Data Model

Tweets Table

CREATE TABLE public.tweetstable

(

t\_id bigint NOT NULL,

u\_id bigint NOT NULL,

msg text NOT NULL,

tweet\_created\_date timestamp,

favorite\_count integer,\*\*\*

retweet\_count integer,\*\*\*

lang character varying,

in\_reply\_t\_id bigint,

in\_reply\_u\_id bigint,

source text,

retweet\_t\_id bigint,

PRIMARY KEY (t\_id)

)

Notes:

\* "retweeted\_status" attribute may not exist in the tweet depending if this tweet is a retweet. If the attribute exists, then extract the original tweet id

\*\* rarely, there is a line "source":"",

\*\*\* Originally, favorite\_count, retweet\_count had NOT NULL constraints, but there were cases when they were null.

Users Table

CREATE TABLE public.userstable

(

u\_id bigint NOT NULL,

u\_name character varying,

u\_screenname character varying,

u\_location character varying,

u\_description text,

u\_friends\_count integer,

u\_followers\_count integer,

u\_listed\_count integer,

u\_fav\_count integer,

u\_status\_count integer,

u\_created\_date timestamp,

u\_lang character varying,

u\_profile\_img\_url character varying,

u\_default\_profile\_img boolean,

u\_tweet\_id bigint,

u\_tweet\_date timestamp,

PRIMARY KEY (u\_id)

)

Notes:

Had "NOT NULL" constraints on many more attributes on both Users Table and Tweets Table. However, ran into problems where there would be null values in favorites\_count and etc. Therefore, took out all NOT NULL on all attributes except the ids

Places Table ("place" attribute)

CREATE TABLE public.placestable

(

t\_id bigint NOT NULL,

u\_id bigint NOT NULL,

p\_id character varying NOT NULL,

p\_type character varying,

p\_name character varying,

p\_fullname character varying,

p\_countrycode character varying,

p\_country character varying,

PRIMARY KEY(t\_id, u\_id, p\_id)

)

Notes on Places Table:

\*\*unlike other BIGINT ids, "53b67b1d1cc81a51"

If "place":null, then it is not added to database

Hashtags, Media, Urls, User\_Mentions are all attributes from Entities attribute.

The reason for tables in 1NF form is each attribute is listed as "Array of Objects"

These attributes (EXCEPT FOR "Media" attribute) have the value of "[]" when there is no information in the json tweet

If Media attribute value is null, then it does not appear in json tweet.

Hashtags\*

CREATE TABLE public.hashtagstable

(

t\_id bigint NOT NULL,

u\_id bigint NOT NULL,

hashtags character varying NOT NULL,

PRIMARY KEY(t\_id, u\_id, hashtags)

)

User Mentions\*

CREATE TABLE public.usermentionstable

(

t\_id bigint NOT NULL,

u\_id bigint NOT NULL,

u\_screenname character varying,

um\_id bigint,NOT NULL^^

um\_screenname character varying,

um\_name character varying,

PRIMARY KEY(t\_id, u\_id, um\_id)

)

Notes

\* Hastags and User Mentions table is 1NF

\*\* user id of the tweet author

\*\*\*User mentions contain all users ids of anyone mention in message with @ symbol. This includes 1) in\_reply\_to 2) Retweets 3) any mentions

^^ for um\_id occasionally, there are jsons lines where um\_id is left "null" even though it should exist, these lines are removed and never added to database

Entities urls \*

CREATE TABLE public.entityurlstable

(

t\_id bigint NOT NULL,

u\_id bigint NOT NULL,

url character varying,

expanded\_url character varying,

PRIMARY KEY(t\_id, u\_id, url)

)

Entities media \*\*

CREATE TABLE public.entitymediatable

(

t\_id bigint NOT NULL,

u\_id bigint NOT NULL,

media\_url character varying NOT NULL,

PRIMARY KEY(t\_id, u\_id, media\_url)

)

Notes:

Both Entities urls and Entities media tables in 1NF

\* (media elements uploaded with the Tweet)

\*\* Represents URLs included in the text of a Tweet or within textual fields of a user object

\*\* An http:// URL pointing directly to the uploaded media file

\*\* media attribute will not exist in tweet if there are no elements