DB Report

The table provided was not normalized and having following columns

created_at, text, tweet_id, in_reply_to_screen_name, in_reply_to_status_id, in_reply_to_user_id, retweet_count, tweet_source, retweet_of_tweet_id, hashtag1, hashtag2, hashtag3, hashtag4, hashtag5, hashtag6, user_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

There are total five subject areas:

1. Tweet_User
2. Tweet_Reply
3. Tweet_Retweet
4. Tweet_source
5. Tweet_hashtag

For each subject area is required individual table as per normalization rule. To remove redundancy.

However, we can create a table for language but in this situation is not required.

After performing 3rd Normalized from tables which will populate looks like this:

Table Tweet_User

(

user_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

)

TABLE Tweet_Reply

(

tweet_Reply_id,

tweet_id ,

in_reply_to_screen_name,

in_reply_to_status_id,

in_reply_to_user_id

)

Table Tweet_Retweet

(retweet_id,
retweet_of_tweet_id ,
retweet_count,
tweet_id
)

TABLE Tweet_source

(
tweet_id ,
tweet_source,
text,
retweet_id,
tweet_Reply_id,
user_id ,
created_at
)

TABLE Tweet_hashtag

(
hashtagid,
hashtag,
tweet_id

)

Functional Dependencies:

user_id -> user_name , user_screen_name

user_lang , user_created_at - > user_id

user_time_zone -> u user_id

tweet_Reply_id -> tweet_id

retweet_id -> retweet_of_tweet_id

hashtagid-> hashtag

Note: Before loading data into final tables I've removed duplicated while populating data into new tables.

Copy the data provided:

Query Editor Query History Scratch Pad

```
COPY bad_giant_table
-- (created_at, text, tweet_id, in_reply_to_screen_name, in_reply_to_status_id, in_reply_to_user_id, retweet_count, tweet_sourc
FROM 'F:\bad_giant_table.csv'
DELIMITER ','
CSV HEADER;
```

Data Output Explain Messages Notifications

```
COPY 110573

Query returned successfully in 17 secs 141 msec.
```

```
SELECT created_at, text, tweet_id, in_reply_to_screen_name, in_reply_to_status_id, in_reply_to_user_id, retweet_count, tweet_so
    FROM public.bad_shorter_table;
```

Data Output Explain Messages Notifications

	created_at timestamp with time zone	text character varying (255)	tweet_id [PK] bigint	in_reply_to_screen_name character varying (255)	in_reply_to_status_id bigint	in_reply_to_user_id bigint	retweet_count integer	tweet_source character varying (255)	retweet_of_tweet_id bigint
1	2012-10-22 08:00:00+05	Everybody cant be spitting rea...	13890351239168	[null]	[null]	[null]	0	<a href=http://twitter.com/do...	[null]
2	2012-10-22 08:00:00+05	Watching Bad Boys 2 love this...	13890363817984	[null]	[null]	[null]	0	<a href=http://www.lg.com rel...	[null]
3	2012-10-22 08:00:00+05	Balik KL for a day before head...	13890355445762	[null]	[null]	[null]	0	<a href=http://seesmic.com/ r...	[null]
4	2012-10-22 08:00:00+05	Too much on my mind cant sl...	13890355429376	[null]	[null]	[null]	0	<a href=http://twitter.com/do...	[null]
5	2012-10-22 08:00:00+05	Just thinking thoe , cause i mi...	13890372210688	[null]	[null]	[null]	0	<a href=http://twitter.com/do...	[null]
6	2012-10-22 08:00:00+05	tava conversando com a julia ...	13890368012289	[null]	[null]	[null]	0	<a href=http://twitter.com/do...	[null]
7	2012-10-22 08:00:00+05	"@j_hopkins07: #NationalFear...	13890372218880	[null]	[null]	[null]	0	<a href=http://twitter.com/do...	[null]
8	2012-10-22 08:00:00+05	"@FUN: If a mosquito is biting...	13890372231170	[null]	[null]	[null]	0	<a href=http://twitter.com/do...	[null]
9	2012-10-22 08:00:00+05	Mereka mirip _. RT @EXOSuh...	13890380619776	[null]	[null]	[null]	0	<a href=http://bit.ly/fVzRUd re...	[null]
10	2012-10-22 08:00:00+05	RT @ohMG91: I'll save tomorr...	13890359640064	[null]	[null]	[null]	0	<a href=http://twitter.com/do...	[null]
11	2012-10-22 08:00:00+05	@Ivania_Gabby im sorry boob...	13890384809984	Ivania_Gabby	260213647538806784	411004777	0	web	[null]
12	2012-10-22 08:00:00+05	And... its gone!.	13890363826176	[null]	[null]	[null]	0	<a href=http://www.facebook....	[null]
13	2012-10-22 08:00:00+05	@Shelby_St_James: Gonna se...	13890355453954	[null]	[null]	[null]	0	<a href=http://twitter.com/dev...	[null]
14	2012-10-22 08:00:00+05	@haziqzeke eh meet me laaa...	13890368040961	haziqzeke	260213610607955968	44981750	0	<a href=http://www.echofon.c...	[null]

```
SELECT created_at, text, tweet_id, in_reply_to_screen_name, in_reply_to_status_id, in_reply_to_user_id, retweet_count, tweet_so
    FROM public.bad_giant_table;
```

