Assignment 1

By

Rachael Joan Dias

0651897

Addressed to,

Professor Brian Srivastava

Trent University

AMOD-5410H-A-2019GW-PTBO Big Data

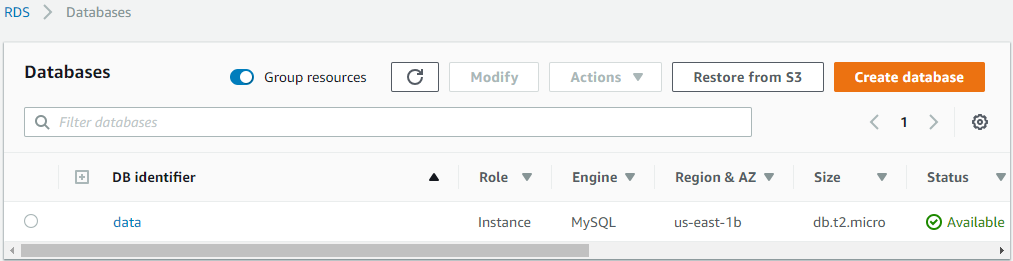
**Objective:**

The main goal of this assignment is to demonstrate data gathering from 2 different source using APIs, storage and retrieval on a MySQL database. In order to demonstrate this, I extracted tweets containing information related to cryptocurrencies which I was then able to link to the actual market price of the cryptocurrencies.

Streaming tweets related to cryptocurrencies such as Bitcoin, Ripple, Ethereum etc. were collected using Twitter’s streaming API and stored on a MySQL database on a Linux server on AWS. Similarly, from another API (coinmarketcap.com) data indicating the latest price and ranking of the cryptocurrencies were stored on the AWS MySQL database.

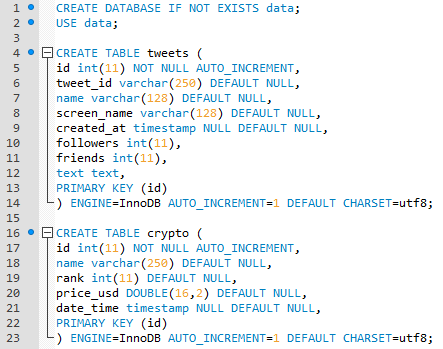
**Step 1: Database Creation**

I first created a database “data” on AWS as you see from the below screenshot.

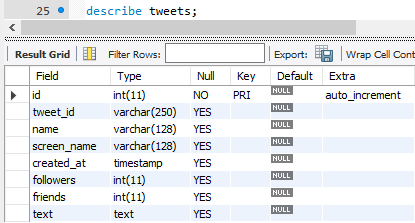


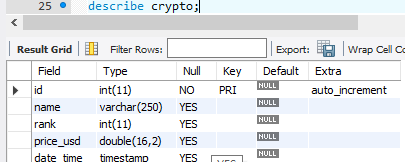
**Step 2: Table Structures/Schemas**

In the next step, I created tables for the data from Twitter and coinmarketcap.com. Below is a screenshot of the DDL statements used to create the tables. “Tweets” for twitter data and “crypto” for coinmarketcap.com data.



