# *Crisis Analysis from Social Media*

## *Using Hadoop MapReduce*

Gaurav Shad
CIDSE, ASU
gshad@asu.edu

Abhishek Vellanki
CIDSE, ASU
avellank@asu.edu

Nitesh Gupta
CIDSE, ASU
ngupta40@asu.edu

***Abstract*—Social Media contains wide variety of data that can be analyzed to get some meaningful information which can be used for predictions, advertisements, marketing etc. The amount of data that needs to be analyzed is so big that its processing is inefficient on a single machine using limited resources. This problem can be addressed by merging cloud computing and Hadoop to do distributed computing of data. The goal of this project is to develop an application that analyzes twitter data related to crisis (including natural disasters, terrorist attacks, etc) and represent in an interactive format the areas that are prone to some crisis. The scope of this project includes setting up Hadoop in a private cloud cluster and implementing MapReduce to process this data. The main tasks include designing of Mapper and Reducer that will help generate some quality output.**

***Keywords—Hadoop, MapReduce, FLUME, Twitter, Data Visualization, Big Insights, Spark***

## I. INTRODUCTION

Social media connects people around the globe and hence turned itself into a platform where people share every event happening in their lives. Twitter's ease and speed of communication has made it the heart of social networking with millions of tweets every day. Because of its popularity, it has led to the development of applications and research in various domains [1]. Disaster detection is one such domain where researchers have already worked in the past and successfully predicted their occurrences."The 2014 earthquake in Napa was detected by USGS in 29 seconds using Twitter data, likely due to the tech savvy population that dominates the area" [2]. In this project we will analyze previous tweets for chosen time period (like 5 years) relating to crisis and return end user, graphical information based on the search criteria (location or crisis or both).

The problems addressed here are computing time and scalability. By processing data in a multimode cluster using MapReduce, we are enabling parallel processing which decreases the computing time and increases the scalability.

The big data that needs to be analyzed can be unstructured which traditional databases cannot handle, query and analyze. There comes a need for a system that can handle and analyze raw unstructured data efficiently. Since this big data is increasing exponentially, if not analyzed, this data remains meaningless.To extract some information from it, requiresthis data to be processed.

**Technologies used:**

Apache Hadoop – It's a framework for processing large data sets.

MapReduce – It's a programming paradigm which performs highly parallelized processing of large data sets over multiple clusters of nodes.

Apache FLUME – This is a service that helps in extracting data with the use of Twitter API and storing it into HDFS.

Data Visualization – IBM Insights or Google Maps API or KDE or Apache Spark will be used on the output set to visualize it.

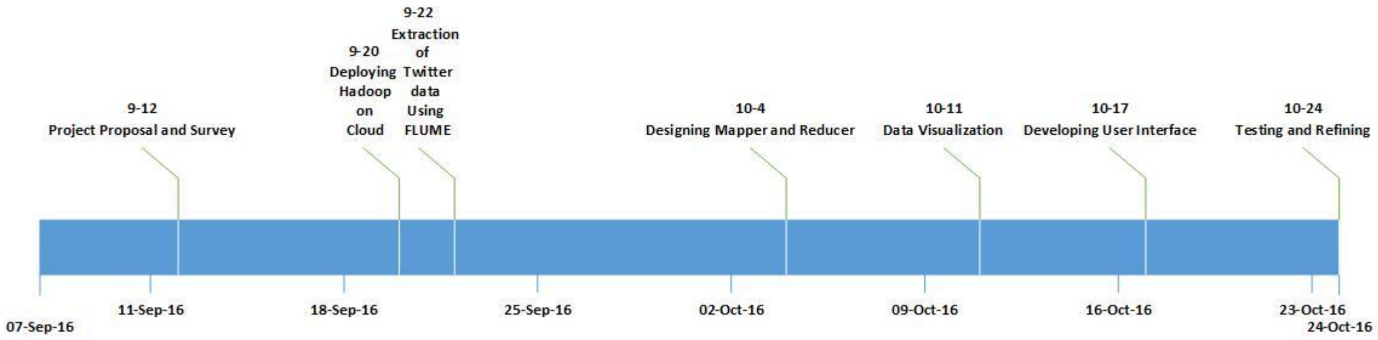
**Expected Outcomes:**

By working on this project, we expect to develop a data analytics application that will help visualize big data. The user interface will be developed to help end user in finding crisis prone areas in particular region.

