Working On R Programming Language For Big Data: A Review

Surbhi Raj[1], Dr. Priya Gupta[2], Onkar Singh[3]

[1]Student, B. Tech (CS), Maharaja Agrasen College, University of Delhi
[2]Asst. Professor, Maharaja Agrasen College, University of Delhi
[3]Asst. Professor, Shaheed Sukhdev College of Business Studies, University of Delhi

ABSTRACT

Data is present everywhere and is increasing day by day from multiple sources. It is generated by everything around us at all times. Analyzing these datasets at a massive scale is one of the biggest challenge. This paper presents the theory for analyzing and visualizing the huge amount of data sets and extracting the meaningful information using R programming language, for predicting the future trend. It introduces the basic structure of R and Hadoop platform for the analysis of big data sets and presents the different stages of data. Open source platform R language and R studio provides users with a big analysis of big data. R package provides an API to use Hadoop which handle large data sets and need more R packages for statistical algorithm computation and large data visualization. This paper focuses on the parallelized method, RHadoop for Big data Analytics to perform computations on big data sets. This paper also focuses on the implementation of tools for exploring data and comprises theory for managing large-scale datasets, which presents the reshape framework for restructuring the data, plotting the data and visualizing the statistical models. It also implements the function melt and cast of reshape package of R for data reshaping and aggregation. It also include the R package, ggplot2 for grammar of graphics for plotting the data. It also discusses strategies for visualizing statistical models.

Keywords: Big Data, R programming, R Hadoop, Mapreduce, rmr, RHBASE, RHDFS, reshape package, Rggobi, Matlab.

I. INTRODUCTION

Big Data is a large sets of ubiquitous data that can be structured and unstructured collected from multiple sources (computers, mobile devices, satellites, cameras, images etc.). It is very difficult to process large and complex sets of data using conventional database systems. Data is too big and cheap. It moves too fast and exceeds the processing capacity of traditional database management systems. The huge amount of fresh data, around 2.5 EBs (Exabytes) are produced by companies, programs and network sensors, every single day and estimated that over 90% of existing data has been generated during the last two years. [11] Big data is valuable for organization falls in two categories: analytical use and enabling new products that is predictive analysis of new products on the basis of past information about that products. It gain a more complex understanding of the relationships between different factors and to uncover previously undetected patterns in data. Big Data sizes range from a few hundred terabytes to many petabytes in a single sets of data. To process, analyze and store large amount of data, an open source platform “Hadoop” is used in a distributed environment. It is designed to scale up from single servers to thousands of machines offering local computation and storage.

R programming is a statistical language and environment for computational statistics, visualizations and data science. It is a scripting language which handle huge amounts of data quickly. It is a free and open source software. It is connected and compatible with almost all types of databases and programming languages. It has robust inbuilt graphical functions. R programming is one of the most powerful language for data mining, information retrieval and data analysis. It also supports machine learning algorithms and many more. It handles vectors, matrices and lists and has the ability to work like a scientific calculator. It is a collection of many inbuilt statistical and mathematical functions and graphical commands. By using R, we analyze the code line by line and save all our work in a file and go back to see what we investigated at a later date. So, It allows to easily share your work with other and see what others are doing with data. R has over 2,000 user contributed packages that increase its functionality. The most famous GUI of R is Rstudio. Rstudio is powerful integrated development environment (IDE) for R. It is used by 2M+ data scientists.

RHadoop is one of the most well-known R packages to support Hadoop functionalities for big data analytics. “R” statistical programming language and big data as a potential solution for the demand of Big Data analytics. Both are open source projects and data driven .R programmers manipulate Hadoop data stores directly from HDFS and HBASE. Also, using Hadoop streaming, they are able to write Map Reduce jobs in R [3]. It is a collection of three R packages: rmr, rhdfs and rhbase.

II. BIG DATA ANALYSIS USING R

Data is increasing exponentially every day from multiple sources. Today, Data is everything. An unstructured data is collected and then organized. Analyze those large data sets and make decision on the basis of this. There are different stages of data that are shown in Fig 1.

