International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2018 IJSRCSEIT | Volume 3 | Issue 5 | ISSN : 2456-3307

A Review Study on Big Data Analysis Using R Studio

Sangita[1], Shagun[2]
[1]MTech Scholar, [2]Asstt. Professor
Department of Computer Science & Engineering Manav Institutes of Technology & Mgt, Haryana, India

ABSTRACT

During the last decade, large statistics evaluation has seen an exponential boom and will absolutely retain to witness outstanding tendencies due to the emergence of new interactive multimedia packages and extraordinarily incorporated systems driven via the speedy growth in statistics services and microelectronic gadgets. up to now, maximum of the modern mobile structures are especially centered to voice communications with low transmission fees. Inside the near destiny, however, huge information access at excessive transmission costs might be. that is a evaluate on available big-records systems that include a hard and fast of tools and approach to load, extract, and enhance distinct data whilst leveraging the immensely parallel processing strength to carry out complicated adjustments and evaluation. “massive-statistics” device faces a series of technical challenges.

Keywords : Huge Statistics

I. INTRODUCTION

The emerging large records science time period, displaying its broader effect on our society and in our enterprise lifestyles cycle, has insightful transformed our society and could hold to draw numerous attentions from technical specialists and in addition to public in standard [1] [2]. It is apparent that we're residing in huge records era, proven with the aid of the sheer quantity of data from a spread of sources and its growing rate of era. as an example, an IDC file predicts that, from 2005 to 2020, the global statistics dimensions will grow with the aid of a issue of three hundred, from a hundred thirty Exabyte’s to 40,000 Exabyte’s, representing a double increase each two years. this is specializes in on hand big-records structures that include a set of equipment and technique to load, extract, and enhance numerous information while leveraging the immensely parallel processing energy to carry out complicated transformations and evaluation. “large-information” machine faces a chain of technical challenges, along with:

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.

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.

A Brief History Of Big Data

Considering the growth and intricacy of “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.

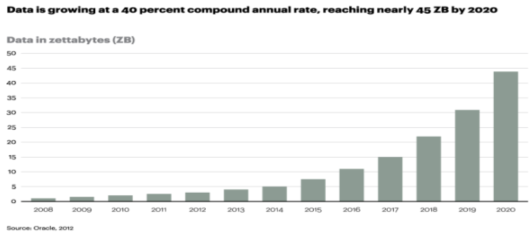

Figure 1: Growth of Big Data.

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.

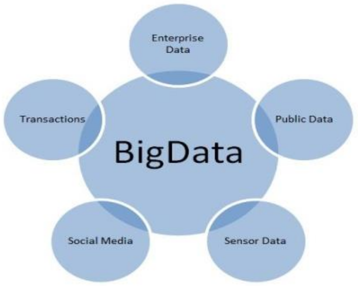

Figure 2: source of big data

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.

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.

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.

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 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.

BIG DATA CLASSIFICATION ALGORITHM

1) Decision Tree
2) Random Forest
3) Support Vector Machine

Decision tree learning uses a decision tree as a predictive version which maps observations approximately an item to conclusions approximately the item's goal fee. Its miles one of the predictive modeling techniques used in statistics, information mining and system getting to know. Tree models where the goal variable can take a finite set of values are referred to as class timber. In these tree systems, leaves constitute magnificence labels and branches constitute conjunctions of functions that cause those magnificence labels. decision bushes in which the target variable can take non-stop values are known as regression timber. In selection analysis, a decision tree can be used to visually and explicitly represent decisions and choice making. In facts mining, a choice tree describes records but now not choices; as an alternative the ensuing classification tree may be an enter for choice making [23].

Random Forests is an ensemble learning method also thought of as a form of nearest neighbor predictor for classification and regression that construct a number of decision trees at training time and outputting the class that is the mode of the classes output by individual trees. Random Forests is a combination of tree predictors where each tree depends on the values of a random vector sampled independently with the same distribution for all trees in the forest. The basic principle is that a group of "weak learners" can come together to form a "strong learner". Random Forests are a wonderful tool for making predictions considering they do not over fit because of the law of large numbers. Introducing the right kind of randomness makes them accurate classifiers and repressors [24].

Literature Review

Shweta Pandey et al. in 2014 shows that Big Data is having challenges related to volume, velocity and variety. Big Data has 3Vs Volume means large amount of data, Velocity means data arrives at high speed, Variety means data comes from heterogeneous resources. In Big Data definition, Big means a dataset which makes data concept to grow so much that it becomes difficult to manage it by using existing data management concepts and tools. Map Reduce is playing a very significant role in processing of Big Data. The main objective of this paper is purposed a tool like Map Reduce is elastic scalable, efficient and fault tolerant for analyzing a large set of data, highlights the features of Map Reduce in comparison of other design model which makes it popular tool for processing large scale data.

