PRINCIPLES OF BIG DATA MANAGEMENT PHASE II REPORT

Team Members:

1. Bhavya Teja Gurijala
2. Dig Vijay Kumar Yarlagadda
3. Yashwanth Manchikatla
4. Chandra Sekhar Janyavula

Phase II Goal:

The objectives of Phase II are:

- Downloading the tweets related to a specific topic.
- Storing the downloaded tweets in a database
- Extracting the relevant data by writing queries and visualizing it.

Data Source Topic:

Accidents are the unintentional and unplanned events that happen which would result in severe injuries. Most of the accidents happen by vehicles i.e., by means of Transportation. The reasons for the occurrence of the accidents may be due to the Bad Weather, Snow, Slippery Roads, Reckless driving, Drink & Drive, Brake Failure etc.,

The Analysis of the occurring of the Accidents will give a clear idea on these unintentional happenings all over the world. We can visualize based on the number of accidents happened in various parts of the world, Visualizing based on the number of accidents in all the days of a week. This Visualization helps us to get an overview on how and when accidents occur.

Technologies:

- Programming Languages: Python, Java Script, Scala
- Web Technologies: HTML & CSS
- Database: MongoDb
- Query Language: Spark SQL
- Visualization Tools: D3js.org, Highcharts.com, Vida.io and wordle.net

Project Implementation:

The main aim of the project is to download the tweets related to ‘Accidents’ and visualize the data by writing useful analytical queries.

Step 1:

Downloaded the tweets related to the keyword ‘Accidents’ from twitter by developing an application written in Python with a library called Tweepy. All the streaming tweets are downloaded and stored in the Json format. We have downloaded 2,62,644 tweets at different timestamps. Though the Tweets are downloaded at various intervals we have integrated them while importing to the database.

Step 2:

The downloaded tweets in order to reduce the noisy data we have uploaded that data to the MongoDb as it accepts the Json data as an input. After importing the data into the MongoDb we extracted the data from the database to process the analytical queries.

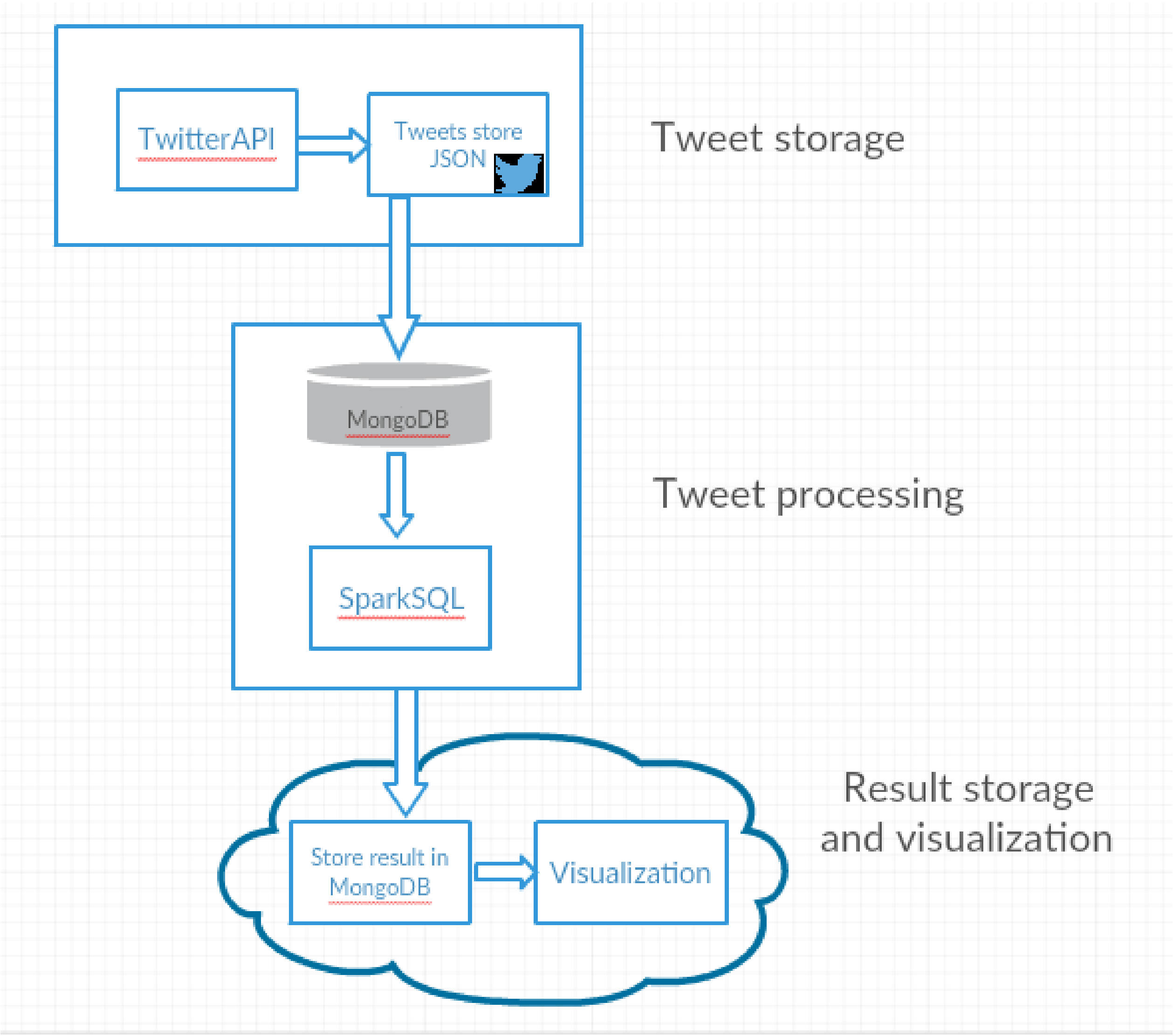
Step 3:

