Advances in Computer Science and Information Technology (ACSIT)
Print ISSN: 2393-9907; Online ISSN: 2393-9915; Volume 2, Number 6; April-June, 2015 pp. 522-526

http://www.krishisanskriti.org/acsit.html

Big Data: Reforming a Research Techniques in Data Science

Jaspreet Kaur[1], Shiwani Aggarwal[2] and Gurpreet Kaur[3]

[1]MBA (IT),M.Tech(CSE),AP(CSE) Ludhiana college of Engineering And Technology, Katani Kalan*
[2]M.Tech(IT),AP(IT) Ludhiana College of Engineering And Technology, Katani Kalan
[3]M.Tech(IT),AP(CSE) PCTE Group of Institutes, Baddowal, Ludhiana*
E-mail: [1]raijaspreet@gmail.com, [2]shiwani.aggarwal20@gmail.com, [3]gurpreet@pcte.edu.in

Abstract—*In Recent years, advancement in technologies have led to deluge of data from distinctive domains like healthcare, businesses, scientific sensors, Financial companies, Information Technology etc over the past 2 decades. Almost every sector in the world is experiencing or anticipating remarkable and exceptional growth in the amount of data in the world as well as leveraging new opportunities in their respective areas. The term "Big Data" was formulated to capture the sheer volume of data and to exhibit its unique characteristics in comparison to traditional data. Big data is commonly unstructured and the analysis of this data requires a real time processing for providing useful information. This paper presents an insightto basics of Big data, and the technologies like HADOOP & MAP-REDUCE, aiming to provide an exposure of big data to all beginners. First, we present the concept of Big Data including its introduction, definition etc. We then identify the significance of Big data in today's era. Next, we discuss various challenges and solutions to these challenges along with opportunities of Big data. Following that, we discuss about HADOOP & MAP-REDUCE required for processing of large amounts of data. Finally, we conclude the paper by presenting some key design Factors for Big data systems.*

Keywords: *Big data, Hadoop, Map-reduce, data science*

1. INTRODUCTION

Big Data is emerging paradigm, which has rapidly developed into a hotspot and has transformed our society. It has attracted diverse attention from people across various domains like academia, science, Businesses, medical, finance and other government organizations. Big data is a data exceeding the traditional database capacity and when applied to datasets; their sizes are beyond the computing power of Hardware and software tools. Internet of things(IoT), weblogs, business transactions, medical records etc; are of large size(Volume), fastdata in/out(velocity) and of high Value(Value). O'Reilly Media even asserted that "the future belongs to the companies and people that turn data into products" [1]. McKinsey, the well-known management and consulting firm, alleged that big data has pen-etrated into every area of today's industry and business functions and has become an important factor in production [2].

What is Big Data? So far, there is no such definition which is universally accepted. As per Wikipedia, "Big Data is defined as the term that encompasses the large and complex datasets that cannot be processed by traditional databases". According to Gartner, "Big Data comprises 3V's i.e. High Volume, High Velocity; High variety and innovative and cost effective processing of Information to yield enhanced decision making[1]."From macro outlook, big data can be regarded as Concoction of human society with physical world .Here physical world consist of Internet of Things, cyberspace and other information technologies, together incorporate Big data. In this sense, big data is categorized into 2 parts, namely, data from physical world, obtained from astronomical data, remote sensing data, scientific data, observations, experiments etc and data from human society which is often acquired from social media, economics, Finance, health etc. Data is profusely increasing day by day and falls into both Structured andunstructured data. In today's time we are living in a data deluge era, evidenced by sheer volume of data and increasing growth rate. For example, [2] Mckinsey report predicts that value of personal location data is estimated to be $100billion in revenue to service providers and around $500 billion to business users and consumers. This huge and tremendous increase in volume of data has attracted interest from diverse sectors, for example government, industry, academia and research. Internet players like Google, facebook, twitter and Amazon were the first facing these huge volumes of data at the internet scale. Analytical Use and empowering new products are the two categories into which the value of big data to an organization falls. Big data analytics can uncover the hidden

patterns, unknown correlations, customer preferences and other useful information to provide more effective marketing, new opportunities, better customer service and other business benefits.

Compared to the traditional data, big data provide the better platform to execute and process huge datasets. Big data wherein defines 3V's to categorize different aspects of big data. At the same time, actual real challenge confines to these diversified data types (variety), timely response (Velocity) and uncertainties in data. Another challenge lies in eliminating the unpredictability of data whether data is reliable or not. Timely response occupies an important place in big data processing as collecting, storing and processing of data requires enough energy resources.

The rest of this paper is organized as follows. Section 2presents the significance of big data. Section 3 provides an overview of opportunities in Big data.The Challenges in bigdata is presented in Section 4. Section 5 presents theMap Reduce Section 6 presents the overview of Hadoop and its use in big data..Section 7presents the conclusions.

