Open Research Issues, Tools andChallenges in Big Data Analytics

Prof. Prachi N. Deshmukh Prof. Rasika S. Badre Prof. Shruti G. Taley

***Abstract*- A huge storage area of terabytes of data is generated each day from modern information systems and digitaltechnologies such as Internet of Things and cloud computing. Study of these massive data requires a lot ofefforts at multiple levels to extract knowledge for decision making. Therefore, big data analysis is a current areaof study and improvement. The basic goal of this paper is to explore the possible impact of big data challenges,open research issues, and various tools related with it. As a result, this article provides a platform to explore bigdata at various stages. Additionally, it opens a new possibility for researchers to extend the result, based on thechallenges and open research issues.**

***Keywords*-Big data analytics; Hadoop; Massive data; Structured data; Unstructured Data.**

I. INTRODUCTION

In digital world, information are generatedfrom different sources and the rapid transition fromdigital technologies has led to growth of big data. Itprovides evolutionary breakthroughs in many fieldswith gathering of large datasets. In general, it refersto the collection of large and difficult datasetswhich are difficult to process using establisheddatabase management tools or data processingapplications. These are obtainable in structured,semi-structured, and unstructured format inpetabytes and outside. Formally, it is distinct from3Vs to 4Vs. 3Vs refers to volume, velocity, andvariety. Volume refers to the large amount of datathat are being generated everyday whereas velocityis the rate of growth and how fast the data arecollected for being analysis. Variety provideinformation about the types of data such asstructured, unstructured, semi-structured etc. Thefourth V refers to veracity that includesaccessibility and responsibility. The prime goal ofbig data analysis is to method data of high volume,velocity, variety, and veracity using various usualand computational intelligent techniques [11].Generally, Data warehouses have been used tocontrol the big dataset. In this case extracting theexact knowledge from the accessible big data is aforemost issue. The key problem in the analysis ofbig data is the need of organization betweendatabase systems as well as with analysis toolssuch as data mining and statistical analysis. Thesechallenges generally occur when we wish toexecute knowledge discovery and demonstrationfor its practical applications. Additionally, thestudy on difficulty theory of big data will helpunderstand required characteristics and formationof difficult patterns in big data, simplify itsrepresentation, gets better knowledge abstraction,and guide the design of computing models andare interested in disseminating the findings of bigdata. This paper focuses on challenges in big dataand its available techniques.

.

II. OPEN RESEARCH ISSUES IN BIG DATA ANALYTICS

Big data analytics and data science arebecoming the research main point in industries andacademia. Data science aims at researching bigdata and knowledge extraction from data.Applications of big data and data science containinformation science, uncertainty modeling,uncertain data analysis, machine learning,statistical learning, pattern recognition, data
warehousing, and signal processing. Main focus ofthis section is to discuss open research issues in bigdata analytics. The research issues pertaining to bigdata analysis are classify into three big categoriesnamely internet of things (IoT), cloud computing,bio inspired computing, and quantum computing.

A. IoT for Big Data Analytics

Internet has simplified globalinterrelations, the ability of businesses, culturalrevolutions and an incredible number of personalcharacteristics. Currently, machines are getting inon the act to control incalculable autonomousgadgets via internet and create Internet of Things(IoT). Thus, appliances are appropriate the user ofthe internet, just like humans with the webbrowsers. Internet of Things is attracting the noticeof recent researchers for its most capableopportunities and

challenges. It has an essentialeconomic and societal impact for the futureconstruction of information, network andcommunication technology. The new regulation offuture will be ultimately, everything will beconnected and intelligently restricted. The idea ofIoT is becoming more important to the practicalworld due to the improvement of mobile devices,fixed and ubiquitous communication technologies,cloud computing, and data analytics.. In a broadersense, just like the internet, Internet of Thingsenables the devices to exist in a myriad of placesand facilitates applications ranging from minor tothe crucial. A number of diversified technologiessuch as computational intelligence, and big-datacan be incorporated simultaneously to improve thedata management and knowledge discovery oflarge scale automation applications. Much researchin this direction has been carried out by Mishra,Lin and Chang [12].Knowledge achievement from IoT data isthe biggest challenge that big data professional arefacing. Therefore, it is essential to developcommunications to analyze the IoT data. An IoTdevice generates constant streams of data and theresearchers can develop tools to extract meaningfulinformation from these data using machine learningtechniques. Accepting these streams of datagenerated from IoT devices and analyzing them toacquire important information is a challengingissue and it leads to big data analytics. Machinelearning algorithms and computational intelligencetechniques is the only clarification to handle bigdata from IoT prospective. Key technologies thatare related with IoT are also discussed in manyresearch papers [17]. Figure 1.depicts an overviewof IoT big data and knowledge discovery process.