Validation of the table structures

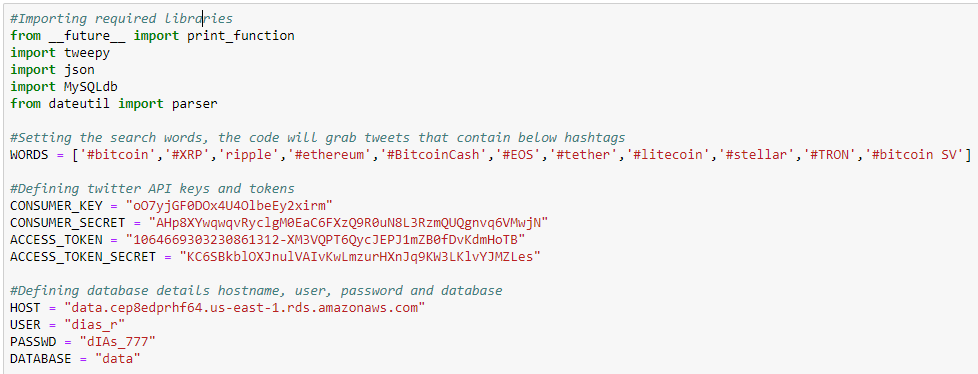


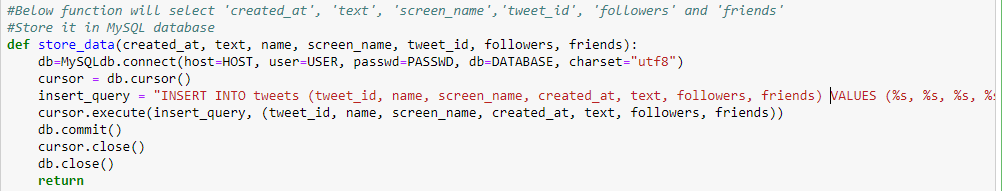
****

**Step 3: Data Gathering**

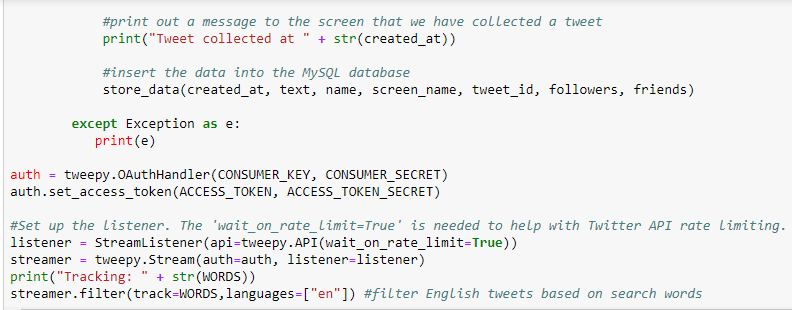
**Twitter Data:**

Streaming tweets from twitter containing words such as #Bitcoin, #Ripple, #Ethereum etc. were collected and inserted into a MySQL database using below Python code

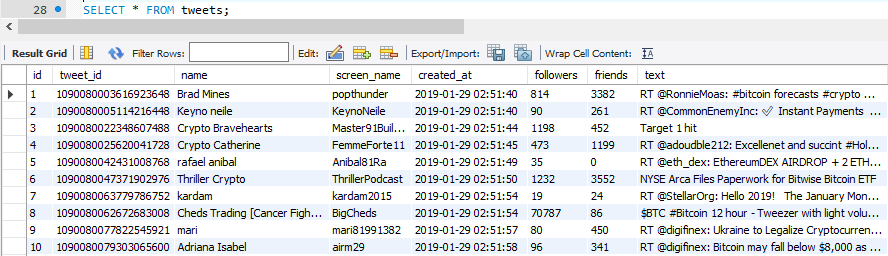








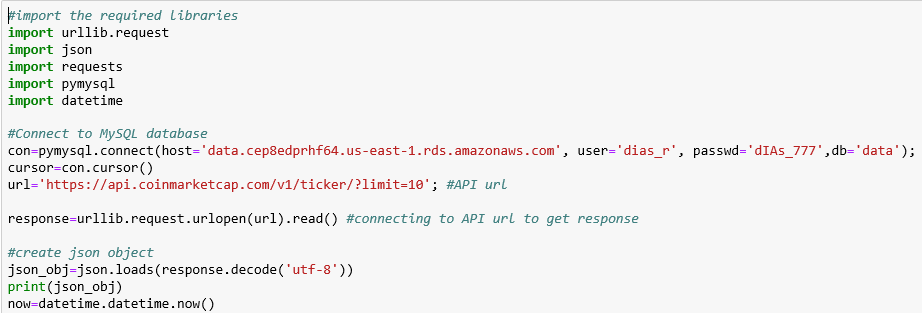
The above Python script was executed for a couple of minutes and tweets were loaded into the “tweets” table as you can see from the below screenshot.



Please see attached script for twitter: twitter.ipynb

**Coinmarketcap:**

Coinmarketcap.com keeps track of the latest value of a cryptocurrency, I extracted the top 10 cryptocurrencies and noted time when they were collected. The code below collects data through the API and inserts the records into the “crypto” table created on the AWS MySQL database.

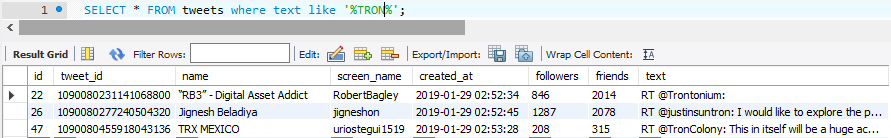


Please see attached script for coinmarketcap.com: coinmarket.ipynb

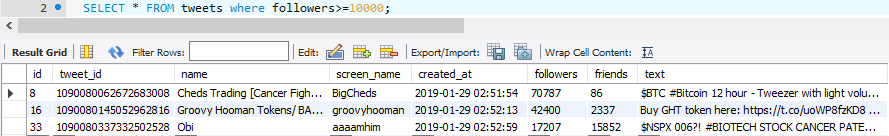
**Step 4: Testing select queries on both the tables “tweets” and “crypto”**

The tweets table contains 8 columns the first column “id” is the primary key which is generated while streaming tweets and loading them into the database.