The Queries are developed by writing a program in scala which include SQL Queries to retrieve the data from the MongoDb. The Retrieved data is stored in Json format which is used for the Visualization.

Step 4:

For the effective visualization of the data we have used several resources. The resources include Highcharts.com, D3js.org, Visa.io and wordle.net. The Json data that is retrieved from the MongoDb is given as an input to these Visualization tools.

Design & Architecture:

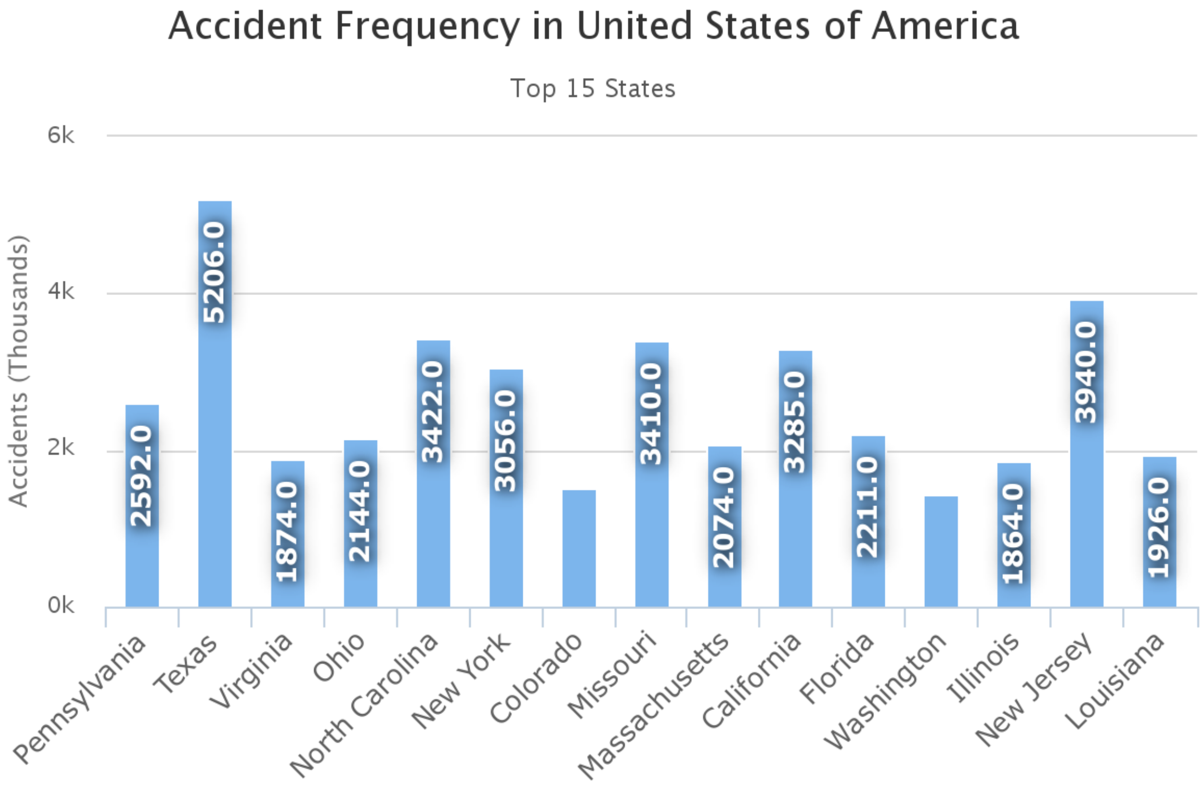


Analytical Queries and Visualization:

QV 1 – Top 15 States of USA:

This Query retrieves the Location attribute of all the tweets. We visualized the top 15 states of United States of America on the basis of the tweet count.

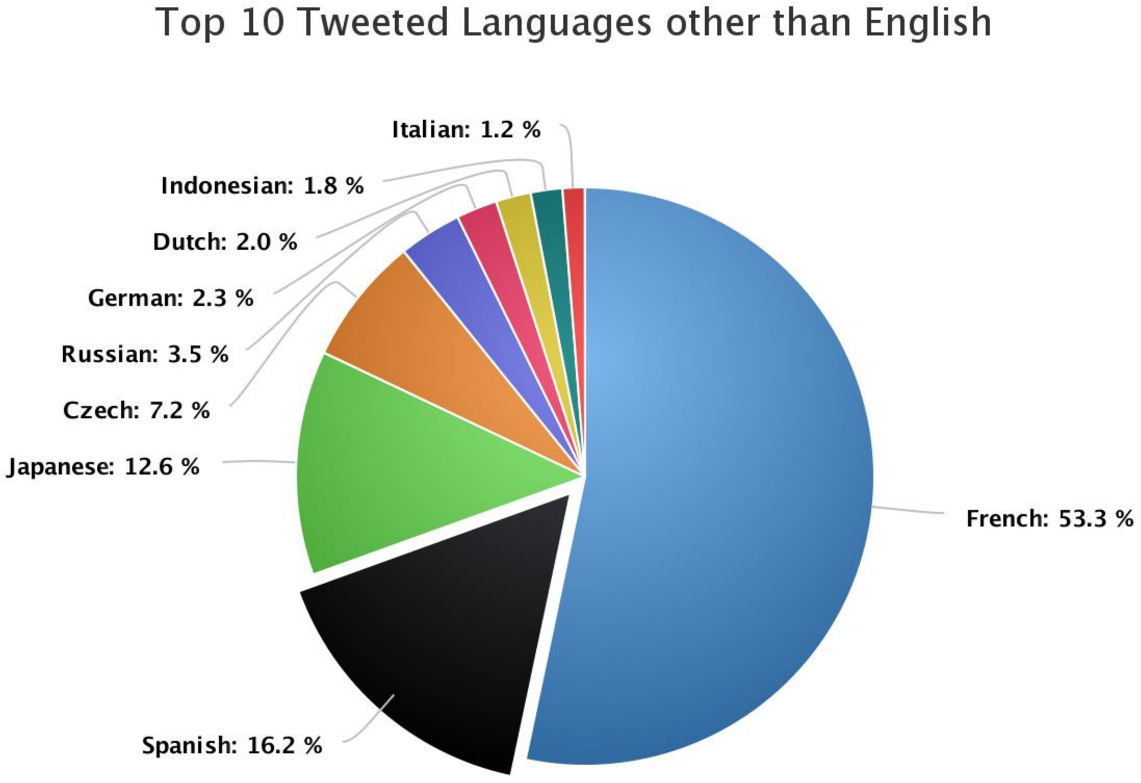
```
sqlContext.sql("SELECT user.location, COUNT(*) AS temptable FROM tweets GROUP BY user.location ORDER BY temptable DESC").collect().foreach(println)
```