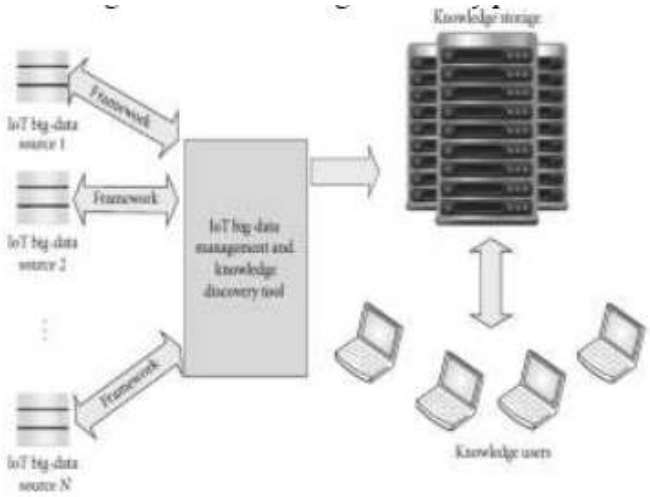

Fig. 1: IoT Big Data Knowledge Discovery

B. Cloud Computing for Big Data Analytics

The improvement of virtualizationtechnologies have complete supercomputing moreaccessible and affordable. Computinginfrastructures that are unknown in virtualizationsoftware make systems to act like a true computer,but with the flexibility of specification data such asnumber of processors, disk space, memory, andoperating system. The use of these virtualcomputers is identified as cloud computing whichhas been one of the most robust big data technique.Big Data and cloud computing technologies areresidential with the importance of developing ascalable and on demand accessibility of resourcesand data. Cloud computing harmonize massive databy on demand contact to configurable computingresources through virtualization techniques. Theprofit of utilizing the Cloud computing includeoffering resources when there is a demand and payonly for the resources which is needed to developthe product. Concurrently, it improves availabilityand cost reduction. So Cloud computing helps indeveloping a business model for all varieties ofapplications with infrastructure and tools.Big data application using cloud computingshould maintain data analytic and development.The cloud environment should propose tools thatallow data scientists and business analysts tointeractively and collaboratively exploreknowledge acquisition data for further processingand extracting fruitful outcome. This can assist tosolve large applications that may arise in variousdomains. In addition to this, cloud computingshould also allow scaling of tools from virtualtechnologies into new technologies similar tospark, R, and other types of big data processingtechniques.

C. Quantum Computing for Big Data Analysis

A quantum computer has memory that isexponentially bigger than its physical size and canmanage an exponential set of inputs simultaneously[10]. This exponential improvement in computersystems might be achievable. If a actual quantumcomputer is accessible now, it could have solvedproblems that are exceptionally difficult on recentcomputers, of course today's large data problems.The main technical difficulty in building quantumcomputer could soon be possible. Quantumcomputing provides a way to join the quantummechanics to method the information. In traditionalcomputer, information is obtainable by long stringsof bits which instruct either a zero or a one. On theother hand a quantum computer uses quantum bitsor qubits. The difference between qubit and bit isthat, a qubit is a quantum system that encodes thezero and the one into two discernible quantumstates. Therefore, it can be capitalized on thephenomena of superposition and embarrassment. Itis because qubits perform quantumly. For example,100 qubits in

quantum systems need to 2100complex values to be stored in a common computersystem. It means that many big data problems canbe solve much more rapidly by superior scalequantum computers compare with typicalcomputers. Hence it is a challenge for thisinvention to built a quantum computer and assistquantum computing to explain big data problems.

III. TOOLS FOR BIG DATA PROCESSING

Large numbers of tools are easy to get toprocess big data. In this section, we discuss somecurrent techniques for analyzing big data withconsequence on three important promising toolsnamely MapReduce, Apache Mahout, and Dryad.

A. Apache Hadoop and MapReduce

The most well-known software stage for big dataanalysis is Apache Hadoop and Mapreduce. Itconsists of hadoop kernel, mapreduce, hadoopdistributed file system (HDFS) and apache hive etc.Map reduce is a programming model for dealingout huge datasets is based on divide and conquermethod. The divide and conquer method isimplemented in two steps such as Map step andReduce Step. Hadoop works on two kinds of nodessuch as master node and worker node. The masternode divides the input into smaller sub problemsand then distributes them to worker nodes in mapstep. Thereafter the master node join the outputs forall the subproblems in reduce step. Moreover,Hadoop and MapReduce ability as a well-builtsoftware support for solving big data problems. It isalso cooperative in fault-tolerant storage and highthroughput data giving out.

