Flaws in Bad_giant_table:

Overview:

The giant table is the collection of tweets that has been tweeted in twitter. It has 27 attributes to describe the tweets and tweet_id as the **primary key**. They attributes are,

1. created_at
2. text
3. tweet_id
4. in_reply_to_screen_name
5. in_reply_to_status_id
6. in_reply_to_user_id
7. retweet_count
8. tweet_source
9. retweet_of_tweet_id
10. hashtag1
11. hashtag2
12. hashtag3
13. hashtag4
14. hashtag5
15. hashtag6
16. user_id
17. user_name
18. user_screen_name
19. user_location
20. user_utc_offset
21. user_time_zone
22. user_followers_count
23. user_friends_count
24. user_lang character
25. user_description
26. user_status_count
27. user_created_at

Flaws in Giant Table:

1.User and tweet dependency

tweet_id - > { user_name, user_screen_name, user_location, user_utc_offset, user_time_zone, user_followers_count, user_friends_count, user_lang, user_description, user_status_count, user_created_at}

{ user_name, user_screen_name, user_location, user_utc_offset, user_time_zone, user_followers_count, user_friends_count, user_lang, user_description, user_status_count, user_created_at} - > user_id

So,

tweet_id -> user_id

Problem and Explanation:

Tweet ID will give the entire details of User along with the user_id.

And the user_id can give the entire detail of Users. So, the User and Tweet has the Reflexivity Functional Depencency.

2.In Reply section with additional user information

tweet_id -> {in_reply_to_screen_name, in_reply_to_status_id ,in_reply_to_user_id}

in_reply_to_user_id -> { in_reply_to_screen_name }

So,

tweet_id -> { in_reply_to_status_id , in_reply_to_user_id }

in_reply_to_user_id -> { in_reply_to_screen_name }

Problem and Explanation:

user_id can give the user screen name which is additionally add in the tweet column. That can be removed. Using Augmentation Rule of Functional dependency.

3. UTC and TimeZone dependency:

user_id -> { user_utc_offset, user_time_zone }

user_utc_offset -> {user_time_zone }

So,

user_id -> {user_utc_offset}

user_utc_offset -> {user_time_zone }

Problem and Explanation:

user_utc_offset can give the user_time_zone which is additionally add in the tweet column. That can be removed using Augmentation Rule of Functional dependency. And have it in the separate table.