2. SIGNIFICANCE OF BIG DATA

A big Data concern has moved from big businesses to state research Centre and mainstream status. It has been transforming our lives, the way we think because of its great value. The emergence of big data has generated new research paradigms, with big data, researchers need to only mine the information which is useful for their work without directly accessing the objects. In 2007, Jim Gray, late turing award winner, in his last speech, illustrated that fourth paradigm separates data science from computational science. [5]He believed our fourth paradigm serves not only as a change in scientific research but also in a way we think.

At present, with the help of big data, world is entering an era of Information age. An extensive use of technologies and Internet of things (IoT) hasresulted into unprecedented growth in data, thereby making data more complex. Depth Analysis plays an important role in increasing the competitiveness of companies and sustaining economic growth rate. Big data is becoming a new point of economic growth and the companies are upgrading themselves to Analysis as a service (AaaS) mode to serve in a better way to their customers and nation, thereby changing the ecology of IT and other industries. Some of the big giant IT industries like Google, IBM, and Amazon have already upgraded themselves in era of Big data. And also are working on cloud computing technologies and cloud-based computing services. Big data and cloud computing is considered as two sides of a coin: big data is an enormous application of cloud computing, whereas cloud computing serves in providing the IT infrastructure to big data. These two nexus combined together are to change the ecosystem of Internet, and even affect the pattern of the entire information industry.

Big data technologies have become a research focus in academia. Many research institutions such as University of California, Berkeley, New York University, and Chinese University of Hong Kong etc has been established across the world to provide a platform to undertake research activities in Big data. Many courses in data analytics have been setup to cultivate the talents among the students. Big data especially Big Network data helps perceive the future better as it contains a societal information and can be viewed as network mapped to a society. Big Data's predictive analysis has been applied to societal information, medical fields, and economic growth rate to generate data for the economic development.

3. CHALLENGES OF BIG DATA

While big data can yield extremely useful information, it also presents a challenge in harnessing the potential of big data with respect to what data to collect, what data to store and what data to process, how much data is secure etc. Aforementioned, processing of big data is beyond the capabilities of traditional Hardware and software tools. So in turn, new hardware and software platforms are required to address these challenges. Some of the challenges faced are discussed below:

a) Data Representation

Many datasets in big data are heterogeneous in nature, structure, granularity, organization and accessibility. A competent data presentation technique should be applied to showcase the structure, hierarchy; diversity of the data, along with implementation of some integration technique to leverage the operations across different datasets.

b) Data Complexity

Due to unprecedented growth of data with the emergence of big data, now we have to face far more complex data objectsas we lack good understanding on addressing the issue of complexity of big data. For instance, knowledge regarding the laws of distribution and association relationship of big data is required. At the same time, in-depth understanding on the inherent relationship between data complexity and computational complexity of big data, as well as domain-oriented big data processing methods are needed. All these greatly confine our capacity to

design highly efficient computational models and methods for solving problems of complexity using big data.

c) Data Privacy & Security

With the proliferation of online services and increasing use of mobile phones, privacy and security concerns regarding the accessing of personal information are increasing. It is vital to understand the support of privacy to eliminate the leakage of private information at various levels.

d) Context awareness

The rate at which mobility of users and devices is increasing, context-awareness holds an important place. An appropriate and efficient content- and context-aware routing of data is required in many cases. Aforementioned, existing infrastructures and Big Data setups focuses on processing and routing all data at once. For example, in constructing, existing data has no relation to the context about the user´s history, location, tasks, habits schedule, etc. A Notion for taking the spatial users into account is a major challenge. The goal is to present the right data to the right people and devices. Therefore,contextual awareness implementation can be suitable approach in improving the quality of solving existing problem. In the context of Big Data, contextualization proves to be an attractive paradigm to combine heterogeneous data streams and thus reducing resource consumption to improve the quality of a mining process.

e) Visual Analytics: How we perceive data

Visual Analytics [6]: A relatively new term, introduced in research agenda books, published in USA & EU. The main concept behind the visual analytics is to build knowledge, methods and combine the strength of human and electronic data processing. It laid emphasis on problem solving, decision making and leveraging data processing and knowledge discovery methods. Simultaneously, there is a need to develop interaction techniques to make easy transition from one scale to another.

f) Computational Complexity

The key features of Big data, that is, three V's makes it difficult for processing of data by traditional computational methods. Older techniques and iterative algorithms designed earlier do not support the complex data processing. When solving big data problems, we need to re-examine the computational techniques and the existing algorithms. To address the computational complexity of big data applications, we need to develop algorithms for distributed and streaming computing of data where communication, storage, and computing are well integrated and optimized. There is a need to develop bootstrapping and approximation methods for big data that are capable of handling large scale computations.

