

# **TREND PREDICTION IN TWITTER**

Project report submitted in partial fulfillment of the requirement for the  
degree of Bachelor of Technology

in

**Computer Science and Engineering**

By

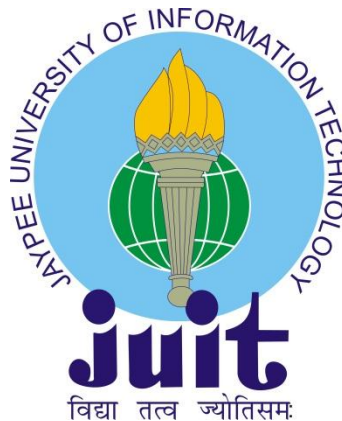
Hempushp Kaushal (141359)

Amber Goel (141373)

Under the supervision of

Ms. Ruhi Mahajan

to



Department of Computer Science and Engineering

**Jaypee University of Information Technology Waknaghat, Solan-  
173234, Himachal Pradesh**

## Candidate's Declaration

We hereby declare that the work presented in this report entitled **“Trend Prediction in Twitter”** in partial fulfillment of the requirements for the award of the degree of **Bachelor of Technology** in **Computer Science and Engineering** submitted in the department of Computer Science & Engineering and Information Technology, Jaypee University of Information Technology Waknaghat is an authentic record of our own work carried out over a period from August 2017 to May 2018 under the supervision of **Ms. Ruhi Mahajan** (Assistant Professor (Grade II) , Department of CSE and IT).

The matter embodied in the report has not been submitted for the award of any other degree or diploma.

Hempushp Kaushal (141359)

Amber Goel (141373)

This is to certify that the above statement made by the candidates is true to the best of my knowledge.

Ms. Ruhi Mahajan

Assistant Professor (Grade II)

Department of CSE and IT

Dated:

## Acknowledgement

It is our privilege to express our sincerest regards to our project supervisor **Ms. Ruhi Mahajan** for their valuable inputs, able guidance, encouragement, whole-hearted cooperation and direction throughout the duration of our project.

We deeply express our sincere thanks to our Head of Department **Prof. Dr. Satya Prakash Ghrera** for encouraging and allowing us to present the project on the topic “Trend Prediction in Twitter” at our department premises for the partial fulfillment of the requirements leading to the award of B-Tech degree.

At the end we would like to express our sincere thanks to all our friends and others who helped us directly or indirectly during this project work.

Date:

Hempushp Kaushal (141359)

Amber Goel (141373)

## TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION	1
1.1 INTRODUCTION	1
1.1.1 Big Data	2
1.1.2 HDFS	2
1.2 PROBLEM STATEMENT	4
1.3 OBJECTIVE	5
1.4 METHODOLOGY	6
CHAPTER 2: LITERATURE SURVEY	7
2.1 “Efficient Analysis of Big Data Using Map Reduce Framework” [1] , Dr. Siddaraju, Sowmya C L, Rashmi K, Rahul M	7
2.2 “Approaches to Analyse Big Data” , Pranay Agarwal [2]	9
2.3 “An Algorithmic Approach to Trending Tweet Prediction and Recommendation” by Bharath Srivatsan and Kelly Zhou [3]	10
2.4 “Data streaming in Hadoop” by Kim Björk, Jonatan Bodvill [4]	11
2.4.1 HDFS	11
2.5 “Apache Flume Architecture” by Deepak Ranjan, Dr. Tripti Arjariya, Dr. Mohit Gangwar [8]	12
2.6 “Processing on Apache Pig” by Ammar Fuad and Alva Erwin [10]	15
CHAPTER 3: SYSTEM DEVELOPMENT	17
3.1 DESIGN	17
3.1.1 Apache Hadoop ecosystem	17
3.1.2 HDFS	18
3.1.3 Stream Data From Twitter api using Flume	21
3.1.4 Using Apache Pig to post process Flume data	21
3.1.5 Display the results by using HTML and Javascript.	22
3.2 Model Development	24
CHAPTER 4: PERFORMANCE ANALYSIS	27
CHAPTER-5 CONCLUSIONS	29
5.1 Conclusion	29
5.2 Future Scope	30
REFERENCES	31

## List of Abbreviations

RT	Retweet
HDFS	Hadoop Distributed File System
BI	Business intelligence
SQL	Structured Query Language
JSON	Javascript object notation
YARN	Yet Another Resource Negotiator
LoC	Lines of code
API	Aapplication programming interface
SPoF	Single purpose of Failure
PHP	Hypertext Preprocessor
HTML	Hypertext Markup Language

## List of Figures

Figure 1:MapReduce Architecture .....	8
Figure 2: Flume .....	12
Figure 3: Flume Agent .....	13
Figure 4: Flume Fetching Data .....	21
Figure 5: Apache Pig Architecture.....	22
Figure 6: Flow Chart .....	23
Figure 7: Creating Twitter Application .....	24
Figure 8:Json Data Stored in HDFS.....	25
Figure 9: Example.....	26
Figure 10: Output.....	26
Figure 11: Test cases .....	27
Figure 12: Test cases graph.....	28

## **Abstract**

Twitter is a very popular on-line social networking and microblogging service, that allows many ample users share short messages in real time concerning events value broad attention expressing popular opinion. These messages in mixture indicate the interests and a focus of the native and world communities, that specially is known as temporal trends in twitter.

So, some time, we have a tendency to simply need to grasp what topics can become hot on Twitter, and why it becomes hot Therefore, we'd like to predict the trend of topics and provide some explanations for the vital variation of trend

A Tweet in twitter will have hashTags and an exact hashTag used most variety of times in tweets globally is claimed to possess highest trend. This data is large and additionally keeps on increasing, thus to process it in traditional manner wouldn't be doable. Hence we would need hadoop framework to fetch and process the data.

# CHAPTER 1: INTRODUCTION

## 1.1 INTRODUCTION

In the modern world of data age, the magnitude of on-line social-media activity has reached new level. a lot of users take part in social awareness streams like microblogging and social networking services to distribute information within the network. Twitter is one such very popular on-line microblogging and social networking service, that allows many ample users share short messages in real time concerning events value broad attention expressing popular opinion. These messages in mixture indicate the interests and a focus of the native and world communities, that specially is known as temporal trends in twitter.

As a microblogging service, it permits users to use a brief text within a limit of one hundred and forty characters as their posts (also referred to as tweets). There are unit 3 styles of tweets: retweet,reply and traditional tweet. Reply may be a message to any user with the “@” sign followed by the username of the person at the start of the tweet. Retweet may be a message that is shared by the user But is originally tweeted by someone else that begins with the sign “RT”. traditional tweet are the tweets except replies and retweets. Users also can add tags to the tweets. On Twitter the tag is named “hashtag”, and it begins with sign “#”. differing types of tweets and hashtags build Twitter information stream more clear. Twitter is incredibly appropriate for news info business enterprise and propagation. The text of tweet is brief, thus it's simple for users to post their tweets with mobile phones or different tools.

There are several events and topics mentioned on Twitter, a number of that get ton of attention to become a “trend” and a few don't. detecting these trends in on-line social websites has become a crucial downside that has attracted the eye of each the trade and also the analysis community in recent years.

So, some time, we have a tendency to simply need to grasp what topics can become hot on Twitter, and why it becomes hot Therefore, we'd like to predict the trend of topics and provide



some explanations for the vital variation of trend. Our main goal is to analyze and discover the rising trends and topics that are there within the stream.

### **1.1.1 Big Data**

“Big data is very a term that is used to represent the availability and growth of data, be it unstructured or structured.” Organisations collect data from various sources .This data comes in various types of formats such as video, audio, email documents sensor data.Big Data is so huge that it's no easy to process if one uses the traditional techniques. Increasing the amount of data increases the accuracy of analysis. It is the process of alalyzing the data to predict the useful info which can help in making better decisions in the future.