Data Output Explain Messages Notifications

	created_at timestamp with time zone	text character varying (255)	tweet_id [PK] bigint	in_reply_to_screen_name character varying (255)	in_reply_to_status_id bigint	in_reply_to_user_id bigint	retweet_count integer	tweet_source character varying (255)	retweet_of_tweet_id bigint
1	2012-10-22 08:00:57+05	0101 @1dorminhoca te amo !	14129447563264	[null]	[null]	[null]	0	web	
2	2012-10-22 08:00:58+05	@Feffiherrero nunca -	14133658619905	Feffiherrero	260212715052756993	165989321	0	web	
3	2012-10-22 08:01:09+05	Voyyy. RT @FeelLikeAKaren: S...	14179766620162	[null]	[null]	[null]	0	web	
4	2012-10-22 08:01:14+05	#GeorgeStrait	14200767500288	[null]	[null]	[null]	0	web	
5	2012-10-22 08:00:00+05	Start looking towards the sky .	13889281708032	[null]	[null]	[null]	0	web	
6	2012-10-22 08:01:29+05	Cool.	14263665274881	[null]	[null]	[null]	0	web	
7	2012-10-22 08:00:50+05	Its not getting easier	14099873505280	[null]	[null]	[null]	0	web	
8	2012-10-22 08:01:50+05	Un suspiro más y lo aspiro.	14351745671168	[null]	[null]	[null]	0	web	
9	2012-10-22 08:01:49+05	HANNAH HAS 800 FOLLOWE...	14347563933697	[null]	[null]	[null]	0	web	
10	2012-10-22 08:01:53+05	hush, hush.	14364307603456	[null]	[null]	[null]	0	web	
11	2012-10-22 08:01:53+05	andre 3000 kill errr thing	14364328558593	[null]	[null]	[null]	0	web	
12	2012-10-22 08:01:57+05	editando. o.o	14381114171393	[null]	[null]	[null]	0	web	
13	2012-10-22 08:02:03+05	RT @Cydiapomelia: http://t.co...	14406271614976	[null]	[null]	[null]	4	<a href="http://blackberry.co...	26021271634459
14	2012-10-22 08:02:08+05	01i01 meu amor ♥	14427259920384	[null]	[null]	[null]	0	web	
15	2012-10-22 08:02:09+05	ya wassalam	14431437422592	[null]	[null]	[null]	0	<a href="http://twitter.com/de...	