QV 2 – Top Tweeted Languages:

This Query retrieve the Language attribute of all the tweets. We have visualized the top 10 Tweeting languages except English.

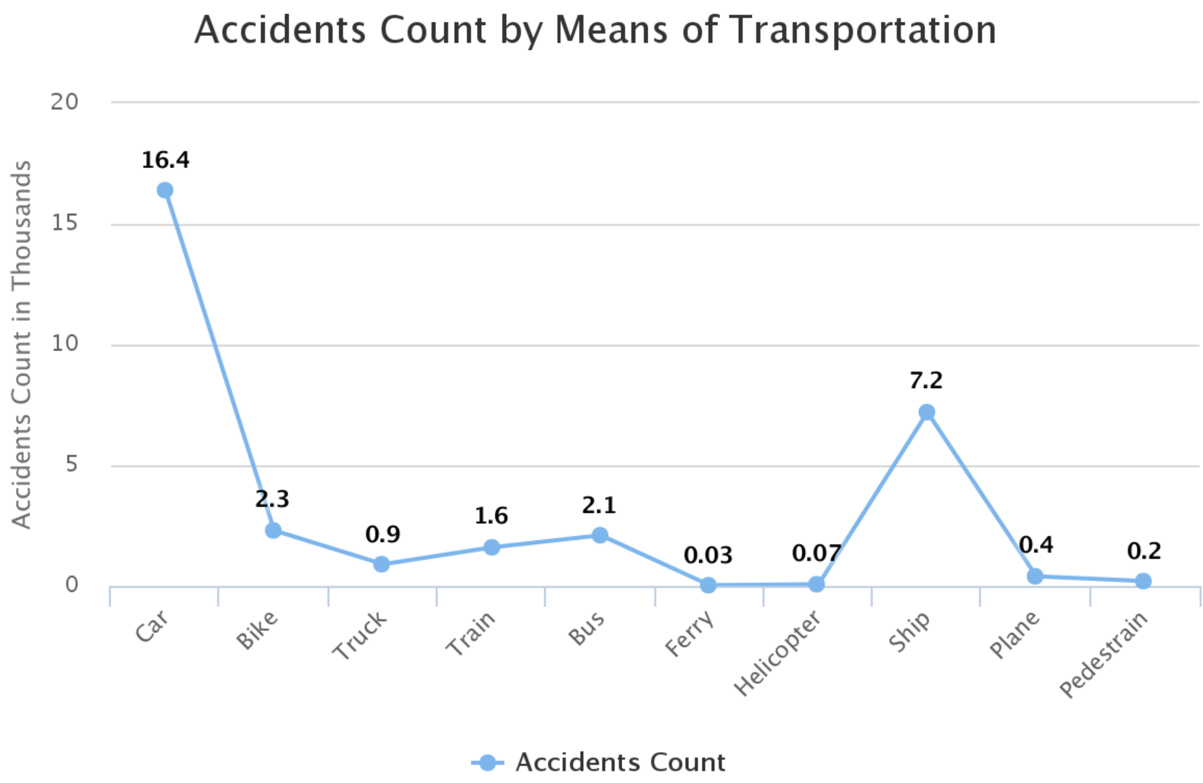
```
sqlContext.sql("SELECT user.lang, COUNT(*) AS temptable FROM tweets GROUP BY user.lang
ORDER BY temptable DESC").collect().foreach(println)
```



QV 3 – Top Means of Transportation:

We have retrieved the tweet count by searching for the keywords like ‘Car’, ‘Bike’,’ Truck’ etc., This gives the types of vehicles involved in the Accidents.

