**CS5540**

**PRINCIPLES OF BIG DATA MANAGEMENT**

**PHASE 2 REPORT**

**Team Member:**

**Anh Tuan Nguyen (18142335)**

**Mohit Sriram Tirumala (16292490)**

**Sai Haneesh Tanneru (16290739)**

**Link**

Tweets data: <https://drive.google.com/file/d/1wrLtoVkWjBIHMjNZ26I7HzOMg9i1AZQA/view?usp=sharing>

Github:

<https://github.com/CS5540/Project>

**OBJECTIVE**

* Stream tweets from Twitter (we collected more than 200,000 tweets).
* Write at least 10 analytic queries to explore and understand the collected tweets.
* Create Visualisations on the results.

**SOFTWARES PLANNING TO USE**

* Apache Spark
  + Spark Streaming to collect tweets.
  + SparkSQL to store and execute queries.
* Apache Zeppelin to execute code and create visualisations.

**IMPLEMENTATION**

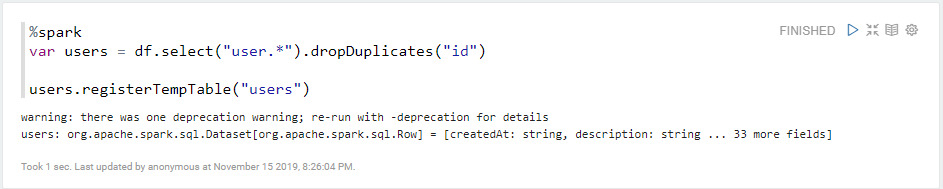
* Initially, we wrote a Spark program to stream the tweets and save them into JSON format, the output of the program contains the tweets with all the details like the IDs, Hashtags, Text, Users’s information, etc.
* The extracted JSON tweets are persisted into the Apache SparkSQL in the form of tables. We used 2 tables to manage the data:
  + Table “Tweets”: store every information of tweets.
  + Table “User”: only contains users’s information who posted the collected tweets.
* We uesd Apache Zeppelin to write queries and visualize the outputs using tables and charts (pie chart and bar chart).

1. Load data

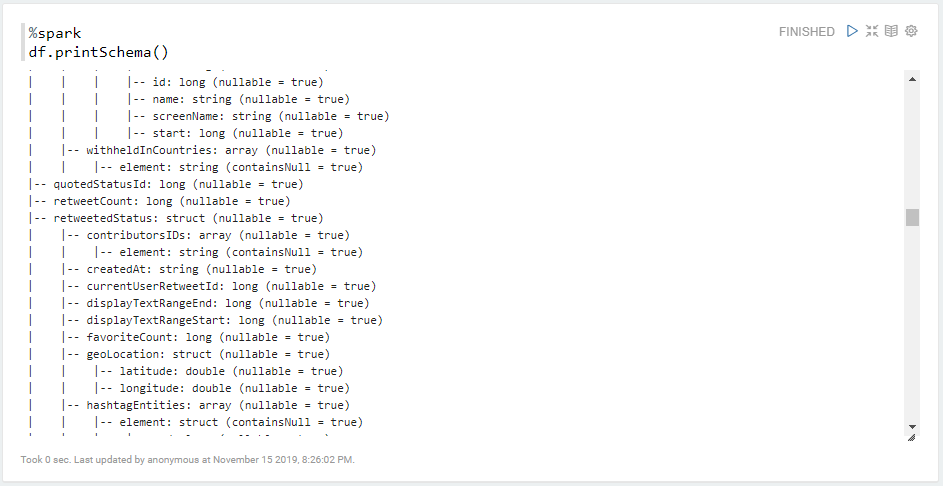
We read tweets from json files. The folders have the name format: “tweets-timestamp” and the files’s name format are “part-\*\*\*\*\*”. Then, we created a table “tweets” to store data.

