Challenges and OpenResearch Issues and Tools on Big Data Analytics

GameilSaadHamzh Ali[1], Dr.A.Nithya[2]

ABSTRACT

Modern information systems and digital technologies such as Internet of Things and cloud computing is generated a huge repository of terabytes of data each day. It requires a lot of efforts to analysis of these massive data at multiple levels to extract knowledge for decision making. Therefore, the current area of research and development isbig data analysis. The aim 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 opens a new horizon for researchers to develop the solution, based on the challenges and open research issues.

Keywords: Big data analytics; Structureddata; Semi Structured data ,Unstructured Data

1- INTRODUCTION

Data are generated in digital world from several sourcesand the fast transition from digital technologies has led togrowth of big data. It provides evolutionary breakthroughs inmany fields with collection of large datasets. In generally,there are a collection ofcomplexand largedatasets whichare difficult to process usingdata processing applications or traditional database managementtools . These are availablein structured, semi-structured, and unstructured format inpetabytes and beyond. Formally, it is defined from four challenges refers tovelocity, volume,variety, and veracity. Velocity is the rate of growth and how fast the data are gathered for being analysiswhereasVolume refers tothe huge amount of data that are being generated every day. Variety provides informationabout the types of data such as structured, unstructured, semi structuredetc. The fourth V refers to veracity that includesavailability and accountability. The prime objective of big dataanalysis is to process data of high volume, velocity, variety, andveracity using various traditional and computational intelligenttechniques [1].Gandomi and Haider [2]discussedSome of these extraction methods for obtaininghelpful information . The following Figure 1 refers to the definition of bigdata. However exact definition for big data is not defined andthere is a believe that it is problem specific. This will support usin obtaining better decision making, insight discovery andoptimization while being innovative and cost-effective.It is expected that the growth of big data is estimated toreach 25 billion by 2015 [3]. From the perspective of

theinformation and communication technology, big data is a robustimpetus to the next generation of information technology industries [4], which are broadly built on the third platform,mainly referring to big data, cloud computing, internet ofthings, and social business. Generally, Data warehouses havebeen used to manage the large dataset. In this case extractingthe exact knowledge from the available big data is a mainissue. Most of the presented approaches in data mining are notusually able to handle the large datasets successfully. In big data analysis the keyproblem is the lack of coordinationbetween database systems as well as with analysis tools such as statistical analysis anddata mining. These challenges generallyarise when we wish to perform knowledge discovery and representationfor its practical applications. A fundamental problemis how to quantitatively describe the essential characteristicsof big data. There is a need for epistemological implicationsin describing data revolution [5]. Additionally, the study oncomplexity theory of big data will help understand essentialcharacteristics and formation of complex patterns in big data,simplify its representation, gets better knowledge abstraction,and guide the design of computing models and algorithmson big data [4]. Much research was carried out by variousresearchers on big data and its trends [6], [7], [8].However, it is to be noted that all data available in theform of big data are not useful for analysis or decision makingprocess. Industry and academia are interested in disseminatingthe findings of big data. This paper focuses on challenges inbig data and its available techniques. Additionally, we stateopen research issues in big data. So, to elaborate this, thepaper is divided into following sections. Sections 2 dealswith challenges that arise during fine tuning of big data.Section 3 provides the open research issues that will helpus to process big data and extract useful knowledge from it.Section 4 furnishes an insight to big data tools and techniques.Conclusion remarks are provided in section 5 to summarizeoutcomes.

2- CHALLENGES IN BIG DATA ANALYTICS

Big data in recent years has been accumulated in severaldomains like health care, public administration, retail, biochemistry,and other interdisciplinary scientific researches.Web-based applications encounter big data frequently, suchas social computing, internet text and documents, and internetsearch indexing. Social computing includes social networkanalysis, online communities, recommender systems,reputation systems, and prediction markets where as internetsearch indexing includes ISI, IEEE Xplorer, Scopus, ThomsonReuters etc. Considering this advantages of big data it providesa new opportunities in the knowledge processing tasks forthe upcoming researchers. However opportunities always followsome challenges.To handle the challenges we need to know various computationalcomplexities, information security, and computationalmethod, to analyze big data. For example, many statisticalmethods that perform well for small data size do not scaleto voluminous data. Similarly, many computational techniquesthat perform well for small data face significant challenges inanalyzing big data. Various challenges that the health sectorface was

being researched by much researchers [9], [10]. Herethe challenges of big data analytics are classified into fourbroad categories namely storage and analysis of data;knowledge discovery computational complexities; visualizationand scalability of data; and information security. We discuss theseissues briefly in the following subsections.

