# Principles of Big Data Management, Spring 2016

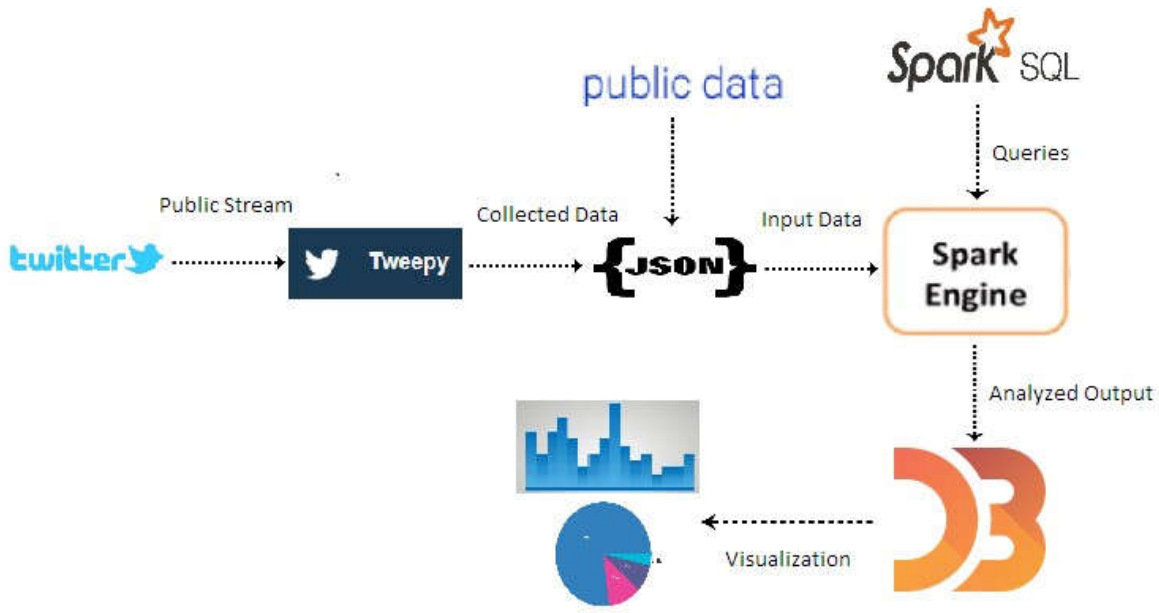
## Project Report

(Sai Venkatesh Gatiganti, Sri Chaitanya Patluri, Meghasai Reddy Bodimani)

**Project Name: Visualization of Twitter Data & Public Dataset using Apache Spark.**

**Introduction**: The Project was done over three phases. The first phase involves collection of Twitter Data using Tweepy, A Python based API used to collect Tweets from public streams & storing it in json format. The second phase involves analysing the collected data using Apache Spark, an Open Source Software used to analyze large amounts of data by running it through custom queries. The third phase involves the visualization of analyzed data through **D3.js**, A JavaScript library for manipulating documents based on data. **D3** helps to bring data to life using HTML, SVG and CSS.

