Research Spectrum @MSIT
June-2018, Vol. 1, Issue 1

BIGDATA WITH R AND HADOOP

Mrs. Manjushree Nayak
Mats School of Information Technology,MATS University, Raipur
nayaksai.sairam@gmail.com

Abstract- **Today we are in the digital world. Everybody uses Internet, smart phones, and wireless sensors devices to produce a massive diversity of digital datasets about all facet of our daily lives from our routine analogue with outside world, medical records &happenings across the globe. So, we need a system which maintains performance of "Big data" and for analyzing this Big data we uses R language. In this paper we have discuss what is analyzing Big data, what is R language, what is Hadoop and how R and Hadoop is implemented in Big data. In last section future work and conclusion are given.**

Keywords-Bigdata, R language, Hadoop, map reduce, R Hadoop

I.INTRODUCTION

R data science is more efficient patent data analysis. R data science is consisted of R project and data science. R project is free and open software for statistical computing and visualization. Data science is to study data as well as big data including data structure, storage, collecting, and analysis. In this paper we are using Big data with R &Hadoop[1]. The volume of data that enterprises acquire every day is increasing exponentially. It is now possible to store these vast amounts of information on low cost platforms such as Hadoop. The conundrum these organizations now face is what to do with all this data and how to glean key insights from this data. Thus R comes into picture. R is a very amazing tool that makes it a snap to run advanced statistical models on data, translate the derived models into colorful graphs and visualizations, and do a lot more functions related to data science.

One key drawback of R, though, is that it is not very scalable. The core R engine can process and work on very limited amount of data. As Hadoop is very popular for Big Data processing, corresponding R with Hadoop for scalability is the next logical step. Operations of R can be made expandable by using a platform such as Hadoop. With this agenda in mind, this book will cater to a wide audience including data scientists, statisticians, data architects, and engineers who are looking for solutions to process and analyze vast amounts of information using R and Hadoop. Using R with Hadoop will provide an adaptable data analytics platform that will extent depending on the size of the dataset to be analyzed. Experienced programmers can then write Map/Reduce modules in R and run it using Hadoop's parallel processing Map/Reduce mechanism to identify patterns in the dataset[2].

Big data is data that have advantage the processing capacity of traditional database systems. The data is more big, moves too fast, or doesn't fit the strictures of your database architectures. To gain value from this data, you must choose an alternative way to process it. The hot IT jargon of 2012, big data has become applicaple as cost-effective approaches have appear to tame the volume, velocity and variability of massive data.

To leading corporations, such as Walmart or Google, this power has been in reach for some time, but at fanciful cost. .Big Data has to deal with large and complex datasets that can be structured, semistructured,or unstructured and will typically not fit in to memory to be processed. They have to be processed in place, which means that computation has to be done where the data resides for processing. When we talk to developers, the people actually building Big Data systems and applications, we get a better idea of what they mean about 3Vs. They typically would mention the 3Vs model of Big Data, which are velocity, volume, and variety.

Getting information about popular organizations that hold Big Data

Some of the popular organizations that hold Big Data are as follows:

- Facebook: It has 40 PB of data and captures 100 TB/day
- Yahoo!: It has 60 PB of data

- Twitter: It captures 8 TB/day
- EBay: It has 40 PB of data and captures 50 TB/day

How much data is considered as Big Data differs from company to company. Though true that one company's Big Data is another's small, there is something common: doesn't fit in memory, nor disk, has rapid influx of data that needs to be processed and would benefit from distributed software stacks. For some companies, 10 TB of data would be considered Big Data and for others 1 PB would be Big Data. So only you can determine whether the data is really Big Data. It is sufficient to say that it would start in the low terabyte range.Also, a question well worth asking is, as you are not capturing and retaining enough of our data do we think we do not have a Big Data problem now? In some scenarios,companies literally discard data, because there wasn't a cost effective way to store and process it.

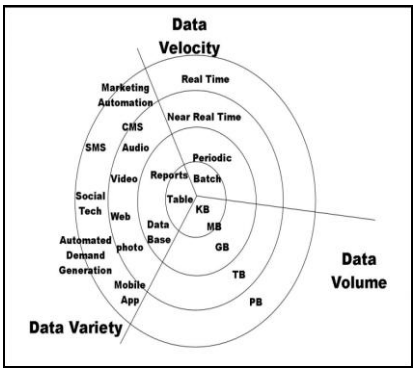

Fig.1 : 3Vs model of Big Data[4]

With platform such as Hadoop, it is possible to start capturing and storing all the Data[4].