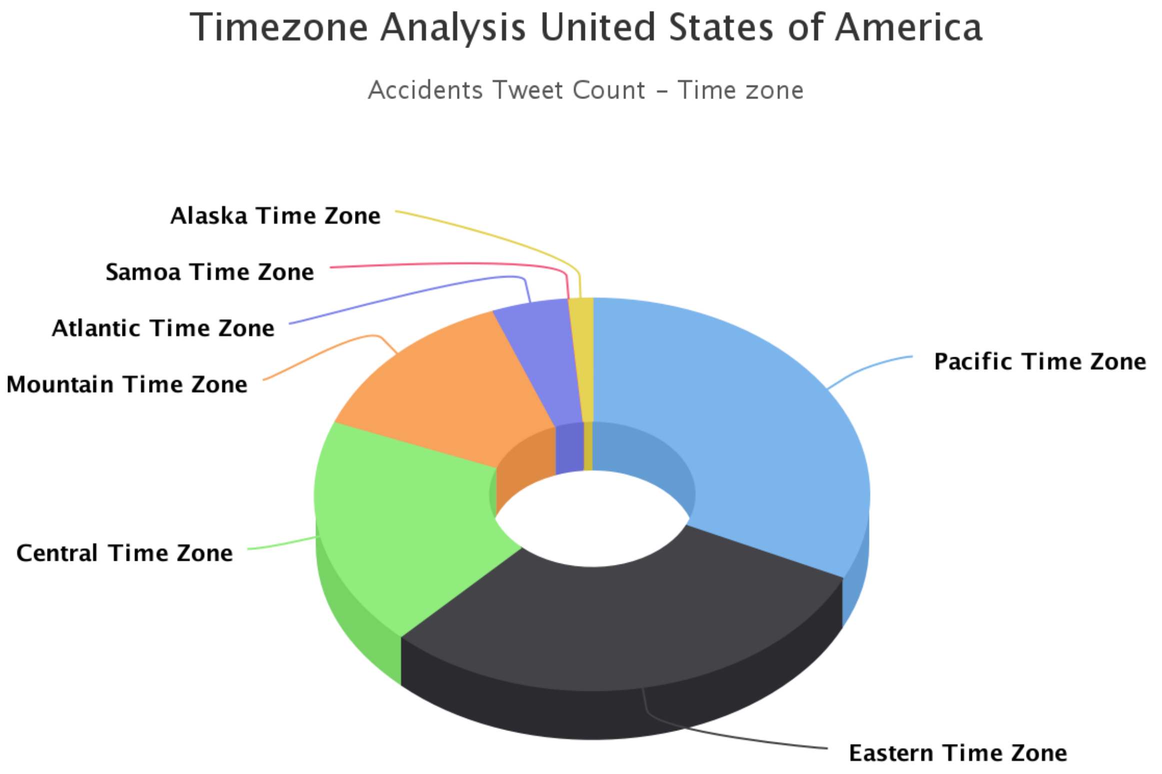
sqlContext.sql("SELECT 'car', COUNT(*) AS temptable FROM tweets WHERE text like '%car%'").collect()



QV 4 – Accident Analysis Based on Time Zones in United States of America:

The Query is to analyze the happenings of the accidents over all the time zones in USA. We have retrieved the time zone attribute of the user information of the tweet.