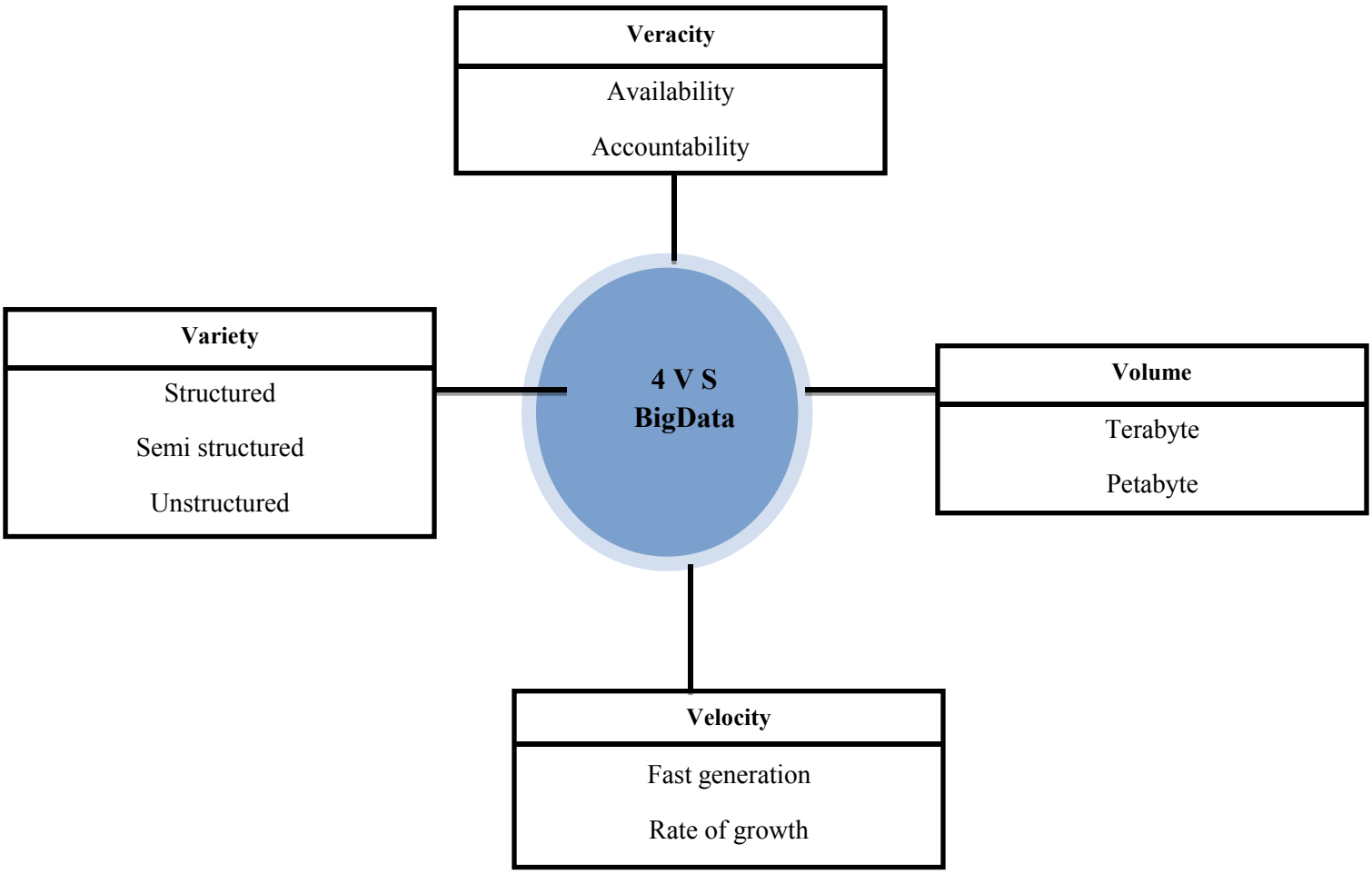

Fig. 1: Characteristics of Big Data[38]

A. Storage and Analysis of Data

In latest years the size of data has grown exponentiallyby severalresources such as aerial sensory technologies, mobile devices, radio frequency identificationreaders, remote sensing etc. These data are stored on spending much costwhereas they ignored or deleted finally because there is noenough space to store them. Therefore, the storage mediums and higher input/output speed is the first challenge forbig data analysis. In such cases, the data accessibility must be on thetop priority for the knowledge discovery and representation.The prime reason is

being that, it must be accessed easily andpromptly for further analysis. In past decades, analyst use harddisk drives to store data but, it slower random input/outputperformance than sequential input/output. To overcome this limitation, the concept of solid state drive (SSD) and phrasechange memory (PCM) was introduced. However the available storage technologies cannot possess the required performancefor processing big data.Another challenge with Big Data analysis is attributedto diversity of data. With the ever growing of datasets, datamining tasks has significantly increased. Additionallydata selection, feature selection, datareduction is an essential taskespecially when dealing with large datasets. This presents anunprecedented challenge for researchers. It is because, existingalgorithms may not always respond in an adequate time whendealing with these high dimensional data. Automation of thisprocess and developing new machine learning algorithms toensure consistency is a major challenge in recent years. Inaddition to all these Clustering of large datasets that helpin analyzing the big data is of prime concern [11]. Recenttechnologies such as hadoop and mapReduce make it possibleto collect large amount of semi structured and unstructureddata in a reasonable amount of time. The key engineeringchallenge is how to effectively analyze these data for obtainingbetter knowledge. A standard process to this end is to transformthe semi structured or unstructured data into structured data,and then apply data mining algorithms to extract knowledge. Aframework to analyze data was discussed by Das and Kumar[12]. Das et al in their paper [13] was discussed the similarly detail explanation of data analysis for public. The major challenge in this case is to pay more attention fordesigning storage systems and to elevate efficient data analysistool that provide guarantees on the output when the datacomes from different sources. Furthermore, design of machinelearning algorithms to analyze data is essential for improvingefficiency and scalability.

