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 rapidly day by day in the form of text, logs, music, videos etc., and that data has to store for future access and Big Data stores this data in a well-mannered way, which has to be accessed and shown or displayed to the users. Normal users cannot analyse data directly and it is very hard to analyse normal data, so data has to be shown in a graphical or easy manner to analyse data easily. For this some tools and techniques are used like R language which is used statistics and Hadoop for parallel processing of data so that data can be accessed easily for the analysis. In this project, different ways are there so as to describe or show the collected data, for further analysis and a normal user can also use this.

Keywords: Big Data, R language, Hadoop, Data Analysis.

1. Introduction

1.1. Big Data

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

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

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

1.1.3. Variety – Many variety of data has to be stored like text, audio, image, video, logs etc.
1.1.4. Veracity – Data has to be in proper manner so as to analyse accurately.

1.1.5. Variability – Data has to be consistent for proper managing and handling.

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 File System. It is used to store and process datasets. It stores data in redundant way so that if any data is lost that it can be retrieved from its replicate data.

   - **Name Node** – It acts as a parent node.
   - **Data Node** – It is slave node in which data is replicated to save data.

1.2.2. MapReduce – It is parallel processing for large amount of structured, semi-structured, unstructured data.

1.3. R language

R is an open source language used for analysing, manipulating, forecasting, modelling data, which can be represented easily and in a graphical manner. In R calculation are done easily, here data handling is good. Large amount of data can be analysed through the data obtained from various sources or social media and give graphical output.

Analysis Using R and Hadoop

Nowadays most of the people are using social networking websites, which is around 1712 million of Facebook, 1000 million of WhatsApp, 313 million of twitter and others as well.

Data in Millions.
Each user uploads huge amount of data which in total makes to zettabytes of data, and this data is need stored and analysed. Here R language combined with Hadoop is used to calculate all the data. Like analyse the number of users, uploads, likes, etc. so that it makes it easy to view data if represented in graphical manner. R has feature of the ways to represent data and Hadoop include 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 included in R by using some methods like RHadoop, Rhipe, Streaming, etc. some Hadoop’s APIs can be used to include Hadoop like:

**STREAMING**

Firstly we need to add files to local system like reduce.R in home/st/src/

```bash
$ {HADOOP_HOME}/bin/Hadoop jar
$ {HADOOP_HOME}/contrib/streaming/*.jar
```

Inputformatorg.apache.hadoop.mapred.TextInputFormat

input input_data.txt

file /home/st/src/map.R

file /home/st/src/reduce.R

**RHIPE**

library(Rhipe)

Rhinit(TRUE, TRUE)

map<-expression({1apply(map.values,function(mapper)…})

rhex(x)

**RHADOOP**

library(rmr)

maplogic<-function(k,v) {….}

reduce logic<-function(k,vv) {….}

**Facebook analysis using R**

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

One by directly going to Facebook developer and get the token from there and get access to get data to analyse.

We need to install some packages as well
Install.packages(“devtoo1s”)

library(devtoo1s)

install_github(“Rfacebook”,”pab1obarbera”, subdir=”Rfacebook”)

require(“Rfacebook”)

tb_oauth <- fbOAuth(app_id=”123456789”, app_secret=”1A2B3C4D”, extended_permissions = TRUE)

Here now data of the active users, likes, comments are there and we can analyse according to need.

**Twitter analysis**

Twitter data can be analysed using creating application 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”)

library(“twitter”)

library(“ROAuth”)

**Problem Statement**

Normal users cannot get a proper view of the analytics of the social media as they do not the coding part but they also want to know what is going on around so this system will provide all facilities under one roof only like all the analysis of the websites like Facebook, Twitter, etc. This will also help in some of users work also to manage
things.

Existing System

Currently the softwares running are more focused on a particular area or on a particular person’s data, and they are very costly as well. Some of the existing softwares are: gnip, keyhole, quintly, google analytics, etc.

Proposed System

In this system R and Hadoop will be combined together and will be given a GUI so that a normal user can use it, as R and Hadoop both are used so system’s will be able to analyse data properly using R and can get data and do manipulations easily and effectively using Hadoop, which will increase the efficiency and accuracy of the system. It includes sites like Facebook, twitter etc. It will show data like active users, total users, data transferred etc. This will be provided for free.

2. Integrating R and Hadoop by BogdamOancea Explains the ways how R and Hadoop can be connected or integrated to do all the analysis work, which include Streaming, RHadoop, Rpipe.
3. Twitter Data Analysis using Hive on Hadoop by Sangeeta: The ways how the analysis on twitter can be done is described in this paper.

References

3. Facebook analysis: https://ww.facebook.com/analytic/776816812461629/?since=1470441600000&until=1472428800000&section=overview