**Architecture:**



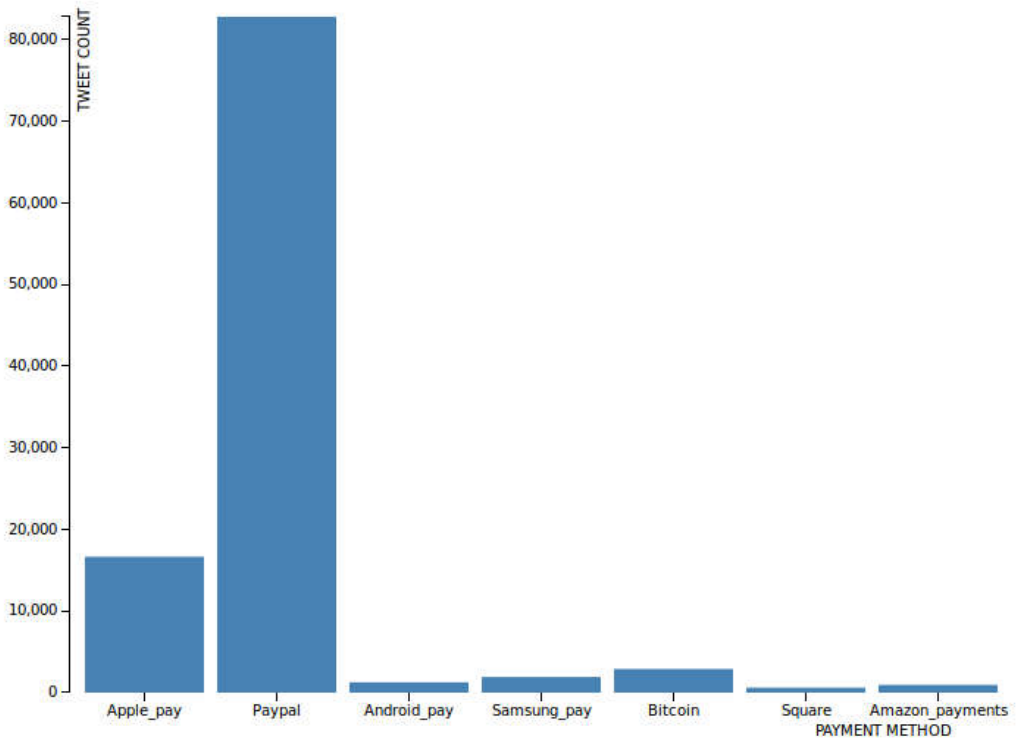
**Queries & Visualization:**

**Dataset – I:**
A Number of tweets was collected over a period of a week about Digital Payment Methods like Apple Pay, Android Pay, Samsung Pay, PayPal, Bit coin, Square, Amazon Payment etc.

**Query I: Number of Tweets for each Payment Method in Dataset I.**

```
sqlContext.sql("SELECT count(text) AS Apple_Pay FROM tweett WHERE text LIKE
'%Apple%Pay%' OR text LIKE '%Apple%pay%'")
sqlContext.sql("SELECT count(text) AS Paypal FROM tweett WHERE text LIKE
'%Paypal%' OR text LIKE '%paypal%'")
sqlContext.sql("SELECT count(text) AS Android_pay FROM tweett WHERE text LIKE
'%Android%Pay%' OR text LIKE '%Google%wallet%'")
sqlContext.sql("SELECT count(text) AS Samsung_Pay FROM tweett WHERE text LIKE
'%Samsung%Pay%' OR text LIKE '%Samsung%pay%'")
sqlContext.sql("SELECT count(text) AS Bitcoin FROM tweett WHERE text LIKE
'%Bitcoin%' OR text LIKE '%bitcoin%'")
sqlContext.sql("SELECT count(text) AS Square FROM tweett WHERE text LIKE
'%Square%' OR text LIKE '%square%'")
sqlContext.sql("SELECT count(text) AS Amazon FROM tweett WHERE text LIKE
'%Amazon%payments%' OR text LIKE '%amazon%payments%'")
```

