ISSN: 2454-1907
DOI: https://doi.org/10.29121/ijetmr.v6.i6.2019.402

IJETMR
International Journal of Engineering Technologies and Management Research
A Knowledge Repository

A REVIEW STUDY ON BIG DATA ANALYSIS USING R STUDIO

Savita [1], Neeraj Verma [2]
[1] MTech Scholar, Dept of Computer Science & Engineering PPIMT, Hisar (Haryana), India
[2] Assistant Professor, Dept of Computer Science & Engineering PPIMT, Hisar (Haryana), India

Abstract:
Big Data Analytics is a way of extracting value from these huge volumes of information, and it drives new market opportunities and maximizes customer retention. The rapid rise of the Internet and the digital economy has fuelled an exponential growth in demand for data storage and analytics, and IT department are facing tremendous challenge in protecting and analyzing these increased volumes of information. The reason organizations are collecting and storing more data than ever before is because their business depends on it. The type of information being created is no more traditional database-driven data referred to as structured data rather it is data that include documents, images, audio, video, and social media contents known as unstructured data or Big Data. This paper primarily focuses on discussing the various technologies that work together as a Big Data Analytics system that can help predict future volumes, gain insights, take proactive actions, and give way to better strategic decision-making.

Keywords: Huge Statistics; Big Data Analysis; R Studio.

Cite This Article: Savita, and Neeraj Verma. (2019). "A REVIEW STUDY ON BIG DATA ANALYSIS USING R STUDIO." *International Journal of Engineering Technologies and Management Research*, 6(6), 129-136. DOI: https://doi.org/10.29121/ijetmr.v6.i6.2019.402.

1. Introduction

Big Data Analytics reflect the challenges of data that are too vast, too unstructured, and too fast moving to be managed by traditional methods. From businesses and research institutions to governments, organizations now routinely generate data of unprecedented scope and complexity. Gleaning meaningful information and competitive advantages from massive amounts of data has become increasingly important to organizations globally. Trying to efficiently extract the meaningful insights from such data sources quickly and easily is challenging.[2] Thus, analytics has become inextricably vital to realize the full value of Big Data to improve their business performance and increase their market share. The tools available to handle the volume, velocity, and variety of big data have improved greatly in recent years. In general, these technologies are not prohibitively expensive, and much of the software is open source.[6] Hadoop, the most commonly used framework, combines commodity hardware with open source software. It takes incoming streams of data and distributes them onto cheap disks; it also provides tools for analyzing the data. However, these technologies do require a skill set that is new to most IT departments, which will need to work hard to integrate all the relevant internal and external sources of data. Although attention to technology isn't sufficient, it is always a necessary

ISSN: 2454-1907
DOI: 10.5281/zenodo.3266158

component of a big data strategy.[9] This paper discusses some of the most commonly used big data technologies mostly open source that work together as a big data analytics system for leveraging large quantities of unstructured data to make more informed decisions.

2. Big Data Concept

Big Data is an important concept, which is applied to data, which does not conform to the normal structure of the traditional database. Big Data consists of different types of key technologies like Hadoop, HDFS, NoSQL, MapReduce, MongoDB, Cassandra, PIG, HIVE, and HBASE that work together to achieve the end goal like extracting value from data that would be previously considered dead. According to a recent market report published by Transparency Market Research, the total value of big data was estimated at $6.3 billion as of 2012, but by 2018, it's expected to reach the staggering level of $48.3 billion that's almost a 700 percent increase [29]. Forrester Research estimates that organizations effectively utilize less than 5 percent of their available data. This is because the rest is simply too expensive to deal with. Big Data is derived from multiple sources. It involves not just traditional relational data, but all paradigms of unstructured data sources that are growing at a significant rate. For instance, machine-derived data multiplies quickly and contains rich, diverse content that needs to be discovered. Another example, human-derived data from social media is more textual, but the valuable insights are often overloaded with many possible meanings. [1]

First, because of the large variety of different records sources and the massive extent, its miles too tough to acquire, combine and evaluation of "huge statistics" with scalability from scattered places.[7]

Second "huge facts" structures want to manage, shop and integrate the amassed massive and varied verity of datasets, whilst provide function and performance warranty [1], in terms of rapid retrieval, scalability and secrecy safety.

Third "large information" analytics have to efficaciously excavation large datasets at one of kind degrees in actual time or close to real time - which includes modeling, visualization [2], prediction and optimization - such that inherent potentials can be discovered to improve choice making and collect similarly advantages. To address these challenges, the researcher IT industry and community has given various solutions for "Big Data" science systems in an ad-hoc manner. Cloud computing can be called as the substructure layer for "Big Data" systems to meet certain substructure requirements, such as cost-effectiveness, resistance[2], and the ability to scale up or down. Distributed file systems and No SQL databases are suitable for persistent storage and the management of massive scheme free datasets [1]. Map Reduce, R is a programming framework, has achieved great success in processing "Big Data" group-aggregation tasks, such as website ranking [10].

