /home/hadoop/
app_events.java  app_labels_new.txt  brand_device.java  events.java  label_categories.csv  train.java
[hadoop@ip-172-31-48-78 ~]$ hive -e 'set hive.cli.print.header=true; select * from mlctest.non_event_data_external' | sed 's/[\t]/,/g' > /home/hadoop/non_eve
nts.csv;

Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j2.properties Async: true
OK
Time taken: 3.263 seconds, Fetched: 74840 row(s)
[hadoop@ip-172-31-48-78 ~]$ hive -e 'set hive.cli.print.header=true; select * from mlctest.events_train_external' | sed 's/[\t]/,/g' > /home/hadoop/events.cs
v;

Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j2.properties Async: true
OK
Time taken: 3.228 seconds, Fetched: 1215598 row(s)
[hadoop@ip-172-31-48-78 ~]$ 
```

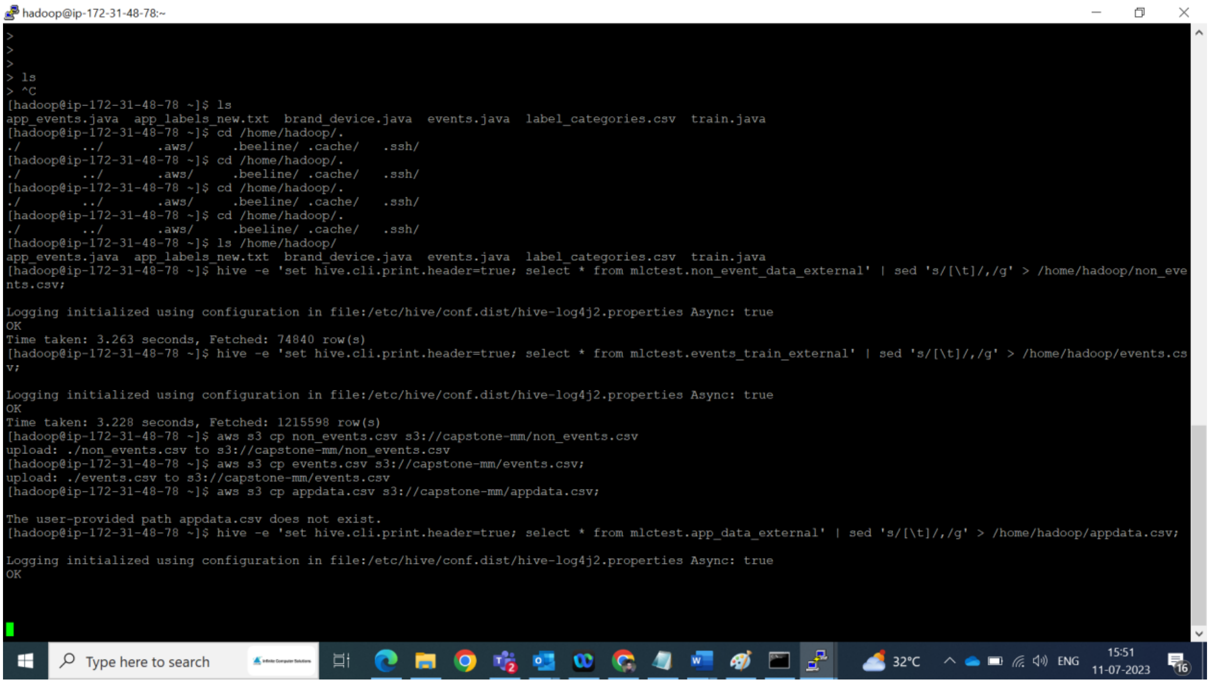
Copying the CSV file to S3 Bucket

================================

aws s3 cp non_events.csv s3://capstone-mm/non_events.csv;

aws s3 cp events.csv s3://capstone-mm/events.csv;

aws s3 cp appdata.csv s3://capstone-mm/appdata.csv;



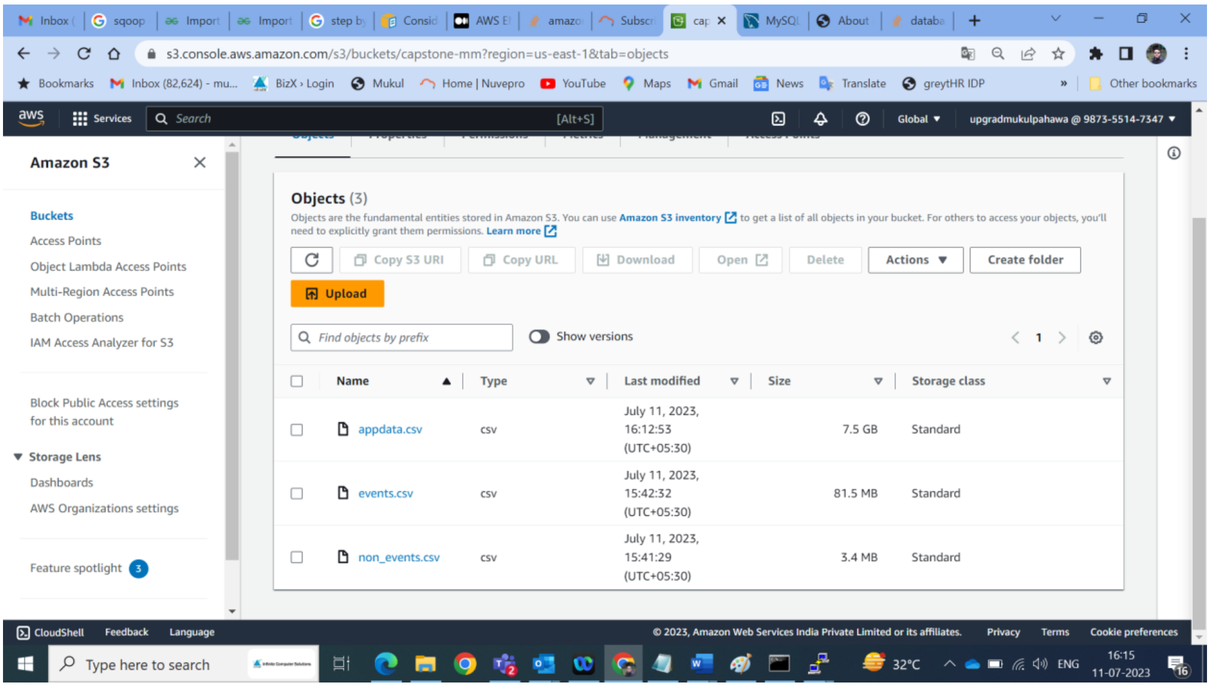
```
>
>
>
> ls
> ^C
[hadoop@ip-172-31-48-78 ~]$ ls
app_events.java  app_labels_new.txt  brand_device.java  events.java  label_categories.csv  train.java
[hadoop@ip-172-31-48-78 ~]$ cd /home/hadoop/.
./          ../          .aws/      .beeline/ .cache/    .ssh/
[hadoop@ip-172-31-48-78 ~]$ cd /home/hadoop/.
./          ../          .aws/      .beeline/ .cache/    .ssh/
[hadoop@ip-172-31-48-78 ~]$ cd /home/hadoop/.
./          ../          .aws/      .beeline/ .cache/    .ssh/
[hadoop@ip-172-31-48-78 ~]$ cd /home/hadoop/.
./          ../          .aws/      .beeline/ .cache/    .ssh/
[hadoop@ip-172-31-48-78 ~]$ ls /home/hadoop/
app_events.java  app_labels_new.txt  brand_device.java  events.java  label_categories.csv  train.java
[hadoop@ip-172-31-48-78 ~]$ hive -e 'set hive.cli.print.header=true; select * from mlctest.non_event_data_external' | sed 's/[\t]/,/g' > /home/hadoop/non_eve
nts.csv;

Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j2.properties Async: true
OK
Time taken: 3.263 seconds, Fetched: 74840 row(s)
[hadoop@ip-172-31-48-78 ~]$ hive -e 'set hive.cli.print.header=true; select * from mlctest.events_train_external' | sed 's/[\t]/,/g' > /home/hadoop/events.cs
v;

Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j2.properties Async: true
OK
Time taken: 3.228 seconds, Fetched: 1215598 row(s)
[hadoop@ip-172-31-48-78 ~]$ aws s3 cp non_events.csv s3://capstone-mm/non_events.csv
upload: ./non_events.csv to s3://capstone-mm/non_events.csv
[hadoop@ip-172-31-48-78 ~]$ aws s3 cp events.csv s3://capstone-mm/events.csv;
upload: ./events.csv to s3://capstone-mm/events.csv
[hadoop@ip-172-31-48-78 ~]$ aws s3 cp appdata.csv s3://capstone-mm/appdata.csv;

The user-provided path appdata.csv does not exist.
[hadoop@ip-172-31-48-78 ~]$ hive -e 'set hive.cli.print.header=true; select * from mlctest.app_data_external' | sed 's/[\t]/,/g' > /home/hadoop/appdata.csv;

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



Objects (3)

Objects are the fundamental entities stored in Amazon S3. You can use Amazon S3 inventory to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. Learn more

□	Name	Type	Last modified	Size	Storage class
□	appdata.csv	csv	July 11, 2023, 16:12:53 (UTC+05:30)	7.5 GB	Standard
□	events.csv	csv	July 11, 2023, 15:42:32 (UTC+05:30)	81.5 MB	Standard
□	non_events.csv	csv	July 11, 2023, 15:41:29 (UTC+05:30)	3.4 MB	Standard

© 2023, Amazon Web Services India Private Limited or its affiliates.