SQL Queries:

1

```
a) How many tweets are there in total?

    Select
    count(*) as total_tweet
    from Tweet_source;
```

Data Output | Explain | Messages | Notifications

	total_tweet bigint
1	43876

2

```
b) How are these tweets distributed across languages? Write a query that shows,
 for every  language ( user_lang ) the number of tweets in that language.

    Select
    user_lang,
    count(*) as total_tweet_Per_Language
     from Tweet_user
    group by 1;
```

Data Output Explain Messages Notifications

	user_lang character varying (10)	total_tweet_per_language bigint
1	fr	221
2	tr	121
3	en	69279
4	fi	1
5	ru	382
6	cs	2
7	ja	9596
8	sv	2
9	fil	9
10	eu	1
11	de	71
12	nl	60
13	id	1360
14	zh-cn	21
15	zh-tw	14
16	ar	1328
17	no	1
18	th	217

3

```
c) compute, for each language, the fraction of total tweets that have that language setting, as well as the fraction of the number of users that have that language

Select
    user_lang,t.user_id , (count(user_lang) * 1.0/ (select count(*) from Tweet_source) ) as Fraction_user_lang,
     (count(t.user_id) * 1.0/ (select count(*) from Tweet_source) ) as Fraction_user
     from Tweet_source t inner join Tweet_user u
     on t.user_id = u.user_id
    group by 1,2;
```

Data Output Explain Messages Notifications

	user_lang character varying (10)	user_id integer	fraction_user_lang numeric	fraction_user numeric
1	en	226803667	0.0000227915033275594858 24	0.0000227915033275594858 24
2	ja	170019464	0.0000227915033275594858 24	0.0000227915033275594858 24
3	pt	310658484	0.0000227915033275594858 24	0.0000227915033275594858 24
4	es	275738606	0.0000227915033275594858 24	0.0000227915033275594858 24
5	en	232351994	0.0000227915033275594858 24	0.0000227915033275594858 24
6	en	51323231	0.0000227915033275594858 24	0.0000227915033275594858 24
7	ja	152568208	0.0000227915033275594858 24	0.0000227915033275594858 24
8	en	26862229	0.0000227915033275594858 24	0.0000227915033275594858 24
9	en	166209840	0.0000227915033275594858 24	0.0000227915033275594858 24
10	en	58362155	0.0000227915033275594858 24	0.0000227915033275594858 24
11	en	361627962	0.0000227915033275594858 24	0.0000227915033275594858 24
12	en	239788600	0.0000227915033275594858 24	0.0000227915033275594858 24
13	en	29499656	0.0000227915033275594858 24	0.0000227915033275594858 24
14	en	228236772	0.0000227915033275594858 24	0.0000227915033275594858 24
15	en	244262497	0.0000227915033275594858 24	0.0000227915033275594858 24
16	en	106826585	0.0000227915033275594858 24	0.0000227915033275594858 24
17	en	59858229	0.0000227915033275594858 24	0.0000227915033275594858 24

4

```sql
a) What fraction of the tweets are retweets ?

Select
    t.tweet_id,
    (count( rt.tweet_id) * 1.0/ (select count(*) from Tweet_source) ) as Fraction_Tweet
     from Tweet_source t
    inner join Tweet_Retweet rt
     on t.retweet_id = rt.retweet_id
    group by 1
```

Data Output Explain Messages Notifications

	tweet_id [PK] bigint	fraction_tweet numeric
1	260214557283336192	0.00002279150332755948582 4
2	260219334582992896	0.00002279150332755948582 4
3	260216721514852352	0.00002279150332755948582 4
4	260219468796530688	0.00002279150332755948582 4
5	260221456879857665	0.00002279150332755948582 4
6	260217409393274880	0.00002279150332755948582 4
7	260217124168036352	0.00002279150332755948582 4
8	260220852929433600	0.00002279150332755948582 4
9	260215245136621568	0.00002279150332755948582 4
10	260219997274656768	0.00002279150332755948582 4
11	260217317135360000	0.00002279150332755948582 4
12	260214066545586176	0.00002279150332755948582 4
13	260217702973571072	0.00002279150332755948582 4
14	260214725038727168	0.00002279150332755948582 4
15	260219540091326464	0.00002279150332755948582 4
16	260214804734685188	0.00002279150332755948582 4

