

Study of the Impact of Big Data on Research and Knowledge Mining for Indian Data Processing Units

Dr Raj Kumar¹, Dr Pankaj Saxena²

Assistant Professor¹, Dean²

Department of Computer Engineering¹, Department of Computer Application²

Agra University¹, RBSMTC, Khandari Campus, Agra²

Corresponding Authors' email id: rajkumar.citd@gmail.com¹

Abstract

The massive repository of terabytes of data is generated each day from modern information systems and digital technologies of our surroundings such as Internet of Things and cloud computing. Analysis of these huge data requires a lot of efforts at multiple levels to extract knowledge for the purpose of decision making to do our daily practices. Therefore, big data analysis is a current area of research and development for the SOP of Indian data processing units like CRA, digi-bank, gene-bank etc. 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 research paper 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: *Big data analytics, Hadoop, Massive data, structured data, Unstructured Data, Drill, and Storm*

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. 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 [1]. Some of these extraction methods for obtaining helpful information were discussed by Gandomi and Haider [2].

Generally, Data warehouses have been used to manage the large dataset. In this case extracting the precise knowledge from the available 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 [5]. 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. 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 phase 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.

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 [23]. 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 flipkart, 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 feed-back, 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.

FOCUS ON RELATED WORKS

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 [27]. Knowledge acquisition from IoT data is the biggest challenge that big data professional are facing. Therefore, it is essential to develop infrastructure to analyze the IoT data. An IoT device generates continuous streams of data and

the 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 [28].

METHODOLOGY

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 [29-30].

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 NetSuite, Cloud9, Jobscience 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. 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.

TOOLS OF BIG DATA ANALYSIS

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 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 allows users to directly interact in real time for their own analysis.

1. Apache Hadoop and MapReduce

The most established software platform for big data analysis is Apache Hadoop and MapReduce. It consists of hadoop kernel, mapreduce, hadoop distributed file system (HDFS) and apache hive etc. Map reduce is a programming model for processing large datasets is based on divide and conquer method. The divide and conquer method is implemented in two steps such as Map step and Reduce Step.

Hadoop works on two kinds of nodes such as master node and worker node. The master node divides the input into smaller sub problems and then distributes them to worker nodes in map step. Thereafter the master node combines the outputs for all the sub-problems in reduce step.

Moreover, Hadoop and MapReduce work as a powerful software framework for solving big data problems. It is also helpful in fault-tolerant storage and high throughput data processing.

2. Apache Spark

It is easy to use and was originally developed in 2009 in UC Berkeleys AMPLab. It was open sourced in 2010 as an Apache project. Spark lets you quickly write applications in java, scala, or python. In addition to map reduce operations, it supports SQL queries, streaming data, machine learning, and graph data processing.

Spark runs on top of existing hadoop distributed file system (HDFS) infrastructure to provide enhanced and additional functionality. Spark consists of components namely driver program, cluster manager and worker nodes.

The driver program serves as the starting point of execution of an application on the spark cluster. The cluster manager allocates the resources and the worker nodes to do the data processing in the form of tasks.

Each application will have a set of processes called executors that are responsible for executing the tasks. The major advantage is that it provides support for deploying spark applications in an existing hadoop clusters.

3. Storm

Storm is a distributed and fault tolerant real time computation system for processing large streaming data. It is specially designed for real time processing in contrasts with hadoop which is for batch processing. Additionally, it is also easy to set up and operate, scalable, fault-tolerant to provide competitive performances.

The storm cluster is apparently similar to hadoop cluster. On storm cluster users run different topologies for different storm tasks whereas hadoop platform implements map reduce jobs for corresponding applications. There are number of differences between map reduce jobs and topologies. The basic 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 starts and terminates 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.

4. 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 metrics for many application, diagnose problems for system and information technology infrastructures, and intelligent support for business operations.

5. Dryad

It is another popular programming model for implementing parallel and distributed programs for handling large context bases on dataflow graph. It consists of a cluster of computing nodes, and a user use the resources of a computer cluster to run their program in a distributed way.

Indeed in dryad user use thousands of machines, each of them with multiple processors or cores.