The team comprises of three students:

1. Gaurav Shad
2. Abhishek Vellanki
3. Nitesh Gupta

We will work together in all the tasks (Tasks 1 -7) so that all team members learn everything involved in the project. Timelines for the project are as follows:



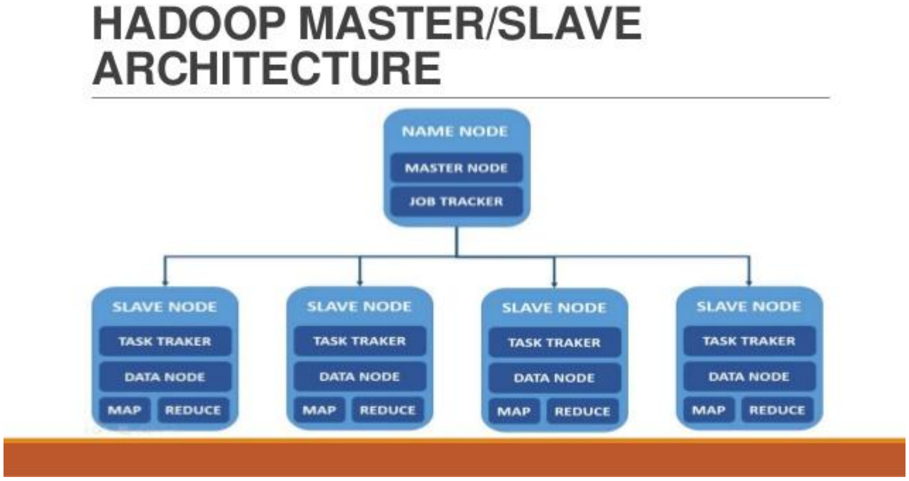
1. Timeline

## II. SYSTEM MODELS

### A. System Model

The model consists of setting up a Hadoop Multi node cluster in a cloud. All the nodes, including the Master Node will reside in the cloud environment. The files need to be configured so as to interact and perform processing over different data nodes present in the cluster.

HDFS Architecture:



2. HDFS Architecture

Master Node:

- It is the master of the system which maintains and manages data on slave nodes.

Slave Node:

- These nodes are deployed for actual storage and data processing.

Job Tracker:

- It resides in the master node which determines the execution plan and manages all the tasks.

Task Tracker:

- It resides in all the slave nodes which keeps track on the task and keeps job tracker updated.

### B. *Software*

1. Apache Hadoop

   It's an open source framework for storage and processing of large data sets. It consists of four modules:

   - Hadoop Common
   - Hadoop Distributed File System (HDFS)
   - Hadoop YARN
   - Hadoop Map Reduce

2. Open Stack

   It's an open source cloud computing platform for public and private clouds. It lets user deploy virtual machines and other instances that handle different tasks for managing the cloud environment.

3. Apache FLUME

   It helps in efficiently collecting and moving large amounts of streaming data directly into HDFS. In our project, it will build a connection by using Twitter API to collect data.

4. Twitter API

   It helps to extract tweets based on a specific query. It can restrict the number of tweets to the given language, location and number of tweets allowed per page.

5. Data Visualization Tool

   We will use Big Insights or Spark to visualize the data that is received from MapReduce.