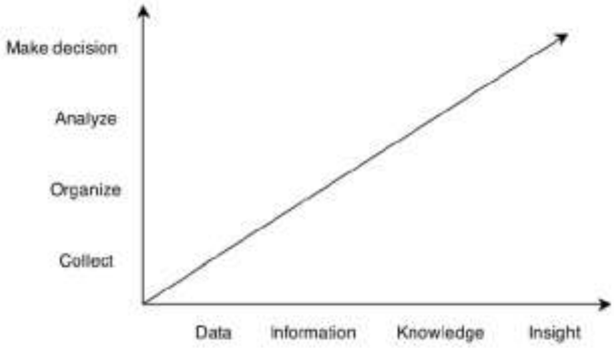


Fig.1. Different Stages of Data [11]

To process and store these large sets of data, a low cost platform Hadoop is used and R is a very amazing programming language to perform statistical models on data and translate the derived models into colorful graphs and visualizations. R and Hadoop are natural match for Big Data analytics and Data Science. R combines with Hadoop for performing Big data analytics (RHadoop).R programming language control and command both Hadoop(HDFS and MAPREDUCE) for storing and analyzing the data and Matlab acts as a server for processing Matlab functions, reading .mat –files, and embedding the functionality and power of Matlab to R via R’s external libraries.[11]

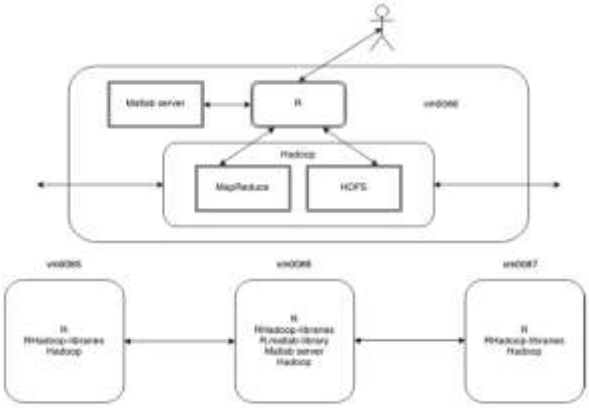


Fig.2. Communication between different components and nodes [11]

A. RHadoop

RHadoop is an open source project that provides client-side integration of R and Hadoop. It was developed by Revolution Analytics. It is a bridge between R, a programming language and environment to statistically explore data sets, and Hadoop, which is a framework that allows for the distributed processing of big data sets across clusters of computers [4]. RHadoop packages for analyzing Big Data are: rmr, rhdfs and rhbase.

B. RHDFS (File Management of the HDFS with R)

RHDFS is an R package which provides the basic connectivity to the Hadoop Distributed File System (HDFS). The main storage mechanism in Hadoop is HDFS (Hadoop Distributed File System). R programmers read, write, modify, and browse files stored in HDFS within R. Then, HDFS API is called in the backend to operate on the data sources stored in the HDFS. Huge amounts of information is stored, scaled up incrementally and survived the failure of significant parts of the storage infrastructure without losing data. The clusters of Machine is created in Hadoop and work is coordinated among them. The memory constraints of R is bounded which allows analyst to easily work with a data subset. It enables the R programmer to store models or other R objects that can be recalled and used in MapReduce jobs. After the execution of MapReduce jobs, it writes the results to HDFS.

C. RHBASE(HBase distributed information with R)

Rhbase is an R interface for operating Hadoop's HBase data source stored at the distributed network. This package is designed with several methods for initialization and read/write and table manipulation operations. Implementing HBase is to providing database a table structures. It opens up the Hadoop framework to R programmer.

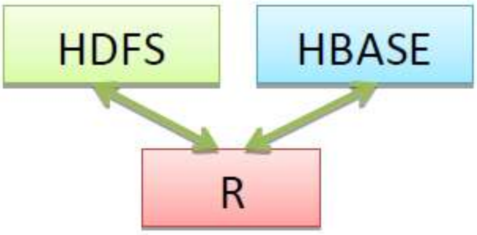


Fig.3 RHDFS and RHBASE [3]

D. RMR(Hadoop MapReduce Functionality in R)

R Programmers easily access the MapReduce programming paradigm using big data sets. It is the fast Parallelized Analytics on large data sets. Using this, it would be easiest, most productive and most elegant way to perform mapreduce jobs. This package allows R programmers to perform statistical analysis in R via Hadoop's MapReduce functionality on a Hadoop cluster. The Job of R programmer is reduced by using this package. They divide their application logic into the map and reduce phases and run it in parallel and submit it with Rmr methods. Then, Rmr calls Hadoop streaming MapReduce API with job parameters such as input directory, output directory, mapper, reducer and so on to perform the R MapReduce Job over the Hadoop cluster in huge data sets [3].The output of these jobs is then stored in traditional Data Warehouse and then the analysis of data can be done by using R.