B. Knowledge Discovery and Computational Complexities

The main issue in big data is Knowledge discovery and representation. It includes a number of sub fields such asinformation retrieval, archiving, authentication, management, preservation, and representation. There are several tools forknowledge discovery and representation such as fuzzy set[14], rough set [15], soft set [16], near set [17], formalconcept analysis [18], principal component analysis [19] etc. toname a few. Additionally many hybridized techniques are alsodeveloped to process real life problems. All these techniquesare problem dependent. Further some of these techniques maynot be suitable for large datasets in a sequential computer. Atthe same time some of the techniques has good characteristicsof scalability over parallel computer. Since the size of bigdata keeps increasing exponentially, the available tools maynot be efficient to process these data for obtaining meaningfulinformation. The most popular approach in case of largedataset management is data warehouses and data marts. Datawarehouse is mainly responsible to store data that are sourcedfrom operational systems whereas data mart is based on a

datawarehouse and facilitates analysis.It requires more computational complexities' to Analysis of large. The major issue is to handle uncertainty and inconsistenciespresent in the datasets. In general, systematicmodeling of the computational complexity is used. It may bedifficult to establish a comprehensive mathematical system thatis broadly applicable to Big Data. But a domain specific dataanalytics can be done easily by understanding the particularcomplexities. A series of such development could simulate bigdata analytics for different areas. Much research and surveyhas been carried out in this direction using machine learningtechniques with the least memory requirements. The basicobjective in this research is to minimize computational costprocessing and complexities [20], [21], [22].However, current big data analysis tools have poor performancein handling computational complexities, uncertainty,

C. Scalability and Visualization of Data

Scalability and security are the most important challenge for big data analysis techniques. In the recent decadesresearchers have paid attentions to accelerate data analysis andits speed up processors followed by Moore's Law. For theprevious, it is necessary to developon-line, sampling, and multiresolutionanalysis techniques. Incremental techniques havegood scalability property in the aspect of big data analysis. Asthe data size is scaling much faster than CPU speeds, there is anatural dramatic shift in processor technology being embeddedwith increasing number of cores [23]. This shift in processorsleads to the development of parallel computing. Real timeapplications like navigation, social networks, finance, internetsearch, timeliness etc. requires parallel computing.The objective of visualizing data is to present them moreadequately using some techniques of graph theory. Graphicalvisualization provides the link between data with proper interpretation.However, online marketplace like flipkart, amazon,e-bay have millions of users and billions of goods to sold eachmonth. This generates a lot of data. To this end, some companyuses a tool Tableau for big data visualization. It has capabilityto transform large and complex data into intuitive pictures. These help employees of a company to monitor latest customer feedback,visualize search relevance,and their sentiment analysis.However, current big data visualization tools mostly have poorperformances in functionalities, scalability, and response intime.We can observe that big data have produced many challengesfor the developments of the hardware and softwarewhich leads tocloud computing, scalability, parallel computing, distributedcomputing, visualization process. To overcomethis issue, we need to correlate more mathematicalmodels to computer science.

D. Information Security

Thehuge amount of data in big data analysis are correlated, analyzed, and mined for meaningful patterns. All organizationshave different policies to safe guard their sensitive information.Themain issue in big dataanalysis is preserving sensitive information. There is a huge security risk associated with big data[24]. Therefore, information security is becoming a big dataanalytics problem. Big data security can be improved byusing the techniques ofauthorization, authentication, and encryption.Various security

measures that big data applicationsface are scale of network, variety of different devices, real timesecurity monitoring, and lack of intrusion system [25], [26].The challenge of security caused by big data has attracted theattention of information security. Therefore, attention has tobe given to develop a multi-level security policy model andprevention system.Although much research has been carried out to securebig data [25] but it requires lot of improvement. The majorchallenge is to develop a multi-level security, privacy preserveddata model for big data.

3. RESEARCH ISSUES IN BIG DATA ANALYTICS

Data science and big data analytics are becoming the focal point of research inacademia and industries. Data scienceaims at researching big data and knowledge extraction fromdata. Big data applications and data science include informationscience, uncertainty modeling, uncertain data analysis,machine learning, statistical learning, pattern recognition, datawarehousing, and signal processing. Effective integration oftechnologies and analysis will result in predicting the futuredrift of events. Main focus of this section is to discuss openresearch issues in big data analytics. The research issuespertaining to big data analysis are classified into three broadcategories namely internet of things (IoT), cloud computing,bio inspired computing, and quantum computing. However itis not limited to these issues. More research issues related tohealth care big data can be found in HusingKuo et al. paper[9].

A. IoT for Big Data Analytics

Internet has rearranged the art ofbusinesses,global interrelations, cultural revolutions and an unbelievable numberof personal characteristics. Currently, machines are getting inon the act to control innumerable autonomous gadgets viainternet and create Internet of Things (IoT). Therefore, the user of the internet is becoming appliances, just like humans withthe web browsers. Internet of Things is attracting the attentionof recent researchers for its most promising opportunitiesand challenges. It hassocietalimpactand imperative economic for the future construction of information, communication technology and network. The new regulation of future willbe eventually; everything will be connected and intelligentlycontrolled. The concept of IoT is becoming more pertinentto the realistic world due to the development of mobile devices,embedded and ubiquitous communication technologies,cloud computing, and data analytics. Moreover, IoTpresentschallenges in combinations of volume, velocity and variety.In a broader sense, just like the internet, Internet of Thingsenables the devices to exist in a myriad of places and facilitatesapplications ranging from trivial to the crucial. Conversely, it isstill mystifying to understand IoT well, including definitions,content and differences from other similar concepts. Severaldiversified technologies such as computational intelligence,and big-data can be incorporated together to improve thedata management and knowledge discovery of large scaleautomation applications. Much research in

this direction hasbeen carried out by Mishra, Lin and Chang [27].Knowledge acquisition from IoT data is the biggest challengethat big data professional are facing. Thus, it isimportant to develop infrastructure to analyze the IoT data. AnIoT device generates continuous streams of data and the researcherscan develop tools to extract meaningful informationfrom these data using machine learning techniques. Understandingthese streams of data generated from IoT devices andanalyzing them to get meaningful information is a challengingissue and it leads to big data analytics. Machine learningalgorithms and computational intelligence techniques is theonly solution to handle big data from IoT prospective. Keytechnologies that are associated with IoT are also discussed inmany research papers [28]. Figure 2 depicts an overview ofIoT big data and knowledge discovery process.