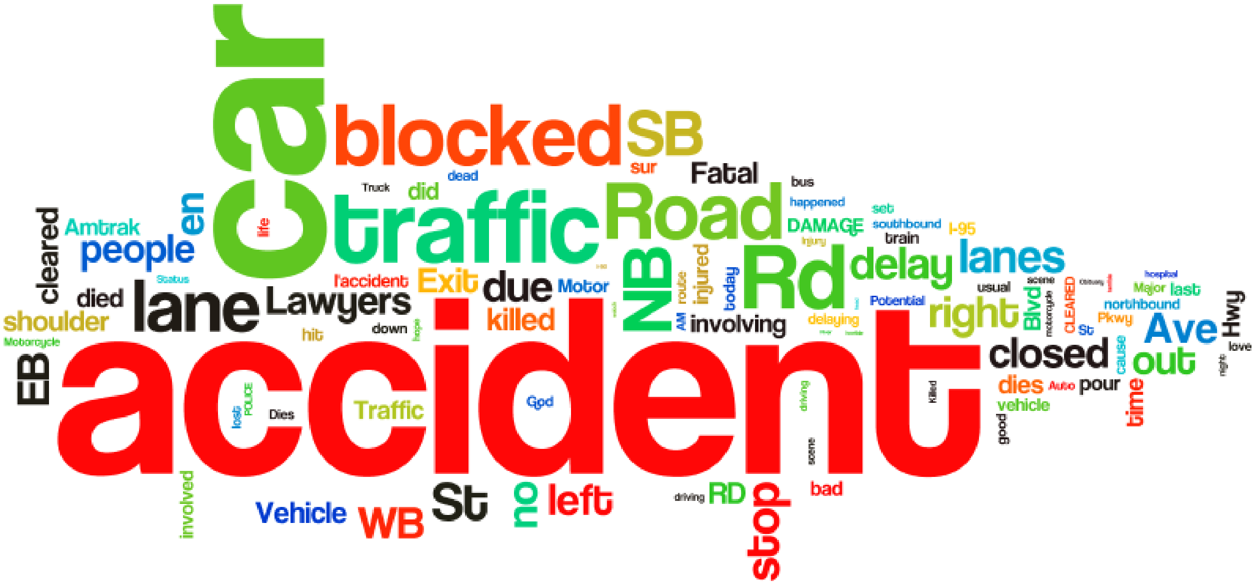
```
sqlContext.sql("SELECT user.time_zone, COUNT(*) AS temptable FROM tweets GROUP BY user.time_zone ORDER BY mytemptable DESC").collect().foreach(println)
```



QV 5 – Top used words in the Tweets:

The Query retrieved the data basing on the text attribute. The Most used words are visualized.

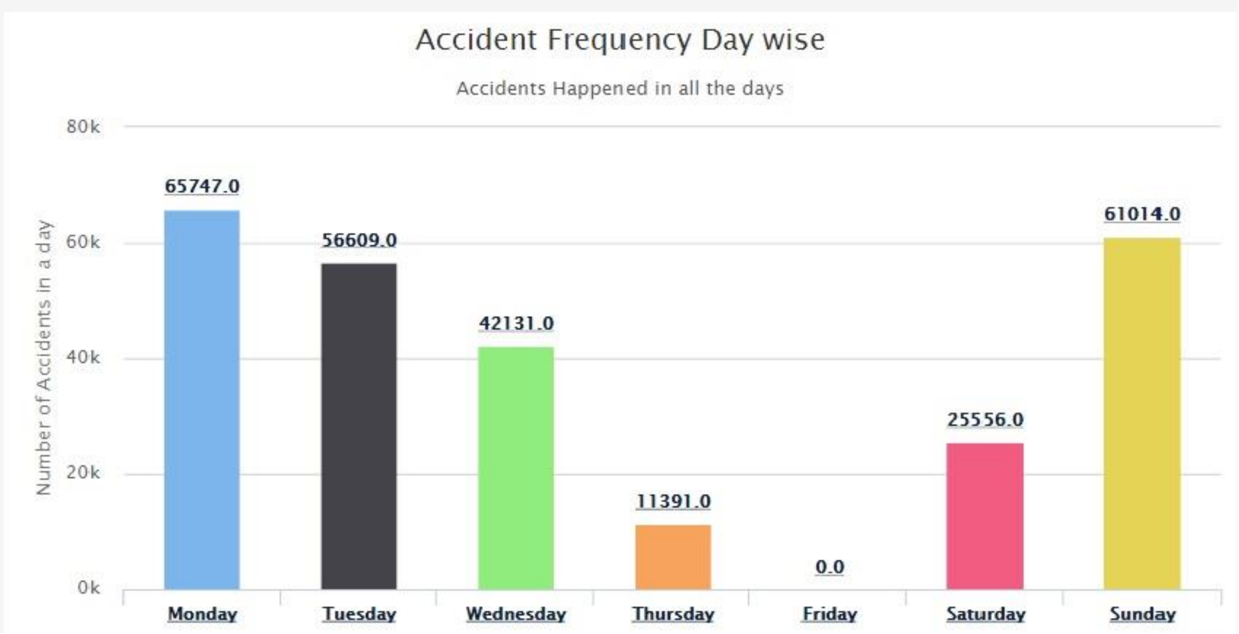
sqlContext.sql("select text, count(*)as temptable from tweets group by text ")



