**Capstone project – Part 1**

**Problem statement**

This Capstone project is based on the domain of the telecom sector. We will be working with a telecom data set and will be required to develop a model to help a company run different marketing campaigns and data monetisation activities around it. We are required to build a model that will predict the demographics of certain users, i.e., their age and gender.

Our aim is to predict the age and the gender of a particular user using this data. Once we have predicted the gender and age for each user, next task will be to map that particular user to the given target-specific ad campaigns. These campaigns aim to improve customer experience and open avenues for revenue generation through cross-selling and upselling activities, and for partnership with other companies for data monetisation activities.

This part of the project looks at the following tasks

1. Data Ingestion and Analysis
2. SQL and Hive Analysis
3. Data Preparation for Modelling

**1. Data ingestion from various data sources**

**Moved the 2 files using Wincp from local machine (part s3) to Hadoop home and then moved them to separate folders**

cp /home/hadoop/app\_labels\_new.txt /home/hadoop/capstonetelcom/stage/applables/app\_labels\_new.txt

cp /home/hadoop/label\_categories.csv /home/hadoop/capstonetelcom/stage/labelcategories/label\_categories.csv

**Using sqoop imported the data from RDS instance for 4 files to 4 external files**

sqoop import --connect jdbc:mysql://mlc-testcapstone.cyaielc9bmnf.us-east-1.rds.amazonaws.com:3306/mlctest --table app\_events --columns "event\_id,app\_id,is\_installed,is\_active" --target-dir /home/hadoop/capstonetelcom/stage/app\_events2 --driver org.mariadb.jdbc.Driver --username student -P -m 1

Text

Description automatically generated

sqoop import --connect jdbc:mysql://mlc-testcapstone.cyaielc9bmnf.us-east-1.rds.amazonaws.com:3306/mlctest --table brand\_device --columns "device\_id,phone\_brand,device\_model" --target-dir /home/hadoop/capstonetelcom/stage/brand\_device --driver org.mariadb.jdbc.Driver --username student -P -m 1

Text

Description automatically generated

sqoop import --connect jdbc:mysql://mlc-testcapstone.cyaielc9bmnf.us-east-1.rds.amazonaws.com:3306/mlctest --table events --columns "event\_id,device\_id,timestamp,longitude,latitude" --target-dir /home/hadoop/capstonetelcom/stage/events --driver org.mariadb.jdbc.Driver --username student -P -m 1

Text

Description automatically generated

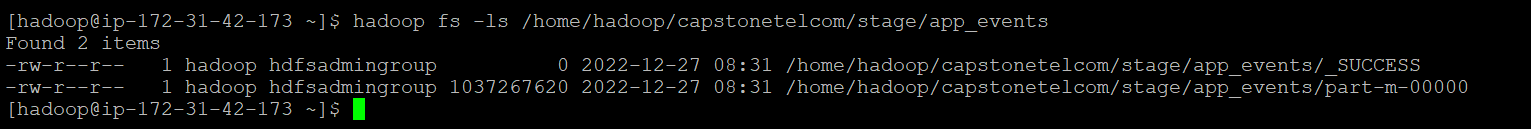
sqoop import --connect jdbc:mysql://mlc-testcapstone.cyaielc9bmnf.us-east-1.rds.amazonaws.com:3306/mlctest --table train --columns "device\_id,gender,age,group\_train" --target-dir /home/hadoop/capstonetelcom/stage/train --driver org.mariadb.jdbc.Driver --username student -P -m 1

Text

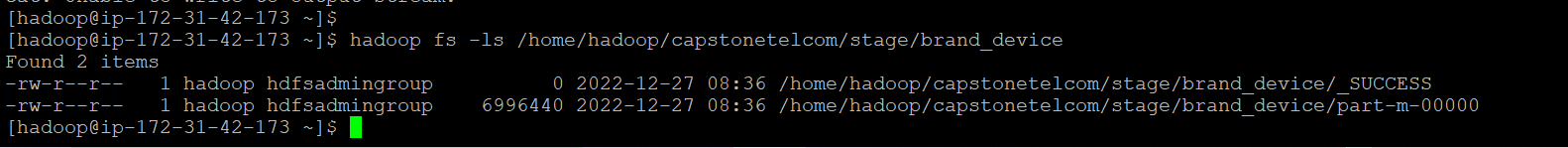
Description automatically generated

**Verified the sqoop has generated the distributed files for each of the data file**

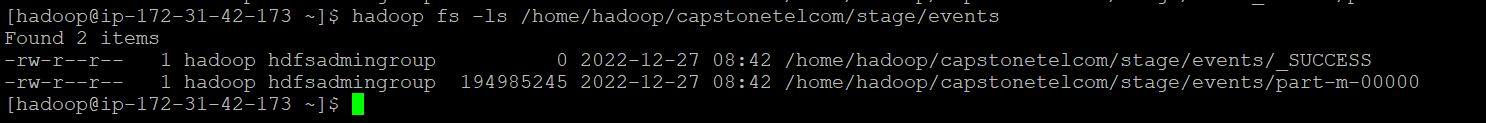
hadoop fs -ls /home/hadoop/capstonetelcom/stage/app\_events2