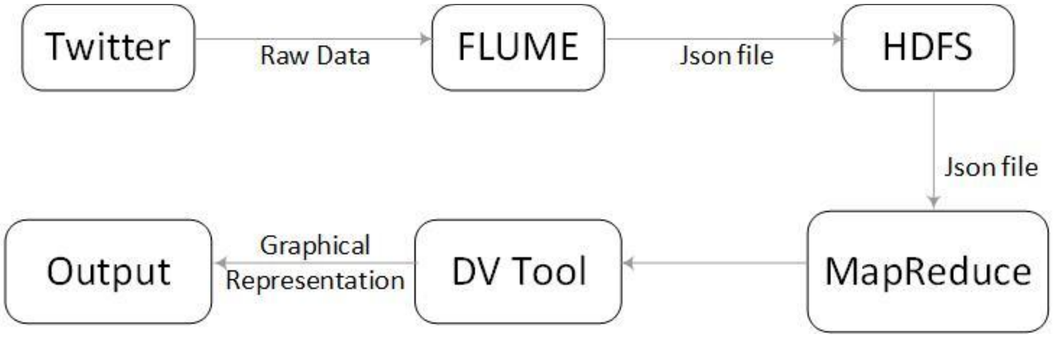
## III. Project Description

The project concentrates on the data set retrieved from Twitter based on time filtering (like last 5 years) and process this data to represent end user how prone a region is to a crisis. Region can be a city, state or country.Performing crisis analysis on tweets can be trickier than normal text or paragraph because tweets are short messages which contain hash tags, slangs etc.

The main task of this implementation is to automatically filter which tweets that contain the crisis information required in this project. The MapReduce will use certain conditions to perform this task. In a research done by USGS National Earthquake Information Center (NEIC), they built similar kind of analytics application by filtering tweets on the following conditions [2]:

1. Remove the tweets containing more than seven words
2. Remove the tweets that contains any links

We will use the similar conditions to perform basic filtering and generalize it for all types of crises. After the basic filtering, we will apply more conditions to remove the tweets that contain crisis information of some other location. Finally, the processing will extract geo locations and data visualization tool will help build a graphical representation of it. And the end user will use the user interface of this application.



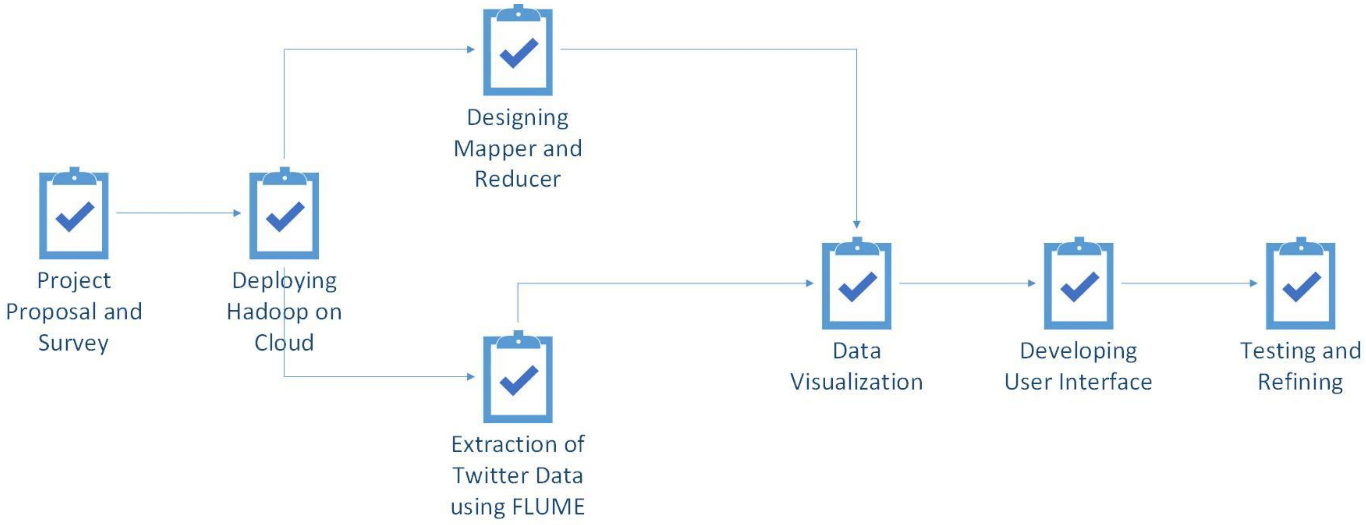
3. Workflow Diagram

*A. Project Overview*

The project is implemented in 6 tasks. **First task** is to install and configure Hadoop in a private cloud. **Second task** is to retrieve data from Twitter using Apache FLUME and sink this data to HDFS. **Third task** is to design Mapper and Reducer to process the data. **Fourth task** is to visualize data generated in the last task. **Fifth task** is to develop a user interface for better communication. **Sixth task** is to test the application and refine it.