II. A BRIEF HISTORY OF BIG DATA

"Big Data"science systems, previous descriptions are based on a one-sided view point, such as chronology or milepost technologies. The history of "Big Data" is presented in terms of the data size of interest. Under this framework, the history of "Big Data" is tied closely to the capability of efficiently storing and managing larger datasets, with size boundaries expanding by orders of degree.

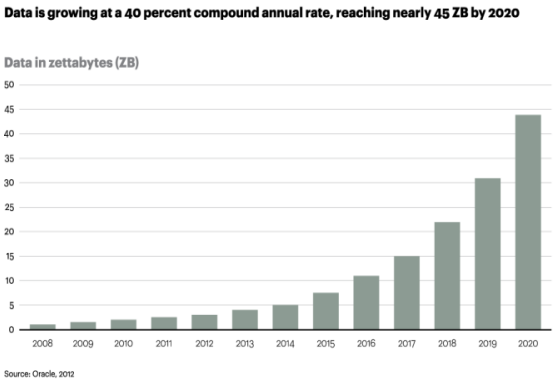

Fig-2 :GROWTH OF BIG DATA.

1) *Megabyte to Gigabyte*: In the 1970s and 1980s, historical business data introduced the earliest "Big Data" challenge in moving from megabyte to gigabyte sizes. [18]

2) *Gigabyte to Terabyte*: In the late 1980s, the popularization of digital technology caused data volumes to expand to several gigabytes or even a terabyte, which is beyond the storage and/or processing capabilities of a single large computer system [2]. Data parallelization was proposed to extend storage capabilities and to improve performance by distributing data and related tasks, such as building indexes and evaluating queries, into disparate hardware.

3) *Terabyte to Petabyte*: During the late 1990s, when the database community was admiring its ,,,,finished"''' work on the parallel database, the rapid development of Web 1.0 led the whole world into the Internet era[2], along with massive semi-structured or unstructured webpages holding terabytes or petabytes (PBs) of data.

Fig.3: Source of Big data[5]

III. INTRODUCING R

R is an open source software package to perform statistical analysis on data. R is a programming language used by data scientist statisticians and others who need to make statistical analysis of data and glean key insights from data using mechanisms, such as regression, clustering, classification, and text analysis. R has various built-in as well as extended functions for statistical, machine learning, and visualization tasks such as:

- Data extraction
- Data cleaning
- Data loading
- Data transformation
- Statistical analysis
- Predictive modeling
- Data visualization, It is one of the most important famous open source statistical analysis packages available on the market today. It is cross platform, has a very wide community support, and a large and ever-growing user community who are adding new packages every day. With its growing list of packages, R can now connect with other data stores, such as MySQL, SQLite, MongoDB, and Hadoop for data storage activities.**[6]**[8]

R's GROWTH

In 2015, IEEE had listed R at 6th position in the top 10 languages of 2015. In addition to this, as the amount of intensive data work increases, demand for tools like R for data-mining, processing and visualization will also increase **[7]**.

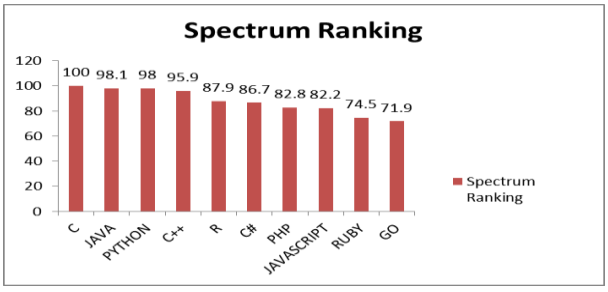

Fig.4: R'S GROWTH

IV. 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 Figure 5.

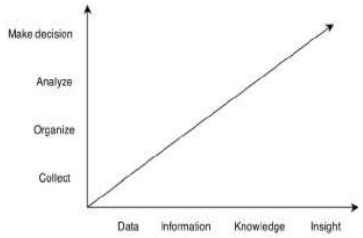

Fig5. Different Stages of Data[10]

To process and store the large sets of data, a low cost platform such as Hadoop is used and R is a very amazing programming language. it performs statistical models on data and translate the derived models into colorful graphs.R and Hadoop are natural match for Big Data analytics and Data Science. R combines with Hadoop for performing Bigdata 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 externallibraries[10].

