Analyzing and Predicting Big Data using the tool "R"

Dr. Arul Murugan[1], Vasantha Kumari N[2]

[1]Presidency University, [2]Presidency College, Bangalore

***Abstract* -** As the technology improves with era, the new trend is Big Data! It is nothing but huge number of data stored in a repository. There is much kind of tools available to process big data. It is stated in this paper that large data is evaluated and processed with the use of the tools called R. Understanding of Data analysis is done through R. It is packaged with algorithms of data modelling and machine learning which supports scholars to carry out the research and develop the product. With the increase in digital systems, managing the big data is a great challenge.

***Keywords* -** Big data, R, data analysis

I. INTRODUCTION

Information or data is a primary thing to go ahead with any research and route it towards the destination. When we say data, it can be of any type, length, etc. and also it can be in structured or unstructured format. Huge amount or volume of data may be text, diagrams, etc. need to be analysed and processed using some of the tools. Big data can also be analysed using different strategies which are implemented in form of algorithms. Big data is not adhered to one particular field, it is massive in nature i.e. it may be existent in fields like agriculture, medicine, education, etc. To discourse the problem of big data, computer science has been evolved with lot of solutions. Many applications developed using big data grounded on technologies in IT field. R has become an important language for analysing the data and also modelling using algorithms. R is used in most of the companies like google, Facebook, etc. [1].

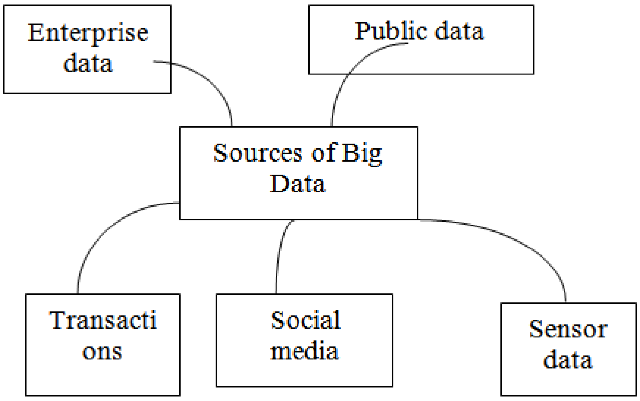
II. ANALYSING BIG DATA

R is not only a language, it is also a free software and statistical tool which supports graphics, pictorial analytics, etc. to enhance decision making and finding the solution [2]. It can be used on different platforms and similar systems.

A. R background - It is a cohesive set of software which facilitates data manipulation, calculation and display of Graphics. Few characteristics of Rare*:*

1. Open source tool.
2. For Visual implement.
3. Effective programming language
4. Accessibility with huge data management.
5. Provisions other programming language.

Many Companies used R for different purposes like Google used for advertising, Facebook make use of R for analysing the updates of face book status.



B. R's Evolution - R has been developed from S language. In 2017, R changes its position ranking of languages to 8th place .Based on IEEE spectrum rankings search, social media and mentions in journal articles.

Rank of the language	Ranking
Python	100.0
C++	99.7
Java	97.5
C	96.7
R	82.9

C. Veracity of the Specifications is maintained - The pattern should be utilized to setup your paper and elegance the script. Do not alter margins, column widths, line spaces, and text fonts. Let it as prescribed. Particularities may be noted. To take as an evidence , the head margin in this pattern measures proportionately more than is customary. This measurement and others are cautious, using specifications that anticipate your paper as one part of the entire proceedings, and not as an independent document.

D. Different diversity of R - With several individuals coming from distinctive proficient credentials, R communal is unlike. The list consists of statisticians, academicians, scientists, and programmers. The Comprehensive R archive Network (CRAN) has a package that is defined by communal participants which is throwing back for this background. The set of available packages for plot stacked areas and confidence bands, to create tables of different types of regression and perform machine learning.

E. R is Thrilling - R is exciting!!Comparison with different types of other languages, R has sophisticated features which enables the tasks to be completed in effective way. Plots and charts can be generated with its features. Coding with R is very fun. R language is primarily important for developing the modelling concepts and graphics. Example, extracting the set of data is very easy with the set of commands. Multiple regression models can be created with very few lines of code.

III. BENEFITS OF USING R TO ANALYZE THE DATA IN RESEARCH

A. R is free software - R tool is free in nature. It has a well-defined package. No restrictions in downloading the package. It has a collection of two packages i.e., R Studio and R which needs to be installed without any license key or other credentials [3]

***B.* R is Portable -** Development team of R has made an effort predominantly to make R package to be installed in computer with different categories of hardware and software. R is accessible to System software's like MAC, UNIX, or WINDOWS.

C. R provisions extensibility - One of the main benefits of R is extensibility. R has different types of extensible functions for statistical modelling, manipulation of data and other graphics. Developers has an option of writing their own set of code and dispense in the form of add on packages.