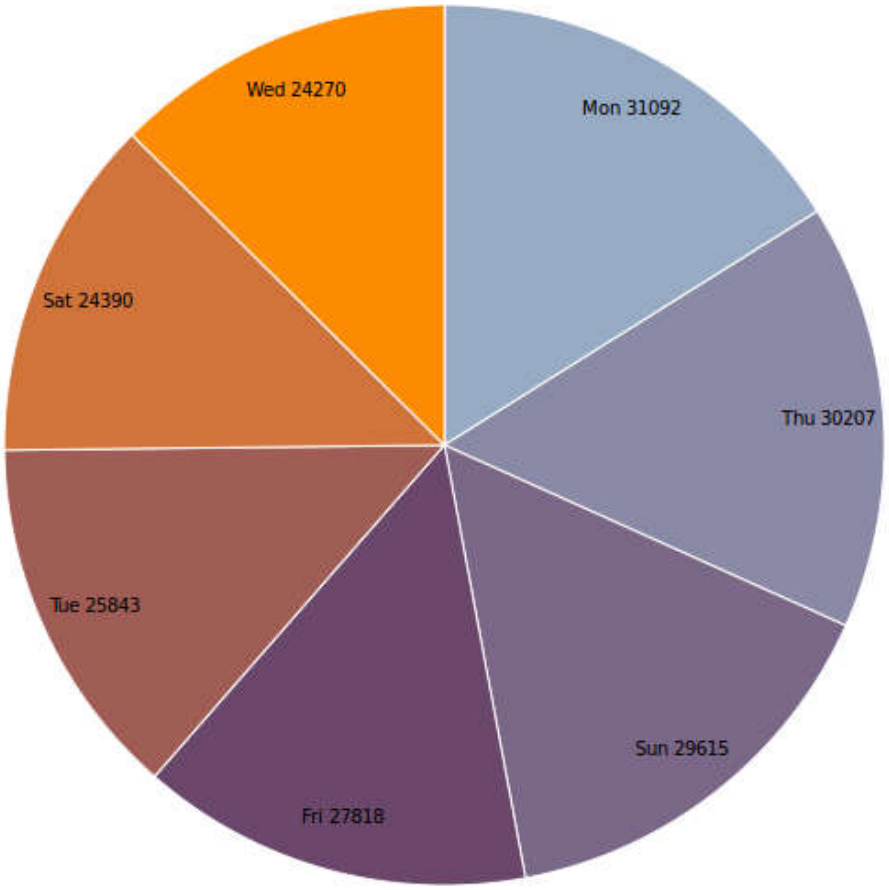
**Visualization:**



**Query II: Tweet count on each day over a period of week for all payment methods in Dataset I.**

```
sqlContext.sql("SELECT SUBSTRING(created_at,1,3) AS
day,count(SUBSTRING(created_at,1,3)) AS daycount FROM tweett GROUP BY
SUBSTRING(created_at,1,3) ORDER BY count(SUBSTRING(created_at,1,3)) DESC")
```

