ISSN 2348-1196 (print)
International Journal of Computer Science and Information Technology Research ISSN 2348-120X (online)
Vol. 7, Issue 4, pp: (71-82), Month: October - December 2019, Available at: www.researchpublish.com

SURVEY ON BIG DATA CLOUD ANALYTICS: CHALLENGES, OPEN RESEARCH ISSUES ON DATA CLOUD AND TOOLS ACCESSIBILITY

Nicholas .O. Okunwe[1*], Dr. Ayodele Fasanya[2], Garba Abdullahi[3]

[1]Department of Computer Science, College of Education Waka-Biu

[2]Department of Physics, College of Education Waka-Biu

[3]Department of Adult Education, College of Education Waka-Biu

* *Email of the Corresponding Author: linearharty@gmail.com*

Abstract: **A huge repository of terabytes of data is generated each day from modern information systems and digital technologies such as Internet of Things and cloud computing. Analysis of these massive data requires a lot of efforts at multiple levels to extract knowledge for decision making. Therefore, big data analysis is a current area of research and development. The basic objective of this paper is to explore the potential impact of big data challenges, open research issues, and various tools associated with it. As a result, this article provides a platform to explore big data at numerous stages. Additionally, it opens a new horizon for researchers to develop the solution, based on the challenges and open research issues.**

Keywords: **Surveying Big data cloud analytics; Had-Oop; Massive data; Structured data; Unstructured Data.**

1. INTRODUCTION

In digital world, data are generated from various sources and the fast transition from digital technologies has led to growth of big data. It provides evolutionary breakthroughs in many fields with collection of large datasets. In general, it refers to the collection of large and complex datasets which are difficult to process using traditional database management tools or data processing applications. These are available in structured, semi-structured, and unstructured format in Peta bytes and beyond. Formally, it is defined from 3Vs to 4Vs. 3Vs refers to volume, velocity, and variety. Volume refers to the huge amount of data that are being generated everyday whereas velocity is the rate of growth and how fast the data are gathered for being analysis. Variety provides information about the types of data such as structured, unstructured, semi structured etc. The fourth V refers to veracity that includes availability and accountability. The prime objective of big data analysis is to process data of high volume, velocity, variety, and veracity using various traditional and computational intelligent techniques, M. K.Kakhani and S. Kakhani etal (2015). Some of these extraction methods for obtaining helpful information was discussed by Gandomi and Haider (2015), The following Figure 1 refers to the definition of big data. However exact definition for big data is not defined and there is a believe that it is problem specific. This will help us in obtaining enhanced decision making, insight discovery and optimization while being innovative and cost-effective. It is expected that the growth of big data is estimated to reach 25 billion by 2015. According to Lynch .c, Big data: How do your data grow Natural, (2008),. From the perspective of the information and communication technology, big data is a robust impetus to the next generation of information technology industries X. Jin, et al, (2015), which are broadly built on the third platform, mainly referring to big data, cloud computing, internet of things, and social business. Generally, Data warehouses have been used to manage the large dataset. In this case extracting the precise knowledge from the available

ISSN 2348-1196 (print)
International Journal of Computer Science and Information Technology Research ISSN 2348-120X (online)
Vol. 7, Issue 4, pp: (71-82), Month: October - December 2019, Available at: www.researchpublish.com

big data is a foremost issue. Most of the presented approaches in data mining are not usually able to handle the large datasets successfully. The key problem in the analysis of big data is the lack of coordination between database systems as well as with analysis tools such as data mining and statistical analysis. These challenges generally arise when we wish to perform knowledge discovery and representation for its practical applications. A fundamental problem is how to quantitatively describe the essential characteristics of big data. There is a need for epistemological implications in describing data revolution R. Kitchin, (2014), Additionally, the study on complexity theory of big data will help understand essential characteristics and formation of complex patterns in big data, simplify its representation, gets better knowledge abstraction, and guide the design of computing models and algorithms on big data, according to R. Kitchin, etal, (2014), C. L. Philip, et al, (2014), K. Kambatla, G. Kollias, etal, (2014), S. Del. Rio, V. Lopez, etal (2014). However, it is to be noted that all data available in the form of big data are not useful for analysis or decision making process. Industry and academia are interested in disseminating the findings of big data. This paper focuses on challenges in big data and its available techniques. Additionally, we state open research issues in big data. So, to elaborate this, the paper is divided into following sections. Sections 2 deals with challenges that arise during fine tuning of big data. Section 3 furnishes the open research issues that will help us to process big data and extract useful knowledge from it. Section 4 provides an insight to big data tools and techniques. Conclusion remarks are provided in section 5 to summarize outcomes.