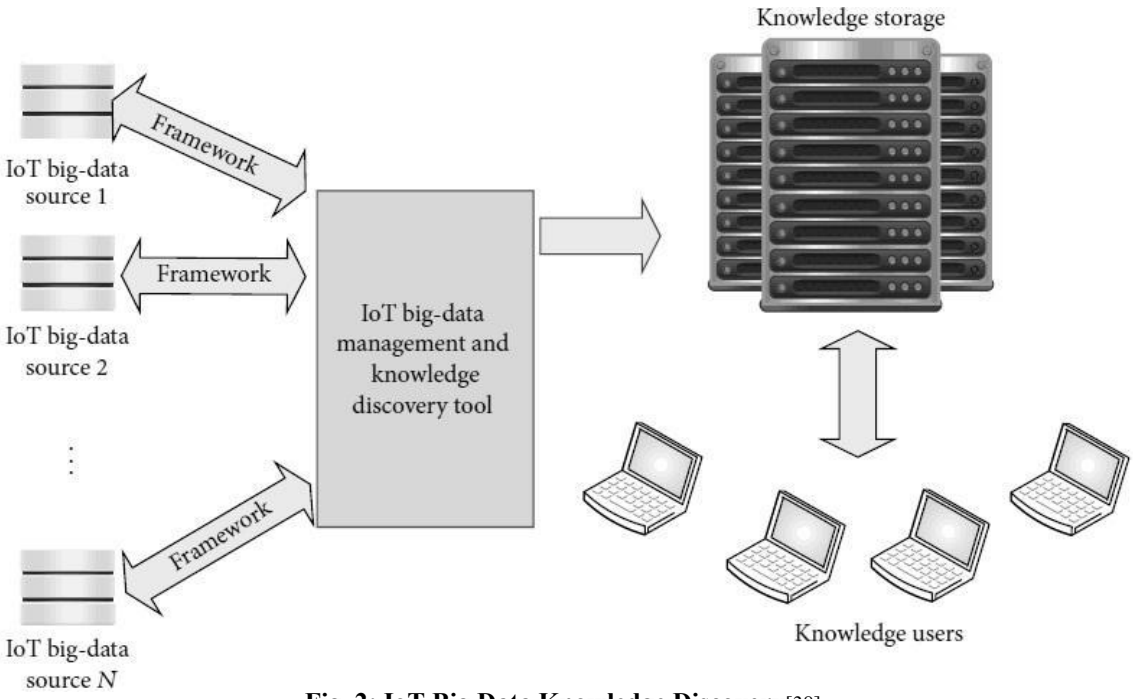

Fig. 2: IoT Big Data Knowledge Discovery[38]

Knowledge exploration system have originated from theoriesof human information processing such as frames, rules,tagging, and semantic networks. In general, it consists offour segments such as knowledge acquisition, knowledgebase, knowledge dissemination, and knowledge application.In phase of knowledge acquisition usedvarious traditional and computational intelligencetechniques to discover theknowledge is discoveredby using various traditional and computational intelligencetechniques. The discovered knowledge is stored in knowledgebases and expert systems are generally designed based on thediscovered knowledge. Knowledge dissemination is importantfor obtaining meaningful information

from the knowledgebase. Knowledge extraction is a process that searches documents,knowledge within documents as well as knowledgebases. The final phase is to apply discovered knowledge invarious applications. It is the ultimate goal of knowledgediscovery. The knowledge exploration system is necessarilyiterative with the judgment of knowledge application. Thereare many issues, discussions, and researches in this area ofknowledge exploration. It is beyond scope of this surveypaper. For better visualization, knowledge exploration systemis depicted in Figure 3.