**Mid-Term goal** for this project is to complete till task 3, designing Mapper and Reducer. We will use these Mapper and Reducer to analyze data and return outputin a readable format.

**Final-Term goal** for this project is to complete all the tasks that involve generating a final graphical representation of the output and a user interface.


Designing Mapper and Reducer
Project Proposal and Survey
Deploying Hadoop on Cloud
Data Visualization
Developing User Interface
Testing and Refining
Extraction of Twitter Data using FLUME

4. Task Dependencies

*B. Task 1 : Deploying Hadoop on cloud*

This task involves setting up a working environment for this project. It involves understanding Hadoop and its configuration files and making a multi node cluster on a private cloud which is the main requirement to process the data.

*C. Task 2: Extracting Twitter data using FLUME*

This task involves collecting tweets using FLUME. FLUME provides this service by establishing a connection with the help of twitter API for collecting streaming data. This data will further be used by Mapper and Reducer.

*D. Task 3: Designing Mapper And Reducer*

This task involves designing of Mapper and Reducer which will filter tweets based on certain conditions similar to the ones used by NEIC (National Earthquake Information centre) to improve the efficiency is identification of crisis. This part will also filter out the tweets that are generated from a location that is different from the crisis location. This will give only those tweets that further needs to be processed for determining the type of crisis and its geo location.

### E. *Task 4: Data Visualization*

This task involves using the data generated in the previous step to be represented on a graph, or a heat map by using some kind of data visualization tools like Big Insights or Spark. The Reducer will generate the data in such a format that can be directly given as an input to the above mentioned tools.

### F. *Task 5: Developing User Interface*

This task involves creation of user interface which will enable the user to just select a region or a crisis and see the end results in graphical format. If the user enters a region, it will display details for all crises that happened in that region. And if the user enters a crisis, it will return regions.

### G. *Task 6: Testing and Refine*

This task involves testing of the developed application for performance and efficiency testing as well as any potential bugs.

### H. *Project Task Allocation*

| | **Gaurav Shad** | **Nitesh Gupta** | **Abhishek Vellanki** |
|---|---|---|---|
| **Deploying Hadoop on Cloud** | 34% | 33% | 33% |
| **Extraction of Twitter Data using FLUME** | 33% | 33% | 34% |
| **Designing Mapper and Reducer** | 33% | 34% | 33% |
| **Data Visualization** | 33% | 33% | 34% |
| **Developing User Interface** | 33% | 34% | 33% |
| **Testing and Refining** | 34% | 33% | 33% |