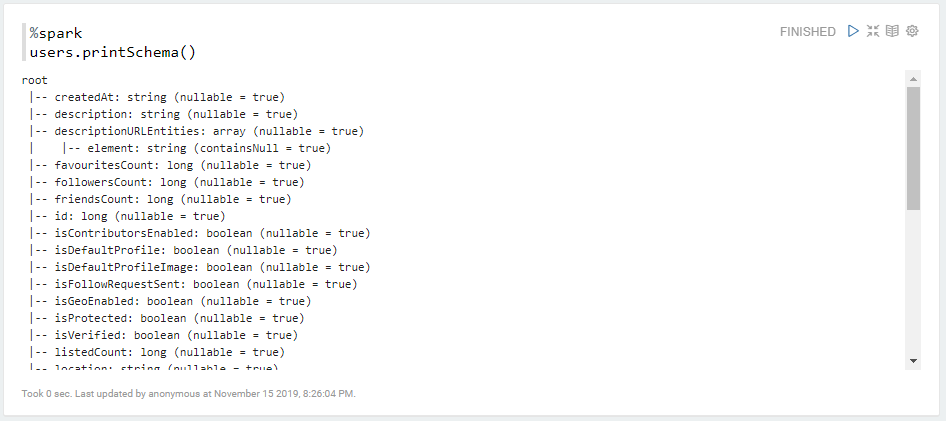


We also created a table “users” to store users’s information.



1. Print the schema

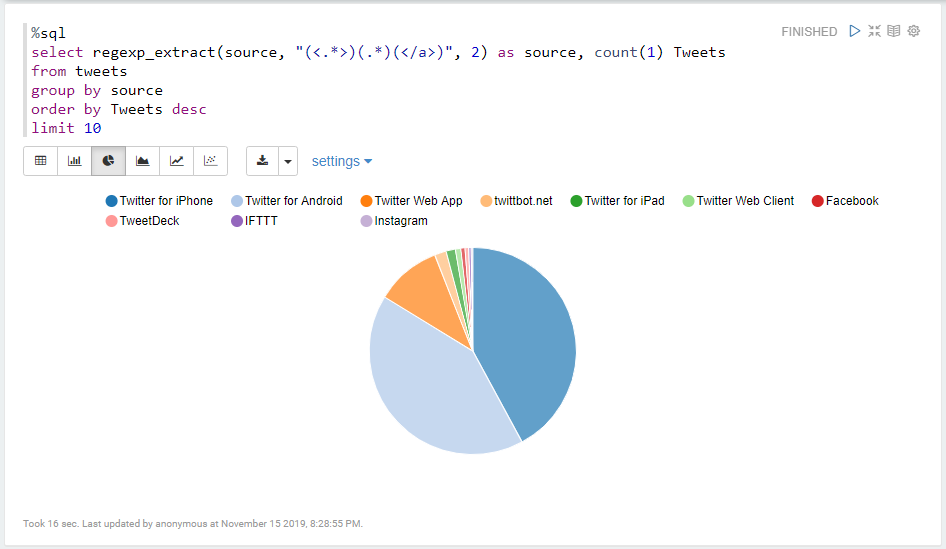




**QUERIES**

We have executed and visualized totally 14 queries.

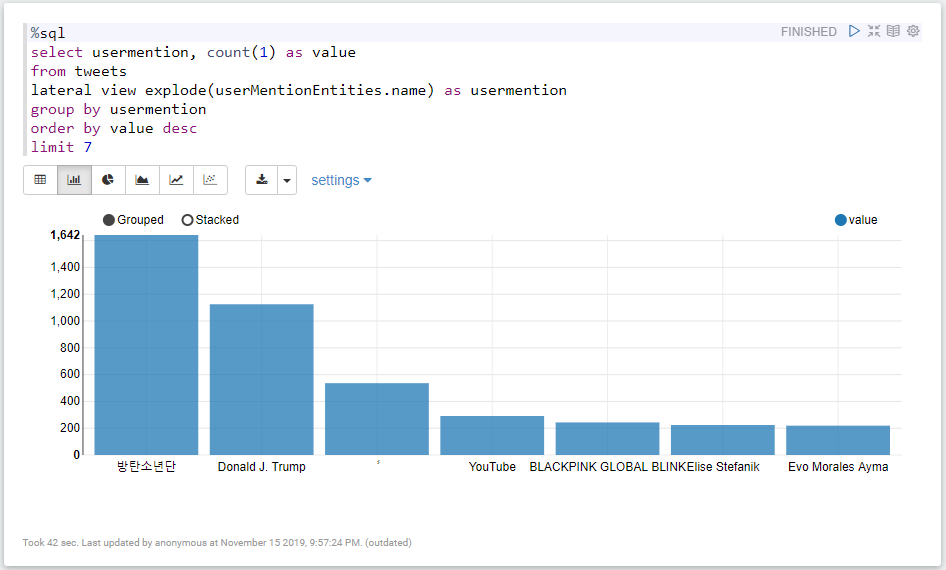
1. Query to get the top 10 sources.