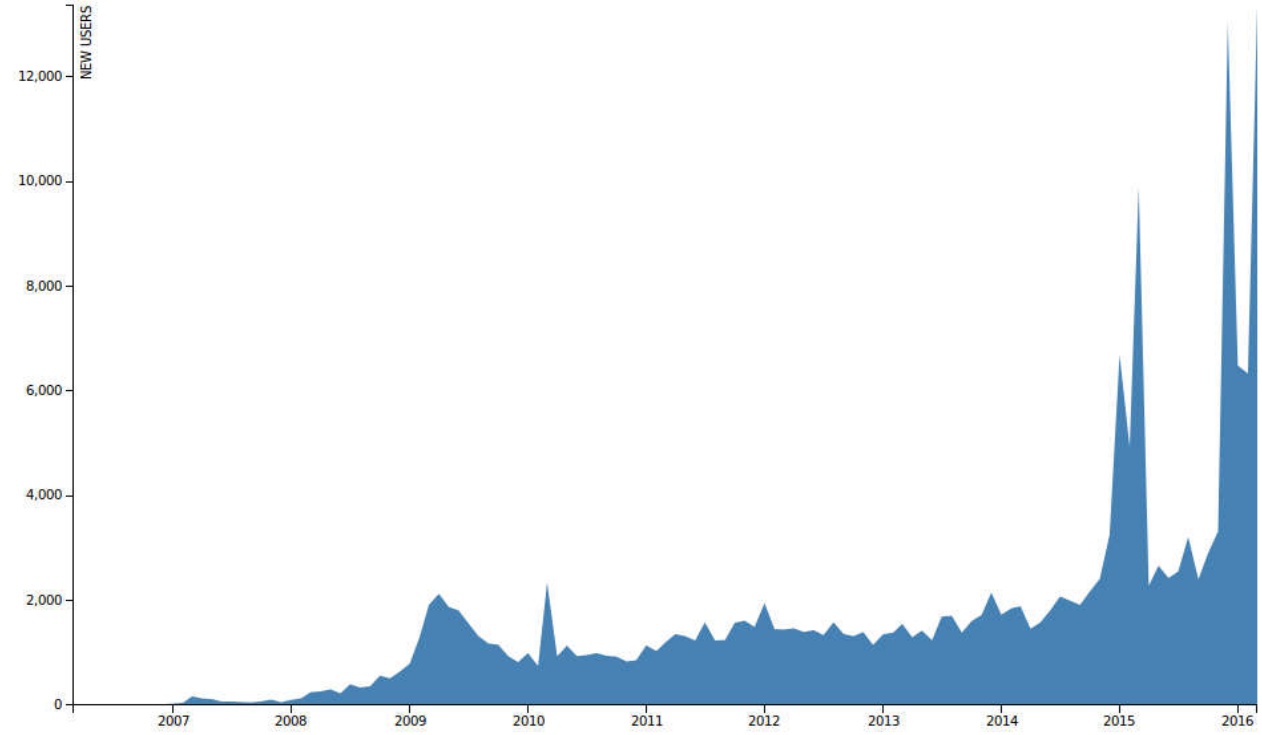
**Visualization:**



## Query III: Number of Twitter accounts created according to month & year from Dataset I.

```
sqlContext.sql("SELECT
CONCAT(SUBSTRING(user.created_at,27,4),SUBSTRING(user.created_at,5,3)) AS
date,count(SUBSTRING(user.created_at,5,3)) AS close FROM tweett GROUP BY
SUBSTRING(user.created_at,27,4),SUBSTRING(user.created_at,5,3) ORDER BY
SUBSTRING(user.created_at,27,4) DESC,CASE SUBSTRING(user.created_at,5,3) when
'Jan' then 1 when 'Feb' then 2 when 'Mar' then 3 when 'Apr' then 4 when 'May'
then 5 when 'Jun' then 6 when 'Jul' then 7 when 'Aug' then 8 when 'Sep' then
9 when 'Oct' then 10 when 'Nov' then 11 when 'Dec' then 12 END DESC")
```

## Visualization:

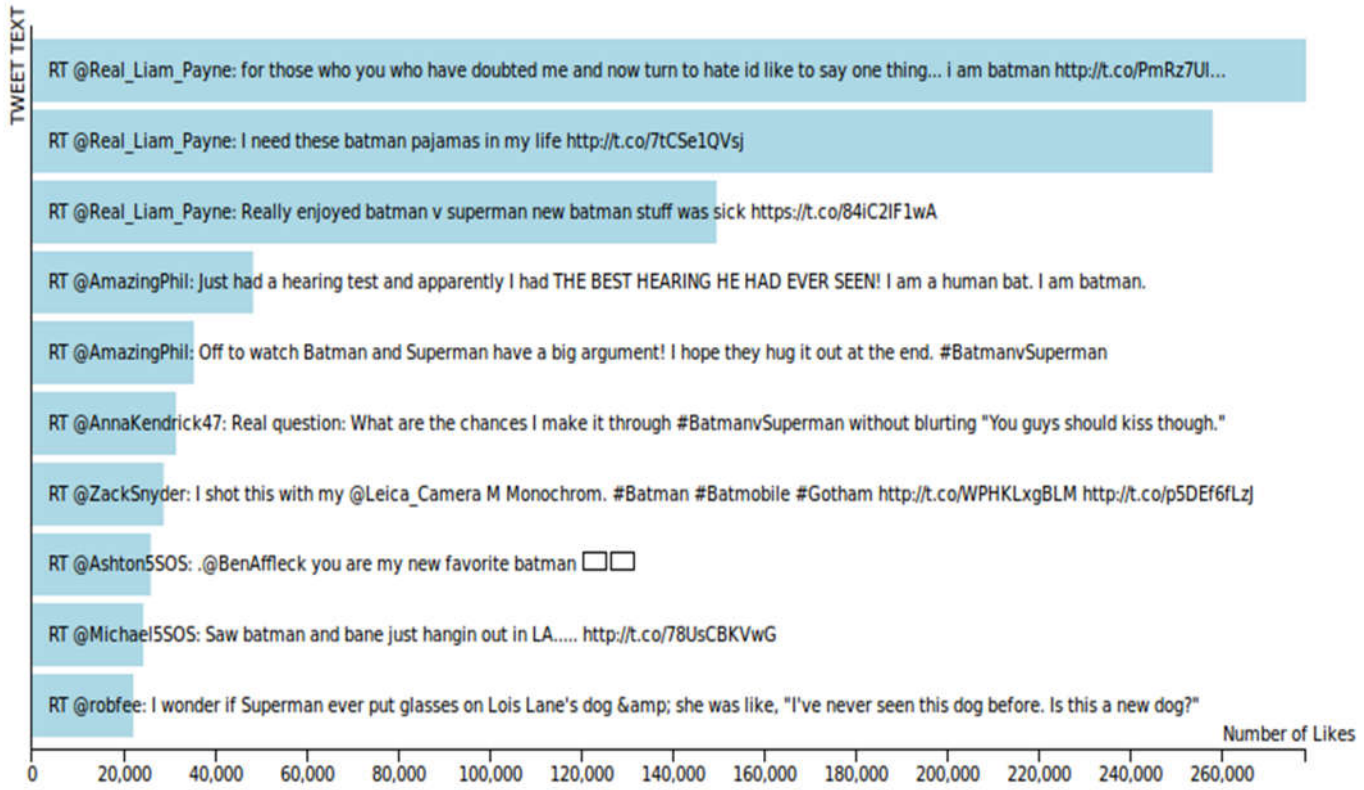


**Dataset II: A Number of Tweets about the movie "Batman v Superman: Dawn of Justice"**

**Query IV: Top ten liked tweets with like count for Dataset II.**

```
sqlContext.sql("SELECT encode(text,'UTF-8') AS
text,MAX(retweeted_status.favorite_count) AS Favorite FROM tweett GROUP BY
text ORDER BY MAX(retweeted_status.favorite_count) DESC LIMIT 10").collect()
```

