International Journal of Innovative Research in Science and Engineering
Vol. No.3, Issue 02, February 2017
www.ijirse.com
ISSN (O) 2454-9665
ISSN (P) 2454-0663

A STUDY ON BIG DATA AND ITS CHALLENGES

A.Gayathry Arun

II B.Tech (IT), Sri Krishna College of Engineering and Technology, Coimbatore, Tamilnadu, INDIA.

ABSTRACT

The rapid developments of WWW, Internet of Things and Cloud Computing have lead to the explosive massive growth of data in almost every organization and business area in recent years. Big Data has rapidly developed in the topics that attract extensive attention from academia, industry, and governments around the world. In this paper the concept of Big Data, including its definition, characteristics, challenges and advantages are briefly discussed. The different perspectives of significance and the grand challenges as well as possible solutions are also described in this paper.

I. INTRODUCTION

Big Data is defined as large amount of data which requires new technologies and architecture to make possible to extract values from it by capturing and analysis process. A new source of big data includes the location specific data arising from traffic management, and from the tracking of personal devices such as smart phones. Big Data has emerged because we are living in a society which makes increasing use of data intensive technologies. Due to such large size of data it becomes very difficult to perform effective analysis using the existing traditional techniques. Since Big Data is a recent upcoming technology in the market which can bring huge benefits. Big Data concept means a datasets which continues to grow so much that it becomes difficult to manage it using existing database management concept and tools. This is an era of Big Data. Big Data is driving radical changes in traditional data analysis platforms. To perform any kind of analysis on such voluminous and complex data, scaling up the hardware platforms becomes imminent and choosing the right hardware/ software platforms becomes a crucial decision if the user's requirements are to be satisfied in a reasonable amount of time. Researchers have been working on building novel data analysis techniques for big data more than ever before which has led to the continuous development of many different algorithms and platforms (Dilpreet Singh and Chandan K Reddy, 2014).

II. BIG DATA CHARACTERISTICS

Mining and extracting meaningful patterns from massive input data for decision making, prediction, and other inferencing is at the core of Big Data Analytics. In addition to analyzing massive volumes of data, Big Data Analytics poses other unique challenges for machine learning and data analysis, including format variation of the raw data, fast moving streaming data, trustworthiness of the data analysis, highly distributed input sources, noisy and poor quality data, high dimensionality, scalability of algorithms, imbalanced input data, unsupervised and un-categorized data, limited supervised/labeled data, etc. Adequate data storage, data indexing/tagging, and fast information retrieval are other key problems in Big Data Analytics (Maryam M Najafabadi, 2015).

Big data analytics

Big Data generally refers to data that exceeds the typical storage, processing, and computing capacity of conventional databases and data analysis techniques. As a resource, Big Data requires tools and methods that can be applied to analyze and extract patterns from large-scale data. The rise of Big Data has been caused by increased data storage capabilities, increased computational processing power, and availability of increased volumes of data, which give organization more data than they have computing resources and technologies to process. In addition to the obvious great volumes of data, Big Data is also associated with other specific complexities, often referred to as the four Vs: Volume, Variety, Velocity, and Veracity (Maryam M Najafabadi, 2015)

Data Volume: The big word in big data itself defines the volume. At present the data existing is in peta bytes (10^15) and is supposed to increase to zeta bytes (10^21) in nearby futures. Data volume measures the amount of data available to an organization, which does not necessarily have to own all of it as long as it can access it.

Data Velocity: Velocity in big data is a concept which deals with the speed of the data coming from various sources. This character is not being limited to the speed of the incoming data but also the speed at which the data flows and aggregated.

Data Variety: Data variety is the measure of the richness of the data representation-text, images, videos, audios etc. Data being produced is not of single category as it not only includes the traditional data but also the semi structured data from all resources.

Data value: Data value measures the usefulness of data in making decisions. Data science is exploratory and useful in getting to know the data, but "ANALYTIC SCIENCE " encompasses the predictive power of big data. User can run certain queries against the data stored and thus can deduct important results.

Complexity : Complexity measures the degree of interconnectedness and interdependence in big data structures such that small chances in one or two elements can yield very large chances or a small chances that ripple across or cascade through the system and substantially affect its behavior, or no chances at all.

II. SIGNIFICANCE

Significance to national development

In the future, big data will become a new point of economic growth. With big data, companies will upgrade and transform to the mode of Analysis as a Service (AaaS), thereby changing the ecology of the IT and other industries. In this context, the global giants of the IT industry (such as IBM, Google, Microsoft, and Oracle) have already begun their technical development planning in the big data era.

Significance to industrial upgrades

Big data is currently a common problem faced by many industries, and it brings grand challenges to these industries' digitization and informationization. Research on common problems of big data, especially on breakthroughs of core technologies, will enable industries to harness the complexity induced by data interconnection and to master uncertainties caused by redundancy and/or shortage of data. Everyone hopes to mine from big data demand-driven in-formation, knowledge and even intelligence and ultimately taking full advantage of the big value of big data.

Significance to scientific research

Big data has caused the scientific community to re-examine its methodology of scientific research and has triggered a revolution in scientific thinking and methods. It is well-known that the earliest scientific research in human history was based on experiments. Later on, theoretical science emerged, which was characterized by the study of various laws and theorems. However, because theoretical analysis is too complex and not feasible for solving practical problems, people began to seek simulation-based methods, which led to computational science.