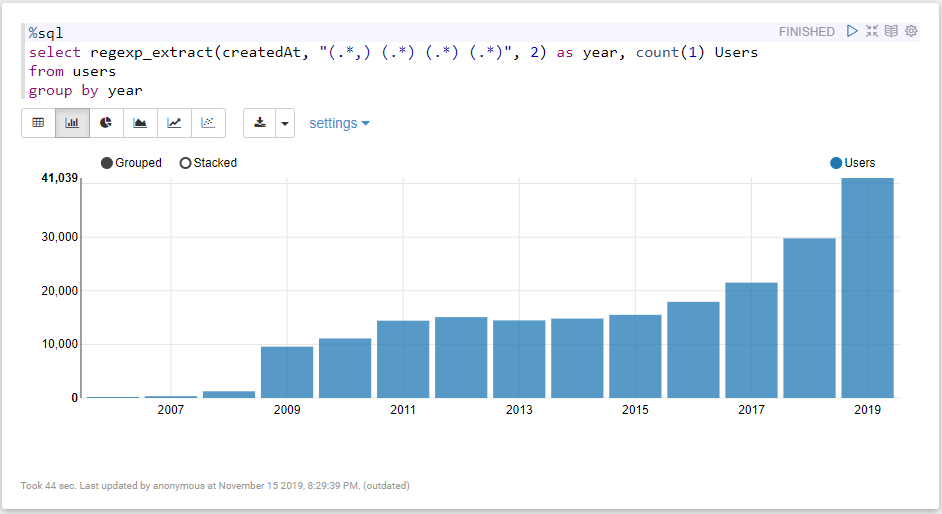
1. Query to get the most popular Hashtags.



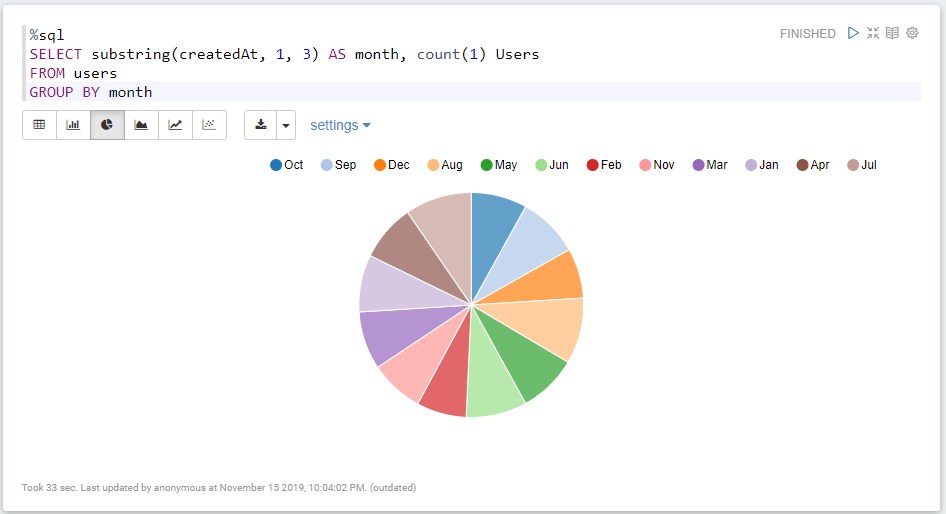
1. Query to get the most mentioned accounts.