“Big-data analytics – It’s the process of analyzing, organizing and collecting huge datasets (big data) to find patterns and some useful data information.” It helps to know about the info contained inside the data,and it also helps to know about the the part of data that is most significant to the business. Big data analysts fundamentally want the knowledge that we get by analyzing data. Analysing data resources and users to make effective decisions using the data that was earlier unusable.

### **1.1.2 HDFS**

“HDFS, the Hadoop Distributed File-System is a distributed file-system which is designed to run on the commodity machines.” HDFS is based on Filing System that war earlier developed by google. This framework works on a mere model that the data will be stored into. HDFS is made to hold huge quantity of data (terabytes or petabytes or even zettabytes), and provide access to this information. In hdfs each file is split into blocks of a preset size. these blocks are then stored across several nodes. Hdfs can be deployed on machine that supports Java.

Hdfs can be deployed on machine that supports Java. Hadoop Map-Reduce is a progarmming model which is used to examine big data. Map-Reduce is a a technique for analyzing the data in

large amount because of its easy programming, fault tolerant and high scalability model. It is a parallel-programming model which is used to get data from the cluster of Hadoop. The term Map-Reduce simply refers to the 2 different tasks named “map” and “reduce” that are performed by the programs in hadoop .It's splits the task and execute them on different nodes parallely and speeds up the computation time .There then the data is fed up to the function called “map function” as the key and value pairs which produces the intermediate keyvalue pairs once the mapping process is done all the reasons from different ports are reduced to create the final result.

## 1.2 PROBLEM STATEMENT

The problem statement can be defined as given below:

- Get the tweets in the form of a stream of tweets and to get the information out of these tweets we have to process these tweets. Further process the “fitered” tweets to detect the “trending” topics.

This problem has three main key aspects :

- To take the input from the Twitter which is in the form of a stream and to store it.
- To create a model to process the tweets feed and to filter-out the information that we don't need and to ouput only relevant information.
- Creating a system to display the trends in a sorted manner in a graphical form

### **1.3 OBJECTIVE**

We need to predict the trend of topics and give some explanations for the important variation of trend. Detecting these online social trends has become a crucial challenge that has got the attention of the industry and the researchers recently. Our Objective is to analyze and detect the trends that appear in the stream, We need to get tweets from the Twitter API and then process these tweets in by using suitable techniques. We may need to filter out some irrelevant information and then we need to display the output in a sorted manner so that we can compare which topic is trending more than other topics.

## 1.4 METHODOLOGY

A Tweet in twitter will have hashTags and an exact hashTag used most variety of times in tweets globally is claimed to possess highest trend.

This data is large and additionally keeps on increasing, thus to process it in traditional manner wouldn't be doable.

Hence we would need hadoop . “Apache Hadoop is an open source framework which is used for distributed storage and processing of datasets.” It uses the Map-Reduce programming model

- “MapReduce is a processing technique and a distributed computing model that supports java.”
- The MapReduce rule contains 2 necessary tasks, particularly Map and reduce. Map takes a information and changes it to another dataset wherever individual parts are broken into key-value pairs. Secondly, the reduce part gets output from map as its input and then it combines these into another set of tuples which is smaller than previous.
- As implied by the name Map-Reduce, the map task and the reduce task are done one after the other .

So, to find the highest n trends in a given time interval, we might want to:

1. Process all the Tweets and parse out the tokens with HashTags.
2. Count the no. of hashTags.
3. Top n hashtags can be found by sorting all the hashtags according to their frequency .

But if we do that we are only considering the the topics that are tweeted with hashtags but what if users tweet anything without using hashtag then we cannot accurately predict what is trending so we need to consider all the words in a tweet rather than only the hashTags.

In order to do this we need to process all the words in a tweet then filter out the common words As the most common words would be the most occurring so it would not give us the result we want.

After filtering out the common words the result we get is the combination of the trending words and the trending hashtags so we can predict what really is trending.

## **CHAPTER 2: LITERATURE SURVEY**

### **2.1 “Efficient Analysis of Big Data Using Map Reduce Framework” [1] , Dr. Siddaraju, Sowmya C L, Rashmi K, Rahul M**

“Big data is a popular term used to describe the exponential growth and availability of data, each structured and unstructured. big data is also important to business and society.”[1] Big data is thus massive that it's troublesome to process using the techniques which being used in the past. A lot of data could result in a lot of correct analyses. a lot of correct analyses could result in a lot of assured higher cognitive process. And higher choices will mean bigger efficiencies, reductions in the value and less risk.

Hadoop relies on an easy information model, any information can match. HDFS is designed to carry terribly giant amounts of data (GB's, TB's). and supply access to the current data. Hadoop “MapReduce” may be a technique that analysis huge information. Map-Reduce has currently a replacement prototype for huge-scale information analysing because of fault tolerance and being an easy model of programming. “MapReduce” truly refers to 2 different tasks “map” and “reduce”. These tasks are executed by the programs in hadoop.

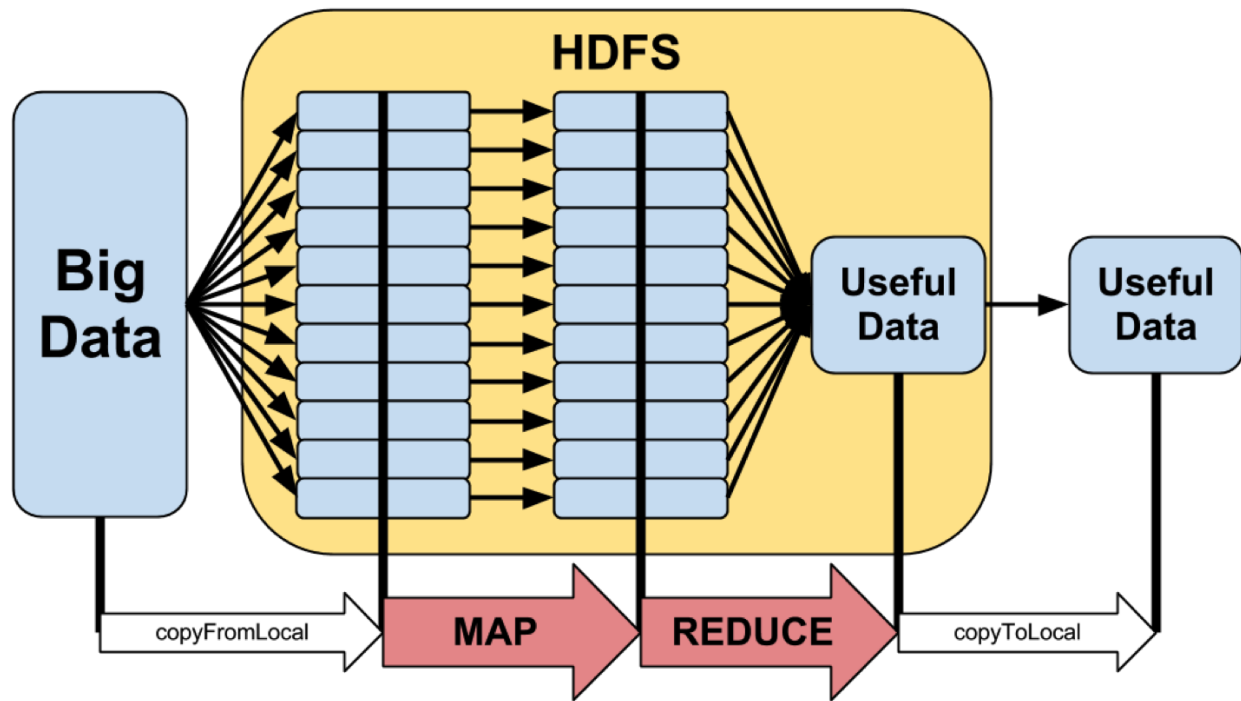


Figure 1:MapReduce Architecture

## **2.2 “Approaches to Analyse Big Data” , Pranay Agarwal [2]**

The conventional BI(Business Intelligence) solutions don't perform as a result of it we need to pre-process the information before it reaches analysing stage. Moreover, the measurability of a number of these dbs providing the volumes of information concerned is doubtful.

