Yokesh Vishwanathan Suresh Kumar
675968182
yvishw2@uic.edu

ADVANCE DATABASE MANAGEMENT SYSTEM – OPTIONAL ASSIGNMENT 2

Deliverable:

1. What is your chosen theme?

A: I have taken popular soccer players as my theme of observation. Please find below details of the 10 soccer players I have taken into observation.

David De Gea - @d_degea => 265982289

Cristiano Ronaldo - @cristiano => 155659213

Neymar Jr - @neymarjr => 158487331

Anthony Martial - @anthonymartial => 1561168178

Romelu Lukaku - @romelulukaku9 => 1137984020

Eden hazard - @hazardeden10 => 366592246

Gerard Pique - @3gerardpique => 224223563

Luiz Suarez - @luissuarez9 => 213745334

Ricardo kaka - @kaka => 60865434

Wayne Rooney - @waynerooney => 285332860

For Question 2 and 3, it can be better understood if we have a look at the schema design of the tables in our usertweetanalysis database.

So basically, the process requires the user ids from userid database [which needs to be fed prior to initiating the process] and accordingly the twitter stream is synced to the local database I had created for this purpose. One more thing to be noted is, different tables are created based on the date of stream sync that is happening.

Following are the data fields in the tables of usertweetanalysis database:

Table: userid

Columns:

userID	int(45) PK

Table: tweetsuser_10_6_2017

Columns:

id	int(11)
textOfTweet	varchar(150)
userScreenName	varchar(30)
followers	varchar(30)
isRetweet	varchar(5)
statusId	varchar(25)
userId	varchar(20)
timestamp	varchar(60)

Yokesh Vishwanathan Suresh Kumar
675968182
yvishw2@uic.edu

2. What data fields about the users are available in the database?

A: With respect to the user, following details are available:

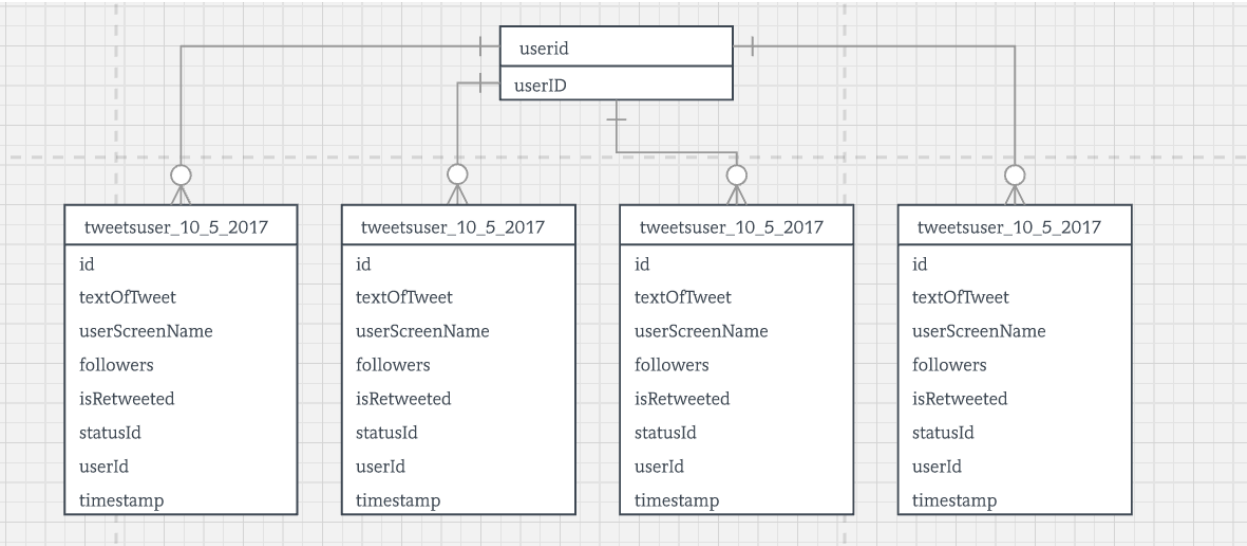
- **userID** – Every twitter user has a twitter ID on top of twitter handle. This information can be found in here. [This attribute can be found in userid table as well as all the tweet tables created]
- **userScreenName** – The screen name of the user is the one that is displayed for every user in twitter. This information is also synced and it can be found in tweet tables that are created.
- **followers** – The count of followers of the user who has tweeted. This can be found in tweet table that gets generated.