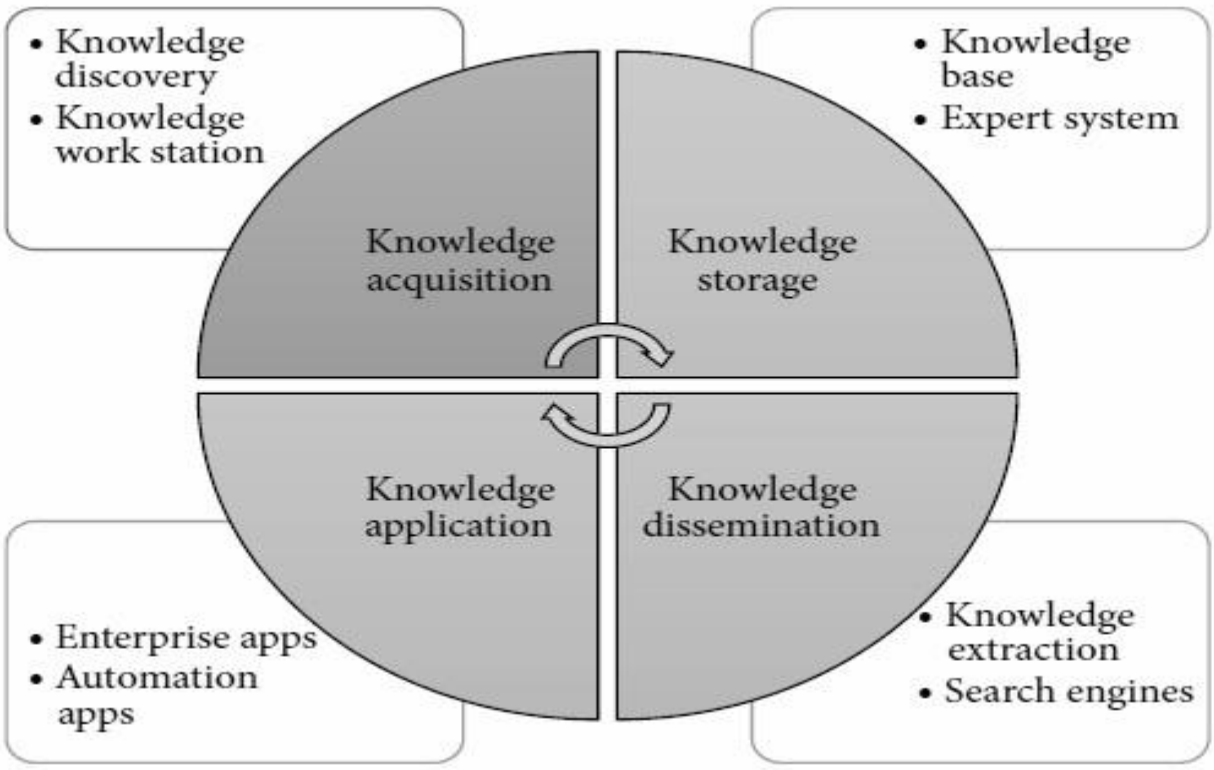

Fig. 3: IoT Knowledge Exploration System[38]

B. Cloud Computing for Big Data Analytics

Virtualization technologies development have madesupercomputing more accessible and affordable. Computing infrastructures that are hidden in virtualization software makesystems to behave like a true computer, but with the flexibilityof specification details such as number of processors, diskspace, memory, and operating system. The use of these virtualcomputers is known as cloud computing which has been oneof the most robust big data techniques. Big Data and cloudcomputing technologies are developed with the importance ofdeveloping a scalable and on demand availability of resource and data.

Cloud computing harmonize massive data by on demandaccess to configurable computing resources throughvirtualization techniques. The benefits of utilizing the Cloudcomputing include offering resources when there is a demandand pay only for the resources which is needed to developthe product. Simultaneously, it improves availability and costreduction. Open challenges and research issues of big dataand cloud computing are discussed in detail by many researcherswhich highlights the challenges in data management,data variety and velocity, data storage, data processing, andresource management [29], [30]. So Cloud computing helpsin developing a business model for all varieties of applicationswith infrastructure and tools.Application ofbig data using cloud computing should supportdata development and analytic. The cloud environment shouldprovide tools that allow data scientists and business analysts tointeractively and collaboratively explore knowledge acquisitiondata for further processing and extracting fruitful results.This can help to solve large applications that may arise invarious domains. In addition to this, cloud computing shouldalso enable scaling of tools from virtual technologies intonew technologies like spark, R, and other types of big dataprocessing techniques.Big data forms a framework for discussing cloud computingoptions. Depending on special need, user can go to themarketplace and buy infrastructure services from cloud serviceproviders such as Google, Amazon, IBM, software as a service(SaaS) from a whole crew of companies such as NetSuite,Cloud9, Jobscience etc. Another advantage of cloud computingis cloud storage which provides a possible way for storing bigdata. The obvious one is the time and cost that are needed toupload and download big data in the cloud environment. Else,it becomes difficult to control the distribution of computationand the underlying hardware. But, the major issues are privacyconcerns relating to the hosting of data on public servers,and the storage of data from human studies. All these issueswill take cloud computing and big data to a high level ofdevelopment.

4. BIG DATA PROCESSING TOOLS

Big 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 MapReduce, 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 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 platforms that support interactive analysis. These tools help us in developing the big data projects. Aextraordinary list of big data tools and techniques is also discussed by much researchers [6], [34]. The typical work flow of big data project discussed by Huang et al is highlighted in this section [35] and is depictedin Figure 4.