B. Apache Mahout

Apache mahout goal to give scalable andcommercial machine learning techniques for largescale and intelligent data analysis applications.Core algorithms of mahout including clustering,classification, pattern mining, regression,dimensionalty reduction, evolutionary algorithms,and batch based mutual filtering execute on top ofHadoop step through map decrease framework. Thegoal of mahout is to build a vibrant, reactive,diverse community to facilitate discussions on theproject and potential use cases. The basic objectiveof Apache mahout is to provide a tool for enrichingbig challenges. The different companies those whohave implemented scalable machine learningalgorithms are Google, IBM, Amazon, Yahoo,Twitter, and facebook [4].

C. Dryad

It is any more popular programmingmodel for implementing parallel and distributedprograms for handling large situation bases ondataflow graph. It consists of a cluster ofcomputing nodes, and an user use the devices of acomputer cluster to run their program in adistributed way. Indeed, a dryad user use thousandsof machines, each of them with several processorsor cores. The main benefit is that users do not needto know anything about correspondingprogramming. A dryad application runs acomputational directed graph that is composed ofcomputational vertices and communicationchannels. Therefore, dryad provides a large numberof functionality including generating of workgraph, scheduling of the machines for the availableprocesses, evolution failure handling in the cluster,set of performance metrics, visualizing the work,invoking user defined policies and dynamicallyupdating the work graph in response to thesestrategy decisions without knowing the semanticsof the vertices [5].

IV. CHALLENGES IN BIG DATA ANALYTICS

Here the challenges of big data analyticsare classify into four large categories namely datastorage space and analysis; knowledge discoveryand computational complexities; scalability andvisualization of data; and information security. Wediscuss these issues for a short time in thefollowing subsections.

A. Data Storage and Analysis

In current years the size of informationhas grown exponentially by various means such asmobile devices, aerial sensory technologies, remotesensing, radio frequency identification readers etc.Thus, the first test for big data analysis is storagespace mediums and high input/output speed. Insuch cases, the data accessibility must be on the topmain concern for the knowledge discovery anddemonstration. The prime reason is being that, itmust be accessed easily and rapidly for furtherstudy. In past decades, analyst use hard disk drivesto accumulate data but, it slower casualinput/output performance than sequentialinput/output. To overcome this limitation, the ideaof solid state drive (SSD) and phrase changememory (PCM) was introduced. However theavailable storage technologies cannot grasp therequired performance for processing big data.Data reduction, data selection, featureselection is a crucial job particularly when dealingwith large datasets. This presents an exceptionalchallenge for researchers. It is because, existingalgorithms may not always respond in an sufficienttime when dealing with these high dimensionaldata. Automation of this technique and developingnew machine learning algorithms to make sureconsistency is a key challenge in recent years.

B. Knowledge Discovery and ComputationalComplexities

Knowledge discovery and demonstrationis a major concern in big data. It includes a numberof sub fields such as authentication, archiving,management, preservation, in order retrieval, andrepresentation. There are numerous tools forknowledge discovery and demonstration such asfuzzy set [9], rough set [19], soft set [3], near set[8], formal model analysis [15], main componentanalysis [7] etc to name a few. Furthermore a lot ofhybridized techniques are also developed toprocess real life problems. All these techniques aretrouble dependent. Further some of thesetechniques may not be suitable for big datasets in asequential computer. At the same time some of thetechniques has good quality characteristics of scalability over similar computer. As the size of bigdata keeps growing exponentially, the existingtools may not be well-organized to method thesedata for obtaining significant information. Themost admired approach in case of big dataset management is data warehouses and data marts.Data warehouse is mostly dependable toaccumulate data that are sourced from operationalsystems while data mart is based on a datawarehouse and facilitates analysis.Analysis of huge dataset requires morecomputational complexities. The mainapprehension is to handle inconsistencies anduncertainty current in the datasets. In general,systematic modeling of the computationalcomplexity is used. It may be difficult to set up acomprehensive mathematical system that is broadlyrelated to Big Data. But a domain specific dataanalytics can be done simply by understanding theparticular complexities. A series of suchdevelopment could suggest big data analytics fordifferent areas. Much research and survey has beencarried out in this direction using machine learningtechniques with the least memory requirements.The basic purpose in these research is to minimizecomputational cost giving out and complexity [13],[2], [14].