- a) “Distributed processing” – “Apache Hadoop” and “MapReduce” provide a similar functionality that suggests that The data is distributed to several nodes and then It is processed at each node simultaneously
- b) “In Memory Databases” – These are the Database Management system which use main-memory for accessing the data. Analysing does not require memory –however data accessibility is a doubt in case of equipment failure / closedown.
- c) “NO SQL Databases” – also known as non relational databases as they don't seem to be supported rigid schemas and may simply comprise new datatypes



### **2.3 “An Algorithmic Approach to Trending Tweet Prediction and Recommendation” by Bharath Srivatsan and Kelly Zhou [3]**

The users of Twitter work to maximize their visibility online, like by building their follower base or by maximizing the impressions on the tweets. If a user joins trending conversations in its early stage he will get more social media recognition, as these trending topics often top the charts on Twitter. So, the ability to detect the trending topics would enable users to engage in conversations before they trend, and thus will put the users at the forefront. “Trend prediction analysis is a promising means of enabling users to maximize their Twitter visibility.”[3] Moreover, Twitter often serves to introduce users to new topics and information sources.

One of the main reasons Twitter identifies trends is to provide a curated list of potentially interesting items to users. Not all trending information aligns with the interests of the different members of the diverse Twitter community, though, and as such, we identify the utility of a recommendation engine. Beyond predicting upcoming trends, recommending potentially trending topics to users would introduce them to ideas that are closely linked to their interests. A recommendation engine can also be used as a tool for users to leave footprints on growing discussions that they have vested interests in, thereby magnifying their impact. With thorough prediction and recommendation analysis, we thus find an opportunity to optimize the user experience on Twitter through maximized visibility.

## **2.4 “Data streaming in Hadoop” by Kim Björk, Jonatan Bodvill [4]**

The Hadoop Framework is based on the google file system which used Map-Reduce model to process data on clusters of machines. This was made achievable by parallel processing. When some data is stored in hadoop. It is then copied over distinct nodes. This eliminates the possibility of data unavailability in case of failure of a node.

### **2.4.1 HDFS**

“HDFS, the Hadoop Distributed File-System is a distributed file-system which is designed to run on the commodity machines.” HDFS is based on Filing System that was earlier developed by google. This framework works on a mere model that the data will be stored into. HDFS is made to hold huge quantity of data (terabytes or petabytes or even zettabytes), and provide access to this information. In hdfs each file is split into blocks of a preset size. these blocks are then stored across several nodes. Hdfs can be deployed on machine that supports Java.

Hdfs can be deployed on machine that supports Java. Hadoop Map-Reduce is a programming model which is used to examine big data. Map-Reduce is a technique for analyzing the data in large amount because of its easy programming, fault tolerant and high scalability model. It is a parallel-programming model which is used to get data from the cluster of Hadoop. The term Map-Reduce simply refers to the 2 different and distinct tasks named “map” and “reduce” that are performed by the programs in hadoop. It's splits the task and execute them on different nodes parallelly and speeds up the computation time. There then the data is fed up to the function called “map function” as the key and value pairs which produces the intermediate keyvalue pairs once the mapping process is done all the reasons from different ports are reduced to create the final result.

## 2.5 “Apache Flume Architecture” by Deepak Ranjan, Dr. Tripti Arjariya, Dr. Mohit Gangwar [8]

The following diagram depicts the fundamental design of Flume. As depicted within the figure below, the data generators (like Twitter, Facebook) create information that gets collected by the Apache Flume agents that are in the cluster. After that, the data that is then collected and transferred to a central store such as HDFS..

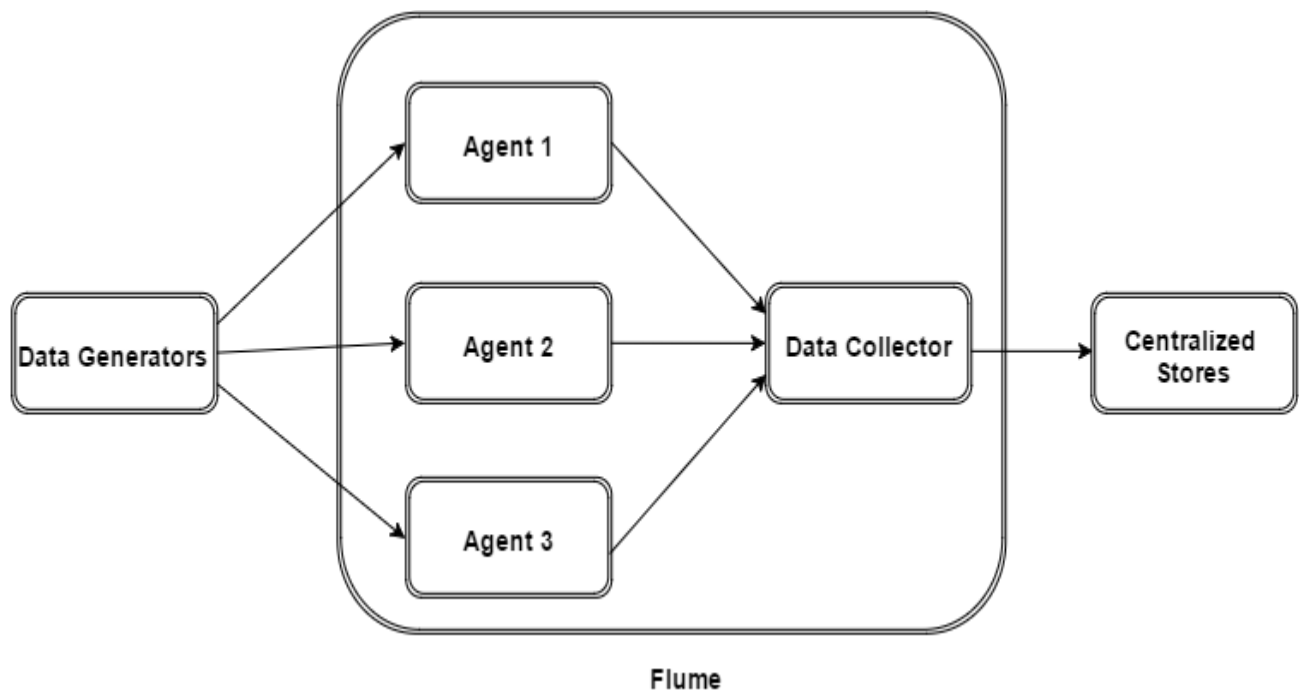


Figure 2: Flume

### Flume Event

“An event refers to basic unit of the data that is transported within Flume.” It comprises of load of computer memory unit array that's to be sent from the source to the destination.

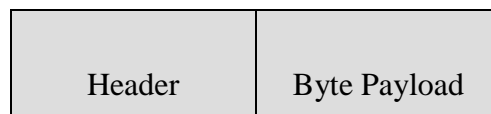


Figure 3: Structure of flume event

## Flume Agent

“A Flume Agent is an independent daemon process in Flume that receives the data from clients or alternative agents and then it forwards the data to the next destination (agent or sink).” There can be more than one Flume agents.

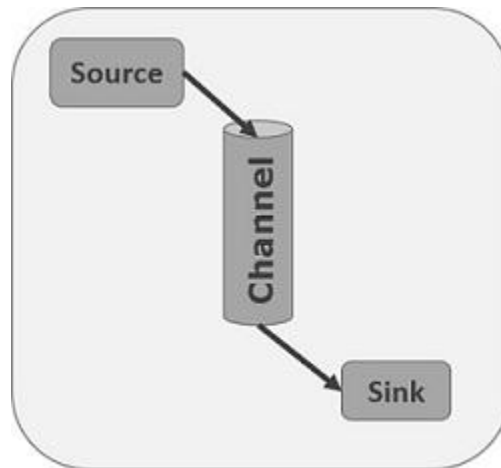


Figure 4: Flume Agent

a “Flume Agent” has 3 main parts which are channel, source, and sink.