HAN HU and YONGGANG WEN in 2014 shows that Recent technological advancements have led to a deluge of data from distinctive domains e.g., health care and scientific sensors, user-generated data, Internet and financial companies, and supply chain systems. Over the past two decades. The term big data was coined to capture the meaning of this emerging trend. In addition to its sheer volume, big data also exhibits other unique characteristics as compared with traditional data. For instance, big data is commonly unstructured and require more real-time analysis. This development calls for new system architectures for data acquisition, transmission, storage, and large-scale data processing mechanisms. In this paper, we present a literature survey and system tutorial for big data analytics platforms, aiming to provide an overall picture for non-expert readers and instill a do-it-yourself spirit for advanced audiences to customize their own big-data solutions.

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.

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.

Big Data Tools: Techniques and Technologies

To capture the value from "Big Data", we need to develop new techniques and technologies for analyzing it. Until now, scientists have developed a wide variety of techniques and technologies to capture, curate, analyze and visualize Big Data.

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.

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]

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.

II. REFERENCES

[1]. CL. Philip Chen, Chun-Yang Zhang, Data intensive applications, challenges, techniques and technologies: A survey on Big Data Information Science 0020-0255 (2014), PP 341-347, elsevier

[2]. Han hu1At. Al. (Fellow, IEEE), Toward Scalable Systems for Big Data Analytics: A Technology Tutorial, IEEE 2169-3536(2014),PP 652-687

[3]. Shweta Pandey, Dr.VrindaTokekar, Prominence of MapReduce in BIG DATA Processing, IEEE (Fourth International Conference on Communication Systems and Network Technologies)978-1-4799-3070-8/14, PP 555-560

[4]. Katarina Grolinger At. Al.Challenges for MapReduce in Big Data, IEEE (10th World Congress on Services)978-1-4799-5069-0/14,PP 182-189

[5]. Zhen Jia1 At. Al.Characterizing and Subsetting Big Data Workloads", IEEE 978-1-4799-6454-3/14, PP 191-201

[6]. AvitaKatal, Mohammad Wazid, R H Goudar, Big Data: Issues, Challenges, Tools and Good

Practices, IEEE 978-1-4799-0192-0/13,PP 404-409

[7]. Du Zhang, Inconsistencies in Big Data, IEEE 978-1-4799-0783-0/13, PP 61-67

[8]. ZibinZheng, Jieming Zhu, and Michael R. Lyu, Service-generated Big Data and Big Data-as-a-Service: An Overview, IEEE (International Congress on Big Data) 978-0-7695-5006-0/13, PP 403-410

[9]. VigneshPrajapati, Big Data Analytics with R and RStudioPackt Publishing

[10]. Lei Wang At. Al., BigDataBench: aBigDataBenchmarkSuitefromInternetServices, IEEE 978-1-4799-3097-5/14.

[11]. AnirudhKadadi At. Al., Challenges of Data Integration and Interoperability in Big Data, IEEE (International Conference on Big Data)978-1-4799-5666-1/14, PP 38-40

[12]. SAS, Five big data challenges and how to overcome them with visual analytics

[13]. HajarMousanif At. Al., From Big Data to Big Projects: a Step-by-step Roadmap, IEEE (International Conference on Future Internet of Things and Cloud) 978-1-4799-4357-9/14, PP 373-378

[14]. Tianbo Lu At. Al., Next Big Thing in Big Data: The Security of the ICT Supply Chain, IEEE (SocialCom/PASSAT/BigData/EconCom/BioMedCom) 978-0-7695-5137-1/13, PP 1066-1073

[15]. Ganapathy Mani, NimaBarit, Duoduo Liao, Simon Berkovich, Organization of Knowledge Extraction from Big Data Systems, IEEE (4 Fifth International Conference on Computing for Geospatial Research and Application) 978-1-4799-4321-0/14, PP 63-69

[16]. Joseph Rickert, Big Data Analysis with Revolution R Enterprise, 2011

[17]. 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

[18]. Ajith Abraham1, Swagatam Das2, and Sandip Roy3, Swarm Intelligence Algorithms for Data Clustering, PP 280-313

[19]. Swagatam Das, Ajith Abraham, Senior Member, IEEE, and Amit Konar, Automatic Clustering Using an Improved Differential Evolution Algorithm, IEEE 2008, PP 218-237

[20]. KarthikKambatla, GiorgosKollias, Vipin Kumar, AnanthGrama, J. Parallel Distrib. Comput, Elsevier 2014, PP 2561-2573

[21]. Yanchang Zhao, R and Data Mining: Examples and Case Studies, www.RDataMining.com,2014

[22]. H. T. Kahraman, Sagiroglu, S., Colak,User Knowledge Modeling Data Set, UCI, vol. 37, pp. 283-295, 2013

[23]. 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.

[24]. Sonja Pravilovic, R language in data mining techniques and statistics, 20130201.12,2013

[25]. 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

[26]. Aditya Krishna Menon, Large-Scale Support Vector Machines: Algorithms and Theory.