The major advantage is that users do not need to know anything about concurrent programming. A dryad application runs a computational directed graph that is composed of computational vertices and communication channels.

Therefore, dryad provides a large number of functionality including generating of job graph, scheduling of the machines for the available processes, transition failure handling in the cluster, collection of performance metrics, visualizing the job, invoking user defined policies and

dynamically updating the job graph in response to these policy decisions without knowing the semantics of the vertices.

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.

REFERENCES

1. M. K.Kakhani, S. Kakhani and S. R.Biradar, Research issues in big data analytics, International Journal of Application or Innovation in Engineering & 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 challenges of big data research, *Big Data Research*, 2(2) (2015), pp.59-64.
5. R. Kitchin, Big Data, new epistemologies and paradigm shifts, *Big Data 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, *Information Sciences*, 275 (2014), pp.314-347.
7. K. Kambatla, G. Kollias, V. Kumar and A. Gram, Trends in big data analytics, *Journal of Parallel and Distributed Computing*, 74(7) (2014), pp.2561-2573.
8. S. Del. Rio, V. Lopez, J. M. Bentez and F. Herrera, On the use of mapreduce for imbalanced big data using random forest, *Information Sciences*, 285 (2014), pp.112-137.
9. MH. Kuo, T. Sahama, A. W. Kushniruk, E. M. Borycki and D. K. Grunwell, Health big data analytics: current perspectives, challenges and 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 at challenges and opportunities of big data analytics in healthcare, *IEEE International Conference on Big Data*, 2013, pp.17-22.
11. Z. Huang, A fast clustering algorithm to cluster very large categorical data sets in data mining, *SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery*, 1997. (IJACSA) *International Journal of Advanced Computer Science and Applications*, Vol. 7
12. T. K. Das and P. M. Kumar, Big data analytics: A framework for

- hr/>
- unstructured data analysis, International Journal of Engineering and Technology, 5(1) (2013), pp.153-156.
- 13.** T. K. Das, D. P. Acharjya and M. R. Patra, Opinion mining about a product by analyzing public tweets in twitter, International Conference on 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 Information Science, 11 (1982), pp.341-356.
- 16.** D. Molodtsov, Soft set theory first results, Computers and Mathematics with Applications, 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 concept and 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 and K. Taha, Efficient machine learning for big data: A review, Big Data Research, 2(3) (2015), pp.87-93.
- 21.** Changwon. Y, Luis. Ramirez and Juan. Liuzzi, Big data analysis using modern statistical and machine learning methods in medicine, International Neurourology Journal, 18 (2014), pp.50-57.
- 22.** P. Singh and B. Suri, Quality assessment of data using statistical and machine learning methods. L. C.Jain, H. S.Behera, J. K.Mandal and D. P.Mohapatra (eds.), Computational Intelligence in Data Mining, 2 (2014), pp. 89-97.
- 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 big data information, International Conference on Information Technology and Management Innovation, 2015, pp.1041-1044.
25. Z. Hongjun, H. Wenning, H. Dengchao and M. Yuxing, Survey of research on information security in big data, Congresso da sociedade Brasileira 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: open problems and future perspectives, BioMed Research International, 2014, (2014), pp.1-13.
27. N. Mishra, C. Lin and H. Chang, A cognitive adopted framework for 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 applications for 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, Journal of Parallel and Distributed Computing, 79 (2015), pp.3-15.
30. I. A. T. Hashem, I. Yaqoob, N. Badrul Anuar, S. Mokhtar, A. Gani and S. Ullah Khan, The rise of big data on cloud computing: Review and open 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 data analytics, H. Yin, K. Tang, Y. Gao, F. Klawonn, M. Lee, T. Weise, B. Li and X. Yao (eds.), Intelligent Data Engineering and Automated Learning, 2013, pp.417-426.

- 33.** M. A. Nielsen and I. L. Chuang,
Quantum Computation and
Quantum Information, Cambridge
University Press, New York, USA
2000.
- 34.** M. Herland, T. M. Khoshgoftaar
and R. Wald, A review of data
mining using 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. Wang Promises and
challenges of big data computing in
health sciences, Big Data Research,
2(1) (2015), pp. 2-11.