## Source

It gets the data from data generator . then it sends it to the channels. “Apache Flume” can support many sources. These sources can be of different types.

Examples – Thrift source , twitter 1%source, Avro source, etc.

## Channel

It is a store that receives flume events and buffers them until these get consumed to the sinks.

e.g. – classification system, Memory ,JDBC channel.

## **Sink**

A sink stores the information in central stores like HDFS. It gets data from channel and sends that for its destination, which may be the central stores or another agent.

e.g – H D F S sink

## **Additional parts of “Flume Agent”**

There are a few more components which play a significant role in transferring of the events from the info generator to the central stores.

## **Interceptors**

Interceptors alter/inspect flume events that are transferred between supply and channel.

## **Channel Selectors:**

These confirm which channel is to be used to transfer the info when there are multiple channels.

There are 2 styles of channel selectors :

- Default channel selectors - These are called replicating channel selectors they replicates all the events in every channel.
- Multiplexing channel selectors -These decides the channel to send an incident supported the address inside the header of that event.

## **Sink Processors**

Their task is to invoke a specific sink from the cluster of sinks. “These are used to create the failover paths for load balance events across multiple sinks from a channel.”

## **2.6 “Processing on Apache Pig” by Ammar Fuad and Alva Erwin [10]**

Apache Pig a tool that is used to research big sets of information. It is used along with “Hadoop” framework; data manipulation operations can be performed in Hadoop by using Apache Pig.

Pig provides a high-level language which is referred to as Pig Latin to write programs. This language has many operators by which the programmers can create functions to read, write & process the data.

Programmers write scripts in PigLatin to analyse the data in Apache Pig. Then all the scripts are converted into “Map” and “Reduce” tasks internally. Apache Pig features a part referred to as Pig-Engine. The pig engine takes the scripts in PigLatin as input . then it converts these to the Map-Reduce jobs.

Programmers can perform Map-Reduce tasks by writing scripts in PigLatin, simply while not having to write complicated code in Java language.

Pig takes advantage of multiquery approach, hence decreasing the size of code to be written. for instance, an operation that will need you to write two hundred lines of code (LoC) in Java is simply done by writing as less as simply ten LoC in Apache Pig. Ultimately, Using Pig cuts the dev. time by nearly sixteen times.

PigLatin is similar to SQL and is very easy to understand. It provides several integral operators to support knowledge operations like filters, joins, ordering, etc. additionally, it also provides you with nested data types such as bags, tuples, and maps that are not in MapReduce.

“Apache Pig” has the subsequent options -

Set of operators - to perform operations such as filter, sort, join, etc.

Easy to program - PigLatin is comparable to S.Q.L.

- simple to write a program.

User Defined Functions - It has the ability to form User-defined Functions

Data Handling – It can handle different types of data, each structured or unstructured and keeps the results in H.D.F.S.

## **CHAPTER 3: SYSTEM DEVELOPMENT**

### **3.1 DESIGN**

#### **3.1.1 Apache Hadoop ecosystem**

“Apache Hadoop is a software Framework that is open source and it is used for distributed storage and processing of the dataset of big data and uses the mapreduce model of programming.” .It consists of clusters of computer. The modules that are there in “Hadoop” are made in a way such that that if there are any hardware failures the these are considered to be common occurrences and hence should be automatically handled by the framework.

The core of Hadoop has a storage part, which is referred to as Hadoop Distributed file-system Also known in short as (HDFS), and it also consists of a processing part that is called the MapReduce programming model. It is highly fault tolerant and is designed to be used on low cost hardware . Hadoop divides the files into giant blocks and sends them to nodes. The packaged code is then transferred to notes and the data there is processed in parallel. This approach permits data processing to be quicker and a lot of with efficiency than it might be in an exceedingly more typical mainframe computer design that depends on a parallel classification system where high-speed networking is used to compute and distribute the data.

The base “Apache Hadoop” framework contains the these modules:

“Hadoop Common” – contains “Shared libraries and utilities required by different Hadoop modules;”

“Hadoop Distributed File system (HDFS)” – “a distributed file-system that stores data on machines;”

“Hadoop YARN” – It allocates resources to the tasks. This allows dstinct users to execute different applications without having to worry about the increase in the workloads.it is a accountable for the task of resources managing and then use these resources for the applications



.We can basically say that it is responsible for dividing the functionalities of resources and to schedule the tasks to be performed.

“Hadoop MapReduce” – It is implementation of the Map-Reduce programming model that processes in task into 2 steps i.e. “map” and “reduce” task. It executes task parallelly by distributing it in small blocks.

The term “Hadoop” does not simply refers to the base modules and submodules, however there are other packages that can be used in hadoop like “Apache Pig, Apache Hive, Apache HBase, Apache Spark, Apache ZooKeeper, Apache Oozie, Apache Phoenix, Apache Flume, Apache Sqoop, Apache Storm and Cloudera impala.”

The Hadoop frame work is written in Java although Map-Reduce Java code is common, Java code is used to carry out the “map” and “reduce” components of the users program

### **3.1.2 HDFS**

“Hadoop” distributed file-system

The “HDFS” is a scalable, distributed, & moveable file-system that is written for “Hadoop” . It is considered to be a datastore however it also allows shell-commands and JavaAPI methods that are like alternative filing systems. TCP/IP sockets are used to to communicate. clients use “remote procedure calls” (“RPC”) to communicate with each another.

A “Hadoop” cluster has one namenode and a multiple datanodes, though redundancy choices are accessible for the namenode owing to its criticality. every datanode operates blocks of information over the n/w employing block protocol for “Hadoop distributed file-system.”

It stores giant files ((GB) to (TB)) in multiple computers. Dependability is achieved by duplicating the data in several hosts, so in theory doesn't need any redundant array of independent disks (R.A.I.D.) storage on.

The default replication value is three which means that the data is kept on three nodes: 2 on a similar rack, and one on a special rack. The data nodes will communicate to each another to keep the duplication of information high, to move data, and to rebalance the data on the nodes. The advantage of not having a completely POSIX-compliant file-system is enhanced performance for data o/p .

It has a secondary namenode. As a result of the namenode is that the sole purpose is to manage and store the data, it will become a problem for supporting a large range of files, particularly an outsized range of small files. However, there are some problems in HDFS like little file problems, measurability issues, “Single purpose of Failure (SPoF)”, data requests problems. However rack awareness is one of the advantages which prevents needless data transfer. once “Hadoop” is employed with alternative filesystems, this advantage isn't invariably obtainable. this could have a major impact on the time to complete a job .