4. Hash Tag:

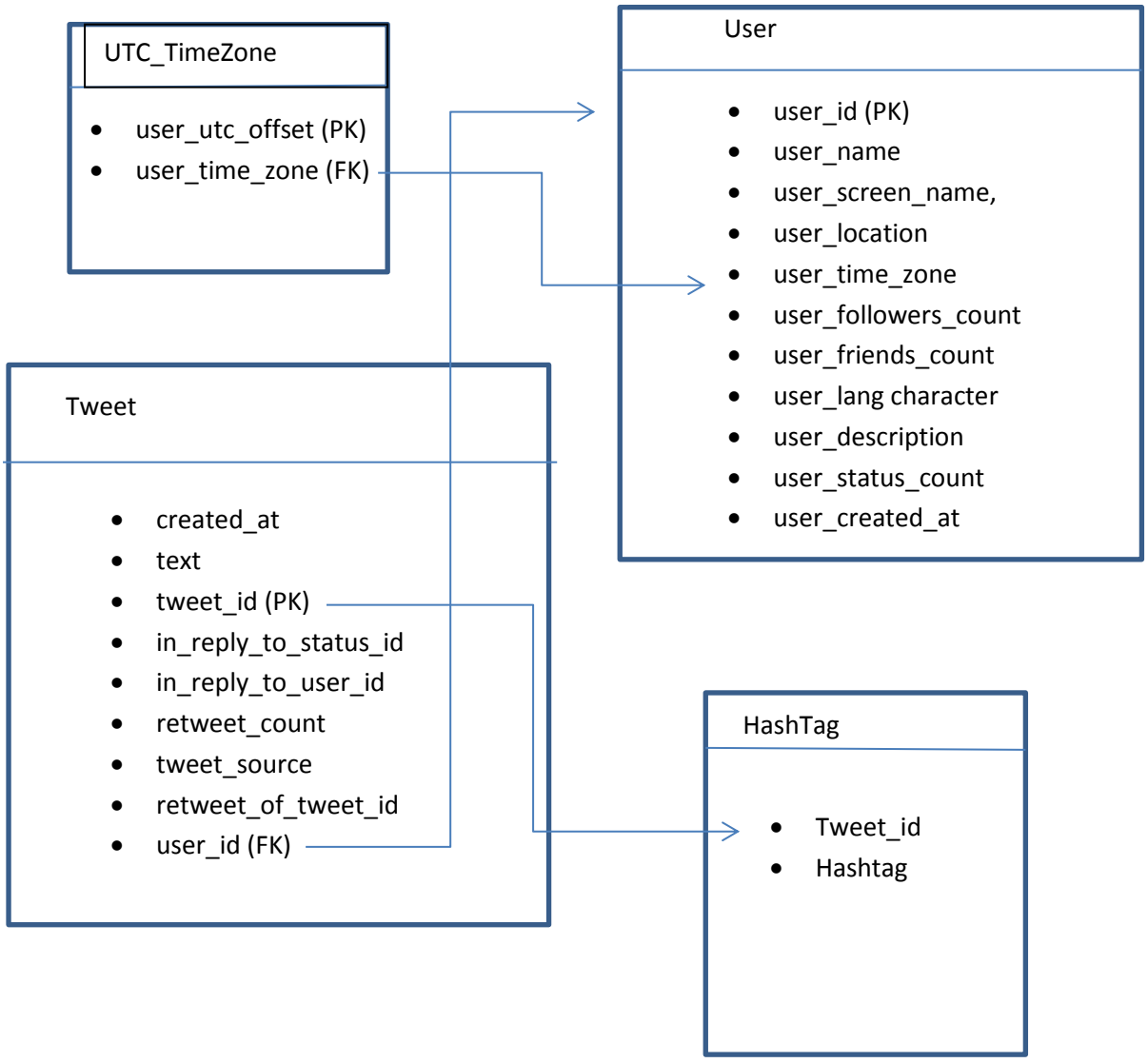
tweet_id -> {hashtag1, hashtag2, hashtag3, hashtag4, hashtag5, hashtag6}

Problem and Explanation:

Here, the hashtag count is limited by only 6. We can't add any further hashtag if needed. And NULL is stored in many cases which may reduce the efficiency of the database.

New Databse Design:

Class Diagram:



Problems Resolved:

Created tables without loosing the functional dependencies and improved efficiciency, which is the way to implement the 3NF.

1. Tweet and User dependency:

As we have created table separately for tweet and user. Now the search can be done efficiently

2. In Reply section with additional user information

Removed the column in_reply_to_screen_name from the tweet table.

3. UTC and TimeZone dependency:

Seperated the table for timezone and UTC column.

4. Hashtag

hashtag table is created separately where we can add any number of hash tags to a tweet and efficiently identify the hashtags using tweet_id.

Solutions to Queries:

Tweets, users and languages

Question 1

"110574"

Explanation:

Total number of tweets combines the normal tweets, replies, retweets

Question 2

User language	Number of tweets
"ar"	"1405"
"ca"	"7"
"cs"	"2"
"de"	"75"
"el"	"2"
"en"	"74077"
"es"	"17910"
"eu"	"1"
"fi"	"1"
"fil"	"9"
"fr"	"231"
"hu"	"1"
"id"	"1423"
"it"	"26"
"ja"	"9861"
"ko"	"691"
"msa"	"14"
"nl"	"61"
"no"	"1"
"pl"	"4"
"pt"	"3977"
"ru"	"396"
"sv"	"2"
"th"	"233"
"tr"	"123"
"ur"	"1"
"zh-cn"	"26"
"zh-tw"	"14"

Explanation:

Each language with number of tweets in each language.