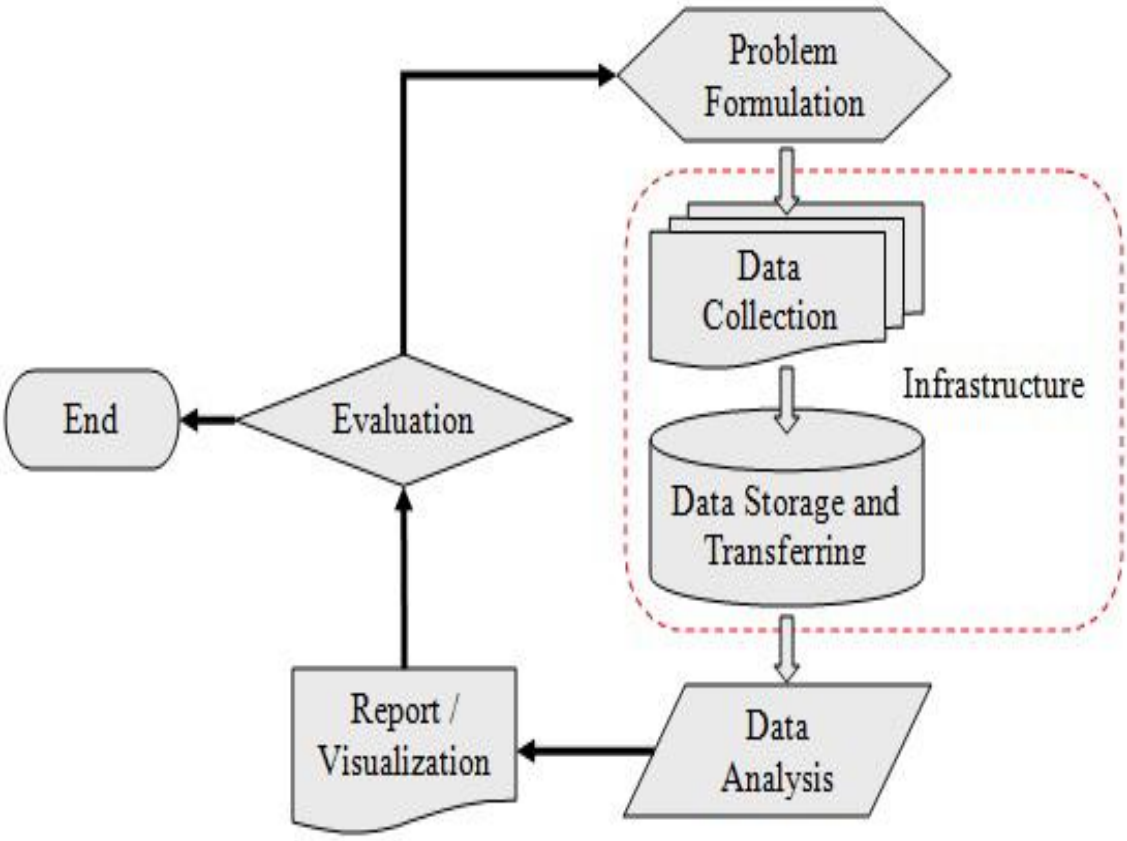

Fig. 4: Workflow of Big Data Project[38]

A. Apache Hadoop and MapReduce

The most recognized software platform for big data analysisis Apache Mapreduce and Hadoop. It involves of hadoopkernel, mapreduce, hadoop distributed file system (HDFS)and apache hive etc. Map reduce is a programming modelfor processing large datasets is based on divide and conquermethod. The method of divide and conquer is implemented in twosteps such as Map step and Reduce Step. Hadoop works ontwo kinds of nodes such as master node and worker node. Themaster node divides the input into smaller sub problems andthen distributes them to worker nodes in map step. Thereafterthe master

node combines the outputs for all the sub problemsin reduce step. Moreover, Hadoop and MapReducework asa powerful software framework for solving big data problems.It is also helpful in fault-tolerant storage and high throughputdata processing.

B. Apache Mahout

The goal of Apache mahout is to provide scalable and commercialmachine learning techniques for large scale and intelligent dataanalysis applications. Core algorithms of mahout includingclustering, classification, pattern mining, regression, dimensionalityreduction, evolutionary algorithms, and batch basedcollaborative filtering run on top of Hadoop platform throughmap reduce framework. The Mahout aims to build aresponsive, vibrant, diverse community to facilitate discussionson the project and potential use cases. The basic objectiveof Apache mahout is to provide a tool for alleviating bigchallenges. The different companies those who have implementedscalable machine learning algorithms are Google, IBM,Amazon, Yahoo, Twitter, and facebook [36].

C. Apache Spark

Apache spark is an open source big data processing frameworkbuilt for speed processing, and sophisticated analytics. It is easy to use and was originally developed in 2009 in UCBerkeleysAMPLab. It was open sourced in 2010 as an Apacheproject. By using Spark you can quickly write applications in java, scala,or python. In addition to map reduce operations, it supportsSQL queries, streaming data, machine learning, and graph dataprocessing. Spark runs on top of existing hadoop distributedfile system (HDFS) infrastructure to provide enhanced andadditional functionality. Spark consists of components namelydriver program, cluster manager and worker nodes. The driverprogram serves as the starting point of execution of an applicationon the spark cluster. The cluster manager allocates theresources and the worker nodes to do the data processing inthe form of tasks. Each application will have a set of processescalled executors that are responsible for executing the tasks.The major advantage is that it provides support for deployingspark applications in an existing hadoop clusters. Figure 5depicts the architecture diagram of Apache Spark. The variousfeatures of Apache Spark are listed below:

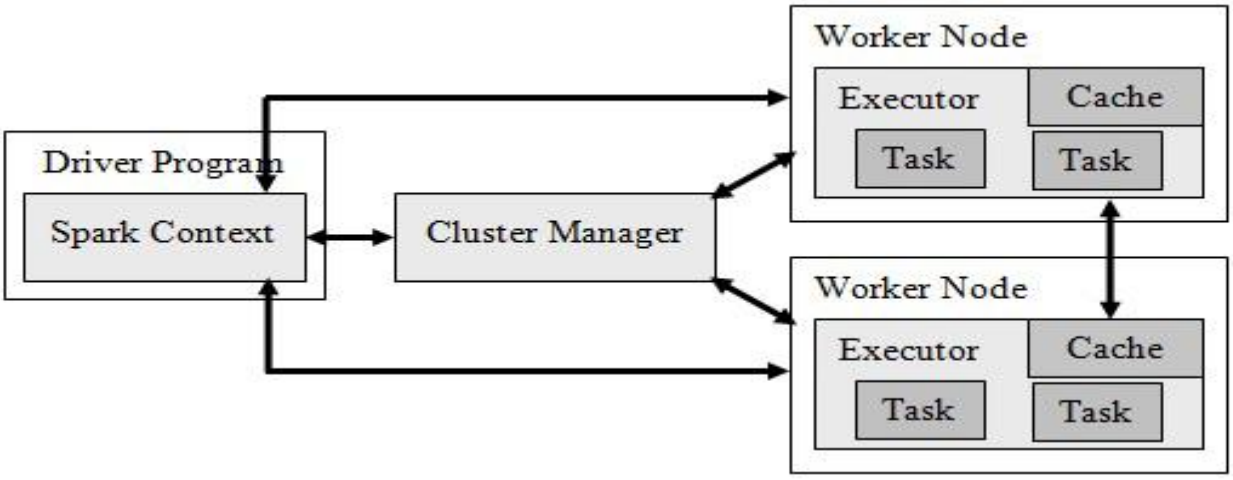

Fig. 5: Architecture of Apache Spark[38]