2. CHALLENGES IN BIG DATA ANALYTICS

Recent year's big data has been accumulated in several domains like health care, public administration, retail, biochemistry, and other interdisciplinary scientific researches. Web-based applications encounter big data frequently, such as social computing, internet text and documents, and internet search indexing. Social computing includes social network analysis, online communities, recommender systems, reputation systems, and prediction markets where as internet search indexing includes ISI, IEEE X-plorer, Scopus, and Thomson.

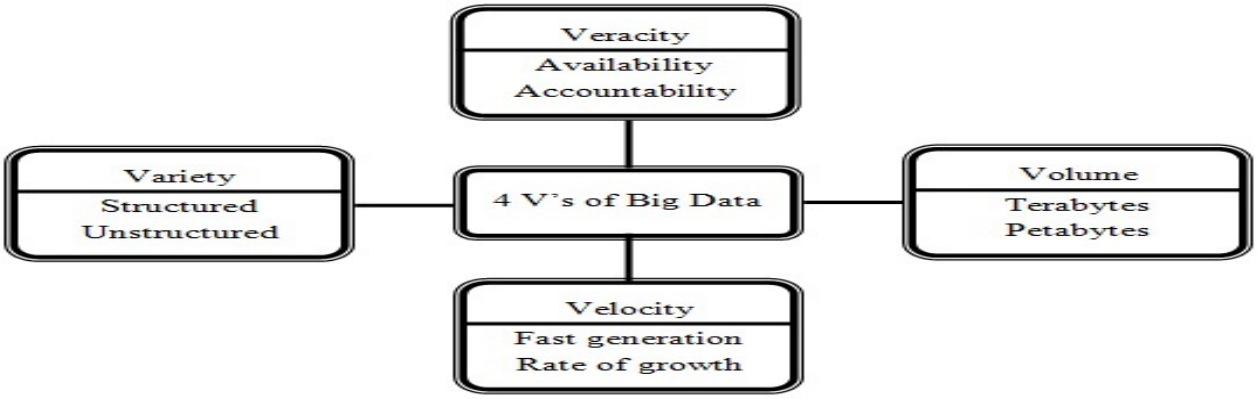

Fig. 1: Characteristics of Big Data

Reuters etc. Considering this advantages of big data it provides a new opportunities in the knowledge processing tasks for the upcoming researchers. However opportunities always follow some challenges. To handle the challenges we need to know various computational complexities, information security, and computational method, to analyze big data. For example, many statistical methods that perform well for small data size do not scale to voluminous data. Similarly, many computational techniques that perform well for small data face significant challenges in analyzing big data. Various challenges that the health sector face was being researched by much researchers MH. Kuo, T. Sahama, A. W. Kushniruk, etal, (2014), R. Nambiar, and R. Vargheese, etal, (2013). Here the challenges of big data analytics are classified into four broad categories namely data storage and analysis; knowledge discovery and computational complexities; scalability and visualization of data; and information security. We discuss these issues briefly in the following subsections.

A. Data Storage and Analysis

In recent years the size of data has grown exponentially by various means such as mobile devices, aerial sensory technologies, remote sensing, radio frequency identification readers etc. These data are stored on spending much cost whereas they ignored or deleted finally because there is no enough space to store them. Therefore, the first challenge for big data analysis is storage mediums and higher input/output speed. In such cases, the data accessibility must be on the

