WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

A Challenging Tool for Research Questions in Big Data Analytics

[1]Dr. Rajasekaran. A, [2]S. Logeshwaran

[1]Assistant Professor, Department of Electronics & Communication Engineering, SCSVMV, Kanchipuram, India
[2]UG Scholar, Department of Computer science and Engineering, PERI Institute of Technology, Chennai, India

ABSTRACT:

In the information age, decision makers have vast amounts of data at their disposal. Big data refers to data sets that are not only large, but also extremely diverse and fast, making them difficult to manipulate with traditional tools and techniques. Such data is growing rapidly, and solutions must be explored and deployed to process and extract value and knowledge from these data sets. Analyzing this data requires significant effort at multiple levels of knowledge extraction in order to make effective decisions. Big data analytics is a current research and development area. Furthermore, it opens up new horizons for researchers to develop solutions based on challenges and open research questions.

Keywords: Big data analytics; Hadoop; Massive data; Structured data; Unstructured Data; IoT

I. INTRODUCTION

In a digital environment, data is generated from various technology sources, leading to the growth of big data. Our large collection of datasets provides evolutionary breakthroughs in many areas. Traditional database management tools and data processing applications have difficulty handling large and complex collections of data sets. They are available in structured, semi-structured, and unstructured formats for petabytes and beyond. Formally defined from 3Vs to 4Vs. 3Vs means volume, speed and variety. Volume refers to the sheer volume of data generated every day, while velocity is the growth rate, the speed at which data is collected for analysis. Diversity refers to data types such as structured, unstructured, and semi-structured. The fourth V relates to authenticity, including availability and accountability. The main goal of big data analytics is to process data with large volume, high speed, diversity and veracity using various conventional computational intelligence techniques [1]. Figure 1 below shows the definition of big data. However, the exact definition of big data has not been defined and is considered problematic. This enables better decision-making, insight generation and optimization while being innovative and cost-effective. The basic purpose of this white paper is to examine the big data challenge, its research challenges, and the potential impact of various tools associated with it. Therefore, this article provides a platform for exploring big data at different stages. In addition, we will name open research questions on big data. This white paper is divided into the following sections: Section 2 addresses big data challenges

II. CHALLENGES

In recent years, big data has been accumulated in multiple fields such as healthcare, retail, and interdisciplinary scientific research such as biochemistry. Web-based applications often encounter big data such as: B. Social Computing, Internet Text and Documents, and Internet Search Indexing. Social computing includes analysis of social networks, online communities, recommendation systems, reputation systems, prediction markets, and Internet search indexes including ISI, IEEE Xplorer, Scopus, Thomson, Reuters, and others.

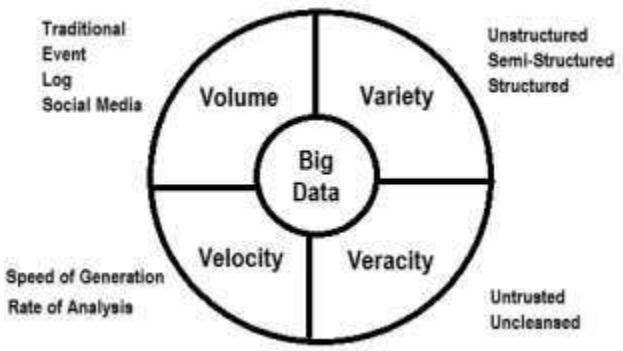

FIG. 1 Characteristics of Big Data.

Considering these advantages of big data, it offers new possibilities for knowledge processing tasks for budding researchers. However, opportunity always comes with some challenges. Addressing challenges requires knowing computational methods for analyzing various computational complexities, information security, and big data. For example, many statistical techniques that work well with small data sets do not scale well with large data sets. Here, big data analytics challenges fall into four broad categories: data storage and analysis. knowledge discovery and computational complexity; data scalability and visualization; and information security. These issues are briefly discussed in the following subsections.

