**Hive analysis**

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

**Loading data into tables from files loaded from RDS**

create external table app\_events\_stg3(

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\_stg3;

Text

Description automatically generated

create table app\_events3

stored as parquet

as

select event\_id,app\_id,is\_installed,is\_active from app\_events\_stg3;

Text

Description automatically generated

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

create external table brand\_device\_stg3(

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\_stg3;

Text

Description automatically generated

create table brand\_device3

stored as parquet

as

select device\_id,phone\_brand,device\_model from brand\_device\_stg3;

Text

Description automatically generated

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

create external table events\_stg3 (

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\_stg3;

Text

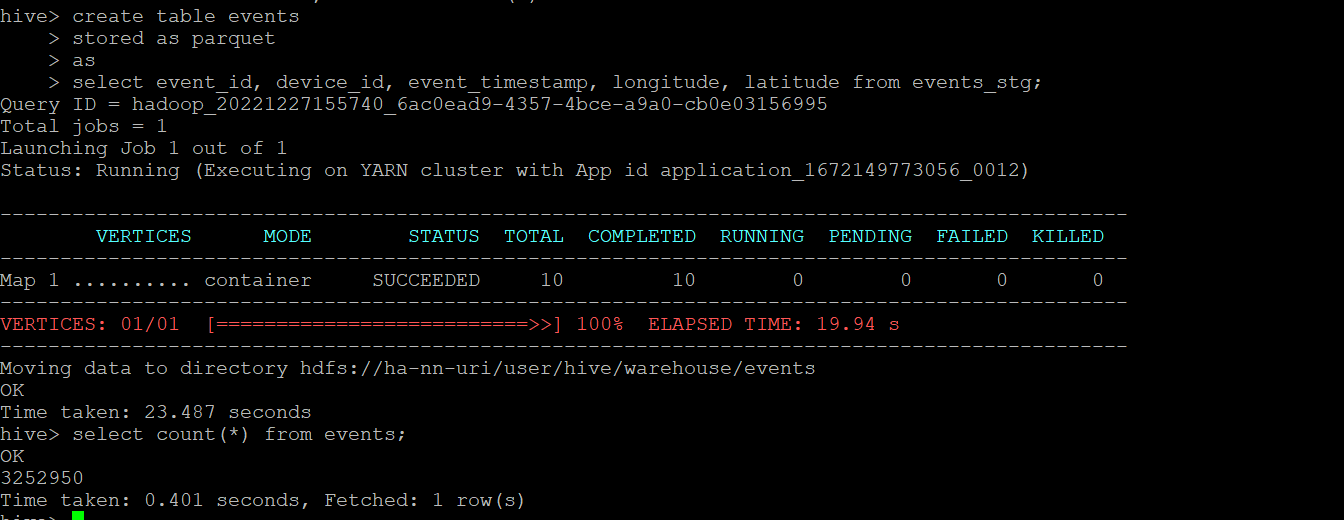
Description automatically generated

create table events3

stored as parquet

as

select event\_id, device\_id, event\_timestamp, longitude, latitude from events\_stg3;



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

create external table train\_stg3 (

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\_stg3;

Text

Description automatically generated

create table train3

stored as parquet

as

select device\_id, gender, age, group\_name from train\_stg3;

Text

Description automatically generated

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

**Loading data into tables from files loaded from S3**

create external table label\_categories\_stg3(

label\_id bigint,

category string

)

ROW FORMAT DELIMITED

FIELDS TERMINATED BY ','

LINES TERMINATED BY '\n';

load data local inpath '/home/hadoop/capstonetelcom/stage/labelcategories/label\_categories.csv' into table label\_categories\_stg3;

create table label\_categories3

stored as parquet

as

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

Text

Description automatically generated

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

create external table app\_labels\_stg3(

app\_id bigint,

label\_id bigint

)

ROW FORMAT DELIMITED

FIELDS TERMINATED BY ','

LINES TERMINATED BY '\n';

load data local inpath '/home/hadoop/capstonetelcom/stage/applables/app\_labels\_new.txt' into table app\_labels\_stg3;

create table app\_labels3

stored as parquet

as

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

Text

Description automatically generated

**Analytics report**

**1. 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.]**

select phone\_brand, ((malecount/totalcount)\*100) as male\_owner\_percentage,

((femalecount/totalcount)\*100) as female\_owner\_percentage

from

(select count(a.device\_id) as cnt1, a.phone\_brand, count(case when gender=='M' then 1 end) as malecount, count(case when gender=='F' then 1 end) as femalecount, count(gender) as totalcount

from

(Select device\_id, count(1) cnt from brand\_device3 group by device\_id having cnt=1) c,

brand\_device3 a,

train3 b

where

a.device\_id=b.device\_id and

a.device\_id=c.device\_id

group by a.phone\_brand

order by cnt1 desc

limit 10) n;

Graphical user interface, text

Description automatically generated

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

with cte as (

select

b.gender,

a.phone\_brand,

count(a.device\_id) as c,

dense\_rank() over (partition by b.gender order by count(a.device\_id) desc) as dr

from

(Select device\_id, count(1) cnt from brand\_device3 group by device\_id having cnt=1) c,

brand\_device3 a,

train3 b

where

a.device\_id=b.device\_id and

a.device\_id=c.device\_id

group by b.gender, a.phone\_brand

)

select \*

from cte

where dr <= 10

order by gender, c desc;

Text

Description automatically generated

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

select male as male\_count, female as female\_count, (male/total)\*100 as male\_percentage, (female/total)\*100 as female\_percentage

from(

select COUNT(case when gender=='M' then 1 end) as male,

COUNT(case when gender=='F' then 1 end) as female,

COUNT(gender) as total

from train3) n;

Text

Description automatically generated

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

select count(a.device\_id) as device\_count, a.phone\_brand, count(device\_model) as model\_count

from

(Select device\_id, count(1) cnt from brand\_device3 group by device\_id having cnt=1) c,

brand\_device3 a,

train3 b

where

a.device\_id=b.device\_id and

a.device\_id=c.device\_id

group by a.phone\_brand

order by device\_count desc

limit 3;

Text

Description automatically generated

**5. 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]**

select (sum(event\_count)/count(device\_id)) as avg\_events\_per\_device

from

(select

a.device\_id, count(a.event\_id) as event\_count

from

events3 a,

train3 b

where

a.device\_id = b.device\_id

group by a.device\_id) as n;

Text

Description automatically generated

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

select count(device\_id) from train3;

Graphical user interface, text

Description automatically generated

Yes the count and percentage of device id column in train table have events data. Below is the query and screenshot

select

count(b.device\_id) as device\_count, ((count(b.device\_id) \* 100.0)/74645) as device\_percentage

from

events3 a,

train3 b

where

a.device\_id = b.device\_id

having count(a.event\_id) >0;

Text

Description automatically generated