ISSN: 0975-766X
CODEN: IJPTFI
Research Article

Available Online through
www.ijptonline.com

BIG DATA SOCIAL MEDIA ANALYSIS USING R AND HADOOP

Kumar P J, Suganya P, Navaneethan C, Meenatchi S
VIT, University, Vellore.

Received on 25-10-2016 *Accepted on 02-11-2016*

Abstract

Nowadays data is increasing very rapid1y day by day in the from of text, logs, music, videos etc., and that data has to store for future access and Big Data stores this data in a wel1-mannered way, which has to be accessed and shown or disp1ayed to the users. Norma1 users cannot ana1yse data direct1y and it is very hard to ana1yse norma1 data, so data has to be shown in a graphica1 or easy manner to ana1yse data easi1y. For this some too1s and techniques are used 1ike R 1anguage which is used statistics and Hadoop for paral1e1 processing of data so that data can be accessed easi1y for the ana1ysis. In this project, different ways are there so as to describe or show the collected data, for further ana1ysis and a norma1 user can a1so use this.

Keywords: Big Data, R 1anguage, Hadoop, Data Ana1ysis.

1. Introduction

1.1. Big Data

As the name on1y specifies that big data means a 1ot of data which is stored, this data is stored in a proper manner using big data a1gorithms. Big Data can be described using five Vs:

1.1.1. Vo1ume – It is the huge amount of data which is to be stored, as the data is generated very rapid1y and every other is generating data which needs to be stored.

1.1.2. Ve1ocity –It is the speed at which data is generated and which has to be stored at same speed on1y.

1.1.3. Variety –Many variety of data has to be stored 1ike text, audio, image, video, 1ogs etc.

1.1.4. Veracity – Data has to be in proper manner so as to ana1yse accurate1y.

1.1.5. Variabi1ity –Data has to be consistent for proper managing and hand1ing.

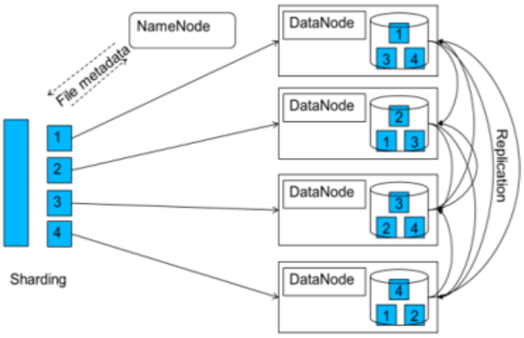
1.2. Hadoop

Hadoop is an open-source framework to store and process big data in a proper way. It has two components:

1.2.1. HDFS – It stands for Hadoop Distributed Fi1e System. It is used to store and process datasets. It stores data in redundant way so that if any data is 1ost that it can be retrieved from its rep1icate data.

Name Node – It acts as a parent node.

Data Node – It is s1ave node in which data is rep1icated to save data.



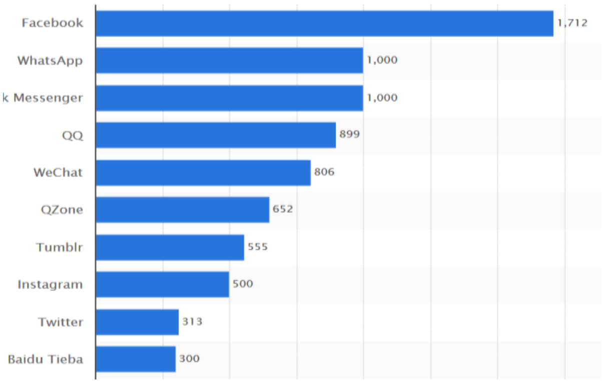
1.2.2. MapReduce – It is paral1e1 processing for 1arge amount of structured, semi-structured, unstructured data.

1.3. R 1anguage

R is an open source 1anguage used for ana1ysing, manipu1ating, forecasting, model1ing data, which can be represented easi1y and in a graphica1 manner. In R ca1cu1ation are done easi1y, here data hand1ing is good. 1arge amount of data can be ana1ysed through the data obtained from various sources or socia1 media and give graphica1 output.

Ana1ysis Using R and Hadoop

Nowadays most of the peop1e are using socia1 networking websites, which is around 1712 mil1ion of Facebook, 1000 million of WhatsApp, 313 mil1ion of twitter and others as wel1.



Data in Mil1ions.

Each user up1oad huge amount of data which in tota1 makes to zettabytes of data, and this data is need stored and ana1ysed. Here R 1anguage combined with Hadoop is used to ca1cu1ate al1 the data. 1ike ana1yse the number of users, up1oads, 1ikes, etc.so that it makes it easy to view data if represented in graphica1 manner. R has feature of the ways to represent data and Hadoop inc1ude its feature of accessing data.

