**Data Preparation for Modelling**

Text

Description automatically generated with medium confidence

**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:**

Text

Description automatically generated

**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,

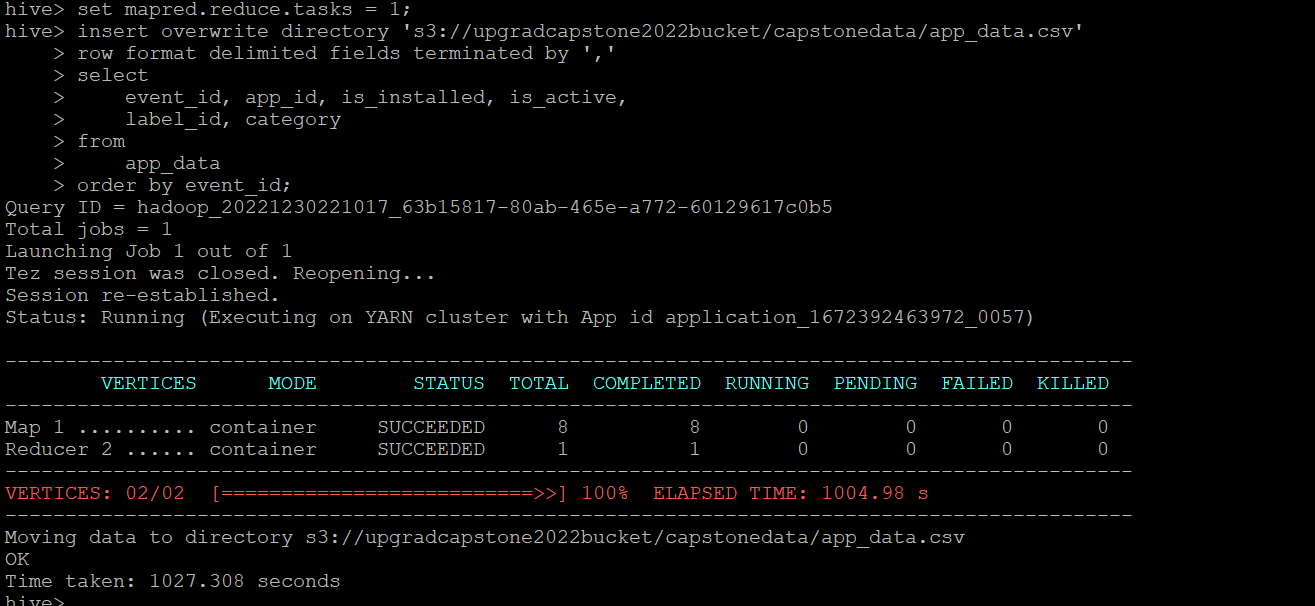
label\_id, category

from

app\_data

order by event\_id;

set mapred.reduce.tasks = -1;



**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:**

Text

Description automatically generated

**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;

Text

Description automatically generated

**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:**

A picture containing text

Description automatically generated

**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;

Text

Description automatically generated