top priority for the knowledge discovery and representation. The prime reason is being that, it must be accessed easily and promptly for further analysis. In past decades, analyst use hard disk drives to store data but, it slower random input/output performance than sequential input/output. To overcome this limitation, the concept of solid state drive (SSD) and phrase change memory (PCM) was introduced. However the available storage technologies cannot possess the required performance for processing big data. Another challenge with Big Data analysis is attributed to diversity of data. with the ever growing of datasets, data mining tasks has significantly increased. Additionally data reduction, data selection, feature selection is an essential task especially when dealing with large datasets. This presents an unprecedented challenge for researchers. It is because, existing algorithms may not always respond in an adequate time when dealing with these high dimensional data. Automation of this process and developing new machine learning algorithms to ensure consistency is a major challenge in recent years. In addition to all these Clustering of large datasets that help in analyzing the big data is of prime concern Z. Huang, (1997). Recent technologies such as had-oop and map Reduce make it possible to collect large amount of semi structured and unstructured data in a reasonable amount of time. The key engineering challenge is how to effectively analyze these data for obtaining better knowledge. A standard process to this end is to transform the semi structured or unstructured data into structured data, and then apply data mining algorithms to extract knowledge. A framework to analyze data was discussed by T. K. Das and P. M. Kumar, (2013). Similarly detail explanation of data analysis for public tweets was also discussed by Das et al in their paper, Acharjya and M. R. Patra, (2014). The major challenge in this case is to pay more attention for designing storage systems and to elevate efficient data analysis tool that provide guarantees on the output when the data comes from different sources. Furthermore, design of machine learning algorithms to analyze data is essential for improving efficiency and scalability.

B. Knowledge Discovery and Computational Complexities

Knowledge discovery and representation is a prime issue in big data. It includes a number of sub fields such as authentication, archiving, management, preservation, information retrieval, and representation. There are several tools for knowledge discovery and representation such as fuzzy set

L. A. Zadeh, (1965), rough set Z. Pawlak, (1982), soft set] D. Molodtsov, (1999), near set J. F.Peters, (2007), R. Wille, (2005), formal concept analysis R. Wille, (2005), principal component analysis I. T.Jolliffe, (2002). etc to name a few. Additionally many hybridized techniques are also developed to process real life problems. All these techniques are problem dependent. Further some of these techniques may not be suitable for large datasets in a sequential computer. At the same time some of the techniques has good characteristics of scalability over parallel computer. Since the size of big data keeps increasing exponentially, the available tools may not be efficient to process these data for obtaining meaningful information. The most popular approach in case of large dataset management is data warehouses and data marts. Data warehouse is mainly responsible to store data that are sourced from operational systems whereas data mart is based on a data warehouse and facilitates analysis. Analysis of large dataset requires more computational complexities. The major issue is to handle inconsistencies and uncertainty present in the datasets. In general, systematic modeling of the computational complexity is used. It may be difficult to establish a comprehensive mathematical system that is broadly applicable to Big Data. But a domain specific data analytics can be done easily by understanding the particular complexities. A series of such development could simulate big data analytics for different areas. Much research and survey has been carried out in this direction using machine learning techniques with the least memory requirements. The basic objective in these research is to minimize computational cost processing and complexities O. Y. Al-Jarrah, P. D. Yoo, et al, (2015), Changwon. Y, Luis.et al, (2014), P. Singh and B. Suri, Quality assessment of data using statistical and machine learning methods. L. C.Jain, H. S.Behera, et al, (2014). However, current big data analysis tools have poor performance in handling computational complexities, uncertainty, and inconsistencies. It leads to a great challenge to develop techniques and technologies that can deal computational complexity, uncertainty, and inconsistencies in a effective manner.

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 re-solution 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 According to A. Jacobs,

(2009). 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. The objective of visualizing data is to present them more adequately using some techniques of graph theory. Graphical visualization provides the link between data with proper interpretation. However, online marketplace like flip-kart, Amazon, e-bay have millions of users and billions of goods to sold each month. This generates a lot of data. To this end, some company uses a tool Tableau for big data visualization. It has capability to transform large and complex data into intuitive pictures. These help employees of a company to visualize search relevance, monitor latest customer feedback, and their sentiment analysis. However, current big data visualization tools mostly have poor performances in functionalities, scalability, and response in time. 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 overcome 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 H. Zhu, Z. Xu and Y. Huang, (2015). 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 Z. Hongjun, M. Yuxing, et al, (2014), I. Merelli, H. Perez-sanchez, S. Gesing et al, Managing, (2014), (2014). The security challenge caused by big data has attracted the attention of information security. Therefore, attention has to be given to develop a multi level security policy model and prevention system. Although much research has been carried out to secure big data Z. Hongjun, M. Yuxing, et al, (2014), but it requires lot of improvement. The major challenge is to develop a multi-level security, privacy preserved data model for big data.