D. Connection of R with other languages - R communicates with other programming languages very easily. While data can be imported not only from Microsoft excels but also from Microsoft Access, R is very friendly with this. Various databases can be connected using ODBC and Oracle package.

IV. APPROACHES TO FOLLOW IN R

There are methods to challenge with R:

1. **Sampling -** Data required for data Analysis is not large amount of data. If so, data need to be reduced by the use of sampling. There is need to check whether performance of the model will be reduced if the data is reduced. If sampling is not required, then it is better to use other strategies.
2. **To use hardware which is large -** There is a need of large memory as R stores each object in the memory. This will be a problem if the data is too huge. As a solution, there is need to increase the size of the memory. If it a 64-bit data machine, then R can handle up to 8 Tera byte of RAM. It is an improvement from 32 bit machine.
3. **Objects storage on hard disk and need to be analyzed -** As a solution to the above, there is need to avoid storing objects into the memory which happens with the help of packages available. Hard disc stores the objects and then need to be analyzed.
4. **Integration of programming languages -** One more alternative is to integrate programming language. The aim is to balance advanced way to deal with the data on one side and other side has higher performance of other languages. Proficiency is mandatory for the developers even in other programming languages.
5. **Big data Analysis opportunity uses R -** R is a free software for processing data analysis for statistics. R played an important role due its free cost. It was not long before in data science, researchers on the go to publish papers in journals along with R code by innovative methods. It is an outstanding interactive development developed for R studio and for R language[4].Many organizations make use of R.

V. PHASES OF DATA ANALYSIS

Based on user‘s utility and proficiency, the type of Analytical process is chosen. There can be variety of users Like engineers, scientist, business analyst, etc. The types of tasks that can be performed varies with data that is simple Questions to data mining, processing of algorithms, fetching Of data, etc.

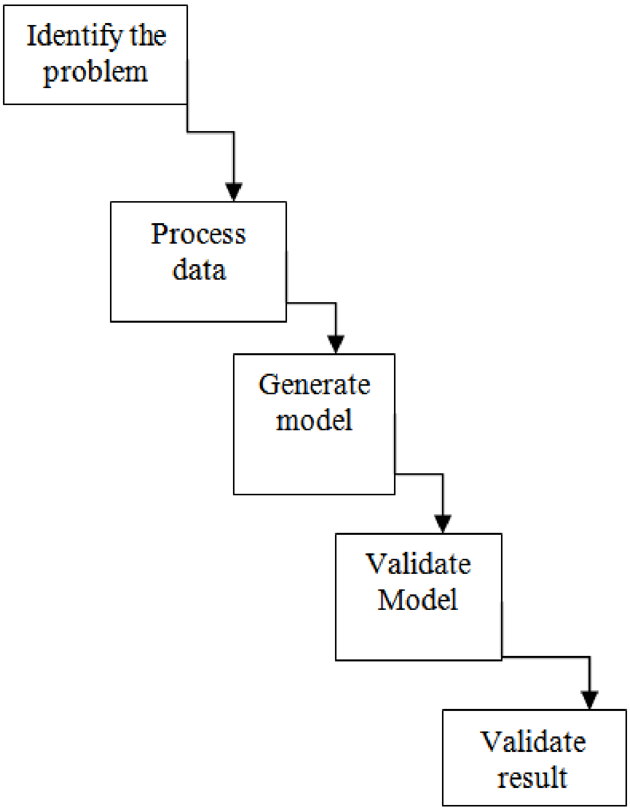


Figure : Steps in Data Analysis

A. Identify the Problem statement - The primary step is to collect the data for the analysis. It goes through planning of data, data selection, etc. collection of data is done either by survey or by means of data that exists which gives the way to collect data for next phase.

B. Pre- processing data -The next phase is to process the data which deals with analysis of syntax and data correction. This stage comprise of cleaning the data, data sieving, data accomplishment, and renovation.

C. Analyzing the data - This stage comes after the pre-processing data where data with similar pattern is brought together. It experiences correlation, visualization, regression and clustering.

D. Validating the result - It is the post processing stage in analysis of data. It includes valuation and validation of the result, interpretation of data. [4]

VI. LITERATURE SURVEY

There was language called S. The tool or language which was created later S is R. It is a free software and language of statistics. Ross Ihaka and Robert Gentleman at University of Auckland established the language of R. R language has IDE called R Studio. It works as a tool for debugging, plotting, and workspace management [5]. R works by formulating mathematical model, which affords better advantage when applied on R [6].

There are many versions of R obtainable till now and type which is latest is R-3.5.2 which was released in the year 2018[7].Versions include both 32 and 64 bit tools.

In last few decades, with academic and industrial sector the requirement of tool has been improved in the field of statistics, science, etc. which has made R a very important tool. All over the world, millions of statisticians and scientists take an advantage of R to solve the tasks in their fields. Many companies make use of R, like Google, Facebook, etc.