A. Data Storage and Analysis

In recent years, the size of data has grown exponentially through various means such as mobile devices, sensor technology, remote sensing, and radio frequency identification readers. This data is either ignored or eventually deleted because there is not enough space to store it, but it is stored with great effort. Therefore, the first issue in big data analysis is to increase the speed of storage media and input/output. In such cases, data accessibility for knowledge discovery and expression must be a top priority. The main reason is the need for easy and quick access for detailed analysis. For decades, analysts used hard drives to store data, but random I/O was slower than sequential I/O. To overcome this limitation, the concepts of Solid State Drives (SSD) and Phrase Change Memory (PCM) were introduced. However, available storage technologies may not provide the performance needed to process big data.

Another challenge in big data analytics is attributed to data diversity. Data mining tasks have grown exponentially with the ever-increasing number of datasets. Additionally, data reduction, data selection, and feature selection are essential tasks, especially when dealing with large data sets. This is an unprecedented challenge for researchers, as existing algorithms do not always respond in time when processing this high-dimensional data. Developing new machine-learning algorithms to ensure the integrity has become a major challenge in recent years. In addition to all these clusters in large datasets useful for big data analysis, the main concern is [7]. New technologies such as Hadoop and MapReduce enable large amounts of semi-structured and unstructured data to be collected in a reasonable amount of time. A major technical challenge is to effectively analyze this data to gain better insights. The main challenge here is to pay more attention to the design of storage systems and develop efficient data analysis tools that guarantee output when data is retrieved from different sources. Moreover, the design of machine learning algorithms for analyzing data is essential to improve efficiency and scalability.

B. Knowledge Discovery and Computational Complexities

Knowledge discovery and representation are major themes of big data. It includes several sub-areas such as authentication, archiving, management, retention, information retrieval, and presentation. Many hybrid techniques have also been developed to handle real-world problems. All of these techniques are problem dependent. Additionally, some of these techniques may not be suitable for large data sets on sequential computers. At the same time, some techniques have good scalability properties for parallel computers. As the size of big data grows exponentially, the available tools may not be efficient at processing this data to get meaningful information. Data warehouses and data marts are the most common approaches to managing large datasets. Data warehouses are primarily responsible for storing data from operational systems, while data marts are based on data warehouses and facilitate analysis.

Analyzing large data sets requires more complex calculations. The main problem is dealing with inconsistencies and uncertainties in the dataset. In general, systematic modeling of computational complexity is used. Establishing a comprehensive mathematical system that is generally applicable to big data can be difficult. However, domain-specific data analysis is easy once you understand certain complexities. Many such developments can simulate big data analytics in various domains. A lot of research and research has been done in this direction using machine learning techniques with minimal memory requirements. A fundamental goal of this work is to minimize computational effort and complexity [8], [9], [10].

However, current big data analytics tools perform poorly when dealing with computational complexity, uncertainty, and discrepancies. Developing methods and techniques that can effectively deal with computational complexity, uncertainties, and contradictions presents a significant challenge.

C. Scalability and Visualization of Data

The most important challenge for big data analysis techniques is its scalability and security. In the last decades researchers have paid attentions to accelerate data analysis and its speed up processors followed by Moore's Law. For the former, it is necessary to develop sampling, on-line, and multi resolution analysis techniques. Incremental techniques have good scalability property in the aspect of big data analysis. As the data size is scaling much faster than CPU speeds, there is a natural dramatic shift in processor technology being embedded with increasing number of cores [11]. This shift in processors leads to the development of parallel computing. Real time applications like navigation, social networks, finance, internet search, timeliness etc. requires parallel computing. We can observe that big data have produced many challenges for the developments of the hardware and software which leads to parallel computing, cloud computing, distributed computing, visualization process, scalability. To over-come this issue, we need to correlate more mathematical models to computer science.

