**PB**

Submitted by: (Team 11)

Moulika Chadalavada (16234180)

Sampath Gattu

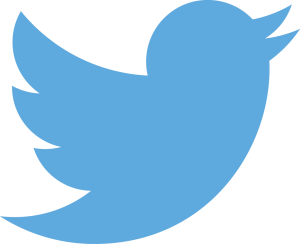
Nikitha Kona

Prathyusha

**Principles of Big Data Management**

**Phase-2**

Fall 2016



**Table of Contents**

**1.** **Introduction** 2

**1.1** **About Twitter** 2

**1.2** **Proposed Project** 2

**2.** **System Requirements** 2

**2.1** **Software Requirements** 2

**2.2** **Language Requirements** 3

**3.** **About the Theme** 3

**4.** **System Architecture** 3

**5.** **Collecting Twitter Tweets** 4

**5.1** **Generating Access Tokens** 4

**5.2** **Streaming Twitter Tweets** 4

**6.** **Analyzing Twitter Data** 4

**6.1** **Query 1** 4

**6.2** **Query 2** 4

**6.3** **Query 3** 4

**6.4** **Query 4** 4

**6.5** **Query 5** 4

**7.** **References** 4

# **Introduction**

## **About Twitter**

Twitter is a massive social networking service that enables users to send data read short 140 character messages called ‘tweets’. More than 140 million active users publish over 400 million tweets every day. Twitter was created in March 2006 by Jack Dorsey, Evan Williams, Biz Stone and Noah Glass and launched in July 2006. Twitters speed and ease of publication have made it an important communication medium for people from all walks of life. This stream of messages from a variety of users contains information on an array of topics, including conventional new stories, events of local interest, opinions, real-time events. Twitter APIs provide access to tweets from a time range, from a user, with a keyword.

## **Proposed Project**

We choose ‘Diseases’ as our topic to do big data analysis. Based on twitter tweets, we predicted some interesting analysis on Diseases using thousands of tweets tweeted by different people. First we collected the tweets from twitter API based on some key words related to Disease. After that, we analyzed the data that we have collected. By using the analysis, we written some interesting SQL queries useful to give a proper result for the analysis.

Here we used Spark to processing the twitter data. Because it has many advantages like

* Speed: Run Programs up to 100x faster than Hadoop Map reduce in memory.
* Ease of Use: Write applications quickly in Java, Scala, Python, R.
* Generality: Combine SQL, streaming, and complex analytics.
* Runs Everywhere: Spark runs on Hadoop, Mesos, standalone, or in the cloud

# **System Requirements**

## **Software Requirements**

* Python 3.5
* JDK 1.8
* Scala 2.11.8
* Intellij IDEA
* Apache Spark

## **Language Requirements**

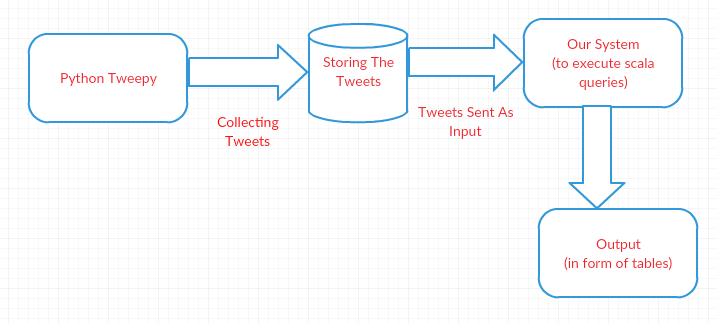
* Python
* Scala
* SQL

# **About the Theme**

We have selected the theme diseases, which provides the analysis on the twitter tweets based on the queries. Initially tweets were collected using hashtags related to disease types such as Heart Stroke, Cancer, Chicken Pox etc. Using this data, the queries are written in Scala which gives analysis on which disease more tweets were done, which user made more number of tweets on which disease, which state people in USA are more involved in tweeting about the diseases, top languages using which tweets were posted. The extracted data is displayed in tables. Based on this data visualization is done in next phase.

# **System Architecture**

First we generated credential for accessing twitter. By using these credentials, we wrote a python program to collect twitter tweets based on keywords related to food. Tweets were stored in a text file in a JSON format. We will give these JSON file to SQL queries for analysis with Spark, Intellij with Scala program with queries.



# **Collecting Twitter Tweets**

## **Generating Access Tokens**

First we generated keys for accessing twitter API. For this we need to register our application by using, [http://apps.twitter.com](http://apps.twitter.com/). We generated access token, access token secret, consumer key, consumer secret

*ACCESS\_TOKEN = "1974127951-N1QRXNCsVCywXl67BU1Wx7VlJ1fw2TlScZDY07s"*

*ACCESS\_SECRET = "zLLfZ4ZF9BxvgzjXAkjJFAVFOL4P1i0fLSaZzc6LrnqZJ"*

*CONSUMER\_KEY = "RfqrQLUtEAvwqSmNpGjZydmOn"*

*CONSUMER\_SECRET = "1XdUvxMDfcL5eE89oAo7ZO8UESv6H7HacNa9B0Y7cGFa5MiPjo"*

## **Streaming Twitter Tweets**

After that we have written a python program for streaming twitter tweets. As our theme is related to ‘Diseases’ we used few hashtags such heartstroke,cancer,malaria,braintumour,chickenpox etc. From twitter, we have streamed almost 100000 instances. Tweets were stored in JSON (JavaScript Object Notation) format in a file.

# **Analyzing Twitter Data**

## **Query 1**

## **Query 2**

## **Query 3**

## **Query 4**

## **Query 5**

# **References**

[https://apps.twitter.com](https://apps.twitter.com/)

[http://www.tutorialspoint.com](http://www.tutorialspoint.com/)

https://www.jetbrains.com/idea/