Twitter Big Data Analysis Using SPARK

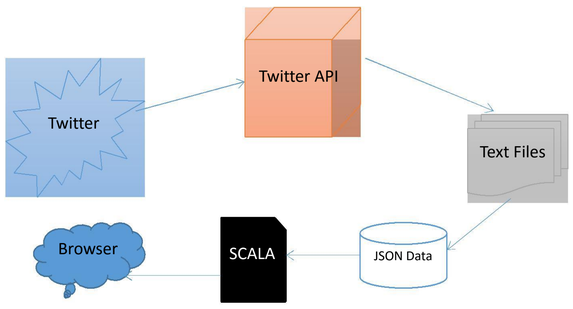
Karthik Ponnuru, Chaitanya Sai Manne, Raghu Pavan Vallamkonda

Computer Science, University of Missouri, Kansas City

INTRODUCTION

- Big data refers to extremely huge data sets that have grown beyond the ability to analyze using traditional data processing tools. It refers to both structured and unstructured data. The capability of storing such large sets of data isn't new. What's new is the ability to analyze this data quickly and effectively.
- Our project emphasizes on analyzing (i.e., visualizing using the charting libraries) such large volumes of data collected from twitter.
- Apache Spark is an open source framework for large scale data processing. Spark SQL supported by spark is used for structured data processing and allows running SQL like queries on Spark data.

ARCHITECTURE



TOPIC

Our topic is Entertanment,which includes the following:

- Sports
- Movies
- Movies, etc.

QUERIES

Country with more number of tweets:

```
SELECT place.country, count(*) as country_count FROM tweets2
where place.country is not null GROUP BY place.country Order By
country_count desc limit 30
```

Top 10 retweeted tweets:

```
SELECT text, count(retweet_count) as c FROM
EntertainmentTable group by text order by c desc limit 10
```

Top places tweeted more about movies in United States

```
SELECT place.full_name, count(*) as c FROM
EntertainmentTable WHERE text like '%movie%' and
place.country = 'United States' GROUP BY place.full_name
ORDER BY c desc limit 15
```

Top 10 most followed users

```
SELECT user.name, user.followers_count FROM tweets2 ORDER
BY user.followers_count desc limit 20
```

High preferable languages used to tweet

```
SELECT lang, count(*) as c FROM EntertainmentTable WHERE
lang is not null GROUP BY lang ORDER BY c desc limit 10
```

More number of tweets are tweeted from which time zone

```
SELECT user.time_zone,count(*) as c FROM EntertainmentTable
WHERE user.time_zone <> 'null' GROUP BY user.time_zone
ORDER BY c desc limit 10
```

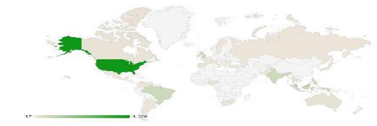
What are the top most popular URL links used in the tweets?

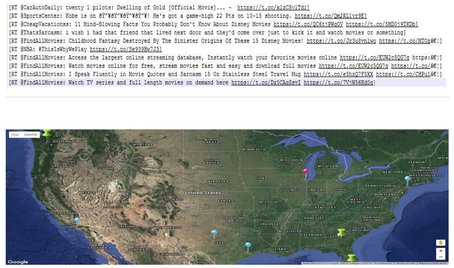
```
SELECT source, count(*) as c FROM EntertainmentTable
WHERE source <> 'null' GROUP BY source ORDER BY c desc
limit 10
```

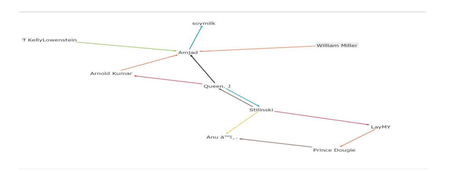
Users who are verified and non-verified

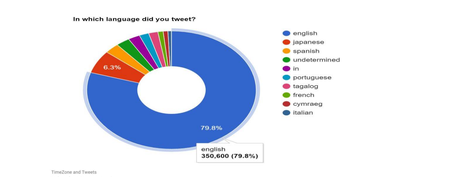
```
SELECT if(user.verified=true,'verified_users',
'nonverified_users') as status,count(*) FROM tweets4 GROUP
BY user.verified LIMIT 2
```

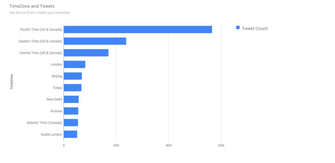
VISUALISATION

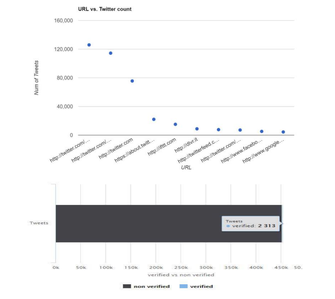












REFERENCES

http://spark.apache.org/docs/latest/sql-programming-guide.html#json-datasets

https://dev.twitter.com/overview/api/tweets

https://github.com/dhotson/springy/

https://developers.google.com/chart/interactive/docs/gallery/map#customizing-markers

http://www.highcharts.com/demo/pie-gradient

http://bl.ocks.org/mbostock/3887235

http://www.highcharts.com/demo/bar-stacked

http://bl.ocks.org/mbostock/3887235

https://developers.google.com/chart/interactive/docs/gallery/barchart#overview

http://www.highcharts.com/demo/bar-stacked

ACKNOWLEDGEMENT

We sincerely thanks University of Missouri Kansas City and Professor, Dr.Praveen Rao for supporting and guiding throughout the project.