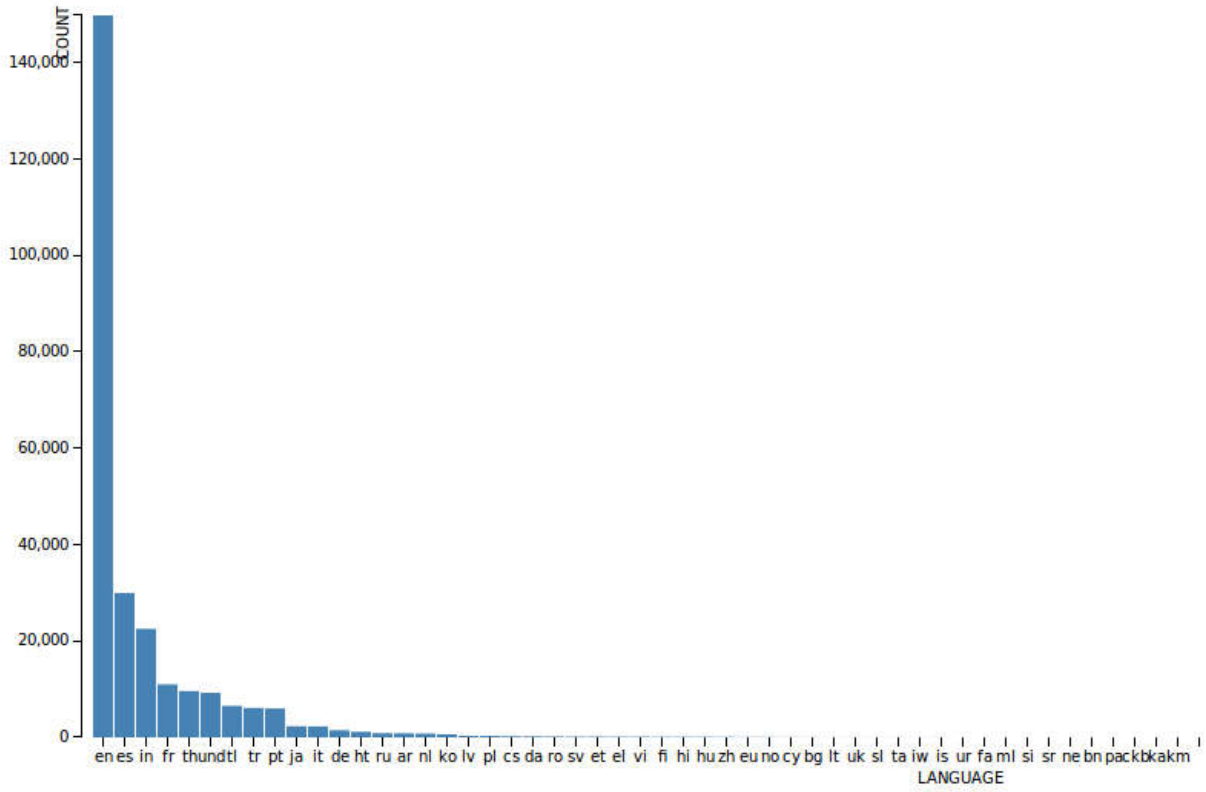
**Visualization:**



**Query V: List of all the Languages that were used to tweet and its count on Dataset II.**

```
sqlContext.sql("SELECT lang AS Language,count(lang) AS count FROM tweett
GROUP BY lang ORDER BY count(lang) DESC")
```