- 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 resourceutilization.
- The foremost advantage is that in addition to MapReduce, 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 andcan be seamlessly combined to create complex workflows.
- Spark helps to run an application in Hadoop 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 todisk.
- It is written in scala programming language and runs on java virtual machine (JVM) environment. Additionally, it supportsjava, python and R for developingapplications using Spark.

5. SUGGESTIONS FOR FUTURE WORK

The amount of data collected from various applicationsall over the world across a wide variety of fields today isexpected to double every two years. It has no utility unlessthese are analyzed to get useful information. This necessitatesthe development of techniques which can be used to facilitatebig data analysis. The development of powerful computers isa boon to implement these techniques leading to automatedsystems. The transformation of data into knowledge is byno means an easy task for high performance large-scale dataprocessing,

including exploiting parallelism of current andupcoming 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 obtainedby combining two or more of these models have beenfound to be fruitful in representing data. These models arealso very much fruitful for analysis. More often than not, bigdata are reduced to include only the important characteristicsnecessary from a particular study point of view or dependingupon the application area. So, reduction techniques have beendeveloped. Often the data collected have missing values. Thesevalues need to be generated or the tuples having these missingvalues are eliminated from the data set before analysis. Moreimportantly, these new challenges may comprise, sometimeseven deteriorate, the performance, efficiency and scalabilityof the dedicated data intensive computing systems. The laterapproach sometimes leads to loss of information and hencenot preferred. This brings up many research issues in theindustry and research community in forms of capturing andaccessing data effectively. In addition, fast processing whileachieving high performance and high throughput, and storingit efficiently for future use is another issue. Further, programmingfor big data analysis is an important challengingissue. Expressing data access requirements of applicationsand designing programming language abstractions to exploitparallelism are an immediate need [38].Additionally, machine learning concepts and tools aregaining popularity among researchers to facilitate meaningfulresults from these concepts. Research in the area of machinelearning for big data has focused on data processing, algorithmimplementation, and optimization. Many of the machinelearning tools for big data are started recently needs drasticchange to adopt it. We argue that while each of the tools hastheir advantages and limitations, more efficient tools can bedeveloped for dealing with problems inherent to big data. Theefficient tools to be developed must have provision to handlenoisy and imbalance data, uncertainty and inconsistency, andmissing values.

6. CONCLUSION

In recent years data are generated at a dramatic pace Analyzing these data is challenging for a general man. Tothis end in this paper, we survey the various research issues,challenges, and tools used to analyze these big data. Fromthis survey, it is understood that every big data platform hasits individual focus. Some of them are designed for batchprocessing whereas some are good at real-time analytic. Eachbig data platform also has specific functionality. Differenttechniques used for the analysis include statistical analysis,machine learning, data mining, intelligent analysis, cloud computing,quantum computing, and data stream processing. Webelieve that in future researchers will pay more attention tothese techniques to solve problems of big data effectively andefficiently.

