Data Preparation for Modelling

Dataset Type	Tables	Primary Key
Non-Event Data	train and brand_devices	device_id
Event Data	events and train	device_id
App Data	app_events , app_labels and label_categories	event_id

Creating table app_data:

Query: create table app_data

stored as parquet

as

select

a.event_id, a.app_id, a.is_installed, a.is_active,

b.label_id, c.category

from

app_events3 a,

app_labels3 b,

label_categories3 c

where

a.app_id = b.app_id and

b.label_id = c.label_id;

Shape of the app_data:

```
Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j2.properties Async: false
hive> select count(*) from app_data;
OK
209355710
Time taken: 4.98 seconds, Fetched: 1 row(s)
hive>
```

Dump the app_data into S3:

set mapred.reduce.tasks = 1;

insert overwrite directory 's3://upgradcapstone2022bucket/capstonedata/app_data.csv'

row format delimited fields terminated by ','

select

event_id, app_id, is_installed, is_active,

label_id, category

from

app_data

order by event_id;

set mapred.reduce.tasks = -1;

```
hive> set mapred.reduce.tasks = 1;
hive> insert overwrite directory 's3://upgradcapstone2022bucket/capstonedata/app_data.csv'
    > row format delimited fields terminated by ','
    > select
    >     event_id, app_id, is_installed, is_active,
    >     label_id, category
    > from
    >     app_data
    > order by event_id;
Query ID = hadoop_20221230221017_63b15817-80ab-465e-a772-60129617c0b5
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_1672392463972_0057)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      8          8        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 1004.98 s
--------------------------------------------------------------------------------
Moving data to directory s3://upgradcapstone2022bucket/capstonedata/app_data.csv
OK
Time taken: 1027.308 seconds
hive>
```

Creating event_data:

Query: create table event_data1

stored as parquet

as

select

a.event_id, a.device_id, a.event_timestamp,a.longitude,a.latitude,

b.gender,b.age,b.group_name

from

events3 a left join train3 b on a.device_id=b.device_id;

Shape of the event_data:

```
hive> create table event_data1
    > stored as parquet
    > as
    > select
    >     a.event_id,a.device_id, a.event_timestamp,a.longitude,a.latitude,
    >     b.gender,b.age,b.group_name
    > from
    >     events3 a left join train3 b on a.device_id=b.device_id;
Query ID = hadoop_20221230233903_d2d80843-1ab3-433c-b004-29ce8cc9ac18
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_1672392463972_0062)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      6          6        0        0       0       0
Map 2 .......... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 40.59 s
--------------------------------------------------------------------------------
Moving data to directory hdfs://ip-172-31-15-164.us-east-2.compute.internal:8020/user/hive/warehouse/event_data1
OK
Time taken: 51.39 seconds
hive> select count(*) from event_data1;
OK
3252950
Time taken: 0.359 seconds, Fetched: 1 row(s)
```

Dump event_data1 into S3

```
set mapred.reduce.tasks = 1;

insert overwrite directory 's3://upgradcapstone2022bucket/capstonedata/event_data.csv'

row format delimited fields terminated by ','

select

   event_id,device_id, event_timestamp,longitude,latitude,

   gender,age,   group_name

from

   event_data1

order by device_id;

set mapred.reduce.tasks = -1;
```

```
hive> set mapred.reduce.tasks = 1;
hive> insert overwrite directory 's3://upgradcapstone2022bucket/capstonedata/event_data.csv'
    > row format delimited fields terminated by ','
    > select
    >     event_id,device_id, event_timestamp,longitude,latitude,
    >     gender,age,group_name
    > from
    >     event_data1
    > order by device_id;
Query ID = hadoop_20221230234355_ae425dd8-a448-4066-97cd-54fdaf974796
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1672392463972_0062)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      6          6        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 67.64 s
----------------------------------------------------------------------------------------------
Moving data to directory s3://upgradcapstone2022bucket/capstonedata/event_data.csv
OK
Time taken: 76.777 seconds
hive> set mapred.reduce.tasks = -1;
```

Creating non_event_data:

Query: create table non_event_data1

stored as parquet

as

select

a.device_id, a.phone_brand, a.device_model,

b.gender,b.age, b.group_name

from

brand_device3 a left join train3 b on a.device_id=b.device_id;

Shape of the non_event_data:

```
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_1672392463972_0063)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Map 2 .......... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 14.76 s
----------------------------------------------------------------------------------------------
Moving data to directory hdfs://ip-172-31-15-164.us-east-2.compute.internal:8020/user/hive/warehouse/non_event_data1
OK
Time taken: 25.66 seconds
hive> select count(*) from non_event_data1;
OK
187245
Time taken: 0.341 seconds, Fetched: 1 row(s)
```

Dump non_event_data to S3:

set mapred.reduce.tasks = 1;

insert overwrite directory 's3://upgradcapstone2022bucket/capstonedata/non_event_data.csv'

row format delimited fields terminated by ','

select

```
    device_id, phone_brand, device_model,
    gender,age, group_name
from
    non_event_data1
order by device_id;
set mapred.reduce.tasks = -1;
```

```
Time taken: 0.341 seconds, Fetched: 1 row(s)
hive> set mapred.reduce.tasks = 1;
hive> insert overwrite directory 's3://upgradcapstone2022bucket/capstonedata/non_event_data.csv'
    > row format delimited fields terminated by ','
    > select
    >     device_id, phone_brand, device_model,
    >     gender,age, group_name
    > from
    >     non_event_data1
    > order by device_id;
Query ID = hadoop_20221230235520_d30d7469-c1ce-4133-bf53-68e5688c20f4
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1672392463972_0063)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 15.10 s
----------------------------------------------------------------------------------------------
Moving data to directory s3://upgradcapstone2022bucket/capstonedata/non_event_data.csv
OK
Time taken: 17.297 seconds
hive> set mapred.reduce.tasks = -1;
hive>
```