Significance to helping people better predict the future

Through effective integration and accurate analysis on multi-source heterogeneous big data, better predictions of future trends of events can be achieved. It is possible for big data analysis to even promote sustainable developments of society and economy and further give birth to new industries related to data services(XiaolongJin, 2015).

III. CHALLENGES OF BIG DATA:

The challenges of big data are usually the real implementation hurdles which require immediate attention. Any implementation without handling these challenges may lead to the failure of the technology implementation and some unpleasant results.

Technical Challenges

Fault Tolerance: With the incoming of the new technologies, big data always intended that whenever the failures occurs the damage done should be within acceptable threshold rather than beginning the whole task from the scratch. Fault-Tolerant computing is extremely hard, involving intricate algorithms. It is simply not possible to device absolutely foolproof, 100% reliable fault tolerance machines or software.

Two methods which seem to increase the fault tolerance in big data are:

First is to divide the whole computation being done into tasks and assign these tasks to different nodes of computation Second is, one node assigned the work of observing that these nodes are working properly. If something happens that particular work is restarted.

Data complexity

The emergence of big data has provided us with unprecedented large-scale samples when dealing with computational problems, although we now have to face far more complex data objects. As aforementioned, the typical characteristics of big data are di-versified types and patterns, complicated inter-relationships, and greatly varied data quality. The inherent complexity of big data (including complex types, complex structures, and complex patterns) makes its perception, representation, understanding and computation far more challenging and results in sharp increases in the computational complexity when compared to traditional computing models based on total data. Traditional data analysis and mining tasks, such as retrieval, topic discovery, semantic analysis, and sentiment analysis, become extremely difficult when using big data. At present, we do not have a good understanding on addressing the complexity of big data.

System complexity

The design of system architectures, computing frameworks, pro-cessing modes, and benchmarks for highly energy-efficient big data processing platforms is the key issue to be addressed in system complexity. Solving these problems can lay the principles for de-signing, implementing, testing, and optimizing big data processing systems. Their solutions will form an important foundation for developing hardware and software system architectures with energy-optimized and efficient distributed storage and processing.(XiaolongJin 2015)

IV. BIG DATA TECHNIQUES

The International Data Corporation (IDC) study says that overall data will be 50 times by 2020, driven in large part of more embedded systems such as sensors in clothing, medical devices and structured like buildings and bridges. The study also determined that unstructured information- such as files, email and videos-will accounts for 90% of all data created over the next decade.

In this the mainly used techniques is Hadoop, which is an open source.

Hadoop is an open source project by Apache Software Foundation. It consists of many small sub projects which belongs to the category of infrastructure for distributed computing. Apache Hadoop is an open source framework for storing and processing large datasets using clusters of commodity hardware. Hadoop is designed to scale up to hundreds and even thousands of nodes and is also highly fault tolerant.

Map Reduce

The programming model used in Hadoop is MapReduce which was proposed by Dean and Ghemawat at Google. MapReduce is the basic data processing scheme used in Hadoop which includes breaking the entire task into two parts, known as mappers and reducers. At a high-level, mappers read the data from HDFS, process it and generate some intermediate results to the reducers. Reducers are used to aggregate the intermediate results to generate the final output which is again written to HDFS. A typical Hadoop job involves running several mappers and reducers across different nodes in the cluster. A good survey about MapReduce for parallel data processing is available in (Lee K-H, 2012).

One of the major drawbacks of MapReduce is its inefficiency in running iterative algorithms. Map Reduce is not designed for iterative processes. Mappers read the same data again and again from the disk. Hence, after each iteration, the results have to be written to the disk to pass them onto the next iteration. This makes disk access a major bottleneck which significantly degrades the performance. For each iteration, a new mapper and reducer have to be initialized. Sometimes the Map Reduce jobs are short-lived in which case the overhead of initialization of that task becomes a significant overhead to the task itself (Dilpreet Singh and Chandan K Reddy, 2014)

Hadoop, mainly consists of:

File system (The Hadoop File System)
Programming paradigm (Map reduce)

Advantages:

Understanding and Targeting Customers, Understanding and Optimizing Business Process, Improving Science and Research, Improving Healthcare and Public Health, Optimizing Machine and Device Performance, Financial trading, Improving Security and Law Enforcement, Good Practices are the selective advantages of Big Data.

V.CONCLUSION

Big data has made a strong impact in almost every sector and industry today. In this paper, we have briefly reviewed the opportunities and significance of big data, as well as some grand challenges that big data brings us. In this report some of the issues are covered that are needed to be analyzed by the organization while estimating the significance of implementing the big data technology and some direct challenges to the infrastructure of the technologies.

REFERENCE

[1] Dilpreet Singh and Chandan K Reddy A survey on platforms for big data analytics, *Journal of Big Data* 2014, **2**:8 doi:10.1186/s40537-014-0008-6

[2] XiaolongJina, Benjamin W.Wah, XueqiChenga, YuanzhuoWang, Significance and Challenges of Big Data Research, Big Data Research 2 (2015) 59–64

[3] Maryam M Najafabadi1, Flavio Villanustre, Taghi M Khoshgoftaar, Naeem Seliya, Randall Wald1and Edin Muharemagic, Deep learning applications and challenges in big data analytics Journal of Big Data (2015) 2:1 DOI 10.1186/s40537-014-0007-7