Question 3

User language	fraction of tweets	fraction of users
"ar"	"0.012706422848047461 43"	"0.013537340900112730 88"
"ca"	"0.000633060213069980 28470"	"0.000067445826548604 353146"
"cs"	"0.000018087434659142 293848"	"0.000019270236156744 100899"
"de"	"0.00067827879971783602"	"0.00072263385587790378"
"el"	"0.000018087434659142 293848"	"0.000019270236156744 100899"
"en"	"0.669931448622641850 71"	"0.713740641891566381 15"
"es"	"0.161972977372619241 41"	"0.172564964783643423 55"
"eu"	"0.000009043717329571 146924"	"0.000009635118078372 050449"
"fi"	"0.000009043717329571 146924"	"0.000009635118078372 050449"
"fil"	"0.000081393455966140 322318"	"0.000086716062705348 454045"
"fr"	"0.002089098703130934 94"	"0.002225712276103943 65"
"hu"	"0.000009043717329571 146924"	"0.000009635118078372 050449"
"id"	"0.012869209759979742 07"	"0.013710773025523427 79"
"it"	"0.000235136650568849 82"	"0.000250513070037673 31"
"ja"	"0.089180096586901079 82"	"0.095011899370826789 48"
"ko"	"0.006249208674733662 52"	"0.006657866592155086 86"
"msa"	"0.000126612042613996 06"	"0.000134891653097208 71"
"nl"	"0.000551666757103839 96"	"0.000587742202780695 08"
"no"	"0.000009043717329571 146924"	"0.000009635118078372 050449"
"pl"	"0.000036174869318284 587697"	"0.000038540472313488 201798"
"pt"	"0.035966863819704451 32"	"0.038318864597685644 64"
"ru"	"0.003581312062510174 18"	"0.003815506759035331 98"
"sv"	"0.000018087434659142 293848"	"0.000019270236156744 100899"
"th"	"0.002107186137790077 23"	"0.002244982512260687 75"
"tr"	"0.001112377231537251 07"	"0.001185119523639762 21"
"ur"	"0.000009043717329571 146924"	"0.000009635118078372 050449"
"zh-cn"	"0.000235136650568849 82"	"0.000250513070037673 31"
"zh-tw"	"0.000126612042613996 06"	"0.000134891653097208 71"

Explanation:

Each language with fraction of tweets in each language and fraction of users in each language.

fraction of tweets in each language = No of tweet in a language / Total tweets

fraction of users in each language = No of users in a language / Total users

Retweeting habits

Question 1

" 0.33079204876372384105"

Explanation:

Fraction of tweets are retweeted.

Fraction of tweets Retweeted = No of tweet Retweeted / Total Tweets

Question 2

"0.00109677012472856706"

Explanation:

Average value for retweet per tweet.

Question 3

"0.669207951236276158 95"

Explanation:

Fraction of tweets are not retweeted.

Fraction of tweets not Retweeted = No of tweet not Retweeted / Total Tweets

Question 4

"0.936196574239875558 45"

Explanation:

Fraction of tweets less than the average value of retweet.

Fraction of tweets not Retweeted = No of tweet less than avg value Retweeted / Total Tweets

Hashtags

Question 1

"10158 "

Explanation:

No of distinct tags

Question 2

Tags	**No of item used**
"ReasonsIFailAtBeingAGirl"	"467"
"RED"	"240"
"oomf"	"190"
"HonestyHour"	"172"
"TeamFollowBack"	"139"
"EresGuapaSi"	"130"
"10PeopleYouTrulyLove"	"126"
"TweetLikeAGirl"	"98"
"ImSingleBecause"	"97"
"WeAllGotThatOneFriend"	"96"

Explanation:

Top 10 hash tags frequently used.

Question 3

Explanation:

No of distinct tags

Replies

Question 1

"25776"

Explanation:

query for finding No of replies

Question 2

"0.48169973511461202817656114 60"

Explanation:

probability of two user having same language

= sum ((No of users in each language C 2) / (Total No of users C 2))

ie. 2 represent user1 and user 2 as mentioned in question.

nCr formula is used

example for nCr:

5C2 = (5 * 4) / (2 * 1)

Question 3

"0.48169973511461202817656114 60"

Explanation:

As per my understanding question 2 and question 3 will produce same result

probability of two user having same language

= sum ((No of users in each language C 2) / (Total No of users C 2))

ie. 2 represent user1 and user 2 as mentioned in question.

nCr formula is used

example for nCr:

5C2 = (5 * 4) / (2 * 1)