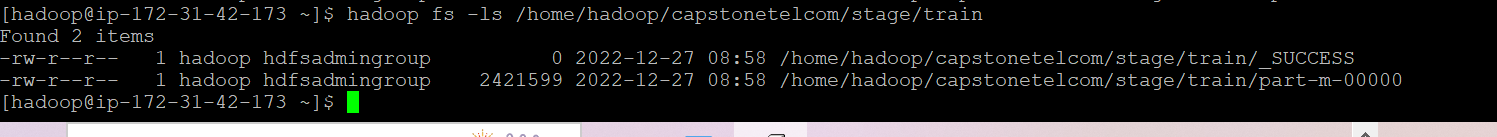
hadoop fs -ls /home/hadoop/capstonetelcom/stage/brand\_device



hadoop fs -ls /home/hadoop/capstonetelcom/stage/events



hadoop fs -ls /home/hadoop/capstonetelcom/stage/train



hadoop fs -cat /home/hadoop/capstonetelcom/stage/app\_events/part-m-00000 | head

A computer screen capture

Description automatically generated with low confidence

hadoop fs -cat /home/hadoop/capstonetelcom/stage/brand\_device/part-m-00000 | head

Text

Description automatically generated

hadoop fs -cat /home/hadoop/capstonetelcom/stage/events/part-m-00000 | head

A computer screen capture

Description automatically generated with medium confidence

hadoop fs -cat /home/hadoop/capstonetelcom/stage/train/part-m-00000 | head

Text

Description automatically generated with low confidence

**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

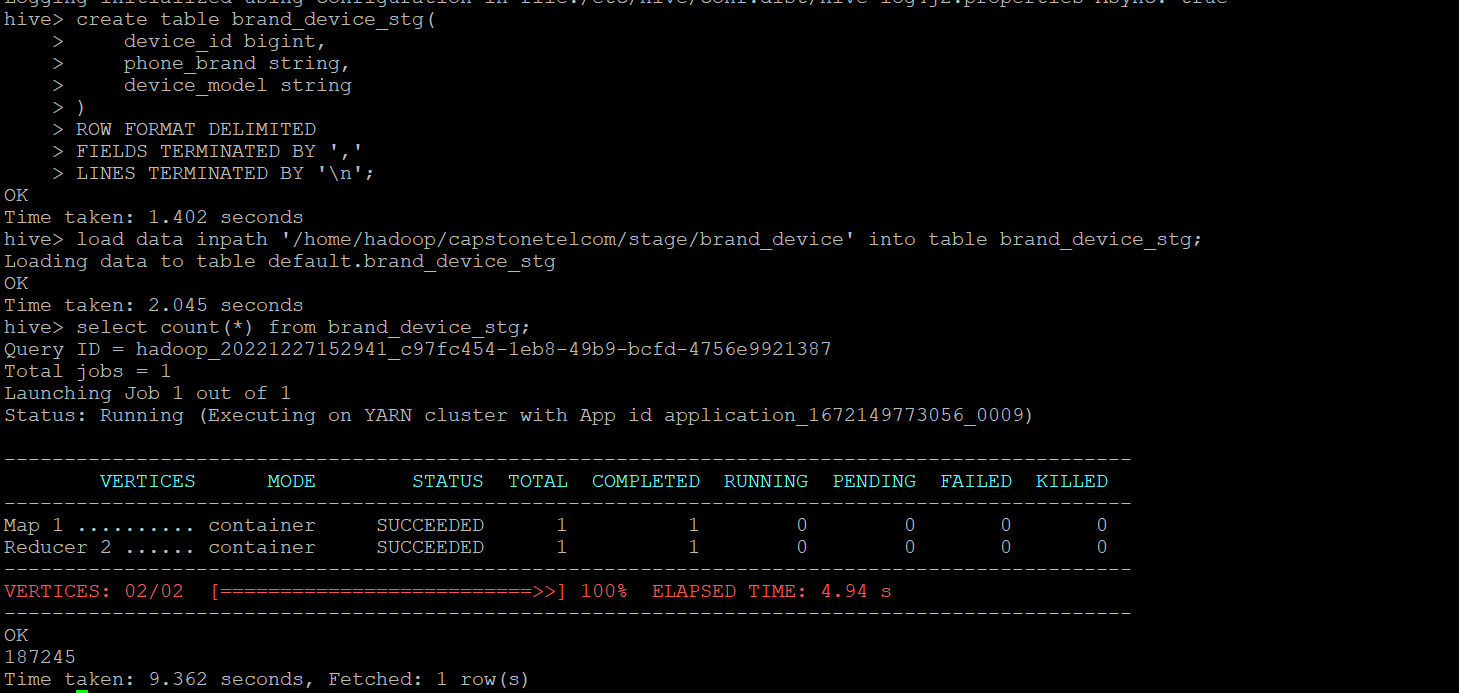
)