Using R, correlation between different data sets are available. Thus it functions as adhesive language. R contains huge number of packages and hence it is easy to work with R and there is no need to be a good R developer. R performs on any platform and hence it is independent [8]. Vendor’s offers support use of R in their software and provides collection of R packages. Users can create models using GUI rather than coding with complex R.

VII. VARIOUS CATEGORIES OF DATA ANALYTICS

The types of Data Analytics are:

1. **Descriptive Analytics:** This type of Analytics describes or summarizes the data that is raw and make it something that is interpretable by humans. It is connected with business intelligence and visibility systems [9].
2. **Predictive Analytics:** It is based on expecting future prospects. For instance, predictive modeling uses statistical techniques [10] like linear and logistic regression to predict future results. It has the ability to predict what will happen. Most people are familiar with one common application with the use of predictive analysis.
3. **Prescriptive Analytics: A**ddresses decision making and efficiency. It allows users to prescribe imaginable number of actions. Big companies are using this for production optimization, scheduling a supply chain, etc. [11]

VIII. BIG DATA CHALLENGES

In recent years, data which is huge is stored in several fields like healthcare, public, retail, and biochemistry and other Researchers. Social Computing includes social network analysis. Key challenges on Big Data:

1. To achieve the life cycle of data
2. To create value from data.
3. Confirm ethical and legal regulations on data.
4. Ensure security of cyber.
5. To master big data professions and skills.

R language has inadequacies in the following like: Memory management, speed and efficiency.

When working with the huge data sets, the design of the language can sometime lead to problems. Data or information has to be stored in physical memory. There is no option of implanting R in a web browser. Challenges on Big data are industry specific. [12]

There are few challenges in Business Enterprise today in which huge amount of data need to be stored in every minute. The amount of data generated every minute makes a very big challenge to store and analyze it. The amount of data has been growing day by day from 40% to 60% per year. Only storing the data is not enough, few analysis tools are required. As there are varieties of data to be stored, analyzing them in the same place is a big challenge. Another important challenge is scarcity of professionals who understand big data and its tools. There is huge amount of data to store and analyze and it’s a challenging for the companies to discover which of the technology will be matched by nonexistent of issues and new tasks. Data has been growing faster. Data which is missing, data that is varying, data that is copied, all results in quality of data challenge. The security and privacy of the data is a great challenge in various companies and organizations.

IX. CONCLUSION

To create a commanding and consistent statistical model, it depends on the big data collection. R language has become very familiar to industries and organizations. Enterprise R gives support to big data, to analyze and predict it. R been used with different data and had become a very important tool to analyze. In this paper, tool called R is analyzed. R is introduced to facilitate the work of large companies and organizations by analyzing the data with modelling, regression, etc. it has more than 2000 packages with a well-defined built in functions. It uses GUI for favoring users to get proper output. R gives better performance than any other analysis tools.

X. REFERENCES

[1]. Patil, S, “Big Data Analytics Using R”. International Research Journal of Engineering and Technology,2016.

[2]. Nasridinov, A., & Park, Y. H, Visual analytics for big data using R. “International Conference on Cloud and Green Computing” (pp. 564-565),2013

[3]. Malviya, A., Udhani, A., & Soni, S, “R-tool: Data analytic framework for big data”. (pp. 1-5). IEEE,2016.\

[4]. http://www.r-statistics.com/tag/hadley-wickham/

[5]. Hu, H., Wen, Y., Chua, T. S., & Li, X, "Toward scalable systems for big data analytics" 2, 652-687,2014.

[6]. Team, R," RStudio: integrated development for R. RStudio." 221,2015

[7]. Kitcharoen, N., Kamolsantisuk, S., Angsomboon, R., & Achalakul, T, "RapidMiner framework for manufacturing data analysis on the cloud".(JCSSE) (pp. 149-154). IEEE,2013.

[8]. Team, R, "RStudio: integrated development for R". , 221, 2015

[9]. Fox, J., & Andersen, R,"Using the R statistical computing environment to teach social statistics courses" 2-4,2005.

[10]. Hu, H., Wen, Y., Chua, T. S., & Li, X, "Toward scalable systems for big data analytics", 2, 652-687,2014.

[11]. Katal, A., Wazid, M., & Goudar, R. H, "Big data: issues, challenges, tools and good practices". (pp. 404-409).,2013.

[12]. Jatain, A., & Ranja, A, "A Review Study on Big Data Anaylsis Using R Studio". International Journal of Computer Science and Mobile Computing, 6(6), 8-13, 2017.

[13]. Priya Parahate, Gaurav Ghogle, Jyoti Bhange, Ashwini Ingle," Review paper on Big data: Challenges and applications", IRJET, Vol-04, e-ISSN: 2395-0056, p-ISSN: 2395-0072, 2017.