5

```
b) Compute the average number of retweets per tweet.


Select
    t.tweet_id,
    avg( rt.retweet_count) as Avg_Retweet
     from Tweet_source t
    inner join Tweet_Retweet rt
     on t.retweet_id = rt.retweet_id
    group by 1
```

Data Output | Explain | Messages | Notifications

	tweet_id [PK] bigint	avg_retweet numeric
1	260214557283336192	1.00000000000000000000
2	260216721514852352	1.00000000000000000000
3	260219468796530688	1.00000000000000000000
4	260221456879857665	0.00000000000000000000
5	260217409393274880	1.00000000000000000000
6	260217124168036352	2.0000000000000000
7	260220852929433600	9.0000000000000000
8	260215245136621568	1.00000000000000000000
9	260219997274656768	1.00000000000000000000
10	260217317135360000	1.00000000000000000000
11	260214066545586176	0.00000000000000000000
12	260217702973571072	1.00000000000000000000
13	260214725038727168	0.00000000000000000000
14	260219540091326464	2.0000000000000000
15	260214804734685188	10.0000000000000000
16	260214544700432384	3.0000000000000000

6

```sql
c)  What fraction of the tweets are never retweeted?

Select
    t.tweet_id,
    (count( t.tweet_id) * 1.0/ (select count(*) from Tweet_source) ) as Fraction_Tweet
     from Tweet_source t
    left  join Tweet_Retweet rt
     on t.retweet_id = rt.retweet_id
     where t.retweet_id is  null
    group by 1
```

Data Output Explain Messages Notifications

	tweet_id [PK] bigint	fraction_tweet numeric
1	260215006052876288	0.0000227915033275594858 24
2	260217652667092993	0.0000227915033275594858 24
3	260216780260274176	0.0000227915033275594858 24
4	260215683760140291	0.0000227915033275594858 24
5	260214263677845504	0.0000227915033275594858 24
6	260214435635933184	0.0000227915033275594858 24
7	260219183583866880	0.0000227915033275594858 24
8	260220056015884288	0.0000227915033275594858 24
9	260219734212091904	0.0000227915033275594858 24
10	260214653714587648	0.0000227915033275594858 24
11	260219686917128193	0.0000227915033275594858 24
12	260217921098362880	0.0000227915033275594858 24
13	260215064773160960	0.0000227915033275594858 24
14	260215127704494080	0.0000227915033275594858 24
15	260221301724168192	0.0000227915033275594858 24

7

```sql
d) What fraction of the tweets are retweeted fewer times than the average number of retweets  (and  what does this say about the distribution)?

select TT.tweet_id,
TT.Fraction_Tweet,
TT.Avg_Retweet from
(
Select
    t.tweet_id,
    (count( rt.tweet_id) * 1.0/ (select count(*) from Tweet_source) ) as Fraction_Tweet,
    avg( rt.retweet_count) as Avg_Retweet
     from Tweet_source t
    inner join Tweet_Retweet rt
     on t.retweet_id = rt.retweet_id
    group by 1
    ) TT
    where TT.Fraction_Tweet < TT.Avg_Retweet

-----------------------------------------------------------------------------------------------------------
```

Data Output Explain Messages Notifications

	tweet_id [PK] bigint	fraction_tweet numeric	avg_retweet numeric
1	260213890351239169	0.00002279150332755948582 4	106.0000000000000000
2	260213890351239170	0.00002279150332755948582 4	1.00000000000000000000
3	260213890351255552	0.00002279150332755948582 4	5027.0000000000000000
4	260213890355453952	0.00002279150332755948582 4	112.0000000000000000
5	260213890359635968	0.00002279150332755948582 4	1.00000000000000000000
6	260213890368012291	0.00002279150332755948582 4	2.0000000000000000
7	260213890368016384	0.00002279150332755948582 4	7.0000000000000000
8	260213890368024576	0.00002279150332755948582 4	1.00000000000000000000
9	260213890368040962	0.00002279150332755948582 4	80.0000000000000000
10	260213890372222976	0.00002279150332755948582 4	30.0000000000000000