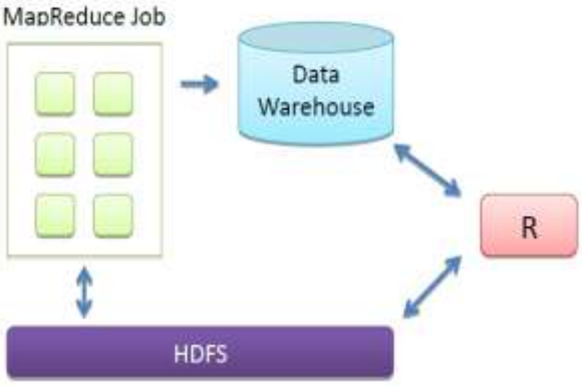


Fig.4: RHadoop Architecture [3]

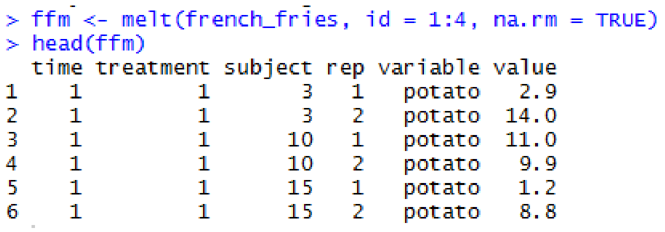
Tools for exploring big data sets
Three families of tools and packages in R for exploring data and models.

1. Reshaping data with the reshape package
2. A layered grammar of graphics for R, ggplot2
3. Visualizing statistical model

A. Reshaping Data with the reshape package

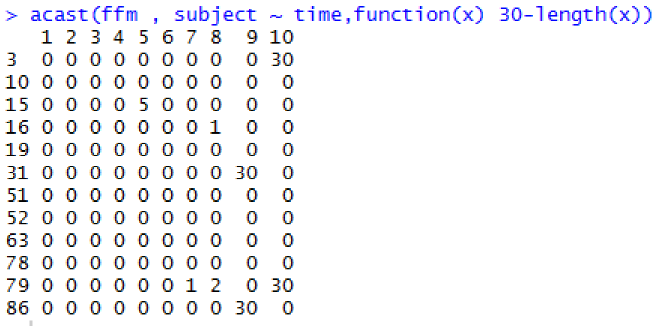
The reshape package for R provides a common framework for many types of data reshaping and aggregation. tapply, by and aggregate are general functions that can aggregate data. It uses just two functions: melt and cast. The many forms of data that melt can consume and cast can produce.

First melt the data to use in analyses:



```
> ffm <- melt(french_fries, id = 1:4, na.rm = TRUE)
> head(ffm)
  time treatment subject rep variable value
1    1         1       3   1   potato   2.9
2    1         1       3   2   potato  14.0
3    1         1      10   1   potato  11.0
4    1         1      10   2   potato   9.9
5    1         1      15   1   potato   1.2
6    1         1      15   2   potato   8.8
```

Cast function cast a molten data frame into an array or data frame. Investigate balance using length as aggregate function [12]



```
> acast(ffm , subject ~ time,function(x) 30-length(x))
   1 2 3 4 5 6 7 8  9 10
3  0 0 0 0 0 0 0 0  0 30
10 0 0 0 0 0 0 0 0  0  0
15 0 0 0 0 5 0 0 0  0  0
16 0 0 0 0 0 0 0 1  0  0
19 0 0 0 0 0 0 0 0  0  0
31 0 0 0 0 0 0 0 0 30  0
51 0 0 0 0 0 0 0 0  0  0
52 0 0 0 0 0 0 0 0  0  0
63 0 0 0 0 0 0 0 0  0  0
78 0 0 0 0 0 0 0 0  0  0
79 0 0 0 0 0 0 1 2  0 30
86 0 0 0 0 0 0 0 0 30  0
```

The Reshape framework divides the task into two components, first describing the structure of input data (melting) and the structure of output data (casting). [12]

B. A layered grammar of graphics

The R package for grammar of graphics, ggplot2 is used for plotting the data which gives a practical perspective and helps to see the bulk of the data and to make sense what's going on. The grammar of graphics is introduced by working through the process of creating a plot and the components that is needed.

Start with ggplot, which creates a new plot object and then add the other components: a single layer, specifying the data, mapping, geom and stat, the two continuous position scales and a Cartesian coordinate system.

This plot shows the relationship between the price and weight (in carats) of 1000 diamonds.