4. OPPORTUNITIES OF BIGDATA

Bigdata's opportunities have increased the demand of data intensive technologies in all developed organizations. Big data has the potential to improve performance & efficiency in the context of massive data of different types. Big data is used to give better understanding of the customer's behavior and preferences to the organization. Various business organizations can get a better sight by implementing a big data properly and using it accordingly. It improves the results of various organizations in all areas including manufacturing product, sales, marketing, and finance. It can be used in the following areas:

- It provides the solution to the data storage challenge& costs by providing efficient database management system and improves data security in information technology business.
- It plays a vital role in social media content by improving services and modifying products.
- An online banking system provides best services to the customers by implementing big data in a detection of fraud during online transactions.
- Big data analytics can transform the medical & healthcare industry. It has evolved as a new promising field for providing an improvedoperational efficiency, reduced medical costs, improved medical plans, improved quality of scanning and monitoring and enhances healthcare spending.
- There are some ways such as online submission system, information mining system, and bioinformatics to get big data opportunities.
- Big data has driven a new phase in scientific research areas.

5. MAP REDUCE

In recent years, the Map Reduce framework has come out as one of the most extensively used parallel and distributed computing platforms for processing data on terabyte and petabyte scales and it helps programmers to generate big data programs. Map Reduce is a problem solving approach developed by Google that allow tasks to be simultaneously performed on distributed computers. It improves the performance of the system, instead of one computer having to loop through a billion words we can now

have 1,000 computers simultaneously looping through only a million words each -- that's a 1000x time improvement.

Map reduce paradigm is persuaded by Yahoo!, Google, Amazon, and Facebook and utilized by various universities in their different research areas; it provides an easy-to-program, reliable parallelization of data intensive computations over many machines.

Basics of Map Reduce

Map Reduce is composed of two key components: Map and Reduce. Map () method transform a set of data into key value pairs and Reduce () method receives all the data for an individual key by performing summary operations from all "mappers". Map Reduce framework allows for distributed processing of both Map and Reduce. Its libraries have been developed using many programming languages. In Map Reduce one node acts as a master node and other nodes act as workers. Map and reduce both tasks are handled by workers node and task is assigned by masters node to individual workers. The output of the map function so obtained serves as the input to reduce function buffered in the memory and generates a set of partitions, equal to the number of reducers.The master node is fully responsible for map reduce process. It gives notification to reduce workers to read the data from local disks of map workers. The output generated by reduce function is attached to output files. These files are used as an input to make another map () & reduce () calls.

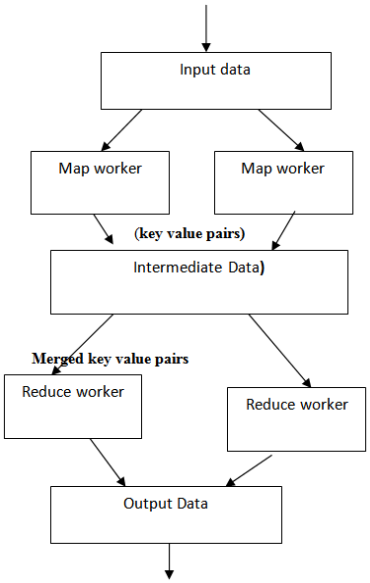

Fig. 5.1: Map Reduce data process

For Instance: Map () takes a word and convert it into a key value pair. For example, there is a file containing the word "Hi John, John is crazy, John studies in 5th grade" and there are 3 occurrences of the word "john" in a file. Here key = word and value=1.we have 3 key pairs that are passed into single reduce ().Then, we have done a loop through 3("john", 1) key-pairs and sum up all the values. The new generated key-value pairs with the same key are merged together and fed into Reduce (). Reduce () then generates another key-value pair, ("john",3) and the total sum of the values associated with that word.

Map reduce provides a platform to scale dataint-ensiveapplications from a single machine to thousand machines in a simple and effortless way, providing high per performance, fault tolerance, parallelism and locality.

Map reduce has become an omnipresent framework for large-scale data processing fields such as data mining, information retrieval, machine learning, and pattern recognition, similarity Finding.

6. OVERVIEW OF HADOOP

Hadoop is an open source software framework developed for storing and processing of larger datasets built from commodity hardware. Also known Apache Hadoop, is a registered trademark ofApache Software Foundation. Because of Google's distributed file system and MapReduce's capability of handling massive data, Hadoop too has attracted a great attention both from academia and industry. Hadoop has become a most talked about technology now days. Due to massive increase in the volume of data, especially from Social media and automated sensors, hadoop has become a major key consideration in many organizations.