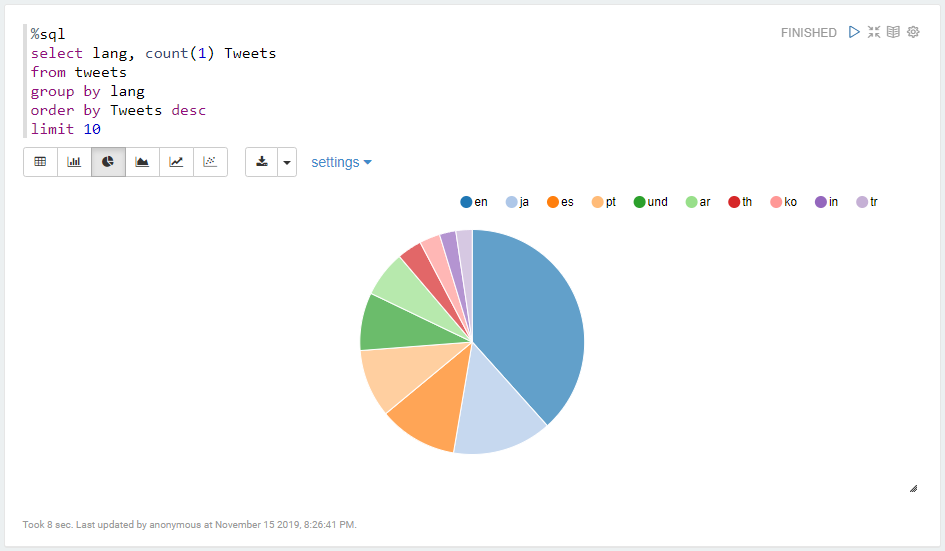
1. Query to get number of Twitter users created in years.



1. Query to get number of Twitter users created in months.



1. Query to get the 10 most used languages.



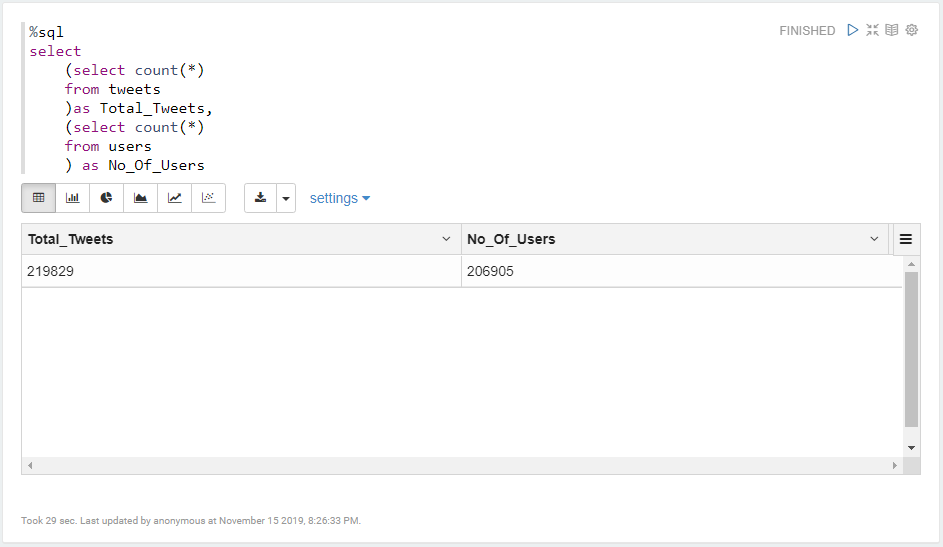
1. Query to get the top users with the most friends.