QV 6 – Accident Count in a Week

This Query retrieves on the basis of the timestamp of the tweet. We have visualized the count on all the days.

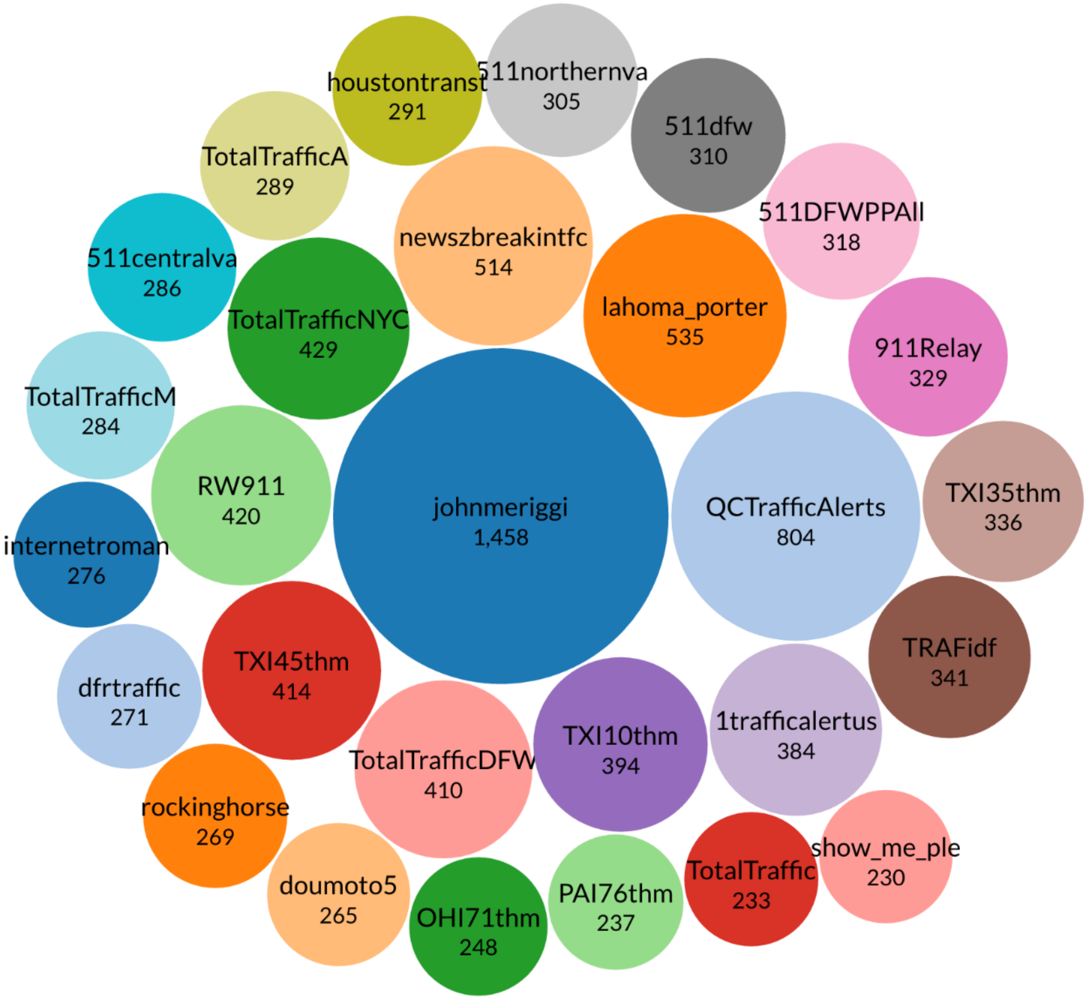
sqlContext.sql("SELECT 'Mon', COUNT(*) AS temptable FROM tweets WHERE created_at like '%Mon%'").collect()