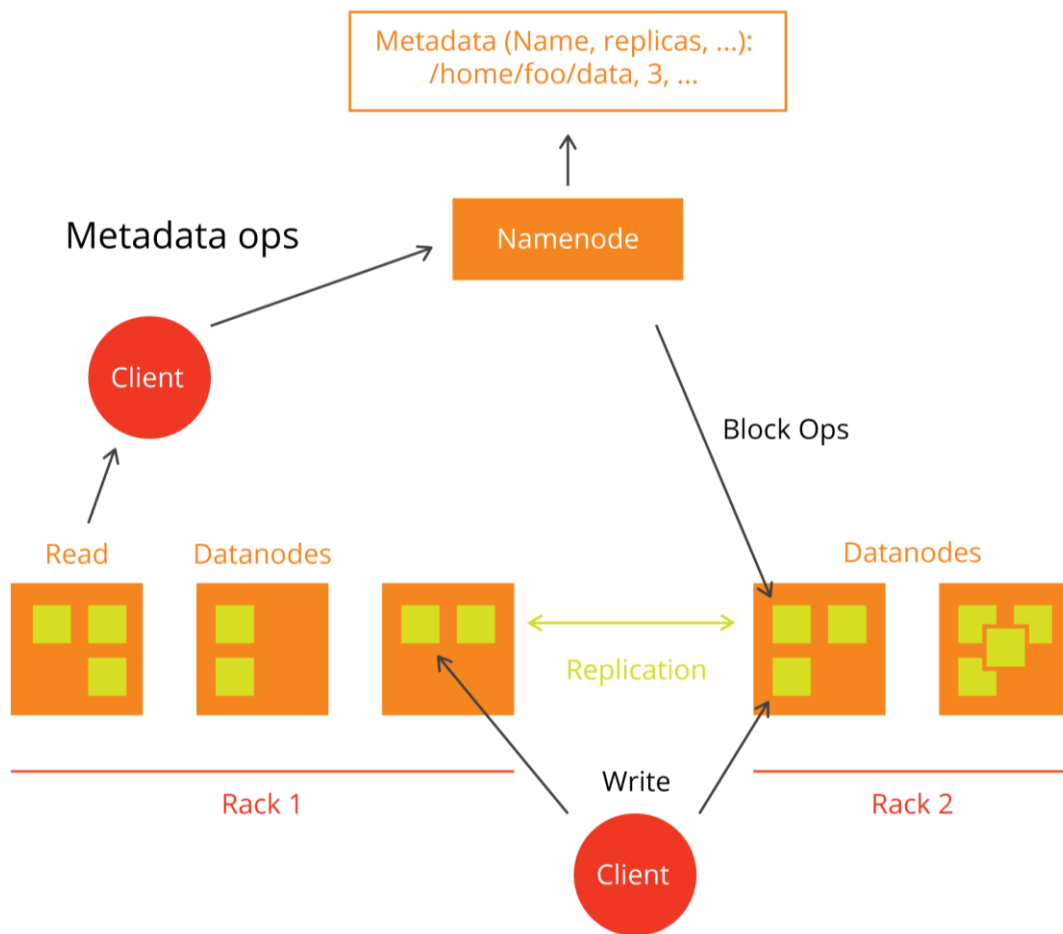


Figure 5: HDFS Architecture

“HDFS” was made for changeless files and should not be appropriate for systems that demand simultaneous writing operations.

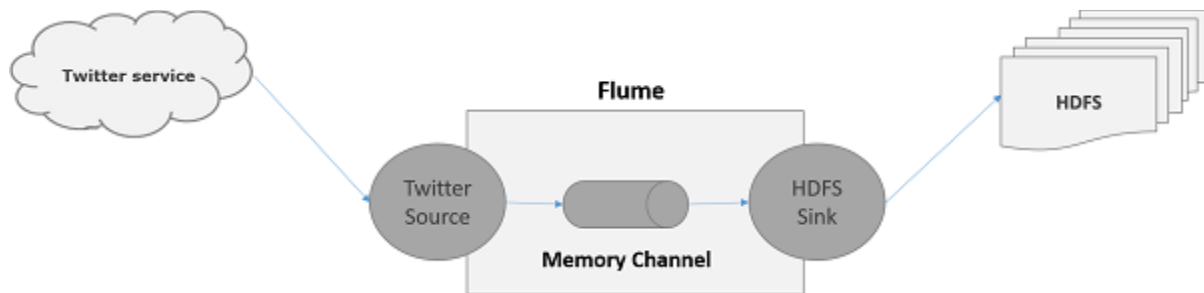
File access can be accomplished by using the JavaAPI, the ThriftAPI the CLI, the HDFS-UI net application over hypertext transfer protocol.

### 3.1.3 Stream Data From Twitter api using Flume

By using “Apache Flume”, we can fetch the data from various services (like twitter) and transfer it to the centralized stores (H.D.F.S. and HBase). We can get data from Twitter API and store it in H.D.F.S. using Apache-Flume.

Log data is generated by a webserver and then the data is collected by a Flume agent. The channel then buffers the data to a sink. The sink then sends it to central stores.

We can fetch the data from Twitter by using the Twitter API as the source and using flume to fetch the data from Twitter API .To fetch the data we first have to create an account with developer privileges on Twitter after that we need to generate the consumer key and consumer access token . These keys and tokens are then used to configure the agent in flume.



### Figure 6: Flume Fetching Data

### 3.1.4 Using Apache Pig to post process Flume data

“Apache Pig” works on top of “Hadoop”. It analyses huge sets of data that are there in “Hadoop Distributed File-System”. Initially we have to load the data into “Pig” to analyze it.

## Preparing HDFS

In Map-Reduce mode, It loads the data from Hadoop Distributed File-System and then it stores the results back into H.D.F.S.

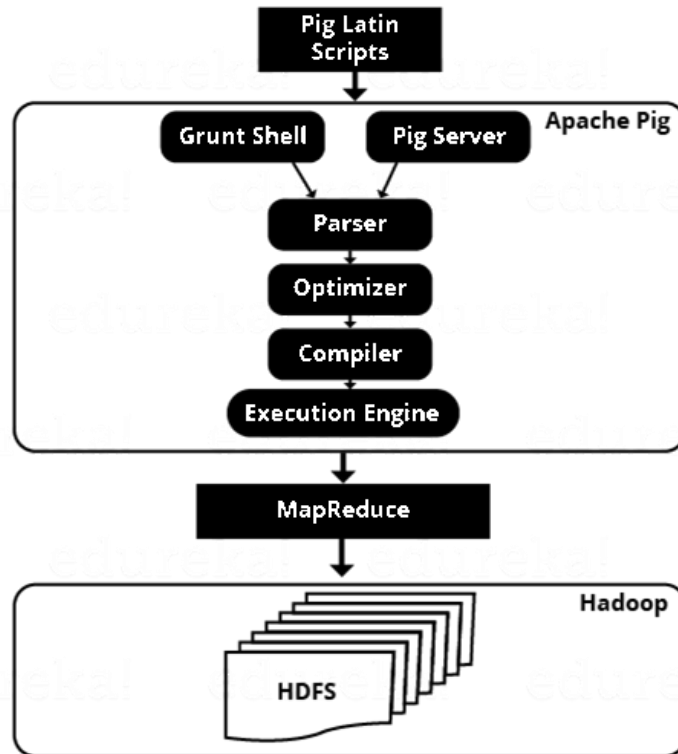


Figure 7: Apache Pig Architecture

### 3.1.5 Display the results by using HTML and Javascript.

We can read the results from the file created by pig after processing the data .

To do this we used the HTML Filereader API to read the data and then storing it in an array of strings .

Then we can display this data in graphical form by using chart.js.