3. OPEN RESEARCH ISSUES IN BIG DATA ACCESS 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 modeling, 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 Health big data analytics: current perspectives, challenges and potential solutions, International Journal of Big Data Intelligence, (2014),

A. I o T 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 devices, 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. Much research in this direction has been carried out by Mishra, Lin and Chang N. Mishra, C. Lin and H. Chang, A, (2015). Knowledge acquisition from IoT data is the biggest challenge that big data professional are facing. Therefore, it is essential to develop

ISSN 2348-1196 (print)
International Journal of Computer Science and Information Technology Research ISSN 2348-120X (online)
Vol. 7, Issue 4, pp: (71-82), Month: October - December 2019, Available at: www.researchpublish.com

infrastructure to analyze the IoT data. An IoT device generates continuous streams of data and the researchers can develop tools to extract meaningful information from these data using machine learning techniques. Understanding 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 X. Y.Chen and Z. G.Jin, (2012).

Figure 2 depicts an overview of IoT big data and knowledge discovery process.

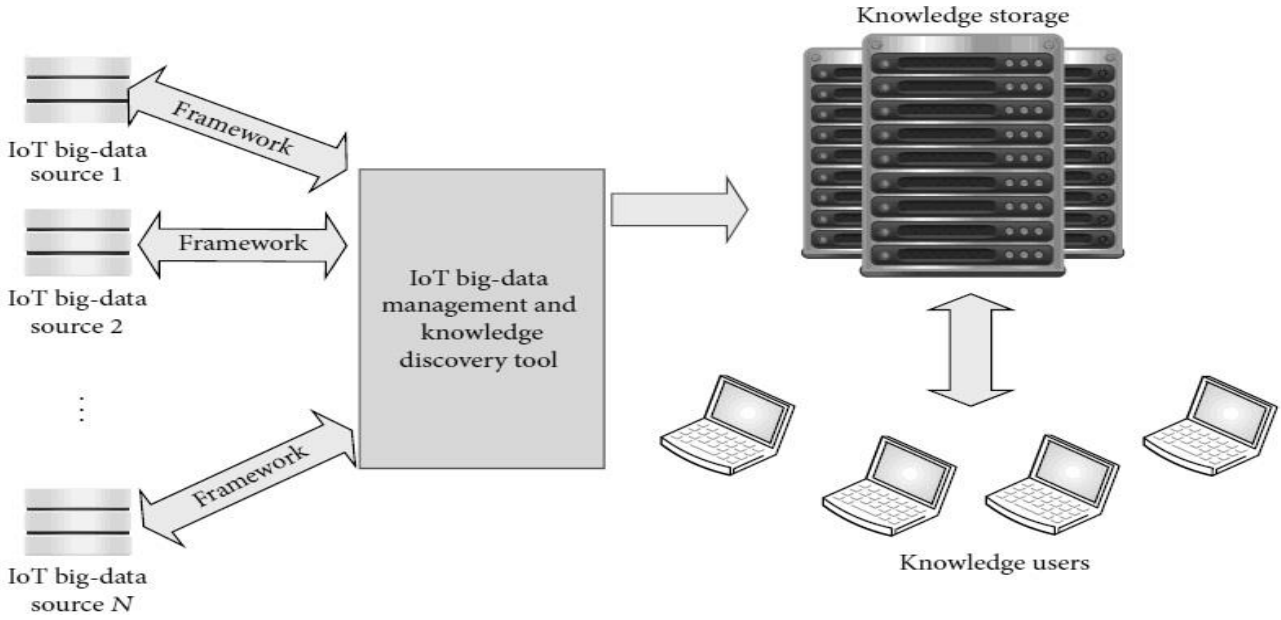

Fig. 2: IoT Big Data Knowledge Discovery

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

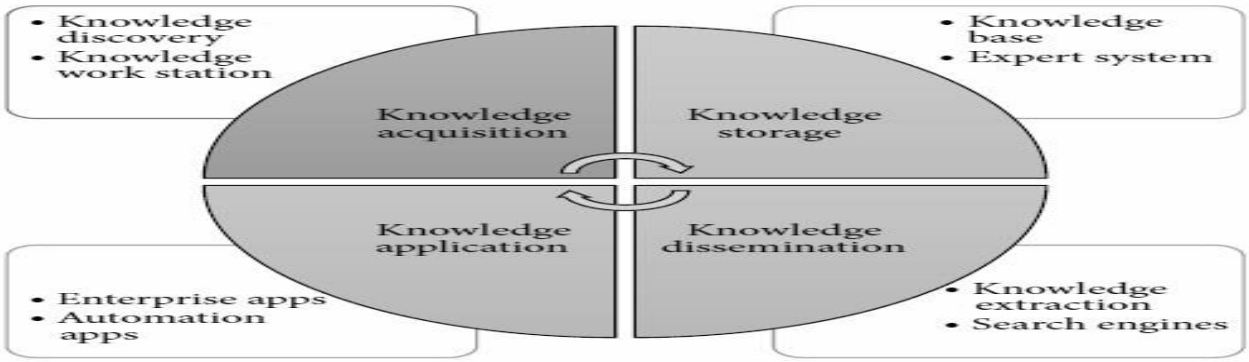

Fig. 3: IoT Knowledge Exploration System

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 researchers which highlights the challenges in data management, data variety and velocity, data storage, data processing, and resource management M. D. Assuno, R. N. Calheiros, S. Bianchi, M. A. S. Netto et al, (2015), I. A. T. Hashem, I. Yaqoob, N. Badrul Anuar, et al, (2014), 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 marketplace 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 Net-Suite, Cloud9, Job science etc. Another advantage of cloud computing is cloud storage which provides a possible way for storing big data. The obvious one is the time and cost that are needed to upload and download big data in the cloud environment. Else, it becomes difficult to control the distribution of computation and the underlying hardware. But, the major issues are privacy concerns relating to the hosting of data on public servers, and the storage of data from human studies. All these issues will take big data and cloud computing to a high level of development.