RStudio integrates data storage, data processing, system management, and other modules to form a powerful system-level solution, which is becoming the mainstay in handling "Big Data" challenges. We can build various "Big Data" application system based on these innovative technologies and platforms. In light of the of big-data technologies, a systematic frame work should be in order to capture the fast evolution of big-data research.

ISSN: 2454-1907
DOI: 10.5281/zenodo.3266158

2.1. Big Data Problem and Challenges

However, considering variety of data sets in "Big Data" problems, it is still a big challenge for us to purpose efficient representation, access, and analysis of shapeless or semi-structured data in the further researches [12]. How can the data be preprocessed in order to improve the quality of data and analysis results before we begin data analysis [1] [2]? As the sizes of dataset are often very large, sometimes several gigabytes or more, and their origin from varied sources, current real-world databases are pitilessly susceptible to inconsistent, incomplete, and noisy data. Therefore, a number of data preprocessing techniques, including data cleaning [11], data integration, data transformation and date reduction, can be applied to remove noise and correct irregularities. Different challenges arise in each sub-process when it comes to data-driven applications.

2.2. Principles for designing Big Data System

In designing "Big Data" analytics systems, we summarize seven necessary principles to guide the development of this kind of burning issues [3]. "Big Data" analytics in a highly distributed system cannot be achievable without the following principles [13]:

1) Good architectures and frameworks are necessary and on the top priority.
2) Support a variety of analytical methods
3) No size fits all
4) Bring the analysis to data
5) Processing must be distributable for in-memory computation.
6) Data storage must be distributable for in-memory storage.
7) Coordination is needed between processing and data units.

2.3. Big Data Opportunities

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 "Big Data" to uncover the next wave of growth in their fields. There are many advantages in business section that can be obtained through harnessing "Big Data" increasing operational efficiency, informing strategic direction, developing better customer service, identifying and developing new products and services, identifying new customers and markets, etc.

2.4. Big Data Analysis

The last and most important stage of the "Big Data" value chain is data analysis, the goal of which is to get useful values, suggest best conclusions and support decision-making system of an organization to stay in competition market. [1]

Descriptive Analytics: exploits historical data to describe what occurred in past. For instance, a regression technique may be used to find simple trends in the datasets, visualization presents data

ISSN: 2454-1907
DOI: 10.5281/zenodo.3266158

in a meaningful fashion, and data modeling is used to collect, store and cut the data in an efficient way. Descriptive analytics is typically associated with business intelligence or visibility systems [2].

Predictive Analytics: focuses on predicting future probabilities and trends. For example, predictive modeling uses statistical techniques [6] such as linear and logistic regression to understand trends and predict future out-comes, and data mining extracts patterns to provide insight and forecasts [4].

Prescriptive Analytics: addresses decision making and efficiency. For example, simulation is used to analyze complex systems to gain insight into system performance and identify issues and optimization techniques are used to find best solutions under given constraints.

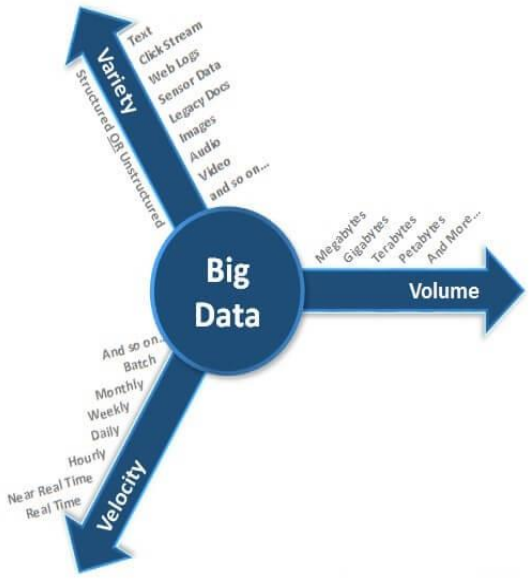
3. Literature Review

Big Data is a data analysis methodology enabled by recent advances in technologies that support high-velocity data capture, storage and analysis. Data sources extend beyond the traditional corporate database to include emails, mobile device outputs, and sensor-generated data where data is no longer restricted to structured database records but rather unstructured data having no standard formatting [20]

3.1. Characteristics of Big Data - The Three V's of Big Data

When do we say we are dealing with Big Data? For some people 1TB might seem big, for others 10TB might be big, for others 100GB might be big, and something else for others. This term is qualitative and it cannot really be quantified. Hence we identify Big Data by a few characteristics which are specific to Big Data. These characteristics of Big Data are popularly known as Three V's of Big Data.[3]

The three v's of Big Data are Volume, Velocity, and Variety as shown below.