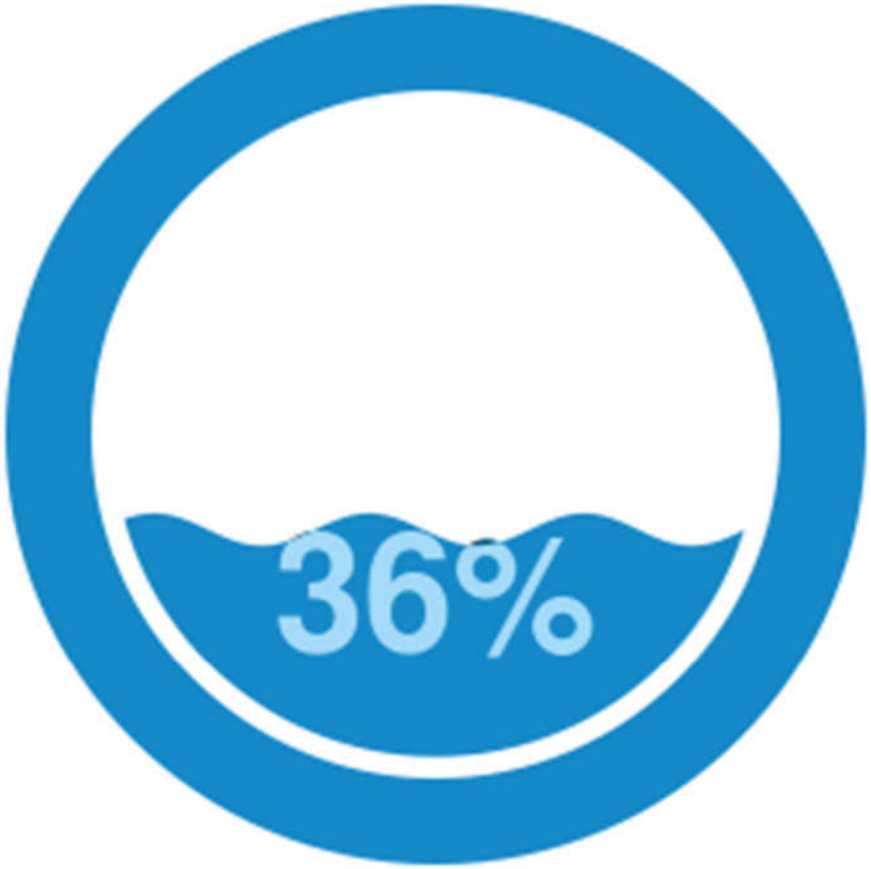
**Visualization:**



**Query VI: Percentage of Tweets with External Links in Tweet Status for Dataset II.**

```
sqlContext.sql("SELECT count(entities.urls[0])*100/count(id) AS percent FROM
tweett")
```