D. Information Security

In big data analysis massive amount of data are correlated, analyzed, and mined for meaningful patterns. All organizations have different policies to safe guard their sensitive information. Preserving sensitive information is a major issue in big data analysis. There is a huge security risk associated with big data [12]. Therefore, information security is becoming a big data analytics problem. Security of big data can be enhanced by using the techniques of authentication, authorization, and encryption. Various security measures that big data applications face are scale of network, variety of different devices,

real time security monitoring, and lack of intrusion system [13], [14]. The security challenge caused by big data has attracted the attention of information security. Therefore, attention has to be given to develop a multilevel security policy model and prevention system.Although much research has been carried out to secure big data [13] but it requires lot of improvement. The major challenge is to develop a multi-level security, privacy preserved data model for big data.

III. RESEARCH ISSUES IN BIG DATA ANALYTICS

Big data analytics and data science are becoming the research focal point in industries and academia. Data science aims at researching big data and knowledge extraction from data. Applications of big data and data science include information science, uncertainty modelling, uncertain data analysis, machine learning, statistical learning, pattern recognition, data warehousing, and signal processing. Effective integration of technologies and analysis will result in predicting the future drift of events. Main focus of this section is to discuss open research issues in big data analytics. The research issues pertaining to big data analysis are classified into three broad categories namely internet of things (IoT), cloud computing, bio inspired computing, and quantum computing. However it is not limited to these issues. More research issues related to health care big data can be found in Husing- Kuo et al. paper [6].

A. IoT for Big Data Analytics

Internet has restructured global interrelations, the art of businesses, cultural revolutions and an unbelievable number of personal characteristics. Currently, machines are getting in on the act to control innumerable autonomous gadgets via internet and create Internet of Things (IoT). Thus, appliances are becoming the user of the internet, just like humans with the web browsers. Internet of Things is attracting the attention of recent researchers for its most promising opportunities and challenges. It has an imperative economic and societal impact for the future construction of information, network and communication technology. The new regulation of future will be eventually, everything will be connected and intelligently controlled. The concept of IoT is becoming more pertinent to the realistic world due to the development of mobile de-vices, embedded and ubiquitous communication technologies, cloud computing, and data analytics. Moreover, IoT presents challenges in combinations of volume, velocity and variety. In a broader sense, just like the internet, Internet of Things enables the devices to exist in a myriad of places and facilitates applications ranging from trivial to the crucial. Conversely, it is still mystifying to understand IoT well, including definitions, content and differences from other similar concepts. Several diversified technologies such as computational intelligence, and big- data can be incorporated together to improve the data management and knowledge discovery of large scale automation applications.

Knowledge acquisition from IoT data is the biggest challenge that big data professional are facing. Therefore, it is essential to develop infrastructure to analyze the IoT data. An IoT device generates continuous streams of data and the re-searchers can develop tools to extract meaningful information from these data using machine learning techniques. Under-standing these streams of data generated from IoT devices and analyzing them to get meaningful information is a challenging issue and it leads to big data analytics. Machine learning algorithms and computational intelligence techniques is the only solution to handle big data from IoT prospective. Key technologies that are associated with IoT are also discussed in many research papers [15]. Figure 2 depicts an overview of IoT big data and knowledge discovery process.

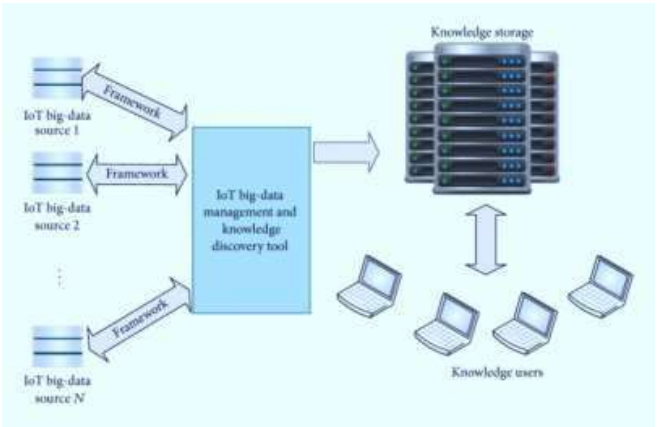

Fig. 2 IoT Big Data Knowledge Discovery

Knowledge exploration system have originated from theories of human information processing such as frames, rules, tagging, and semantic networks. In general, it consists of four segments such as knowledge acquisition, knowledge base, knowledge dissemination, and knowledge application.