ISSN: 2454-1907
DOI: 10.5281/zenodo.3266158

3.2. Volume

Volume refers to the size of data that we are working with. With the advancement of technology and with the invention of social media, the amount of data is growing very rapidly. This data is spread across different places, in different formats, in large volumes ranging from Gigabytes to Terabytes, Petabytes, and even more. Today, the data is not only generated by humans, but large amounts of data is being generated by machines and it surpasses human generated data. This size aspect of data is referred to as Volume in the Big Data world.[10]

3.3. Velocity

Velocity refers to the speed at which the data is being generated. Different applications have different latency requirements and in today's competitive world, decision makers want the necessary data/information in the least amount of time as possible. Generally, in near real time or real time in certain scenarios. In different fields and different areas of technology, we see data getting generated at different speeds. A few examples include trading/stock exchange data, tweets on Twitter, status updates/likes/shares on Facebook, and many others. This speed aspect of data generation is referred to as Velocity in the Big Data world.[10]

3.4. Variety

Variety refers to the different formats in which the data is being generated/stored. Different applications generate/store the data in different formats. In today's world, there are large volumes of unstructured data being generated apart from the structured data getting generated in enterprises. Until the advancements in Big Data technologies, the industry didn't have any powerful and reliable tools/technologies which can work with such voluminous unstructured data that we see today. In today's world, organizations not only need to rely on the structured data from enterprise databases/warehouses, they are also forced to consume lots of data that is being generated both inside and outside of the enterprise like clickstream data, social media, etc. to stay competitive. Apart from the traditional flat files, spreadsheets, relational databases etc., we have a lot of unstructured data stored in the form of images, audio files, video files, web logs, sensor data, and many others. This aspect of varied data formats is referred to as Variety in the Big Data world.[5]

Katarina et al. in 2014 discussed challenges for Map Reduce in Big Data. In the Big Data community, Map Reduce has been seen as one of the key enabling approaches for meeting the continuously increasing demands on computing resources imposed by massive data sets. At the same time, Map Reduce faces a number of obstacles when dealing with Big Data including the lack of a high-level language such as SQL, challenges in implementing iterative algorithms, support for iterative ad-hoc data exploration, and stream processing. The identified Map Reduce challenges are grouped into four main categories corresponding to Big Data tasks types: data storage, analytics, online processing, security and privacy. The main objective of this paper is identifies Map Reduce issues and challenges in handling Big Data with the objective of providing an overview of the field, facilitating better planning and management of Big Data projects, and identifying opportunities for future research in this field.

ISSN: 2454-1907
DOI: 10.5281/zenodo.3266158

Zhen Jia1, Jianfeng Zhan, Lei Wang, Rui Han, Sally A. McKee, Qiang Yang, Chunjie Luo, and Jingwei Li in 2014 discussed Characterizing and Subsetting Big Data Workloads. A large number of benchmarks pose great challenges, since our usual simulation-based research methods become prohibitively expensive. They use hardware performance counters to analyze micro architectural behaviors of those scale out workloads. They compare the scale-out workloads and traditional benchmarks to identify the key contributors to the micro architecture inefficiency on modern processors. They conclude that mismatches exist between the needs of scale-out workloads and the capabilities of modern processors. Much work focuses on comparing the performance of different data management systems. For OLTP or database systems evaluation, TPC-C is often used to evaluate transaction-processing system performance in terms of transactions per minute. Cooper defines a core set of benchmarks and report throughput and latency results for five widely used data management systems.

3.5. Big Data Tools: Techniques and Technologies

We need tools (platforms) to make sense of "Big Data". Current tools concentrate on three classes, namely, batch processing tools, stream processing tools, and interactive analysis tools. Most batch processing tools are based on the Apache RStudio infrastructure, such as Map reduce [4], R Programming and Dryad. The interactive analysis processes the data in an interactive environment, allowing users to undertake their own analysis of information.

3.6. R Programming

The R language is well mounted as the language for doing data, facts evaluation, information-mining algorithm improvement, stock buying and selling, credit score danger scoring, marketplace basket evaluation and all [9] way of predictive analytics. However, given the deluge of information that must be processed and analyzed nowadays, many groups had been reticent about deploying R past studies into production programs. [16]

3.7. Comparisons of Classification for Big Data Science

To apply different classification technique I have chosen a real dataset about the student's knowledge status about the subject of Electrical DC Machines. Distribution of every numeric variable can be checked with function summary (), which returns the minimum, maximum, mean, median, and the first (25%) and third (75%) quartiles. For factors (or categorical variables), it shows the frequency of every level.