C. Scalability and Visualization of Data

The mainly central challenge for big dataanalysis techniques is its scalability and security. Itis necessary to grow sampling, on-line, andmultiresolution analysis techniques. Incrementaltechniques have excellent scalability property inthe side of big data analysis. As the informationsize is scaling much quicker than CPU speeds,there is a natural dramatic shift in processor technology being fixed with growing number ofcores [1]. This move in processors leads to thegrowth of parallel computing. The reason ofvisualizing data is to current them additionaleffectively using some techniques of graph theory.Graphical visualization provides the link betweendata with proper explanation. However, onlinemarketplace like flipkart, amazon, e-bay havemillions of users and billions of generate to soldeach month. This generates a lot of information.

D. Information Security

In big data analysis massive sum of dataare connected, analyzed, and mined for meaningfulpatterns. Preserving responsive information is amost important issue in big data analysis [6].Therefore, information protection is becoming abig data analytics problem. Security of big data canbe improved by using the techniques ofauthentication, authorization, andencryption.Different protection measures that bigdata applications appearance are scale of network,variety of different devices, real time securitymonitoring, and need of interruption system [18],[26]. The security challenge caused by big data hasattracted the attention of information security.Therefore, interest has to be particular to develop amulti level protection rule model and avoidancesystem.

V. CONCLUSION

In current years information are generatedat a amazing pace. Analyzing these information ischallenging for a general man. To this conclusionin this paper, we study the various research issues,challenges, and tools worn to analyze these bigdata. From this study, it is appreciate that every bigdata stage has its individual focus. Some of themare intended for batch processing whereas some arehigh-quality at real-time analytic. Each big datastage also has exact functionality. Differenttechniques available for the analysis includestatistical analysis, machine learning, data mining,intelligent analysis, cloud computing, quantumcomputing, and data stream processing. We regardas in prospect researchers will give more focus tothese techniques to explain troubles of big datasuccessfully and resourcefully.

REFERENCES

1 A. Jacobs, The pathologies of big data, Communications ofthe ACM, 2(8) (2009), pp.36-44.

2 Changwon. Y, Luis. Ramirez and Juan. Liuzzi, Big dataanalysis using modern statistical and machine learning methodsin medicine, International Neurourology Journal, 18 (2014),pp.50-57.

3 D. Molodtsov, Soft set theory first results, Computers andMathematics with Aplications, 37(4/5) (1999), pp.19-31.

4 G. Ingersoll, Introducing apache mahout: Scalable,commercial friendly machine learning for building intelligentapplications, White Paper, IBM Developer Works, (2009), pp.1-18.

5 H. Li, G. Fox and J. Qiu, Performance model for parallelmatrix multiplication with dryad: Dataflow graph runtime,Second International Conference on Cloud and GreenComputing, 2012, pp.675-683.

6 H. Zhu, Z. Xu and Y. Huang, Research on the securitytechnology of big data information, International Conference onInformation Technology and Management Innovation, 2015,pp.1041-1044.

7 I. T.Jolliffe, Principal Component Analysis, Springer, NewYork, 2002.

8 J. F.Peters, Near sets. General theory about nearness ofobjects, Applied Mathematical Sciences, 1(53) (2007), pp.2609-2629.

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

10 M. A. Nielsen and I. L.Chuang, Quantum Computation andQuantum Information, Cambridge University Press, New York,USA 2000.

11 M. K.Kakhani, S. Kakhani and S. R.Biradar, Researchissues in big data analytics, International Journal of Applicationor Innovation in Engineering & Management, 2(8) (2015),pp.228-232.

12 N. Mishra, C. Lin and H. Chang, A cognitive adoptedframework for iot big data management and knowledgediscovery prospective, International Journal of DistributedSensor Networks, 2015, (2015), pp. 1-13

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

14 P. Singh and B. Suri, Quality assessment of data usingstatistical 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.

15 R. Wille, Formal concept analysis as mathematical theoryof concept and concept hierarchies, Lecture Notes in ArtificialIntelligence, 3626 (2005), pp.1-33.

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

17 X. Y.Chen and Z. G.Jin, Research on key technology andapplications for internet of things, Physics Procedia, 33, (2012),pp. 561-566.

18Z. Hongjun, H. Wenning, H. Dengchao and M. Yuxing,Survey of research on information security in big data,Congresso da sociedadaBrasileira de Computacao, 2014, pp.1-6.

19 Z. Pawlak, Rough sets, International Journal of ComputerInformation Science, 11 (1982), pp.341-356.