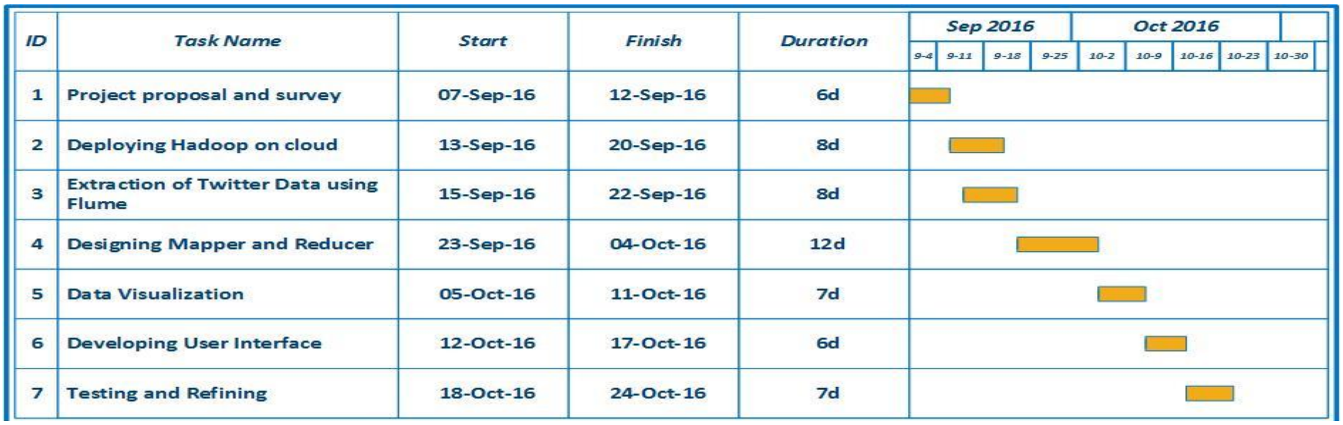
5. Task Allocation

### I. *Deliverables*

Outcomes of the project

- Working Hadoop Map Reduce Environment in a private cloud.
- User Interface designing to receive input from user.
- An application of big data analysis using twitter data.
- Data Visualization on processed data

### J. *Project Timeline*



| ID | Task Name | Start | Finish | Duration |
|---|---|---|---|---|
| 1 | Project proposal and survey | 07-Sep-16 | 12-Sep-16 | 6d |
| 2 | Deploying Hadoop on cloud | 13-Sep-16 | 20-Sep-16 | 8d |
| 3 | Extraction of Twitter Data using Flume | 15-Sep-16 | 22-Sep-16 | 8d |
| 4 | Designing Mapper and Reducer | 23-Sep-16 | 04-Oct-16 | 12d |
| 5 | Data Visualization | 05-Oct-16 | 11-Oct-16 | 7d |
| 6 | Developing User Interface | 12-Oct-16 | 17-Oct-16 | 6d |
| 7 | Testing and Refining | 18-Oct-16 | 24-Oct-16 | 7d |

6. Gantt Chart

## IV. Risk Management of the Project

| **Risk** | **Mitigation Strategy** |
|---|---|
| Hadoop installation and configuration is a tricky. Issue can come up with the configuration of xml files. | To solve this issue, installation must be done with proper time and care. |
| Lack of large data sets from Twitter because of rate limit of Twitter API | Twitter API must be run in parallel and flume must be used to get more data |
| Distinguishing between normal tweets and useful tweets (crisis information) | Conditions needs to be improved so as filter out the data precisely |

## V. Conclusion

To summarize, this project results in data visualization from the output of MapReduce algorithm. It returns us a graphical representation of the crisis that already happened in a region, queried in by the user. The tweets are streamed using Hadoop, Flume and Twitter API. Mapper and Reducer will be used to filter out the tweets based on certain conditions and process the required data. A user interface will help user interact with this big data analysis application to gain knowledge on this topic.

Futurework:
In the Mapper and Reducer, K-means or some other clustering method can be implemented to improve the quality of data and efficiency of the application. Additionally, data from other social media like Facebook can also be extracted and processed to return more meaningful data.

## Acknowledgment

We would like to express our gratitude towards our professor, Dr. Dijiang Huang and Teaching Assistant Mr. Ankur Chowdhary for assisting and giving us valuable inputs so that we could deliver it in time.

## References

[1] Shamanth Kumar, Fred Morstatter, Huan Liu "Twitter Data Analytics" http://tweettracker.fulton.asu.edu/tda/TwitterDataAnalytics.pdf

[2] “How the USGS uses Twitter data to track earthquakes”
https://blog.twitter.com/2015/usgs-twitter-data-earthquake-detection

[3] “How-to: Analyze Twitter Data with Apache Hadoop”
http://blog.cloudera.com/blog/2012/09/analyzing-twitter-data-with-hadoop/

[4] Manoj Kumar Danthala, Dr. Siddhartha Ghosh "Bigdata Analysis: Streaming Twitter Data with Apache Hadoop and Visualizing using BigInsights"
https://www.ijert.org/view-pdf/13162/bigdata-analysis-streaming-twitter-data-with-apache-hadoop-and-visualizing-using-biginsights

[5] “Hadoop” http://hadoop.apache.org/

[6] “FLUME” http://flume.apache.org/

[7] “Open Stack” http://www.openstack.org/

[8] “Twitter Developers API” https://dev.twitter.com/docs/api/1.1/get/search/tweets