REFERENCES

[1] M. K.Kakhani, S. Kakhani and S. R.Biradar, Research issues in bigdata analytics, International Journal of Application or Innovation inEngineering & Management, 2(8) (2015), pp.228-232.
[2] A. Gandomi and M. Haider, Beyond the hype: Big data concepts, methods,and analytics, International Journal of Information Management,35(2) (2015), pp.137-144.
[3] C. Lynch, Big data: How do your data grow?, Nature, 455 (2008),pp.28-29.
[4] X. Jin, B. W.Wah, X. Cheng and Y. Wang, Significance and challengesof big data research, Big Data Research, 2(2) (2015), pp.59-64.
[5] R. Kitchin, Big Data, new epistemologies and paradigm shifts, BigData Society, 1(1) (2014), pp.1-12.
[6] C. L. Philip, Q. Chen and C. Y. Zhang, Data-intensive applications,challenges, techniques and technologies: A survey on big data, InformationSciences, 275 (2014), pp.314-347.
[7] K. Kambatla, G. Kollias, V. Kumar and A. Gram, Trends in big dataanalytics, Journal of Parallel and Distributed Computing, 74(7) (2014),pp.2561-2573.
[8] S. Del. Rio, V. Lopez, J. M. Bentez and F. Herrera, On the use ofmapreduce for imbalanced big data using random forest, InformationSciences, 285 (2014), pp.112-137.
[9] MH. Kuo, T. Sahama, A. W. Kushniruk, E. M. Borycki and D. K.Grunwell, Health big data analytics: current perspectives, challengesand potential solutions, International Journal of Big Data Intelligence,1 (2014), pp.114-126.
[10] R. Nambiar, A. Sethi, R. Bhardwaj and R. Vargheese, A look atchallenges and opportunities of big data analytics in healthcare, IEEEInternational Conference on Big Data, 2013, pp.17-22.
[11] Z. Huang, A fast clustering algorithm to cluster very large categoricaldata sets in data mining, SIGMOD Workshop on Research Issues onData Mining and Knowledge Discovery, 1997.[12] T. K. Das and P. M. Kumar, Big data analytics: A framework forunstructured data analysis, International Journal of Engineering andTechnology, 5(1) (2013), pp.153-156.
[13] T. K. Das, D. P. Acharjya and M. R. Patra, Opinion mining about aproduct by analyzing public tweets in twitter, International Conferenceon Computer Communication and Informatics, 2014.
[14] L. A. Zadeh, Fuzzy sets, Information and Control, 8 (1965), pp.338-353.
[15] Z. Pawlak, Rough sets, International Journal of Computer InformationScience, 11 (1982), pp.341-356.
[16] D. Molodtsov, Soft set theory first results, Computers and Mathematicswith Aplications, 37(4/5) (1999), pp.19-31.
[17] J. F.Peters, Near sets. General theory about nearness of objects,Applied Mathematical Sciences, 1(53) (2007), pp.2609-2629.
[18] R. Wille, Formal concept analysis as mathematical theory of conceptand concept hierarchies, Lecture Notes in Artificial Intelligence, 3626(2005), pp.1-33.
[19] I. T.Jolliffe, Principal Component Analysis, Springer, New York, 2002.[20] O. Y. Al-Jarrah, P. D. Yoo, S. Muhaidat, G. K. Karagiannidis andK. Taha, Efficient machine learning for big data: A review, Big DataResearch, 2(3) (2015), pp.87-93.
[21] Changwon. Y, Luis. Ramirez and Juan.Liuzzi, Big data analysisusing modern statistical and machine learning methods in medicine,International Neurourology Journal, 18 (2014), pp.50-57.
[22] P. Singh and B. Suri, Quality assessment of data using statistical andmachine learning methods. L. C.Jain, H. S.Behera, J. K.Mandal andD. P.Mohapatra (eds.), Computational Intelligence in Data Mining, 2(2014), pp. 89-97.
[23] A. Jacobs, The pathologies of big data, Communications of the ACM,52(8) (2009), pp.36-44.
[24] H. Zhu, Z. Xu and Y. Huang, Research on the security technology of bigdata information, International Conference on Information Technologyand Management Innovation, 2015, pp.1041-1044.
[25] Z. Hongjun, H. Wenning, H. Dengchao and M. Yuxing, Survey ofresearch on information security in big data, Congresso da sociedadaBrasileira de Computacao, 2014, pp.1-6.
[26] I. Merelli, H. Perez-sanchez, S. Gesing and D. D.Agostino, Managing,analysing, and integrating big data in medical bioinformatics: openproblems and future perspectives, BioMed Research International, 2014,(2014), pp.1-13.
[27] N. Mishra, C. Lin and H. Chang, A cognitive adopted frameworkfor iot big data management and knowledge discovery prospective,International Journal of Distributed Sensor Networks, 2015, (2015), pp.1-13
[28] X. Y.Chen and Z. G.Jin, Research on key technology and applicationsfor internet of things, Physics Procedia, 33, (2012), pp. 561-566.
[29] M. D. Assuno, R. N. Calheiros, S. Bianchi, M. a. S. Netto and R. Buyya,Big data computing and clouds: Trends and future directions, Journalof Parallel and Distributed Computing, 79 (2015), pp.3-15.
[30] I. A. T. Hashem, I. Yaqoob, N. BadrulAnuar, S. Mokhtar, A. Gani andS. Ullah Khan, The rise of big data on cloud computing: Review andopen research issues, Information Systems, 47 (2014), pp. 98-115.
[31] L. Wang and J. Shen, Bioinspired cost-effective access to big data,International Symposium for Next Generation Infrastructure, 2013, pp.1-7.
[32] C. Shi, Y. Shi, Q. Qin and R. Bai Swarm intelligence in big dataanalytics, H. Yin, K. Tang, Y. Gao, F. Klawonn, M. Lee, T. Weise,B. Li and X. Yao (eds.), Intelligent Data Engineering and AutomatedLearning, 2013, pp.417-426.
[33] M. A. Nielsen and I. L.Chuang, Quantum Computation and QuantumInformation, Cambridge University Press, New York, USA 2000.
[34] M. Herland, T. M. Khoshgoftaar and R. Wald, A review of data miningusing big data in health informatics, Journal of Big Data, 1(2) (2014),pp. 1-35.
[35] T. Huang, L. Lan, X. Fang, P. An, J. Min and F. WangPromises andchallenges of big data computing in health sciences, Big Data Research,2(1) (2015), pp. 2-11.
[36] G. Ingersoll, Introducing apache mahout: Scalable, commercial friendlymachine learning for building intelligent applications, White Paper,IBM Developer Works, (2009), pp. 1-18.
[37] H. Li, G. Fox and J. Qiu, Performance model for parallel matrix multiplicationwith dryad: Dataflow graph runtime, Second InternationalConference on Cloud and Green Computing, 2012, pp.675-683.
[38] D. P. Acharjya ,Kauser Ahmed P , Survey on Big Data Analytics: Challenges, Open Research Issues and Tools, International Journal of Advanced Computer Science and Applications, Vol. 7(2), 2016

GameilSaadHamzh Ali: MSc Computer Science , Ph.D. Research Scholar , Department of Computer Science, RathnavelSubramaniam college of Arts & Science, India

Dr.A. Nithya : MSc ,MPhil ,Ph.D. , Research Supervisor, Department of Computer Science, RathnavelSubramaniam collegeof Arts &Science,India