3.8. The importance of Big Data

The main importance of Big Data consists in the potential to improve efficiency in the context of use a large volume of data, of different type. If Big Data is defined properly and used accordingly, organizations can get a better view on their business therefore leading to efficiency in different areas like sales, improving the manufactured product and so forth.

Big Data can be used effectively in the following areas:

ISSN: 2454-1907
DOI: 10.5281/zenodo.3266158

- In information technology in order to improve security and troubleshooting by analyzing the patterns in the existing logs;
- In customer service by using information from call centers in order to get the customer pattern and thus enhance customer satisfaction by customizing services;
- In improving services and products through the use of social media content. By knowing the potential customers preferences the company can modify its product in order to address a larger area of people;
- In the detection of fraud in the online transactions for any industry;
- In risk assessment by analyzing information from the transactions on the financial market

4. Conclusions

Today's technology landscape is changing fast. Organizations of all shapes and sizes are being pressured to be data driven and to do more with less. Even though big data technologies are still in a nascent stage, relatively speaking, the impact of the 3V's of big data. That is why this article presents the Big Data concept and the R technologies associated in order to understand better the multiple benefices of this new concept ant technology.

In the future we propose for our research to further investigate the practical advantages that can be gain through R Studio.

References

[1] Carson Kai-Sang Leung, Richard Kyle MacKinnon, Fan Jiang, "Reducing the Search Space for Big Data Mining for Interesting Patterns from Uncertain Data", IEEE 2014, PP 315-322
[2] Ajith Abraham1, Swagatam Das2, and Sandip Roy3, "Swarm Intelligence Algorithms for Data Clustering", PP 280-313
[3] Swagatam Das, Ajith Abraham, Senior Member, IEEE, and Amit Konar, "Automatic Clustering Using an Improved Differential Evolution Algorithm", IEEE 2008, PP 218-237
[4] KarthikKambatla, GiorgosKollias, Vipin Kumar, AnanthGrama, "J. Parallel Distrib. Comput", Elsevier 2014, PP 2561-2573
[5] Yanchang Zhao, "R and Data Mining: Examples and Case Studies", www.RDataMining.com,2014
[6] H. T. Kahraman, Sagiroglu, S., Colak, "User Knowledge Modeling Data Set", UCI, vol. 37, pp. 283-295, 2013
[7] Mrigank Mridul, Akashdeep Khajuria, Snehasish Dutta, Kumar N, "Analysis of Bidgata using Apache Hadoop and Map", Volume 4, Issue 5, May 2014 Reduce, PP. 555-560.
[8] Sonja Pravilovic," R language in data mining techniques and statistics", 20130201.12,2013
[9] Vrushali Y Kulkarni," Random Forest Classifiers: A Survey and Future Research Directions", International Journal of Advanced Computing, ISSN:2051-0845, Vol.36, Issue.1, April 2013
[10] Aditya Krishna Menon," Large-Scale Support Vector Machines: Algorithms and Theory".
[11] Arcot Rajasekar, Sharlini Sankaran, Howard Lander, Tom Carsey, Jonathan Crabtree, Hye-Chung Kum, Merce Crosas, Gary King, Justin Zhan, "Sociometric methods for relevancy analysis of Long Tail Science Data"
[12] danah boyd, Kate Crawford, "Six Provocations for Big Data", Oxford Internet Institute's ,2011
[13] Tony Hey, Anne Trefethen, "The Data Deluge: An e-Science Perspective", Grid Computing – Making the Global Infrastructure a Reality", Wiley, January 2003

ISSN: 2454-1907
DOI: 10.5281/zenodo.3266158

[14] Arcot Rajasekar, Sharlini Sankaran, Howard Lander, Tom Carsey, Jonathan Crabtree, Hye-Chung Kum, Merce Crosas, Gary King, Justin Zhan, "Sociometric methods for relevancy analysis of Long Tail Science Data"
[15] Yanchang Zhao, "R and Data Mining: Examples and Case Studies", www.RDataMining.com, 2014
[16] H. T. Kahraman, Sagiroglu, S., Colak,"User Knowledge Modeling Data Set", UCI, vol. 37, pp. 283-295, 2013
[17] Mrigank Mridul, Akashdeep Khajuria, Snehasish Dutta, Kumar N, "Analysis of Bidgata using Apache RStudio and Map", Volume 4, Issue 5, May 2014 Reduce, PP. 555-560.
[18] Sonja Pravilovic," R language in data mining techniques and statistics", 20130201.12,2013
[19] Vrushali Y Kulkarni," Random Forest Classifiers: A Survey and Future Research Directions", International Journal of Advanced Computing, ISSN: 2051-0845, Vol.36, Issue.1, April 2013
[20] Aditya Krishna Menon," Large-Scale Support Vector Machines: Algorithms and Theory".