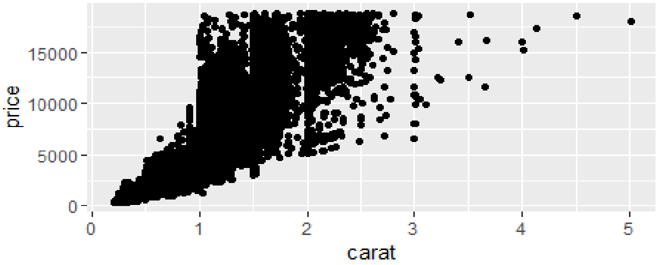
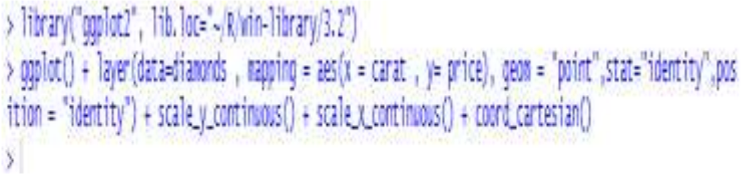


Fig.5. Scatterplot of price vs carat



```
> library("ggplot2", lib.loc="~/R/win-library/3.2")
> ggplot() + layer(data=diamonds , mapping = aes(x = carat , y= price), geom = "point",stat="identity",pos
ition = "identity") + scale_y_continuous() + scale_x_continuous() + coord_cartesian()
> 
```

This is the full specification of the scatterplot of price vs carat.

C. Visualizing statistical model

Visualizing statistical model is based on three strategies: display the model in the data space; look all members of a collection; and explore the process of model fitting, not just the end result [12]. It gives the mathematical summary, describing the main features of the data. Rggobi, an R package which connects the statistical programming environment of R. It is used with three R packages: classifly , clusterfly and meifly.

- **Display the model in data-space**

Visualizing the model in the context of data, displaying the model in the high-dimensional data space, which used to see how well the model responds to the data. This approach generate the full regression surface and visualize it in the context of data.

- **Collections are more informative than singletons**

This approach is to visualize all the collection of models simultaneously. This is the strategy of trace plot, which show hundreds of tree models simultaneously. It is possible to see the space of models, showing both the common and the unusual. Tools is used to select the smaller subsets of interesting models. It can be done by calculating the descriptive statistics on multiple levels, then linked brushing is used to connect these statistics to one another and to displays of the model.

- **Don’t just look at the final result; explore how the algorithms works**

In this approach, iterations is observed which helps to understand how the algorithm works and can reveal potential pitfalls. Also, suggest possible avenues for improvement.

III. LITERATURE REVIEW

Harshawardhan S. Bhosle & Prof. Devendra P. Gadekar stated the techniques and technology to store, manage, capture, analyze the Big data of petabyte- or huge-sized data sets which is structured or unstructured form and in capable to fit in traditional Database management methods. To process huge sets of data, described an inexpensive and efficient way, parallelism is used. They illustrated the Big Data, its definition and 3 V’s of Big Data are Volume, Variety and Velocity Also the Problems with Big Data Processing. The problems are Heterogeneity and Incompleteness of Information, rapidly increasing volume of data, its fast moving speed, Human collaboration and the security of Big Data is huge concern. They provided Hadoop as a solution of Big Data Processing and explained Hadoop Architecture, HDFS Architecture and Mapreduce Architecture. And also focused on comparison of different components of Hadoop [5].

Ulla Gain & Virpi Hotti, explained the Big Data Analytics which is the examination of Big Data. They presented the current reported knowledge in terms of definition of Big Data Analytics. They launched a term Data-miling to explain an effort to uncover the information nuggets and explained data- miling as an examination of heterogeneous data or as a part of competitive advantage. Also gave the example which is concerned about the investments of coal power in Europe. Finally, described the current status of Big Data phenomenon [6].