In knowledge acquisition phase, knowledge is discovered by using various traditional and computational intelligence techniques, the discovered knowledge is stored in knowledge bases and expert systems are generally designed based on the discovered knowledge. Knowledge dissemination is important for obtaining meaningful information from the knowledge bases and expert systems are generally designed based on the discovered knowledge. Knowledge dissemination is important for obtaining meaningful information from the knowledge base. Knowledge extraction is a process that searches documents, knowledge within documents as well as knowledge bases. The final phase is to apply discovered knowledge in various applications. It is the ultimate goal of knowledge discovery. The knowledge exploration system is necessarily iterative with the judgement of knowledge application. There

are many issues, discussions, and researches in this area of knowledge exploration. It is beyond the scope of this survey paper. For better visualization, knowledge exploration system is depicted in Figure 3.

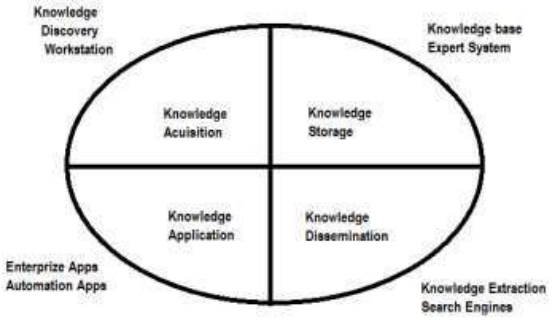

FIG. 3 IoT Knowledge Exploration Syste*m*

B. Cloud Computing for Big Data Analytics

The development of virtualization technologies have made Supercomputing more accessible and affordable. Computing infrastructures that are hidden in virtualization software make systems to behave like a true computer, but with the flexibility of specification details such as number of processors, disk space, memory, and operating system. The use of these virtual computers is known as cloud computing which has been one of the most robust big data technique. Big Data and cloud computing technologies are developed with the importance of developing a scalable and on demand availability of resources and data. Cloud computing harmonize massive data by on demand access to configurable computing resources through virtualization techniques. The benefits of utilizing the Cloud computing include offering resources when there is a demand and pay only for the resources which is needed to develop the product. Simultaneously, it improves availability and cost reduction. Open challenges and research issues of big data and cloud computing are discussed in detail by many re- searchers which highlights the challenges in data management, data variety and velocity, data storage, data processing, and resource management [16]. So Cloud computing helps in developing a business model for all varieties of applications with infrastructure and tools. Big data application using cloud computing should support data analytic and development. The cloud environment should provide tools that allow data scientists and business analysts to interactively and collaboratively explore knowledge acquisition data for further processing and extracting fruitful results. This can help to solve large applications that may arise in various domains. In addition to this, cloud computing should also enable scaling of tools from virtual technologies into new technologies like spark, R, and other types of big data processing techniques. Big data forms a framework for discussing cloud computing options. Depending on special need, user can go to the Market place and buy infrastructure services from cloud service providers such as Google, Amazon, IBM, software as a service (SaaS) from a whole crew of companies such as NetSuite, Cloud9, Job science, etc.

C. Bio-inspired Computing for Big Data Analytics

Bio inspired computing is a technique inspired my nature to address complex real world problems. Biological systems are self- organized without a central control. A bio-inspired cost minimization mechanism search and find the optimal data service solution on considering cost of data management and service maintenance. These techniques are developed by biological molecules such as DNA and proteins to conduct computational calculations involving storing, retrieving, and processing of data. A significant feature of such computing is that it integrates biologically derived materials to perform computational functions and receive intelligent performance. These systems are more suitable for big data applications. Huge amount of data are generated from variety of resources across the web since the digitization. Analyzing these data and categorizing into text, image and video, etc. will require lot of intelligent analytics from data scientists and big data professionals. Proliferations of technologies are emerging like big data, IoT, cloud computing, bio inspired computing etc.

