**Hive analysis**

**Created external tables and data is stored in parquet format which is faster in accessing**

create external table app\_events\_stg2(

event\_id bigint,

app\_id bigint,

is\_installed int,

is\_active int

)

ROW FORMAT DELIMITED

FIELDS TERMINATED BY ','

LINES TERMINATED BY '\n';

load data inpath ‘/home/hadoop/capstonetelcom/stage/app\_events2’ into table app\_events\_stg2

Text

Description automatically generated

create table app\_events

stored as parquet

as

select event\_id,app\_id,is\_installed,is\_active from app\_events\_stg2;

Text

Description automatically generated

--------------------------------------------------------------------------------------------------------------------------------------

create external table brand\_device\_stg(

device\_id bigint,

phone\_brand string,

device\_model string

)

ROW FORMAT DELIMITED

FIELDS TERMINATED BY ','

LINES TERMINATED BY '\n';

load data inpath ‘/home/hadoop/capstonetelcom/stage/brand\_device’ into table brand\_device\_stg

Text

Description automatically generated

create table brand\_device

stored as parquet

as

select device\_id,phone\_brand,device\_model from brand\_device\_stg;

Text

Description automatically generated

--------------------------------------------------------------------------------------------------------------------------------------

create external table events\_stg (

event\_id bigint,

device\_id bigint,

event\_timestamp timestamp,

longitude decimal(10,2),

latitude decimal(10,2)

)

ROW FORMAT DELIMITED

FIELDS TERMINATED BY ','

LINES TERMINATED BY '\n';

load data inpath ‘/home/hadoop/capstonetelcom/stage/events’ into table events\_stg;

Text

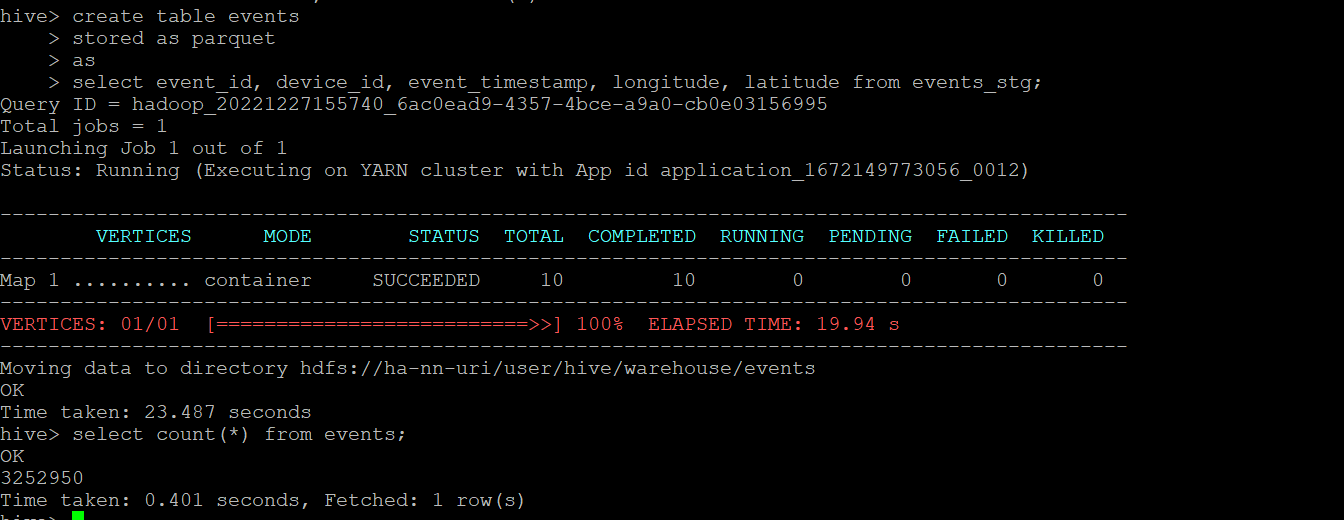
Description automatically generated

create table events

stored as parquet

as

select event\_id, device\_id, event\_timestamp, longitude, latitude from events\_stg;



--------------------------------------------------------------------------------------------------------------------------------------

create external table train\_stg (

device\_id bigint,

gender string,

age int,

group\_name string

)

ROW FORMAT DELIMITED

FIELDS TERMINATED BY ','

LINES TERMINATED BY '\n';

load data inpath ‘/home/hadoop/capstonetelcom/stage/train’ into table train\_stg;

Text

Description automatically generated

create table train

stored as parquet

as

select device\_id, gender, age, group\_name from train\_stg;

Text

Description automatically generated

--------------------------------------------------------------------------------------------------------------------------------------

create external table label\_categories\_stg3(

label\_id bigint,

category string

)

ROW FORMAT DELIMITED

FIELDS TERMINATED BY ','

LINES TERMINATED BY '\n'

stored as textfile

location '/home/hadoop/capstonetelcom/stage/labelcategories/'

tblproperties ('skip.header.line.count'='1');

create table label\_categories

stored as parquet

as

select label\_id,category from label\_categories\_stg3;

--------------------------------------------------------------------------------------------------------------------------------------

create external table app\_labels\_stg3(

app\_id bigint,

label\_id bigint

)

ROW FORMAT DELIMITED

FIELDS TERMINATED BY ','

LINES TERMINATED BY '\n'

stored as textfile

location '/home/hadoop/capstonetelcom/stage/applabels'

tblproperties ('skip.header.line.count'='1');

create table app\_labels

stored as parquet

as

select app\_id,label\_id from app\_labels\_stg3;

**Analytics report**

The 10 most popular brands and the percentage of the respective Male and Female owners of these brands [Handle the device id duplicates from brand\_device table.]

The 10 most popular brands for Male and Female? [Handle the device id duplicates from the brand\_device data set.]

The count and percentage analysis of the Gender in the train data set

The top mobile phone brands offering the highest number of models [Provide details about the top three brands.]

The average number of events per device id [Applicable to the device\_id column from the train table, which has at least one associated event in the event table]

Whether the count and percentage of the device\_id column in the train table have corresponding events data available