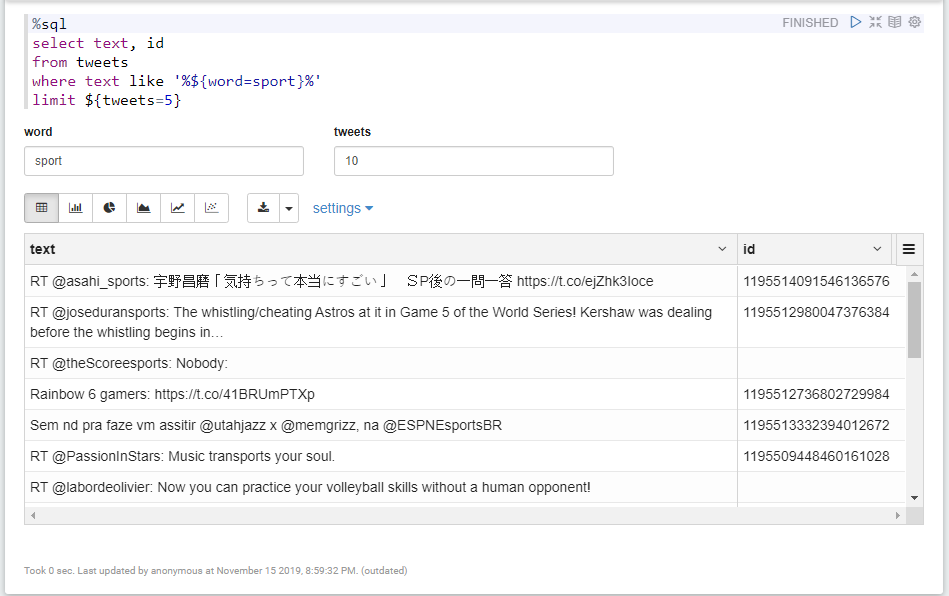
1. Query to get the top users with the most followers.



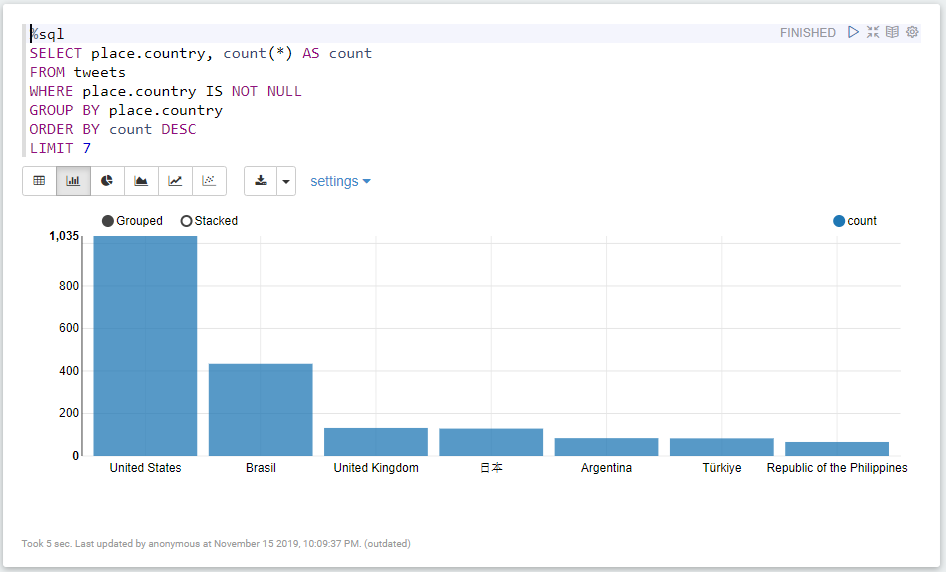
1. Query to get number of tweets and users collected.



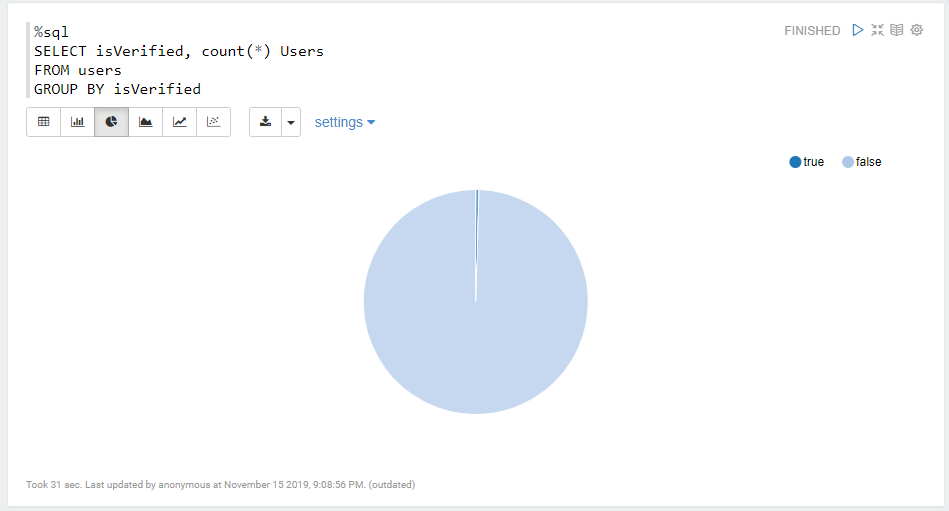
1. Query to get some tweets which contain the keyword.