C. Bio-inspired Computing for Big Data Analytics

Bio-inspired computing is a technique inspired ny 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 whereas equilibrium of data can be done only by selecting right platform to analyze large and furnish cost effective results. Bio inspired computing techniques serve as a key role in intelligent data analysis and its application to big data. These algorithms help in performing data mining for large datasets due to its optimization application. The most advantage is its simplicity and their rapid convergence to optimal solution, Wang .L and J. Shen, Bio inspired cost-effective access to big data, International, (2013), while solving service provision problems. Some applications to this end using bio inspired computing was discussed in detail by Cheng, C. Shi, et al, (2013). From the discussions, we can observe that the bio-inspired computing models provide smarter interactions, inevitable data losses, and help is handling ambiguities. Hence, it is believed that in future bio-inspired computing may help in handling big data to a large extent.

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 by M. A. Nielsen and I. L.Chuang, (2000). 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. It is because qubits behave quantumly. 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 built a quantum computer and facilitate quantum computing to solve big data problems.

4. TOOLS FOR BIG DATA ACCESS PROCESSING

Large numbers of tools are available to process big data. In this section, we discuss some current techniques for analyzing big data with emphasis on three important emerging tools namely Map Reduce, Apache Spark, and Storm. Most of the available tools concentrate on batch processing, stream processing, and interactive analysis. Most batch processing tools are based on the Apache Had-oop 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 allows users to directly interact in real time for their own analysis. For example Dremel and Apache Drill are the big data platforms 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 C. L. Philip, Q. Chen and C. Y. Zhang, (2014), M. Herland, T. M. Khoshgoftaar and R. Wald, (2014). The typical work flow of big data project discussed by Huang et al is highlighted in this section T. Huang, and F. Wang, Promises and challenges of big data computing in health sciences, Big Data Research (2015), and is depicted in Figure 4.