3. What data fields about individual tweets are available?

A: The following data files are available with respect to the tweets that are collected –

- **textOfTweet** – The entire text message for that tweet that has been synced.
- **isRetweet** – This is a flag that states where that tweet is just a retweet or a new tweet message.
- **statusId** – The status id is a unique id for a user that gets generated whenever a user makes a tweet. By using status id and user id, we can directly open a tweet using the link below:
  Format - https://twitter.com/<userID>/status/<statusId>
  E.g.: https://twitter.com/50907102/status/916359875470479360
- **timestamp** – This attribute is associated with the time the tweet has been posted.

4. E-R Diagram of User Tweet Analysis DB –



Let's look at the tweet user tables that gets generated each day for the tweets that are streamed. Each row has a unique id which is stored in id column of the table. The other attributes like textOfTweet, userScreenName, followers, isRetweeted, statusId, timestamp are all dependent on userId column. Here id column can be the primary key and since other non-prime attributes are dependent on userId which is another non-prime attribute, this table is in 2NF.

Yokesh Vishwanathan Suresh Kumar
675968182
yvishw2@uic.edu

5. Summary –

Informative Queries:

1. To get the number of tweets about Gerard Pique on complete available data set,

   ```
   select count(*) from (SELECT * FROM usertweetanalysis.tweetsuser_10_6_2017 union select * from usertweetanalysis.tweetsuser_10_5_2017 union SELECT * FROM usertweetanalysis.tweetsuser_10_7_2017 union select * from usertweetanalysis.tweetsuser_10_8_2017) s where s.textOfTweet like '%pique%';
   ```

2. To display all the tweets posted by sports persons we are trying to sync the tweet data,

   ```
   select  * from (SELECT * FROM usertweetanalysis.tweetsuser_10_6_2017 union select * from usertweetanalysis.tweetsuser_10_5_2017 union SELECT * FROM usertweetanalysis.tweetsuser_10_7_2017 union select * from usertweetanalysis.tweetsuser_10_8_2017) s where s.userId in (select * from userid);
   ```

3. To check how many tweets are just retweeted and how many are actually posted about a particular player,

   ```
   select  s.isRetweet, count(*) from (SELECT * FROM usertweetanalysis.tweetsuser_10_6_2017 union select * from usertweetanalysis.tweetsuser_10_5_2017 union
   SELECT * FROM usertweetanalysis.tweetsuser_10_7_2017 union select * from usertweetanalysis.tweetsuser_10_8_2017) s where s.textOfTweet like '%cristiano%' group by s.isRetweet;
   ```

To analyse the twitter data I did the following steps:

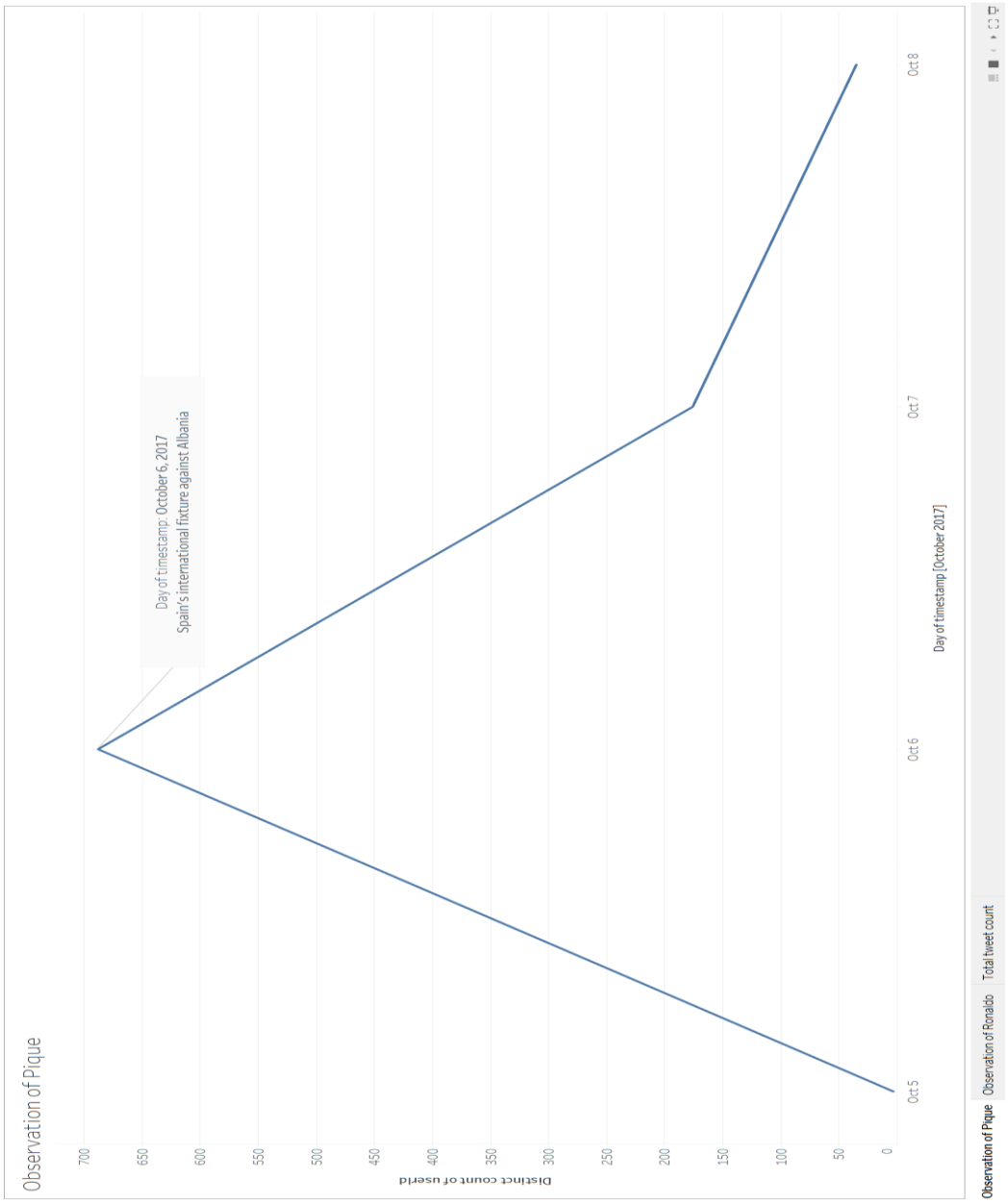
1. Using the java application provided by professor, I configured it to point to my local database system.
2. Then for the next four days, I streamed the twitter chatters for the selected 10 sports person and it was stored in usertweetanalysis DB.
3. Each day, the system was able to stream the twitter chatters for a time frame of 1 – 2 hours [varying time slot].
4. Using tableau, I was able to visually present some of the analysis done on the available data.
5. Three different sheets are created in tableau to show some of the observations made with the available data.

Yokesh Vishwanathan Suresh Kumar
675968182
yvishw2@uic.edu

Observation 1 –

The user set consists of famous soccer players with few of them involving in International soccer match scheduled between the time frame I have selected [Oct. 5 – Oct. 8] and some of them not involved in the international fixtures and few of them who are already retired from International games. This data set was to better understand how the players are on trend based on their profession.

Recently Pique was involved in a controversy as he raised his voice regarding Catalonia referendum which would have cost his Spain career. On Oct. 6th Spain was involved in International game against Albania and Pique was part of the game. This led to a spike of tweets regarding him.