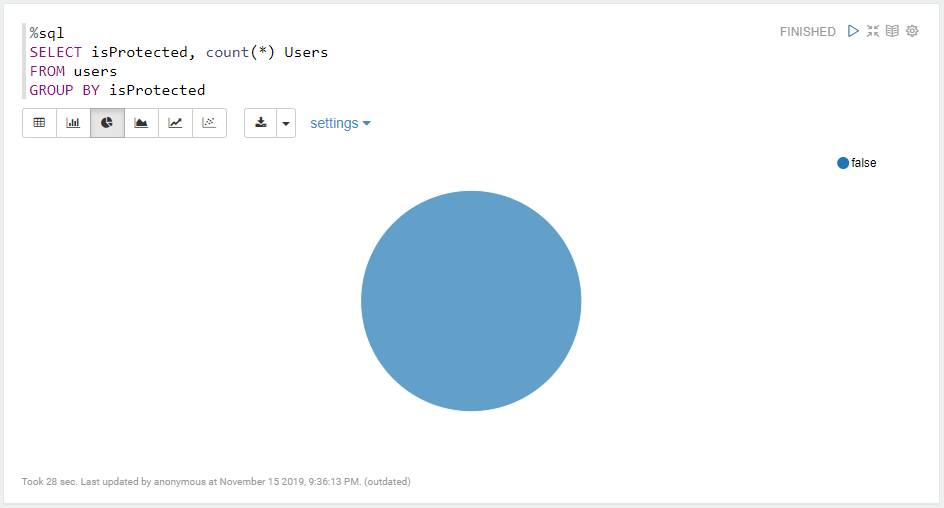
1. Query to get the top countries where tweets came from.



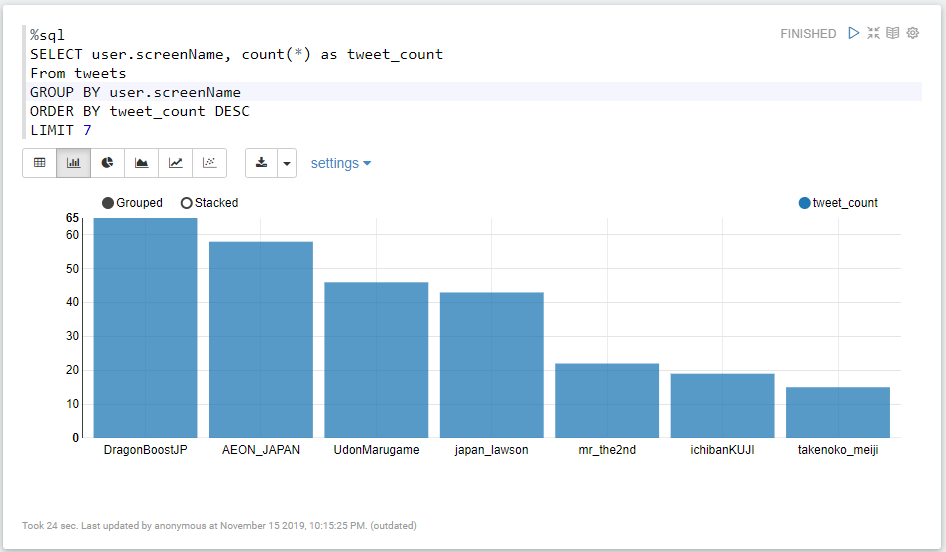
1. Query to get the verified status of users.



1. Query to get the protected status of users.



1. Query to get the top users by Tweets.



**CODE TESTING**

* We can test the code by passing in the actual collected tweets to the program.