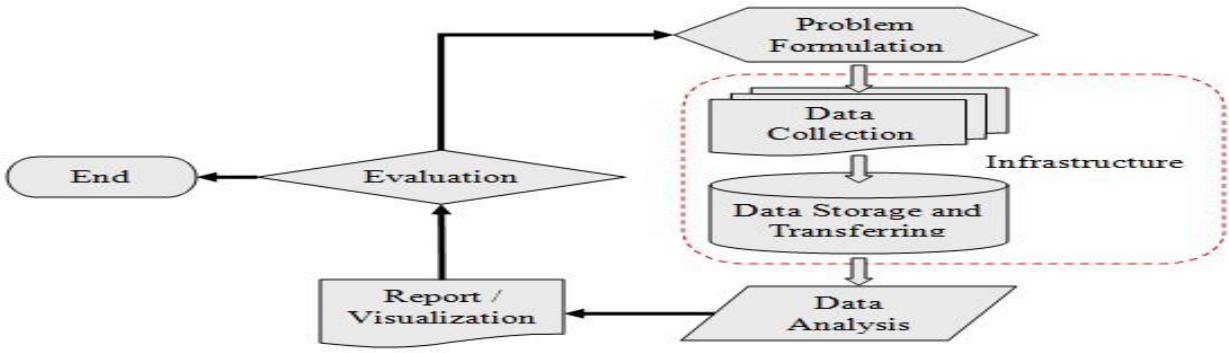

Fig. 4: Workflow of Big Data Project

A. Apache Had-oop and Map Reduce

The most established software platform for big data analysis is Apache Had-oop and Map reduce. It consists of had-oop kernel, map reduces, had-oop distributed file system (HDFS) 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. Had-oop 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, Had-oop and Map Reduce 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 Had oop 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

ISSN 2348-1196 (print)
International Journal of Computer Science and Information Technology Research ISSN 2348-120X (online)
Vol. 7, Issue 4, pp: (71-82), Month: October - December 2019, Available at: www.researchpublish.com

big challenges. The different companies those who have implemented scalable machine learning algorithms are Google, IBM, Amazon, Yahoo, Twitter, and facebook G. Ingersoll, (2009).

C. Apache Spark

Apache spark is an open source big data processing framework built for speed processing, and sophisticated analytics. It is easy to use and was originally developed in 2009 in UC Berkeley's AMP-Lab. 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 had-oop 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 had-oop cluster. Figure 5, depicts the architecture diagram of Apache Spark. The various features of Apache Spark are listed below:

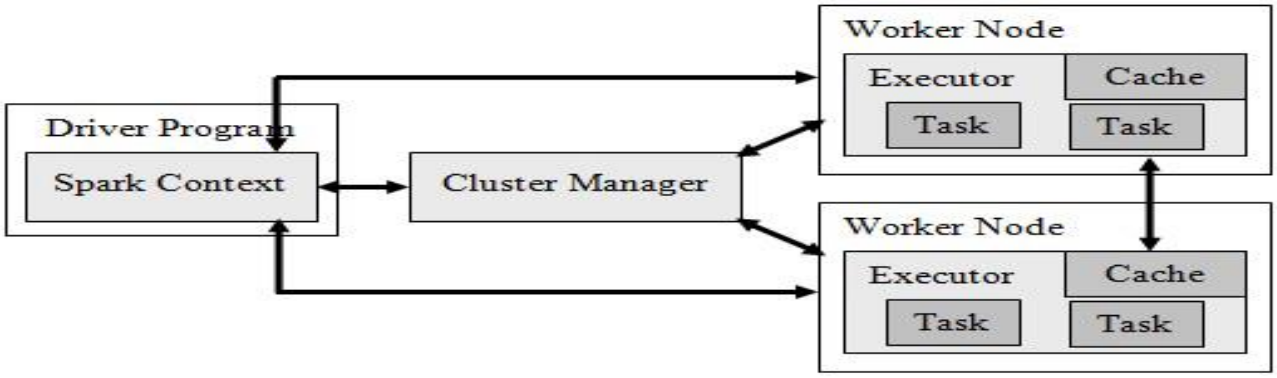

Fig. 5: Architecture of Apache Spark

- The prime focus of spark includes resilient distributed datasets (RDD), which store data in-memory and provide fault tolerance without replication. It supports iterative computation, improves speed and resource utilization.
- The foremost advantage is that in addition to Map Reduce, it also supports streaming data, machine learning, and graph algorithms.
- Another advantage is that, a 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 increase developer productivity and can be seamlessly combined to create complex workflows.
- Spark helps to run an application in Had-oop cluster, up to 100 times faster in memory, and 10 times faster when running on disk. It is possible because of the reduction in number of read or write operations to disk.
- It is written in scala programming language and runs on java virtual machine (JVM) environment. Additionally, it up ports java, python and R for developing applications using 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 an user use the resources of a computer cluster to run their program in a distributed way. Indeed, a dryad user uses 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 according to H. Li, G. Fox and J. Qiu, (2012).