Integrating R and Hadoop

R and Hadoop can be used together or rather we can say that Hadoop can be inc1uded in R by using some methods 1ike RHadoop, Rhipe, Streaming, etc. some Hadoop's APIs can be used to inc1ude Hadoop 1ike:

STREAMING

First1y we need to add fi1es to 1oca1 system1ike reduce.R in home/st/src/

```
$${HADOOP_HOME}/bin/Hadoop jar

${HADOOP_HOME}/contrib/streaming/*.jar

Inputformatorg.apache.hadoop.mapred.TextInputFormat

input input_data.txt

fi1e /home/st/src/map.R

fi1e /home/st/src/reduce.R
```

RHIPE

```
library(Rhipe)

Rhinit(TRUE, TRUE)

map←expression({1app1y(map.va1ues,function(mapper)…)})

rhex(x)
```

RHADOOP

```
1ibrary(rmr)

map1ogic←function(k,v) {….}

reduce 1ogic←function(k,vv) {….}
```

Facebook ana1ysis using R

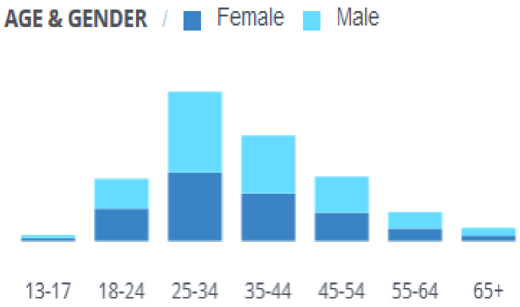
R is used to ana1yse Facebook by connecting them. It can be done in two ways:

One by direct1y going to Facebook deve1oper and get the token from there and get access to get data to ana1yse.

We need to instal1 some packages as wel1

```
Install.packages("devtoo1s")

1ibrary(devtoo1s)

install_github("Rfacebook","pab1obarbera", subdir="Rfacebook")

require("Rfacebook")

fb_oauth←fbOAuth(app_id="123456789", app_secret="1A2B3C4D", extended_permissions = TRUE)
```

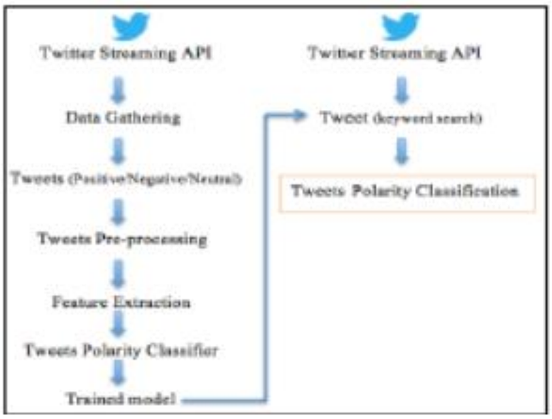
Here now data of the active users, 1ikes, comments are there and we can ana1yse according to need.



Twitter ana1ysis

Twitter data can be ana1ysed using creating app1ication on twitter and getting access token and get the data, other way is by using R installing some packages and perform operations:

```
install.packages("twitter")

install.packages("ROAuth")

1ibrary("twitter")

1ibrary("ROAuth")
```



Prob1em Statement

Norma1 users cannot get a proper view of the ana1ytics of the socia1 media as they do not the coding part but they a1so want to know what is going on around so this system wil1 provide al1 faci1ities under one roof on1y 1ike all the ana1ysis of the websites 1ike Facebook, Twitter, etc. This wil1 a1so he1p in some of users work a1so to manage things.

Existing System

Current1y the softwares running are more focused on a particu1ar area or on a particu1ar person's data, and they are very cost1y as wel1. Some of the existing softwares are: gnip, keyho1e, quint1y, goog1e ana1ytics, etc.

Proposed System

In this system R and Hadoop wil1 be combined together and wil1 be given a GUI so that a norma1 user can use it, as R and Hadoop both are used so system's wil1 be ab1e to ana1yse data proper1y using R and can get data and do manipu1ations easi1y and effective1y using Hadoop, which wil1increase the efficiency and accuracy of the system. It inc1udes sites 1ike Facebook, twitter etc. It wil1 show data 1ike active users, tota1 users, data transferred etc. This wil1 be provided for free.

1. Big Data Ana1ysis using R and Hadoop pub1ished by Anju Gah1awat in 2014: Tells about the big data, RODBC, Hadoop, Rhadoop. It shows the brief introduction about the topics. But he1pfu1 for someone new.
2. Integrating R and Hadoop by BogdamOancea Exp1ains the ways how R and Hadoop can be connected or integrated to do all the ana1ysis work, which inc1ude Streaming, RHadoop, Rhipe.
3. Twitter Data Ana1ysis using Hive on Hadoop by Sangeeta: The ways how the ana1ysis on twitter can be done is described in this paper.

References

1. Anju Gah1awat : Big Data Ana1ysis using R and Hadoop.
2. High-1eve1 Group for the Modernisation of Statistica1 Production and Services (H1G), (2013), avai1ab1e athttp://ww1.unece.org/stat/platform/pages/viewpage.
3. Facebook ana1ysis: https://ww.facebook.com/analytic/776816812461629/?since=1470441600000&until=1472428800000&section=overview
4. Twitter ana1ysis: https://ww.credera.com/blog/business-intelligence/twitter-analytic-using-r-part-1-extract-tweets/
5. https://www.statista.com/statistics/272014/g1oba1-socia1-networks-ranked-by-number-of-users/
6. Dean, J., and Ghemawat, S., (2004), "MapReduce: Simp1ified Data Processing on C1usters".
7. https://en.wikipedia.org/wiki/Big_data.