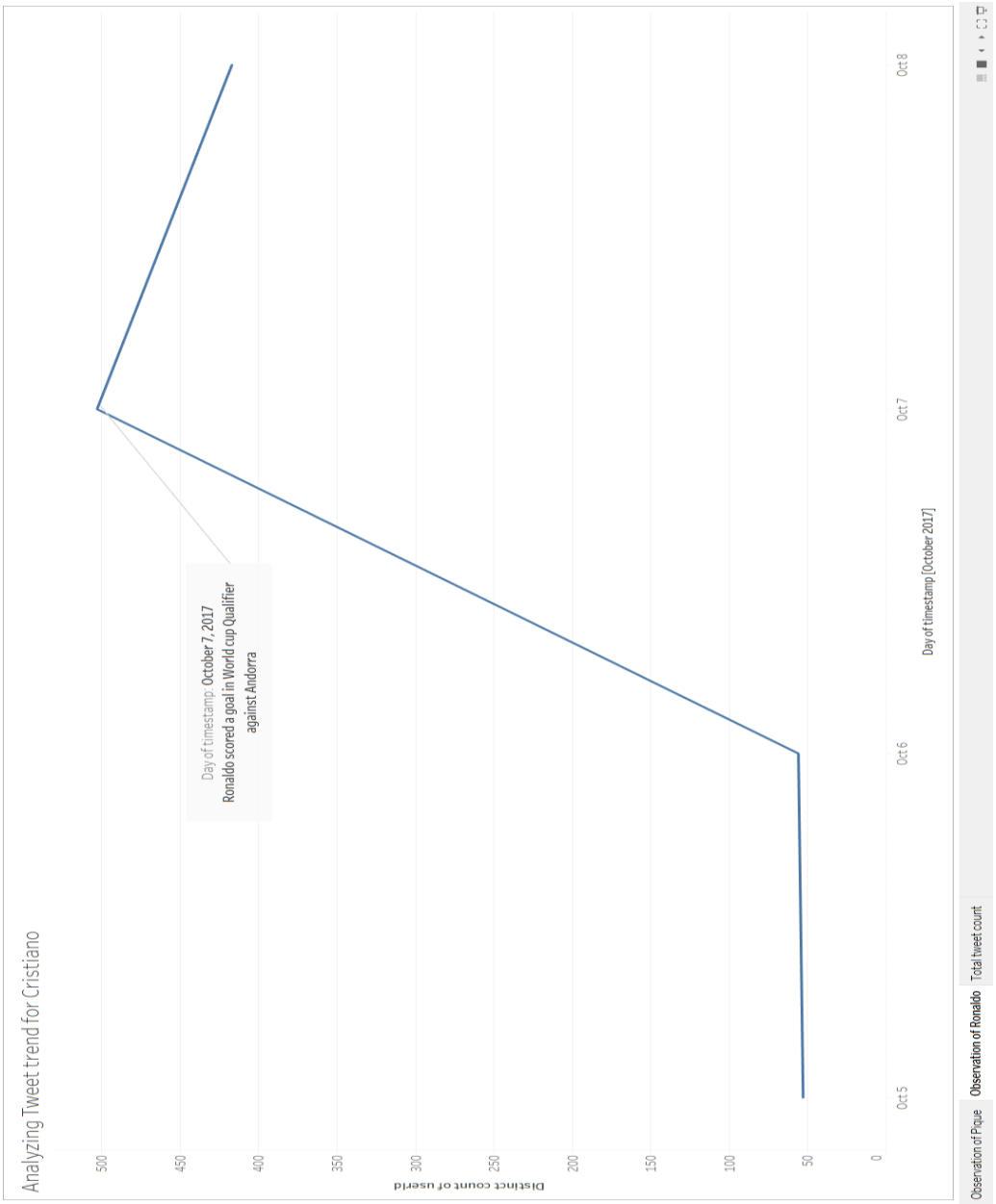


Yokesh Vishwanathan Suresh Kumar
675968182
yvishw2@uic.edu

Observation 2 –

Another observation is of World's famous footballer Cristiano Ronaldo who was also involved in International fixture against Andorra for Portugal on Oct.7th which led to a spike in his tweet chatter count.

Ronaldo scored a goal in this game and won the game for Portugal and this can be seen spike continuing to next day as well.



Yokesh Vishwanathan Suresh Kumar
675968182
yvishw2@uic.edu

Overall Observation –

I have plotted the tweet count for 10 different players on 4 sample days and we can see that, the players are on trend if they are on a game, score a goal or involve in some controversy. People talk about them. Some of the footballers who were once world famous, are not trending on a present day if they are not involved in a game or in some interesting activity. People need some activity happening for them to talk about a person, a celebrity, a sports person.