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 had-oop 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 had-oop cluster. On storm cluster users run different topologies for different storm tasks whereas had-oop platform implements map reduce jobs for corresponding applications. There are number of differences between map reduce jobs and topologies. The basic difference is that map reduce job eventually finishes whereas a topology processes messages all the time, or until user terminate it. A storm cluster consists of 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. The supervisor complies tasks as assigned to them by nimbus. In addition, it start and terminate the process as necessary based on the instructions of nimbus. The whole computational technology is partitioned and distributed to a number of worker processes and each worker process implements a part of the topology.

F. Apache Drill

Apache drill is another distributed system for interactive analysis of big data. It has more flexibility to support many types of query languages, data formats, and data sources. It is also specially designed to exploit nested data. Also it has an objective to scale up on 10,000 servers or more and reaches the capability to process pate-bytes of data and trillions of records in seconds. Drill use HDFS for storage and map reduce to perform batch analysis.

G. Jasper soft

The Jasper soft package is an open source software that produce reports from database columns. It is a scalable big data analytical platform and has a capability of fast data visualization on popular storage platforms, including Mango DB, Cassandra, and Redis etc. One important property of Jasper soft is that it can quickly explore big data without extraction, transformation, and loading (ETL). In addition to this, it also has an ability to build powerful hypertext markup language (HTML) reports and dashboards interactively and directly from big data store without ETL requirement. These generated reports can be shared with anyone inside or outside user's organization.

H. Splunk

In recent years a lot of data are generated through machine from business industries. Splunk is a real-time and intelligent platform developed for exploiting machine generated big data. It combines the up-to-the-moment cloud technologies and big data. In turn it helps user to search, monitor, and analyze their machine generated data through web interface. The results are exhibited in an intuitive way such as graphs, reports, and alerts. Splunk is different from other stream processing tools. Its peculiarities include indexing structured, unstructured machine generated data, real-time searching, reporting analytical results, and dashboards. The most important objective of Splunk is to provide matrices for many applications, diagnose problems for system and information technology infrastructures, and intelligent support for business operations.

5. SUGGESTIONS FOR RECOMMENDATION

The amount of data collected from various applications all over the world across a wide variety of fields today is expected to double every two years. It has no utility unless these are analyzed to get useful information. This necessitates the development of techniques which can be used to facilitate big data analysis. The development of powerful computers is a boon to implement these techniques leading to automated systems. The transformation of data into knowledge is by no means an easy task for high performance large-scale data processing, including exploiting parallelism of current and upcoming computer architectures for data mining. Moreover, these data may involve uncertainty in many different forms. Many different models like fuzzy sets, rough sets, soft sets,

neural networks, their generalizations and hybrid models obtained by combining two or more of these models have been found to be fruitful in representing data. These models are also very much fruitful for analysis. More often than not, big data are reduced to include only the important characteristics necessary from a particular study point of view or depending upon the application area. So, reduction techniques have been developed. Often the data collected have missing values.

These values need to be generated or the tuples having these missing values are eliminated from the data set before analysis. More importantly, these new challenges may comprise, sometimes even deteriorate, the performance, efficiency and scalability of the dedicated data intensive computing systems. The later approach sometimes leads to loss of information and hence not preferred. This brings up many research issues in the industry and research community in forms of capturing and

accessing data effectively. In addition, fast processing while achieving high performance and high throughput, and storing it efficiently for future use is another issue. Further, programming for big data analysis is an important challenging issue. Expressing data access requirements of applications and designing programming language abstractions to exploit parallelism are an immediate need Acharjya D.P, and S. Dehuri et al, (2015). Additionally, machine learning concepts and tools are gaining popularity among researchers to facilitate meaningful results from these concepts. Research in the area of machine learning for big data has focused on data processing, algorithm implementation, and optimization. Many of the machine learning tools for big data are started recently needs drastic change to adopt it. We argue that while each of the tools has their advantages and limitations, more efficient tools can be developed for dealing with problems inherent to big data. The efficient tools to be developed must have provision to handle noisy and imbalance data, uncertainty and inconsistency, and missing values.

6. CONCLUSION

In recent years data are generated at a 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. We believe that in future researchers will pay more attention to these techniques to solve problems of big data effectively and efficiently.