ROW FORMAT DELIMITED

FIELDS TERMINATED BY ','

LINES TERMINATED BY '\n';

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



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

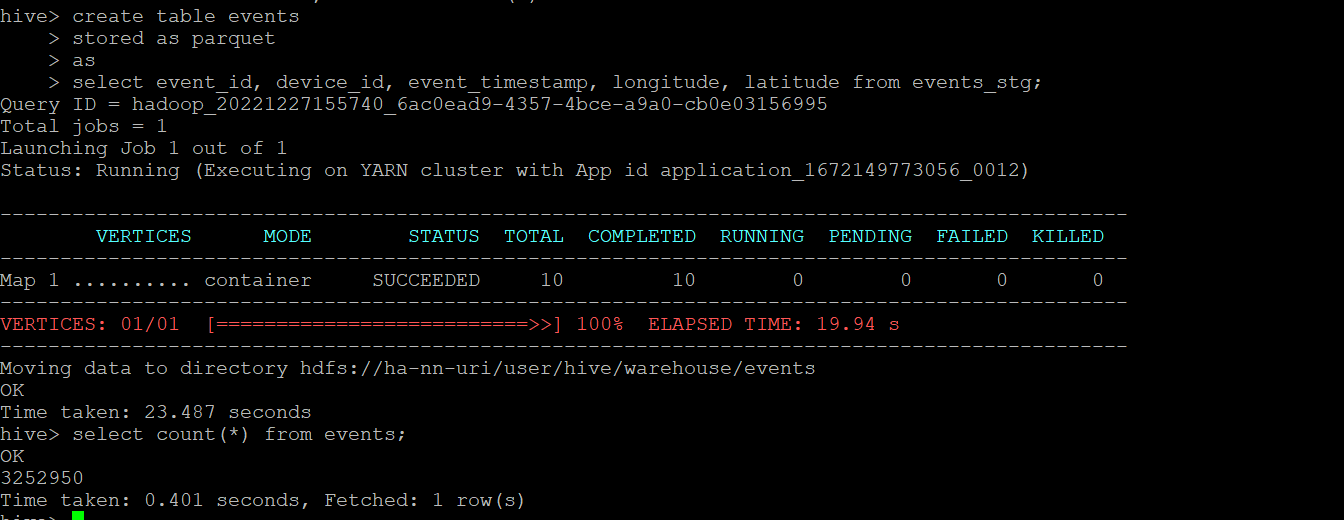
Description automatically generated

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

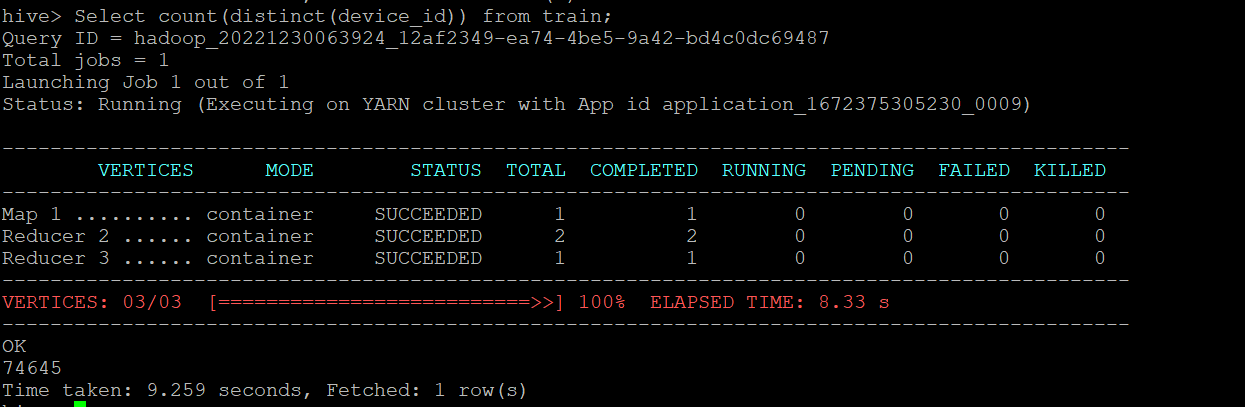
select app\_id,label\_id from app\_labels\_stg3;

**2. Basic SQL analysis**

**Count of unique device ids in the train table**

Select count(distinct(device\_id)) from train;

Identified 74645 unique device ids.