The core components of Apache Hadoop consists of Hadoop distributed file system (HDFS) and the MapReduce. HDFS is a Java based DFS for storing all kinds of data and Mapreduce is a programming model for processing large datasets. There are many advantages of hadoop but the following features make them best suitable for data management perspective:

- **Scalability** : Hadoop's feature of scaling Hardware infrastructure up and down without changing the formats helps easily redistribute data to accommodate those changes.
- **Low Storage Cost:**Hadoop brings massively parallel computation to commodity servers, leading to a sizeable decrease in cost per terabytes, thereby making the processing of volume of big data affordable.

- **Flexibility:**Hadoop provides a flexibility to absorb any kind of data from multiple sources, providing solutions to many problems.
- **Analysis:**Hadoop's capability of processing variety of data in any form helps carry out analysis of data thereby helping organization uncover the hidden patterns, derive new opportunities and providing perfect opportunity with minimal investment.

7. CONCLUSION

Big data now has a great impact on each and every industry. There are no doubts, that big data's research area and study is far more beyond than our academia. In this paper, we have discussed an introduction to big data along with significance, opportunities and challenges of big data. Big data has been leveraging many opportunities now and then to every industry in the market to expand their business and understand their customers better. Also, an overview of Hadoop and Map Reduce has been discussed. We close by listing few suggestions on how to make the Big data projects a success. Following points listed below are helpful in successful completion of a Big data project:

a) Clear and concise requirement specification of technical, social and economic needs.
b) Finding of Kernel data and processing it isnecessary for proper execution and efficient working of big data.
c) A top-down or bottom up approach should be followed to solve niche problems and to put them together in a complete solution.

Finally, an integrated approach should be followed for providing a better solution to big data problems.

8. ACKNOWLEDGEMENT

We extend our heartfelt gratitude to all the members who helped us in this research paper. It would not have been possible to complete this paper without their help and guidance. Want to give special thanks to Our Parents for their continuous help and motivation. Lastly, we would also like to thank almighty for showering his blessings and giving us a strength.

REFERENCES

[1] XiaolongJina, Benjamin W.Waha,,XueqiCheng, YuanzhuoWang, "Significance and challenges of Big Data", Elseveir, 2015
[2] GangChen, SaiWu, YuanWang, " The Evolvement of Big Data Systems: From the Perspective of an Information Security Application"Elseveir, 2015,
[3] J. Manyika et al., Big data: The Next Frontier for Innovation, Competition,and Productivity. San Francisco, CA, USA: McKinsey Global Institute, 2011, pp. 1_137
[4] HAN HU, YONGGANG WEN,TAT-SENG CHUA, AND XUELONG LI, "Toward Scalable Systems for Big Data Analytics: A Technology Tutorial", IEEEAccess,July 2014
[5] Katina Michael, Keith W. Miller, "Big Data: NewOpportunitiesand New Challenges".
[6] A. Cuzzocrea, Privacy and security of big data: current challenges and future research perspectives, in: Proceedings of the First International Workshop on Privacy and Securityof Big Data, PSBD '14, 2014.
[7] C. O'Neil, R. Schutt, Doing Data Science: Straight Talk from the Frontline, O'Reilly Media, Inc., 2013.
[8] Big data, http://en.wikipedia.org/wiki/Big_data, 2014.
[9] Data science, http://en.wikipedia.org/wiki/Data_science, 2014.
[10] Declaration to be the world's most advanced IT nation, availableat: http://japan.kantei.go.jp/policy/it/2013/0614_declaration.pdf, June 2013.
[11] D. Fisher, R. DeLine, M. Czerwinski, and S. Drucker, ``Interactions withbig data analytics," Interactions, vol. 19, no. 3, pp. 50_59, May 2012.
[12] Feng Li, Beng Chin Ooi, M. Tamer Özsu, Sai Wu, Distributed data management using mapreduce, ACM Comput. Surv.46(3) (2014) 31.
[13] Li Feng, M. Tamer Özsu, Gang Chen, Beng Chin Ooi, R-store: a scalable dis-tributed system for supporting real-time analytics, in: ICDE, 2014, pp.40–51
[14] Makoto Onizuka, Hiroyuki Kato, Soichiro Hidaka, Keisuke Nakano, Zhenjiang Hu, Optimization for iterative queries on mapreduce, Proc. VLDB Endow. 7(4) (2013) 241–252.
[15] Jens Dittrich, Jorge-ArnulfoQuiané-Ruiz, Efficient big data processing in hadoopmapreduce, Proc. VLDB Endow. 5(12) (2012) 2014–2015.
[16] T. Kalil, Big data is a big deal, available at: http://www.whitehouse.gov/blog/2012/03/29/big-data-big-deal, 2012