REFERENCES

[1] Acharjya D. P, Das T. K, & Patra M.R, (2014). Opinion mining about a product by analyzing public tweets in twitter, International Conference on Computer Communication and Informatics,

[2] Acharjya D. P, Dehuri .S, & Sanyal .S, Computational Intelligence for Big Data Analysis, Springer International Publishing AG, Switzerland, USA, ISBN 978-3-319-16597-4, 2015.

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

[4] Assuno M. D, Calheiros R. N, Bianchi .S, Netto M.A.S, & Buyya .R, Big data computing and clouds: Trends and future directions, Journal of Parallel and Distributed Computing, 79 (2015), pp.3-15.

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

[6] Chen X.Y and Jin Z.G, Research on key technology, & applications for internet of things, Physics Procedia, 33, (2012), pp. 561-566.

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

[8] Gandomi .A, & Haider .M, Beyond the hype: Big data concepts, methods, and analytics, International Journal of Information Management, 35(2) (2015), pp.137-144.

[9] Hashem I. A. T, Yaqoob .I, Badrul Anuar N, Mokhtar S, Gani .A, & Ullah Khan S, The rise of big data on cloud computing: Review and open research issues, Information Systems, 47 (2014), pp. 98-115.

[10] Herland M, Khoshgoftaar T.M, & Wald .R, A review of data mining using big data in health informatics, Journal of Big Data, 1(2) (2014), pp. 1-35.

[11] Huang T, Lan .L, Fang .X, An .P, Min .J, & Wang .F, Promises and challenges of big data computing in health sciences, Big Data Research, 2(1) (2015), pp. 2-11.

[12] Huang .Z, 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).

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

[14] Ingersoll .G, Introducing apache mahout: Scalable, commercial friendly machine learning for building intelligent applications, White Paper, IBM Developer Works, (2009), pp. 1-18.

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

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

[17] Jolliffe I .T, Principal Component Analysis, Springer, New York, 2002.

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

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

[20] Kitchin R, Big Data, new epistemologies and paradigm shifts, Big Data Society, 1(1) (2014), pp.1-12.

[21] Kumar P. M, & Das T. K, Big data analytics: A framework for unstructured data analysis,International Journal of Engineering and Technology, 5(1) (2013), pp.153-156.

[22] Kuo M.H, Sahama .T, Kushniruk A. W, Borycki E.M and Grunwell D.K, Health big data analytics: current perspectives, challenges and potential solutions, International Journal of Big Data Intelligence, 1 (2014), pp.114-126.

[23] Li .H, Fox .G, & Qiu .J, Performance model for parallel matrix multiplication with dryad: Dataflow graph runtime, Second International Conference on Cloud and Green Computing, 2012, pp.675-683.

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

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

[26] Mishra .N, Lin .C, & Chang .H, A cognitive adopted framework for iot big data management and knowledge discovery prospective, International Journal of Distributed Sensor Networks, 2015, (2015), pp. 1-13

[27] Molodtsov D, Soft set theory first results, Computers and Mathematics with Aplications, 37(4/5) (1999), pp.19-31.

[28] Nambiar .R, Sethi .A, Bhardwaj .R, & Vargheese R, A look at challenges and opportunities of big data analytics in healthcare, IEEE International Conference on Big Data, 2013, pp.17-22.

[29] Nielsen M. A, & Chuang I. L, Quantum Computation and Quantum Information, Cambridge University Press, New York, USA 2000.

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

[31] Peters J. F, Near sets. General theory about nearness of objects, Applied Mathematical Sciences, 1(53) (2007), pp.2609-2629.

[32] Pawlak Z, Rough sets, International Journal of Computer Information Science, 11 (1982), pp.341-356.

[33] Shi .C, Shi .Y, Qin .Q, & Bai R, Swarm intelligence in big data analytics, H. Yin, K. Tang,

[34] Y. Gao, F. Klawonn, M. Lee, T. Weise, B. Li and X. Yao (eds.), Intelligent Data Engineering and Automated Learning, 2013, pp.417-426.

[35] Singh .P, & Suri B, 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.

[36] Wang .L, & Shen .J, Bioinspired cost-effective access to big data, International Symposium for Next Generation Infrastructure, 2013, pp.1-7.

[37] Wille R, Formal concept analysis as mathematical theory of concept and concepthierarchies, Lecture Notes in Artificial Intelligence, 3626 (2005), pp.1-33.

[38] Zadeh L. A, Fuzzy sets, Information and Control, 8 (1965), pp.338- 353.

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