Shilpa and Manjit Kaur, similarly described the Big Data, its parameters and its various challenges. And established Hadoop which uses MapReduce paradigm for Big Data processing, again. This review paper reports the Real Time Literature Review about Big Data, According to 2013, facebook’s users 1.11 billion accounts is active from which 751 million users using facebook from mobile. Flicker having feature of unlimited photo uploads (50MB per photo), Unlimited video uploads (500 MB per video), unlimited storage, Unlimited Bandwidth. The total of 87 million registered members for flicker, more than 3.5 million images uploaded daily. [7]

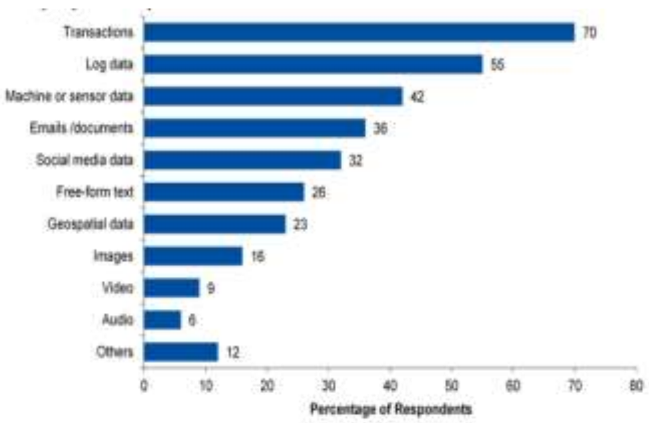


Fig.6. Big Data Sources [7]

Harish D, Anusha M.S, Dr. Daya Sagar K.V describe the techniques and tools to process Big Data as to analyze, visualize and predict the future trend of market by extracting the meaningful information from the huge amount of data. They also presented the data generated by leading companies and focused on Big Data Analysis in R using ff, fbase packages and their functions and Data storage with ff object. And explained RODBC connectivity and RHadoop architecture.

Hitesh Goyal, Surrender Singh presented Big Data Analysis using R. They explained the open source platform R language and R studio which provides breakthrough performance, scale, portability and innovation for Big Data Analytics. Similarly, It described the Big Data, History and Paradigms, Problems and challenges, principles for designing Big Data, its classification algorithm and concluded that Big Data Analytics is still in initial stage.

FUTURE SCOPE

Better tools is to develop to deal with unstructured data for data analysis tasks. In future, Better framework is to develop to improve the limitations of layered grammar, which is purely static.

CONCLUSION

This paper presents the theoretical and an efficient way for Big Data Analytics in RHadoop which is the fast parallelized analytics for processing huge data sets. It describes that How R programming language combine with an inexpensive platform Hadoop to perform complex algorithms on large data sets. This is the easiest, efficient and more productive way to write map reduce jobs. It is one-two orders of magnitude less code than java. It described three practical tools for exploring data and models. It presented the practical nature of data analysis. It would be very useful for predictive analysis.

REFERENCES

[1]. International Journal of Innovative Research in Advance Engineeering(IJIRAE) ISSN: 2349-2163 Issue 4, volume 2 (April 2015)

[2]. http://thenewstack.io/data-visualization-basics-r-programming-language/ updated on 4th April 2016

[3]. Advanced ‘Big Data Analytics’ with R and Hadoop- REVOLUTION ANALYTICS WHITE PAPER

[4]. http://www.adaltas.com/blog/2012/05/19/hadoop-and-r-is-rhadoop/Adaltas- Hadoop and R with Rhadoop on May 19th, 2012

[5]. Harshawardhan S. Bhosle & Prof. Devendra P. Gadekar; “Big Data And Hadoop”. International Journal of Scientific and Research Publications, Volume 4,Issue 10, October 2014.

[6]. Ulla Gain & Virpi Hotti;”Big Data Analytics for Professionals, Data- milling for Laypeople”, World Journal of Computer Application and Technology I(2): 51-57,2013

[7]. Shilpa and Manjit Kaur;”Big Data and Methodology- A review”, International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 10,October 2013.

[8]. K.Arun, Dr.L.Jabasheela; ”Big Data: Review, Classification and Analysis Survey”, IJIRIS,volume 1 ,Issue 3, September 2014.

[9]. Harish D, Anusha M.S, Dr. Daya Sagar K.V; “ BIG DATA ANALYSIS USING RHADOOP”,IJIRAE ISSN:2349-2163, Issue 4, Volume 2,April 2015

[10]. Hitesh Goyal, Surrender Singh;” Big Data Analysis Using R(Big Data Analysis Applications, challenges, Techniques) ” International Journal of Advanced Research in Computer Science and Software Engineering, Volume 5, Issue 9, September 2015.

[11]. LAMPI J.(2014)Large-Scale Distributed Data Management and Processing Using R, Hadoop and MapReduce. University of Oulu, Department of Computer Science and Engineering. Master’s Thesis

[12]. Practical Tools for exploring data and models, Hadley Alexander Wickham