8

```
a) What is the number of distinct hashtags found in these tweets?


Select distinct hashtag
from  Tweet_hashtag
```

Data Output | Explain | Messages | Notifications

	hashtag character varying (150)
1	dontwakemeup
2	spyair
3	غادة_شبير
4	IDX
5	burkesquad
6	stress
7	believetour
8	TmIt
9	DOIT
10	10FactsAboutMe
11	cmnews
12	GospelRap
13	Madvanna
14	Swacked
15	nottacos
16	Tyga

```
b) What are the top ten most popular hashtags, by number of usages?


Select hashtag,
count( hashtag ) as Top_Hashtags
from  Tweet_hashtag
group by 1
order by 2 desc limit 10
```

Data Output | Explain | Messages | Notifications

	hashtag character varying (150)	top_hashtags bigint
1	ReasonsIFailAtBeingAGirl	467
2	RED	205
3	oomf	188
4	HonestyHour	172
5	TeamFollowBack	139
6	EresGuapaSi	130
7	10PeopleYouTrulyLove	126
8	TweetLikeAGirl	98
9	ImSingleBecause	97
10	WeAllGotThatOneFriend	96

10

```
c) Write a query giving, for each language, the top three most popular hashtags in that language.

Select hashtag,user_lang, count( hashtag )
from  Tweet_hashtag
join tweet_source
on  Tweet_hashtag.tweet_id = tweet_source.tweet_id
join tweet_user
on tweet_source.user_id = tweet_user.user_id
group by 1,2
order by 3 desc limit 3
```

Data Output Explain Messages Notifications

	hashtag character varying (150)	user_lang character varying (10)	count bigint
1	ReasonsIFailAtBeingAGirl	en	184
2	RED	en	84
3	MagicMike	en	78

11

```
a)  How many tweets are neither replies, nor replied to?

select count(*) as  count_Noreply
from tweet_source
where tweet_Reply_id is null
```

Data Output Explain Messages Notifications

	count_noreply bigint
1	21438

```
b) If a user user1 replies to another user user2 , what is the probability that they have the same  language setting?

Select   user_lang,
COUNT(user_lang)*1.0/(SELECT count(tweet_id) FROM tweet_source) as Prob_UselangSetting
from tweet_source as  ts
JOIN tweet_reply as tu
on ts.tweet_Reply_id= tu.tweet_Reply_id
join tweet_user as tu1
on ts.user_id=tu1.user_id
group by 1
```

Data Output Explain Messages Notifications

	user_lang character varying (10)	prob_uselangsetting numeric
1	ar	0.00423921961892606436
2	ca	0.000022791503327559485824
3	de	0.00027349803993071383
4	en	0.32728598778375421643
5	es	0.08601513355820949950
6	fil	0.000045583006655118971647
7	fr	0.00109399215972285532
8	id	0.00556112681192451454
9	it	0.00013674901996535691
10	ja	0.06212963807092715836
11	ko	0.00394293007566779105
12	msa	0.000045583006655118971647
13	nl	0.00025070653660315434
14	no	0.000022791503327559485824

13

```
c) How does this compare to the probability that two arbitrary users have the same language  setting?
Throughout, you may create views that support your queries.

WITH lang_setting AS(
Select   ts.user_id,
COUNT(user_lang)*1.0/(SELECT count(tweet_id) FROM tweet_source) as Prob_UselangSetting
from tweet_source as  ts
JOIN tweet_reply as tu
on ts.tweet_Reply_id= tu.tweet_Reply_id
join tweet_user as tu1
on ts.user_id=tu1.user_id
AND tu1.user_lang = tu1.user_lang
group by 1
)
Select * from lang_setting ORDER BY RANDOM() LIMIT 2
```

Data Output Explain Messages Notifications

	user_id integer	prob_uselangsetting numeric
1	548028773	0.00002279150332755948582 4
2	746909557	0.00002279150332755948582 4