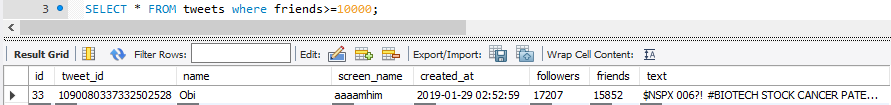
Selecting the tweets which contain “TRON” the text column contains the tweet message using the below query we can filter out tweets based on certain keywords. Only 3 people tweeted about the TRON cryptocurrency

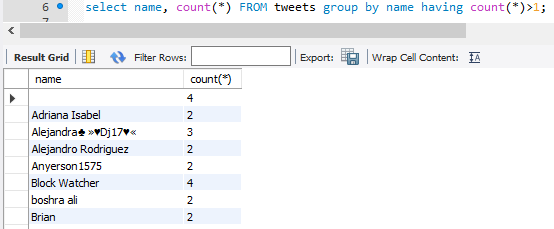
****

From the data collected there are only 3 users which have more than 10000 followers

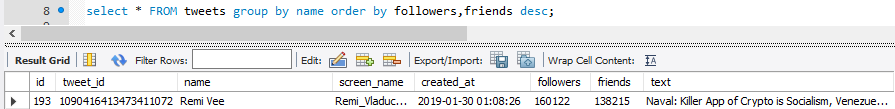


Only one user has more than 10000 friends

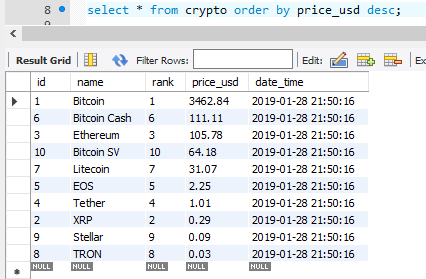


Below is a list of users that tweeted more than once from the data gathered

User Remi Vee has the greatest number of friends and followers

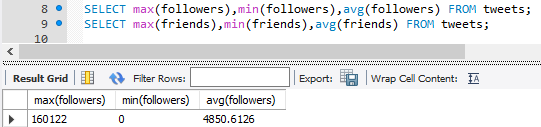


Below is the list of the top 10 cryptocurrencies as you can see Bitcoin has the highest price followed Bitcoin Cash and Ethereum. The rank column indicates the popularity of the cryptocurrency.