D. Quantum Computing for Big Data Analysis

A quantum computer has memory that is exponentially larger than its physical size and can manipulate an exponential set of inputs simultaneously. This exponential improvement in computer systems might be possible. If a real quantum computer is available now, it could have solved problems that are exceptionally difficult on recent computers, of course today's big data problems. The main technical difficulty in building quantum computer could soon be possible. Quantum computing provides a way to merge the quantum mechanics to process the information. In traditional computer, information is presented by long strings of bits which encode either a zero or a one. On the other hand a quantum computer uses quantum bits or qubits. The difference between qubit and bit is that, a qubit is a quantum system that encodes the zero and the one into two distinguishable quantum states. Therefore, it can be capitalized on the phenomena of superposition and entanglement. For example, 100 qubits in quantum systems require 2100 complex values to be stored in a classic computer system. It means that many big data problems can be solved much faster by larger scale quantum computers compared with classical computers. Hence it is a challenge for this generation to build a quantum computer and facilitate quantum computing to solve big data problems.

IV. TOOLS FOR BIG DATA PROCESSING

Large numbers of tools are available to process big data. In this section, we discuss some current techniques for analysing big data with emphasis on three important engineering tools namely MapReduce, Apache Spark, and Storm. Most of the available tools concentrate on batch processing, steam processing and interactive analysis. Most batch processing tools are based on the Apache Hadoop infrastructure such as Mahout and Dryad. Stream data applications are mostly used for real time analytic. Some examples of large scale streaming platform are Strom and Splunk. The interactive analysis process allow users to directly interact in real time for their own analysis. For example Dremel and Apache Drill are the big data plat-forms that support interactive analysis. These tools help us in developing the big data projects. A fabulous list of big data tools and techniques is also discussed by much researchers [6]. The typical work flow of big data project discussed by Huang et al is highlighted in this section and is depicted in Figure 4.

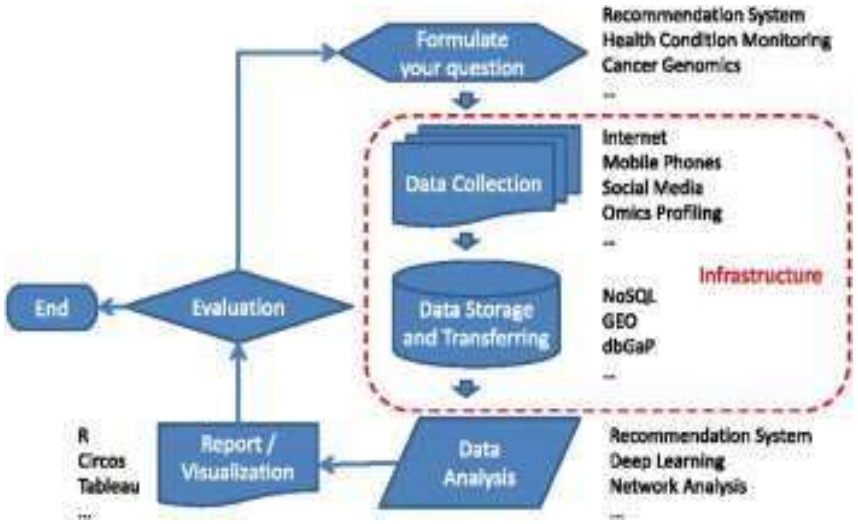

Fig. 4 Work Flow of Big Project

A. Apache Hadoop and MapReduce

The most established software platform for big data analysis is Apache Hadoop and mapreduce. It consists of hadoop kernel, mapreduce, hadoop distributed file system (HDFS) and apache hive etc. Map reduce is a programming model and apache hive etc. Map reduce is a programming model for processing large datasets is based on divide and conquer method. The divide and conquer method is implemented in two steps such as Map step and Reduce Step. Hadoop works on two kinds of nodes such as master node and worker node. The master node divides the input into smaller sub problems and then distributes them to worker nodes in map step. Thereafter the master node combines the outputs for all the sub problems in reduce step. Moreover, Hadoop and MapReduce works as a powerful software framework for solving big data problems. It is also helpful in fault-tolerant storage and high throughput data processing.