QV 7 – Top Tweeters related to Accidents

The Query processes the data on the basis of the Screen name of the user and counting the number of tweets by them.

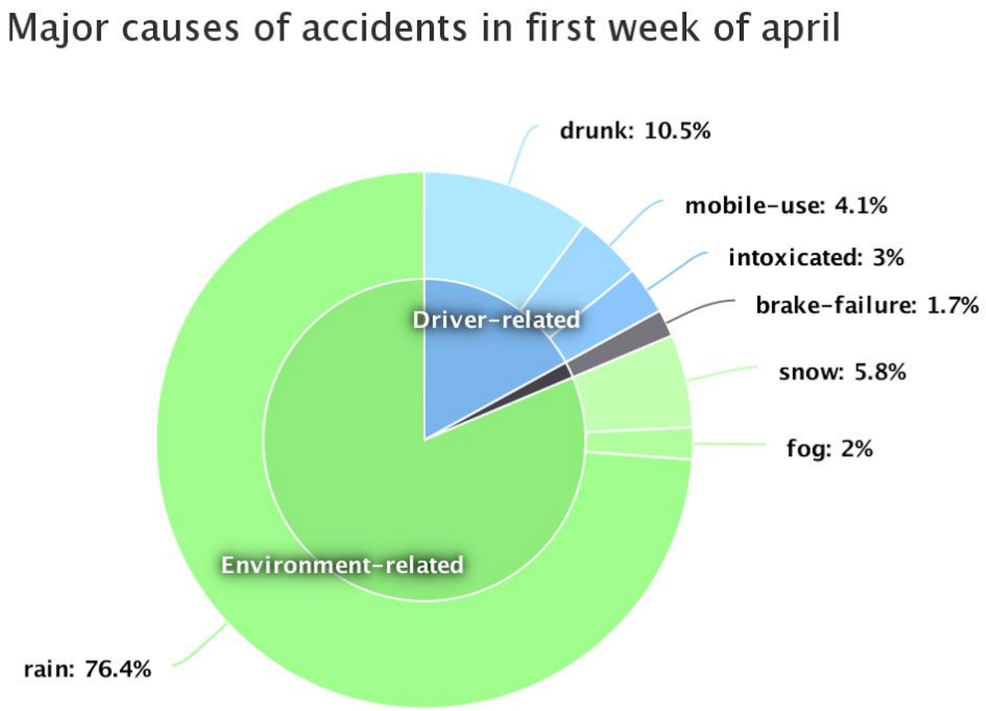
SELECT user.screen_name FROM tweets



QV 8 – Most Caused of the Accidents in the first week of April

Query optimizes the tweet results based on the reasons for the accidents like snow, brake failure etc., And we have visualized the data in a pie diagram.

```
sqlContext.sql("SELECT 'snow', COUNT(*) AS temptable FROM tweets WHERE text like
'%snow%'").collect()
```



GitHub URL:

https://github.com/digvijaykumaryarlagadda/ReachMe

Google drive URL:

https://drive.google.com/drive/folders/0BynseSOfmTFGVVdlQlBsakQ1cGM

References:

1. www.highcharts.com
2. www.wordle.net
3. www.vida.io
4. www.d3js.org
5. https://github.com/Stratio/Spark-MongoDB