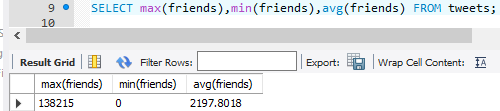


**Step 5: Basic functions (MIN, MAX, AVG)**

Finding the maximum, minimum and average number of followers

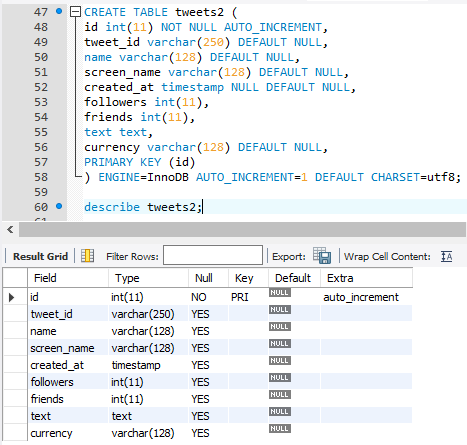
****

Finding the maximum, minimum and average number of friends

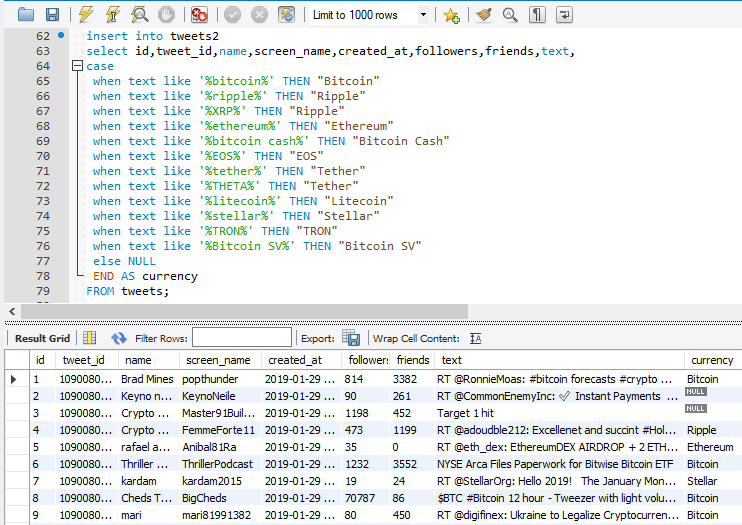
****

**Step 6: JOINS (INNER, LEFT, RIGHT, FULL)**

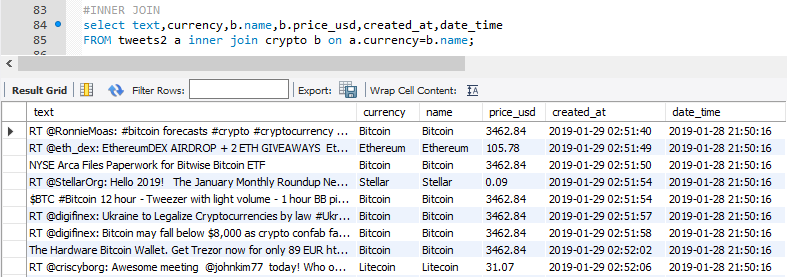
In order to link the “tweets” and “crypto” table I created a new table “tweets2” which indicates the crypto currency that was tweeted



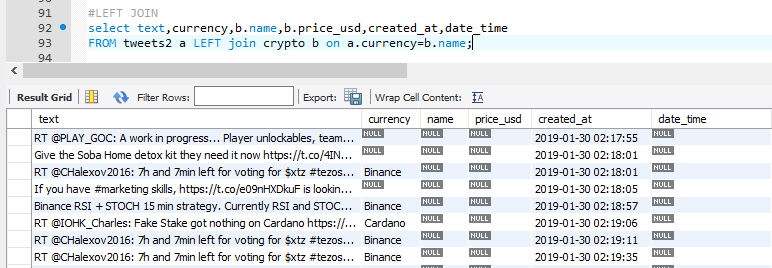
Now the currency column will contain the currency depending on the currency tweeted. Below query was executed to identify and insert the currency based on tweets.



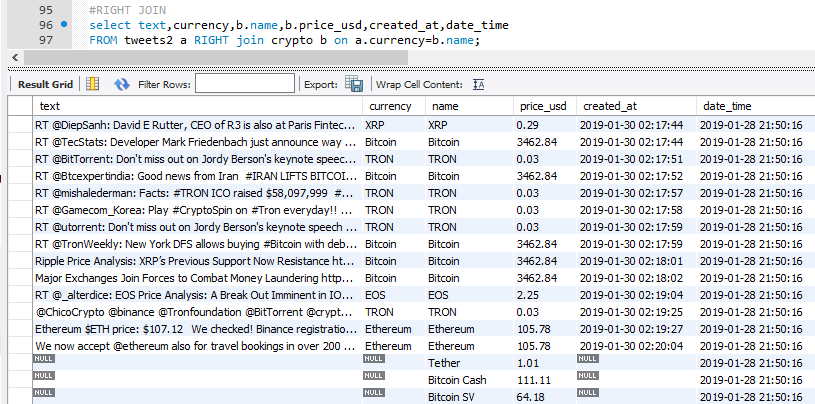
**Inner Join:** Inner join tweets which contain same currencies in the crypto table. All the matching rows between “tweets2” and “crypto” are returned in other words if the currency column of “tweets2” matches the name of the “crypto” table those tweets are returned.



**Left Join:** To demonstrate a left join I pulled tweets related to currencies such as Binance and Cardano. These currencies are not present in the crypto table since they are not in the top 10. The query will return all the rows present in the left table “tweets2” even if they are not present in the “crypto” table. As we can see from the below screenshot for Cardano and Binance the text column and currency columns are returned from the “tweets2” table while all the columns of the crypto table are populated with Null values.



**Right Join:** Like the left join query the result of this query returned all the records from the right table “crypto” and only records that matched from the left table “tweets2”. We can see that none of the tweets contained Tether, Bitcoin Cash and Bitcoin SV hence those values were populated with Null for the left table.



**Conclusion:** To summarize, I was able to extract tweets which contained keywords related to cryptocurrencies and load them into a MySQL table “tweets”. I also extracted data from coinmarketcap.com which publishes the latest value of a cryptocurrency and inserted the top 10 cryptocurrencies into the table “crypto”. Since it was difficult to link the cryptocurrencies with the tweets because neither of the table have much in common other than the keywords, I created a new table “tweets2”. Tweets2 identifies tweets based on the cryptocurrency in the tweet, I then joined the tables based on the cryptocurrency tweeted.