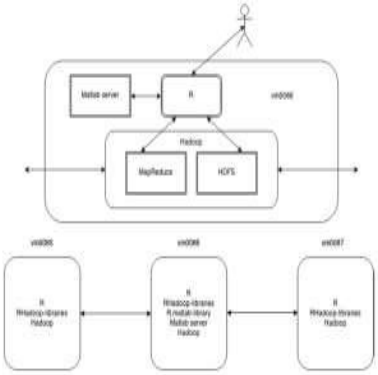

Fig.6:. Communication between different components and nodes [10]

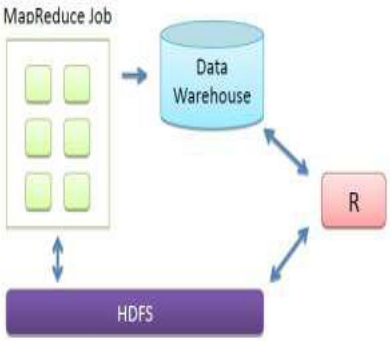

Fig-7:RHadoop Architecture [10]

V. FUTURE WORK

In future we propose to anayze the potencial of Big Data and the power that can be enabled through Big Data Analysis.The bonds between “Big Data” and knowledge hidden in it are highly crucial in all areas of national priority. This initiative will also lay the groundwork for complementary “Big Data” activities, such as “Big Data”substructure projects, platforms development, and techniques in settling complex, data-driven problems in sciences and engineering. Researchers, policy and decision makers have to recognize the potential of harnessing “BigData” to uncover the next wave of growth in their fields.

VI. CONCLUSION

Official statistics is increasingly considering big data for building new statistics because its potential to produce more relevant and timely statistics than traditional data sources. The software tools successfully used for storage and processing of big data sets on clusters is Hadoop. In this paper we have presented three ways of integrating R and Hadoop for processing large scale data sets. To create a powerful and reliable statistical model, data transformation, evaluation of multiple model options, and visualizing the results are essential. T

his is the reason why the R language has proven so p admired: its interactive language uplifts exploration, clarification and presentation. Revolution R Enterprise gives the big-data support and speed to allow the data scientist to repeat through this process quickly.

References

[1] Sunghae Jun”*Patent Big Data Analysis by R Data Language for Technology Management”, VOLUME- International Journal of Software Engineering and Its Applications Vol. 10, No. 1 (2016), pp. 69-78 http://dx.doi.org/10.14257/ijseia.2016.10.1.08

[2]https://mail.google.com/mail/u/0/#search/r++programming/159b693f0b578bca?projector=1

[3] https://www.oreilly.com/ideas/what-is-big-data

[4]https://mail.google.com/mail/u/0/#search/r++programming/159b693f0b578bca?projector=

[5]Hitesh Goyal,et.al, ” Big Data Analysis Using R (Big Data Analysis Applications, Challenges, Techniques), Volume 5, Issue 9, September 2015 ISSN: 2277 128X, International Journal of Advanced Research in Computer Science and Software Engineering

[6]http://197.14.51.10:81/pmb/INFORMATIQUE

[7]Sanchita Patil, ” Big Data Analytics Using R”VOLUME-International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056 Volume: 03 Issue: 07 | July-2016 www.irjet.net p-ISSN: 2395-0072

[8]Manjushree Nayak”prelims of R”volume- International Journal of Technical Research and Applications e-ISSN: 2320-8163, www.ijtra.com Volume 5, Issue 1 (Jan-Feb 2017), PP.35-

[9]Hitesh Goyal et.al ,” Key Capabilities for Big Data Analytics using R”https: //blogs.oracle.com /R/entry/key_capabilities_for_big_data

[10]Surbhi Raj et.al,” Working On R Programming Language For Big Data: A Review”,volume-International Journal of Enhanced Research in Science, Technology & Engineering ISSN: 2319-7463, Vol. 5 Issue 7, July-2016

[11]http://www.revistadestatistica.ro/wp-content/uploads/2014/07/rrs_2_2014_a08.pdf