B. Apache Mahout

Apache mahout aims to provide scalable and commercial machine learning techniques for large scale and intelligent data analysis applications. Core algorithms of mahout including clustering, classification, pattern mining, regression, dimensionality reduction, evolutionary algorithms, and batch based collaborative filtering run on top of Hadoop platform through map reduce framework. The goal of mahout is to build a vibrant, responsive, diverse community to facilitate discussions on the project and potential use cases. The basic objective of Apache Mahout is to provide a tool for elevating big challenges. The different companies those who have implemented scalable machine learning algorithms are Google, IBM, Amazon, Facebook, and Twitter.

C. Apache Spark

Apache spark is an open source big data processing frame- work built for speed processing, and sophisticated analytics. It is easy to use and was originally developed in 2009 in UC Berkeleys AMPLab. It was open sourced in 2010 as an Apache project. Spark lets you quickly write applications in java, scala, or python. In addition to map reduce operations, it supports SQL queries, streaming data, machine learning, and graph data processing. Spark runs on top of existing hadoop distributed file system (HDFS) infrastructure to provide enhanced and additional functionality. Spark consists of components namely driver program, cluster manager and worker nodes. The driver program serves as the starting point of execution of an application on the spark cluster. The cluster manager allocates the resources and the worker nodes to do the data processing in the form of tasks. Each application will have a set of processes called executors that are responsible for executing the tasks. The major advantage is that it provides support for deploying spark applications in an existing hadoop clusters. Figure 5 depicts the architecture diagram of Apache Spark. The various features of Apache Spark are listed below:

Another advantage is that user can run the application program in different languages such as Java, R, Python, or Scala. This is possible as it comes with higher-level libraries for advanced analytics. These standard libraries increases developer productivity and can be seamlessly combined to create complex work- flows

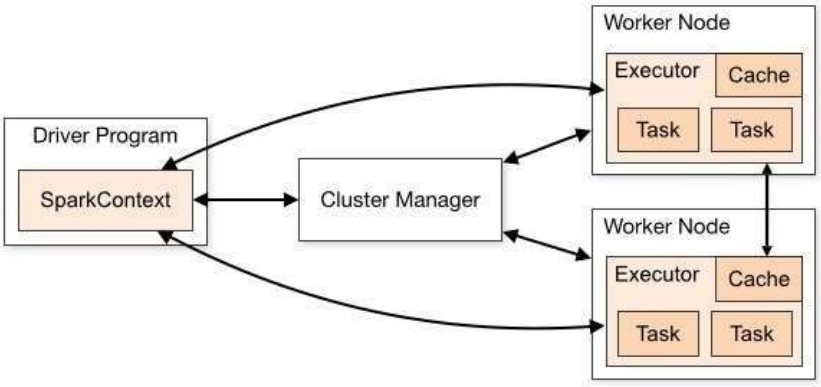

Fig. 5 Architecture of Apache Spark

D. Dryad

It is another popular programming model for implementing parallel and distributed programs for handling large context bases on dataflow graph. It consists of a cluster of computing nodes, and a user use the resources of a computer cluster to run their program in a distributed way. Indeed a dryad user use thousands of machines, each of them with multiple processors or cores. The major advantage is that users do not need to know anything about concurrent programming. A dryad application runs a computational directed graph that is composed of computational vertices and communication channels. Therefore, dryad provides a large number of functionality including generating of job graph, scheduling of the machines for the available processes, transition failure handling in the cluster, collection of performance metrics, visualizing the job, invoking user defined policies and dynamically updating the job graph in response to these policy decisions without knowing the semantics of the vertices.