### 3.1.6 System Flow chart

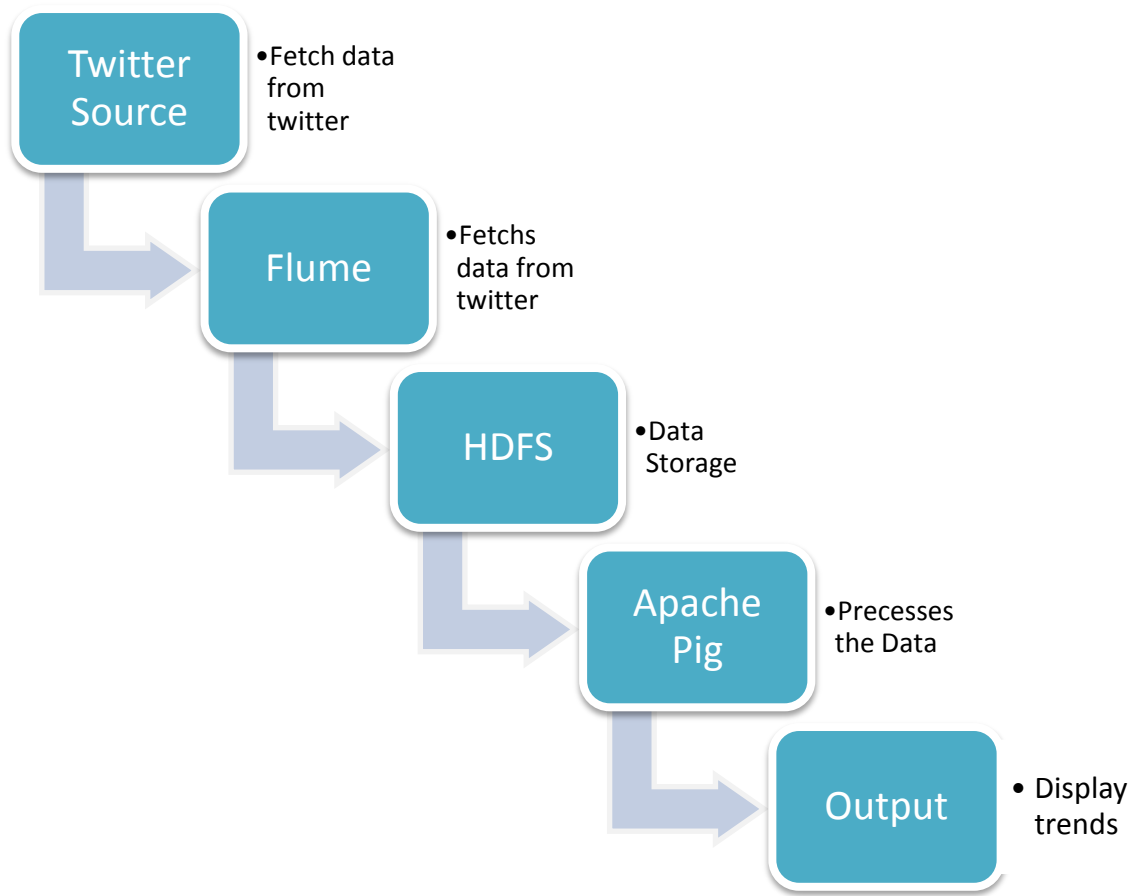
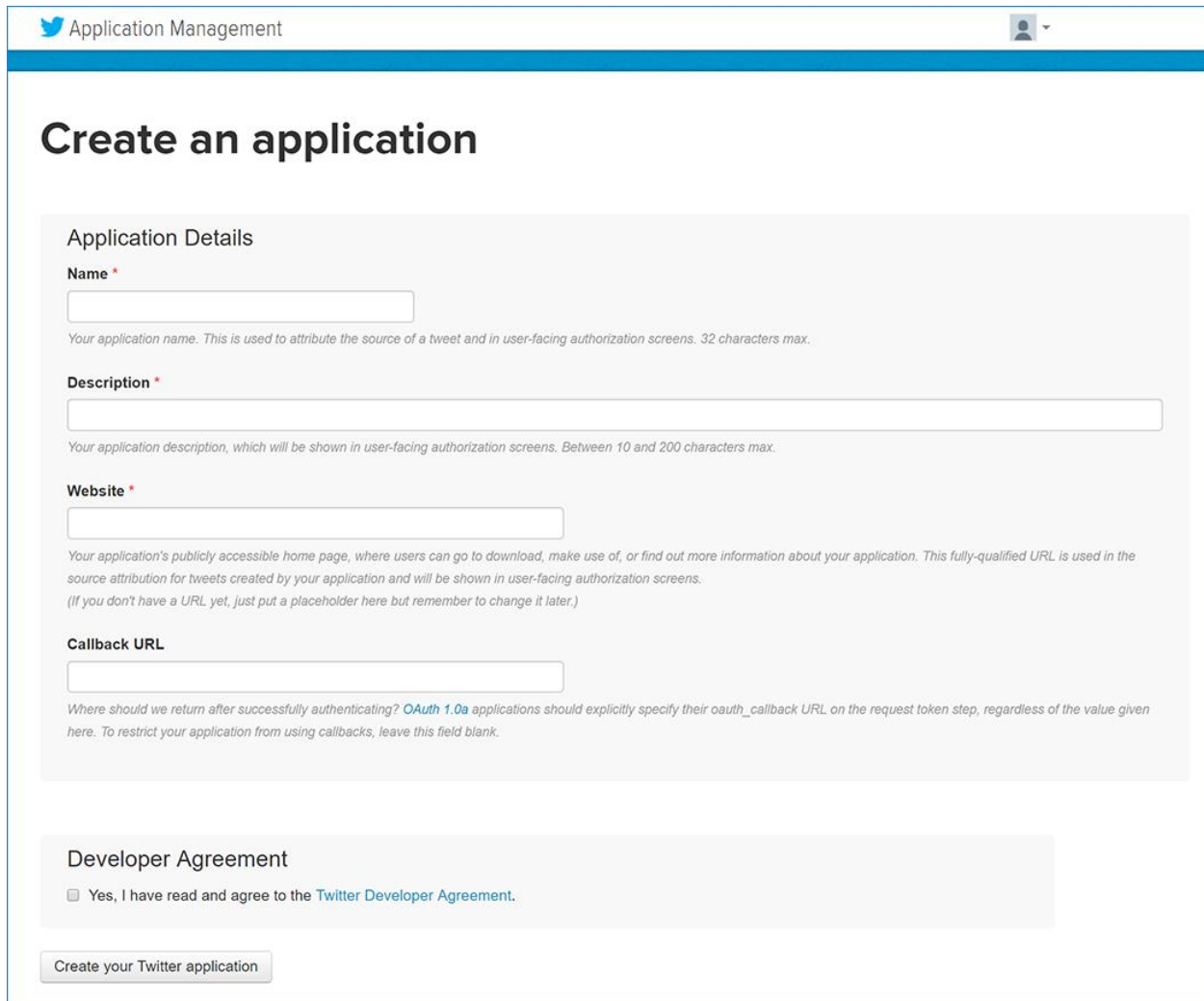


Figure 8: Flow Chart

## 3.2 Model Development

A Tweet in a twitter can have hashTags (#'s) and a particular hashTag that is used most number of times in tweets worldwide is said to have the highest trend.

1.First we have to create an application twitter and then we can use Twitter API's for fetching the real-time social data and then store it into HDFS.



The screenshot displays the 'Application Management' interface on Twitter. The main heading is 'Create an application'. Below this, there is a section titled 'Application Details' which contains four input fields: 'Name', 'Description', 'Website', and 'Callback URL'. Each field has a corresponding text box and a small asterisk indicating it is required. Below each input field is a line of explanatory text. The 'Name' field is followed by the text: 'Your application name. This is used to attribute the source of a tweet and in user-facing authorization screens. 32 characters max.' The 'Description' field is followed by: 'Your application description, which will be shown in user-facing authorization screens. Between 10 and 200 characters max.' The 'Website' field is followed by: 'Your application's publicly accessible home page, where users can go to download, make use of, or find out more information about your application. This fully-qualified URL is used in the source attribution for tweets created by your application and will be shown in user-facing authorization screens. (If you don't have a URL yet, just put a placeholder here but remember to change it later.)' The 'Callback URL' field is followed by: 'Where should we return after successfully authenticating? OAuth 1.0a applications should explicitly specify their oauth\_callback URL on the request token step, regardless of the value given here. To restrict your application from using callbacks, leave this field blank.' Below the 'Application Details' section is a 'Developer Agreement' section with a checkbox and the text: 'Yes, I have read and agree to the Twitter Developer Agreement.' At the bottom of the form is a button labeled 'Create your Twitter application'.

Application Management

## Create an application

**Application Details**

**Name \***

Your application name. This is used to attribute the source of a tweet and in user-facing authorization screens. 32 characters max.

**Description \***

Your application description, which will be shown in user-facing authorization screens. Between 10 and 200 characters max.

**Website \***

Your application's publicly accessible home page, where users can go to download, make use of, or find out more information about your application. This fully-qualified URL is used in the source attribution for tweets created by your application and will be shown in user-facing authorization screens. (If you don't have a URL yet, just put a placeholder here but remember to change it later.)

**Callback URL**

Where should we return after successfully authenticating? OAuth 1.0a applications should explicitly specify their oauth\_callback URL on the request token step, regardless of the value given here. To restrict your application from using callbacks, leave this field blank.