**Visualization:**

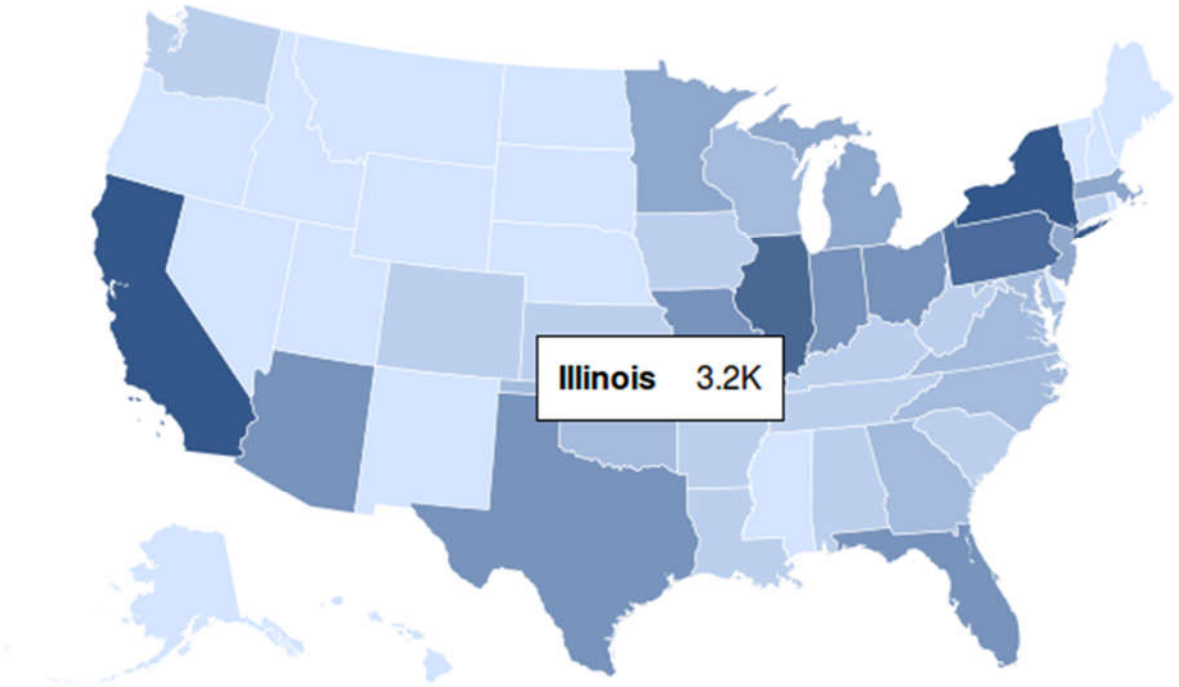


**Dataset – III: List of all Accredited Colleges in U.S.A.** (Source: U.S. Department of Education)

**Query – VII: Number of Colleges in each state for Dataset III.**

```
sqlContext.sql("SELECT Institution_State AS States,count(Institution_State)
AS Colleges FROM tweett GROUP BY Institution_State ORDER BY
count(Institution_State) DESC")
```

**Visualization:**



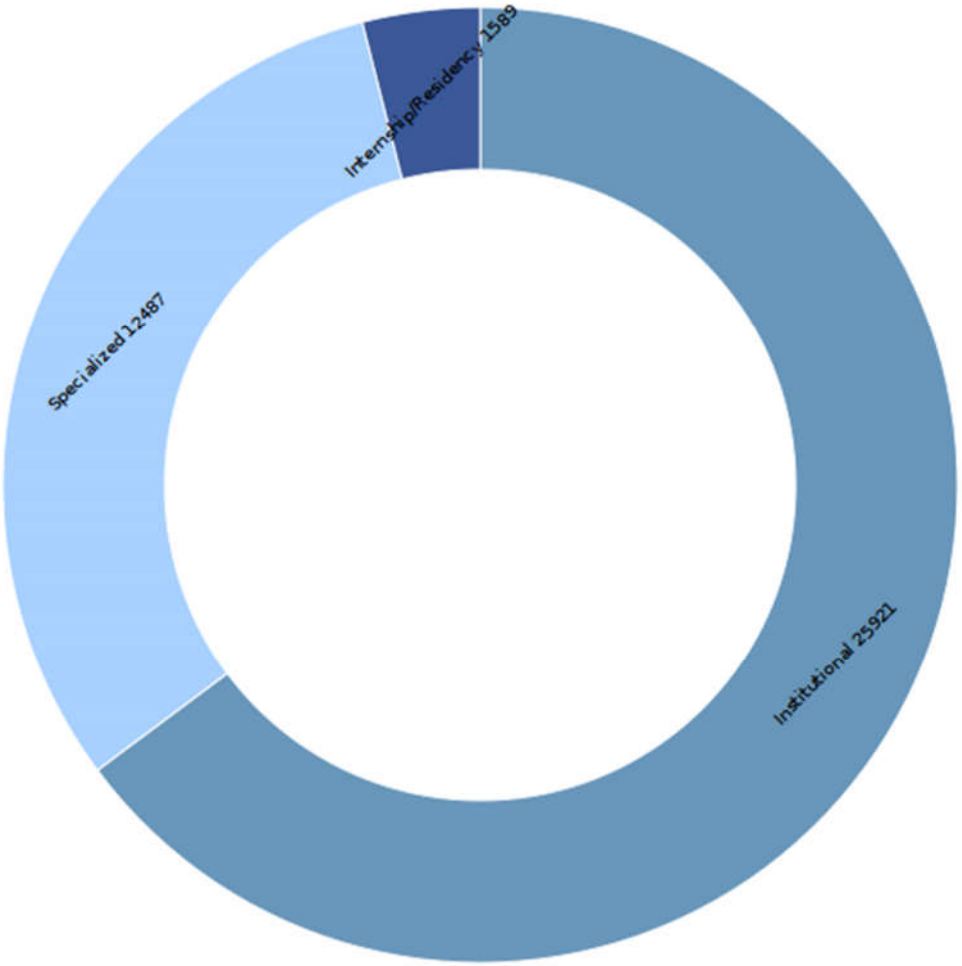

**Sample Output:**

IL,3222
CA,2737
NY,2569
PA,2175
TX,1625

**Query VIII: Ratio of Accreditation Types of all Colleges in Dataset III.**

```
sqlContext.sql("SELECT Accreditation_Type,count(Accreditation_Type) AS count
FROM tweett GROUP BY Accreditation_Type ORDER BY count(Accreditation_Type)
DESC")
```