**Check whether there are any duplicate device ids present in the brand\_device table. If yes, how many duplicates?**

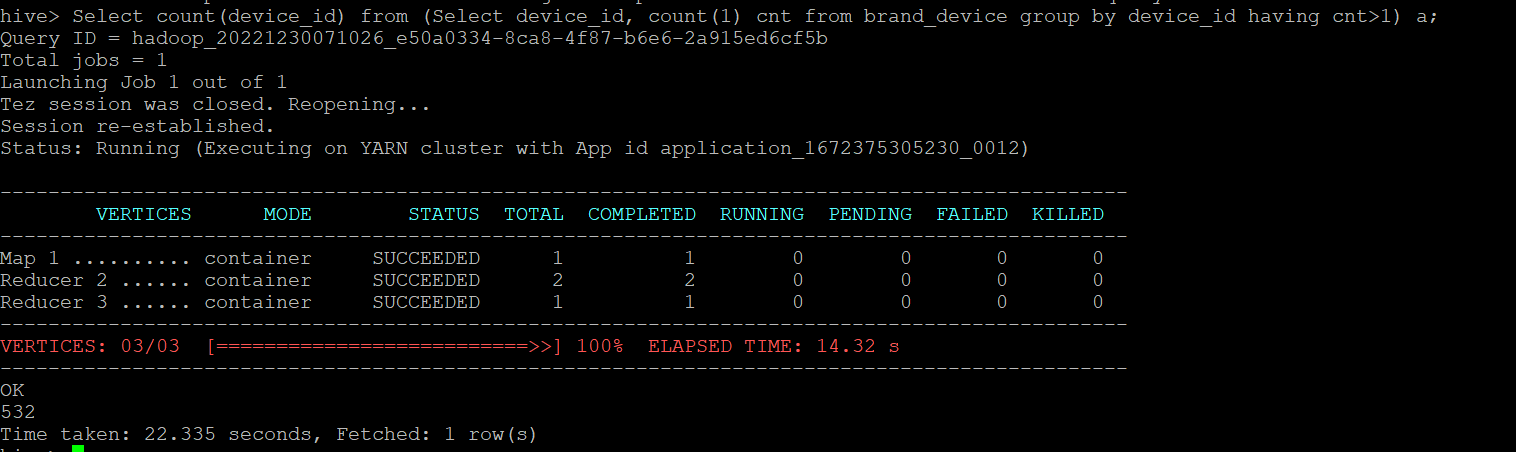
Select device\_id, count(1) cnt from brand\_device group by device\_id having cnt>1;

There are duplicates. Text

Description automatically generated

Select count(device\_id) from (Select device\_id, count(1) cnt from brand\_device group by device\_id having cnt>1) a;

Identified 532 duplicates



**Number of unique phone brands from the brand\_device table**

Select count(distinct(phone\_brand)) from brand\_device;

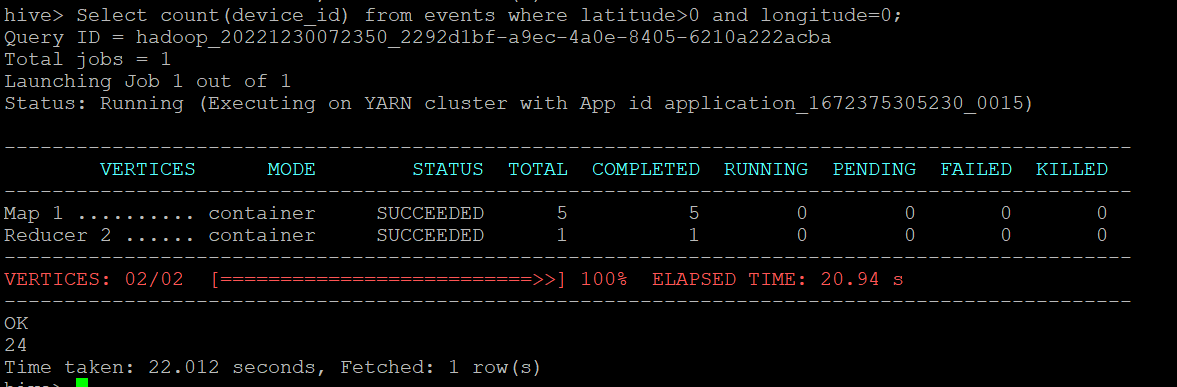
There are distinct 97 phone brands; Text

Description automatically generated

**Count of device ids where the latitude and longitude detail are zero, from the events table**

Select count(device\_id) from events where latitude>0 and longitude=0;

There are 24 rows with lat and long =0



**3. Hive for 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

**4. Data Preparation for Modelling**