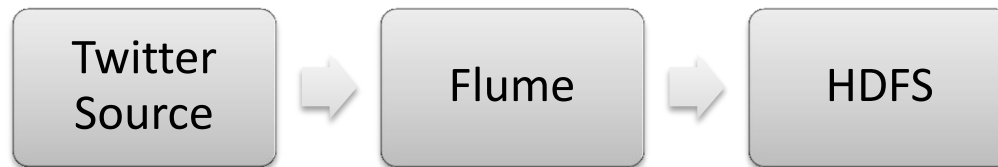
**Developer Agreement**

☐ Yes, I have read and agree to the [Twitter Developer Agreement](#).

Create your Twitter application

Figure 9: Creating Twitter Application

2. To fetch the data we can use tools like apache flume with which we can authenticate our keys and start fetching data from twitter.



3. After fetching the data is stored into H.D.F.S. which is used for storing such huge amount of data.

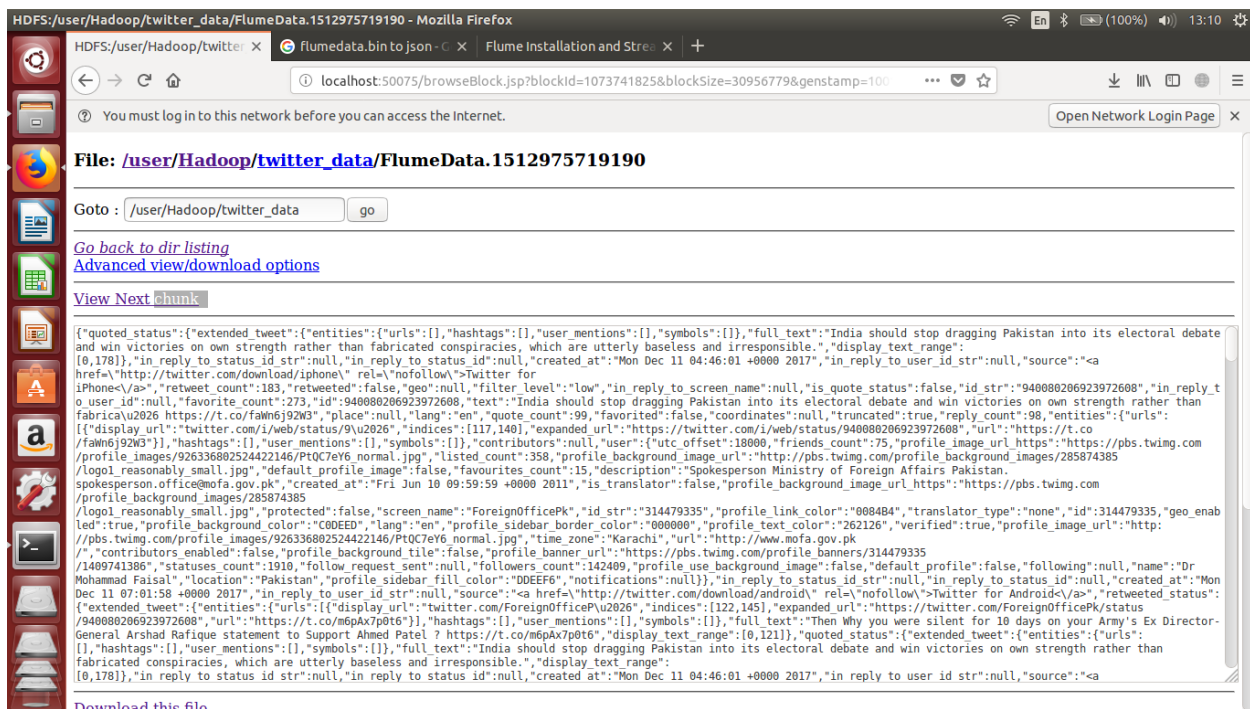


Figure 10:Json Data Stored in HDFS

4. After storing data into H.D.F.S, we can process it.

5. Then we can start analyzing such large amount of data using tools such as Apache Pig.

6. We need to process all the data in a tweet then filter out the unnecessary data .

- Join the Data with left outer join.
- Filter out where the stopwords is null.



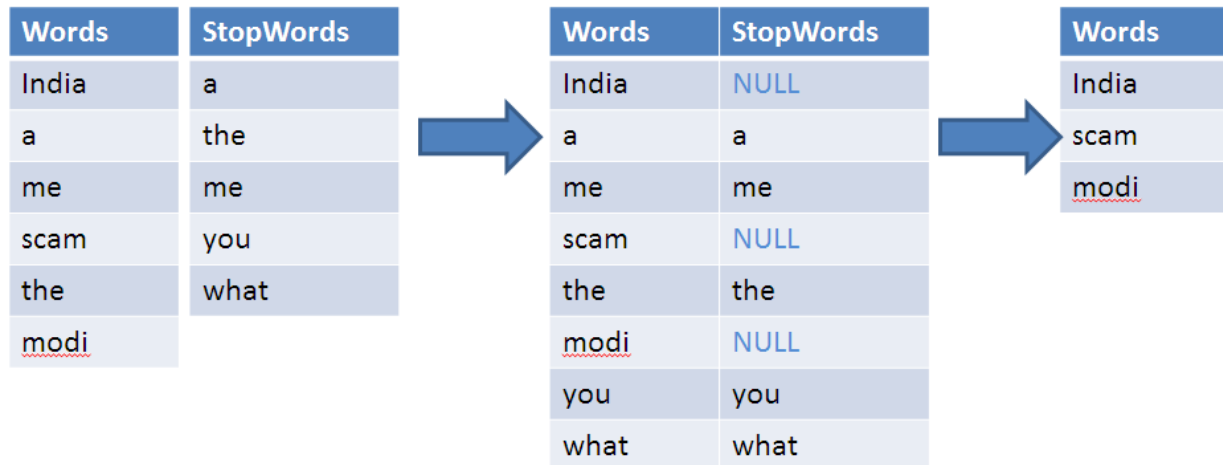


Figure 11: Example

7. Then we can display the trending topics by sorting the topics according to their no. of times of occurrence .

8. Then we can make a user interface which can show us in visual form that what are the current trends. With the help of HTML and JavaScript we can Read the output produced by pig and then use that information to displayThe output in the form of Bar graph.By using HTML file reader API we can read the contents of the file and then store those contents in an array of Strings which are later used to create the graph.To create graph we can use chart.js which is the JavaScript library that uses HTML5 Canvas element to draw different types of charts.

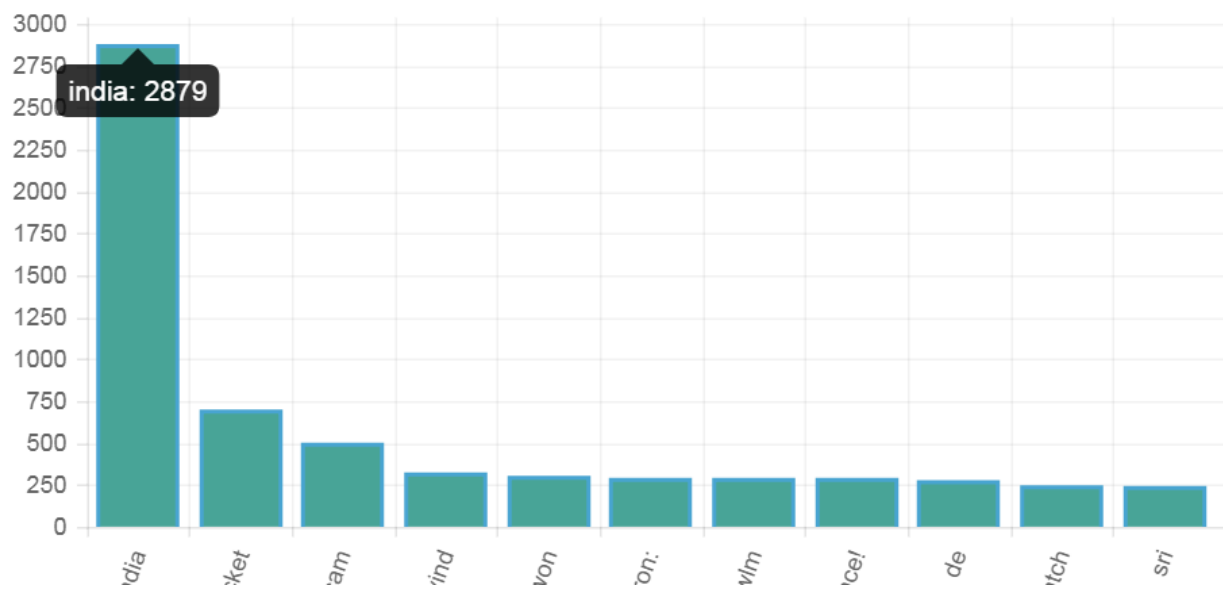


Figure 12: Output

## CHAPTER 4: PERFORMANCE ANALYSIS

### Performance Analysis of MapReduce to find frequency of words

Test Cases:

Test File	No. of Words	Time Taken (ms)
test1.txt	28,554	8450
test2.txt	84,773	9480
test3.txt	3,05,984	10550
test4.txt	6,11,968	12499
test5.txt	12,28,936	16670

Figure 13: Test cases

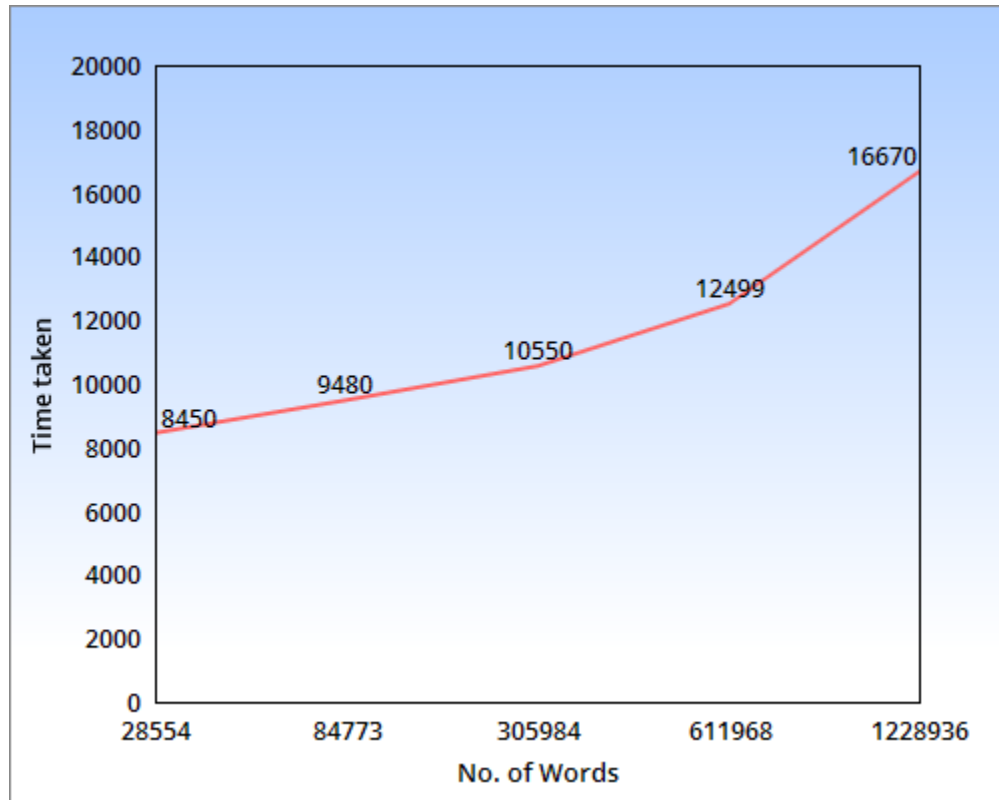


Figure 14: Test cases graph

## **CHAPTER-5 CONCLUSIONS**

### **5.1 Conclusion**

As we are in an age of big data, to process the huge amount of information has not been this easy. By using better tools like MapReduce with “Hadoop” and “HDFS”, assures quicker advancement in several areas and increasing the benefits and success of the many organizations.

Map-Reduce has got plenty of attention in various areas, such as data processing, image retrieval, pattern recognition, data retrieval, and machine learning. Still the quantity of information that require to be processed grows, several processing ways have not become appropriate

It was quite complex and much research was needed to be done in this project because the field of distributed systems and Hadoop is very huge. These kind of projects have advantages for researchers students.

The main contribution of this project is to analyze the trends in online social websites. This model can help in classifying “trends” and “non-trends” in very early stages. We have also developed a highly efficient model to filter noise from tweets. Also this prediction model is generic enough to be applied on any social media network, which has connection among users.

## 5.2 Future Scope

We have certainly established the benefits of constructing social graph of twitter. So resolving more relations for the users in the graph can be useful and can improve the performance of the model. The model's success depends on the topic wise clustering of tweets. Currently we have used simple clustering which not very strict", other better clustering algorithms can be used. Storing and processing graphs have been the real challenge and bottleneck of the whole pipeline. It's necessary to improve this step by exploring better ways to do the same. Other data structures and algorithms can be explored to process the data faster.

**Predict Pre-trending Tweets Based on Currently Trending Tweets** After compiling a list of common tweets, we can compare this list of possible topics with the list of currently trending topics. We can thus compare possible pre-trending tweets with currently trending tweets to predict accordingly.

## REFERENCES

- [1] Dr. Siddaraju, Sowmya C L, Rashmi K, Rahul M “Efficient Analysis of Big Data Using Map Reduce Framework” ISSN 2347-6435(Online) Volume 2, Issue 6, June 2014
- [2] Pranay Agarwal “Prediction of Trends in Online Social Netwok” ,April 2013
- [3] Bharath Srivatsan and Kelly Zhou “An Algorithmic Approach to Trending Tweet Prediction ” March 2016
- [4] Z. X. Rong Lu and Q. Yang, “Trends predicting of topics on twitter based on macd,” in National Laboratory of Pattern Recognition, 2012.
- [5] Lecture 6 – “Trend Detection In Twitter Social Data” (Analyzing Big Data With Twitter), Berkeley School of Information ,2015
- [6] Shilpi Taneja, ManishTaneja “Big data and Twitter” , International Journal Of Research In Computer Applications And Robotics, Vol.2 Issue.5, Pg.: 144-150 ,May 2014
- [7] Kim Björk, Jonatan Bodvill “Data streaming in Hadoop-A Study Of Real Time Data Pipeline Integration Between Hadoop Environments And External Systems” Stockholm, Sweden 2015
- [8] Deepak Ranjan, Dr. Tripti Arjariya, Dr. Mohit Gangwar “Trend Analysis Using Hadoop and Its Ecosystems” International Journal of Innovative Research in Computer and Communication Engineering Vol. 5, Issue 5, May 2017
- [9] Evan Miller, Kiran Vodrahalli, Albert Lee “Estimating Trending Topics on Twitter with Small Subsets of the Total Data”. ,January 2013[10] <https://deliciousbrains.com/using-javascript-file-api-file-upload/> , “Using Javascript file api”

[10] <https://deliciousbrains.com/using-javascript-file-api-file-upload/> , “Using Javascript file api”

[11] <https://flume.apache.org/releases/content/1.6.0/FlumeUserGuide.html> , “Flume User guide”

[12] <https://www.scribd.com/document/313266640/Apache-Flume-Fetching-Twitter-Data> , “Fetching Data From Flume Using Apache Pig”

[13] <http://bigdatadimension.com/fetching-streaming-data-using-apache-flume/> , “Fetching Streaming Data Using Apache Flume”

[14] [https://hadoop.apache.org/docs/r1.2.1/hdfs\\_design.html](https://hadoop.apache.org/docs/r1.2.1/hdfs_design.html) , “HDFS Architecture”

[15] <http://a4academics.com/tutorials/83-hadoop/835-hadoop-architecture> , “Hadoop Architecture HDFS and MapReduce”