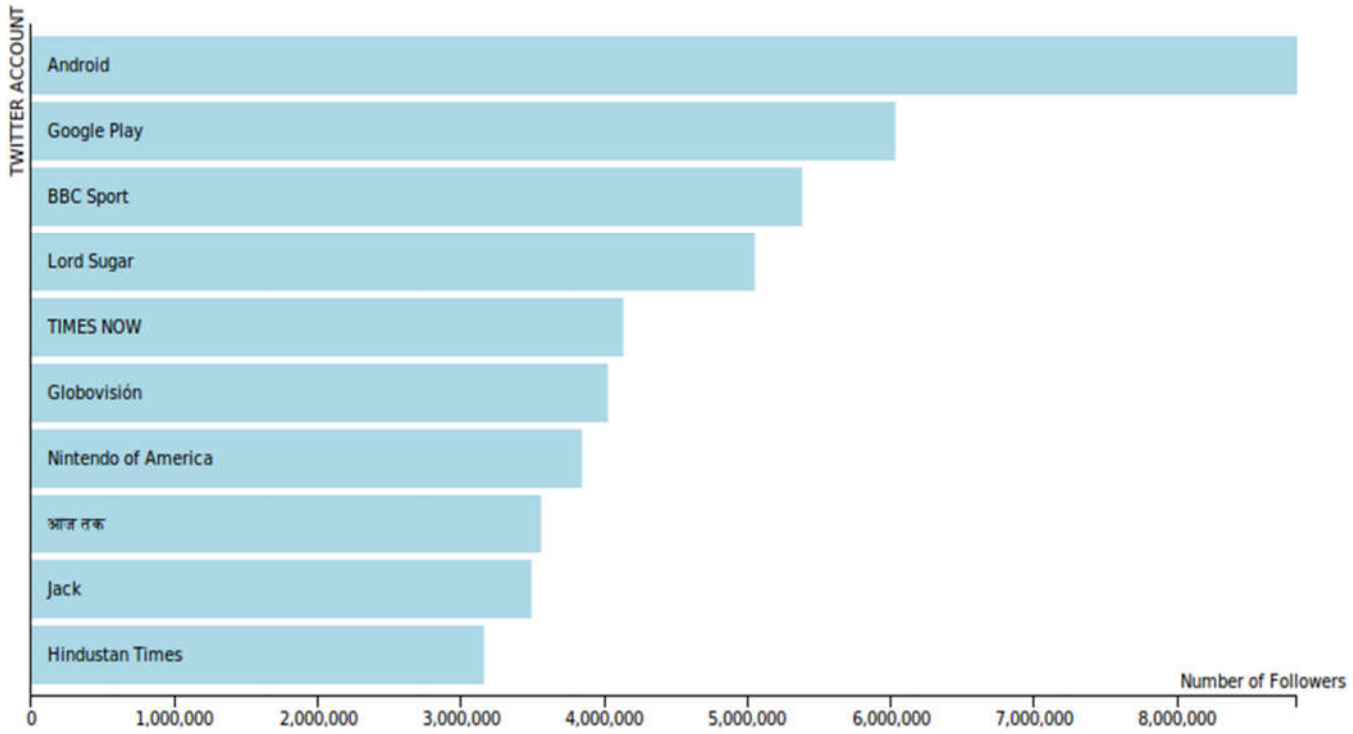
**Visualization:**



## Query IX: Top Follower Count for Verified Users in Dataset I.

```
sqlContext.sql("SELECT encode(user.name,'UTF-8') AS
Name,max(user.followers_count) AS Followers_Count FROM tweett WHERE
user.verified=True GROUP BY user.name ORDER BY max(user.followers_count) DESC
LIMIT 10")
```

## Visualization:



**System Requirement:**

Environment: PySpark(Apache Spark)
Programming Language: Python
Data Source: Twitter (Datasets I & II) and Public Dataset(U.S. Department of Education)
Datasize: 2.3 GB (approx.)
Data Format: JSON, CSV
Visualization: D3.js

**Team Members:**

Sai Venkatesh Gatiganti
Sri Chaitanya Patluri
Meghasai Reddy Bodimani

**References:**

1. https://d3js.org/
2. http://www.tweepy.org/
3. https://databricks.com/