E. Storm

Storm is a distributed and fault tolerant real time computation system for processing large streaming data. It is specially designed for real time processing in contrasts with hadoop which is for batch processing. Additionally, it is also easy to set up and operate, scalable, fault-tolerant to provide competitive performances. The storm cluster is apparently similar to hadoop cluster. On storm cluster users run different topologies for different storm tasks whereas hadoop platform implements map reduce jobs for corresponding applications. There are number of differences between map reduce jobs and topologies. The basic differences is that map reduce job eventually finishes whereas a topology process messages all two kinds of nodes such as master node and worker node. The master node and worker node implement two kinds of roles such as nimbus and supervisor respectively. The two roles have similar functions in accordance with job tracker and task tracker of map reduce framework. Nimbus is in charge of distributing code across the storm cluster, scheduling and assigning tasks to worker nodes, and monitoring the whole system.

V. CONCLUSION

In recent years data are generated at dramatic pace analyzing these data is challenging for a general man. To this end in this paper, we survey the various research issues, challenges, and tools used to analyze these big data. From this survey, it is understood that every big data platform has its individual focus some of them are designed for batch processing whereas some are good at real-time analytic. Each big data platform also has specific functionality. Different techniques used for the analysis include statistical analysis, machine learning, data mining, intelligent analysis, cloud computing, quantum computing, and data stream processing.

References

[1]. M. K. Kakhani, S. Kakhani and S. R.Biradar, Research issues in big data analytics, International Journal of Application or Innovation inEngineering &Management, 2(8) (2015), pp.228-232.

[2]. Lynch, Big data: How do your data grow? Nature, 455 (2008), pp.28-29.

[3]. X. Jin, B. W.Wah, X. Cheng and Y. Wang, Significance and challenges or big data research, Big Data Research, 2(2)(2015), pp.59-64.

[4]. C. L. Philip, Q. Chen and C. Y. Zhang, Data-intensive applications, challenges, techniques and technologies: A survey on big data, Information Sciences 275 (2014), pp.314-347.

[5]. K. Kambatla, G. Kollias, V. Kumar and A. Gram, Trends in big data analytics, Journal of Parallel and Distributed Computing, 74(7) (2014), pp.2561-2573.

[6]. S. Del. Rio, V. Lopez, J. M. Bentez and F. Herrera, on the use of mapreduce for imbalanced big data using random forest, Information Sciences, 285 (2014), pp.112-137.

[7]. Z. Huang, A fast clustering algorithm to cluster very large categorical data sets in data mining, SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery, 1997.

[8]. O.Y. AL-Jarrah, P.D. Yoo, S.Muhaidat, G. K. Karagiannidis and K. Taha, Efficient machine learning for big data: A review, Big Data Research 2(3) (2015), pp 87-93.

[9]. Changwon. Y, Luis. Ramirez and Juan. Liuzzi, Big data analysis using modern statistical and machine learning methods in medicine, International Neurourology Journal, 18 (2014), pp.50-57.

[10]. P. Singh and B. Suri, Quality assessment of data using statistical and machine learning methods. L. C.Jain, H. S.Behera, J. K.Mandal and D. P.Mohapatra (eds.), Computational Intelligence in Data Mining, 2 (2014), pp. 89-97.

[11]. A. Jacobs, The pathologies of big data, Communications of the ACM, 52(8) (2009), pp.36-44.

[12]. H. Zhu, Z. Xu and Y. Huang, Research on the security technology of big data information, International Conference on Information Technology Management Innovation, 2015, pp.1041-1044.

[13]. Z. Hongjun, H. Wenning, H. Dengchao and M. Yuxing, Survey of research on information security in big data, Congress da sociedada Brasileira de Computacao, 2014, pp.1-6.

[14]. I. Merelli, H. Perez-sanchez, S. Gesing and D. Agostino, Managing, analyzing, and integrating big data in medical bioinformatics: open problems and future perspectives, BioMed Research International, 2014, (2014), pp.1-13.