ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 3, March 2018

Big Data – Technologies, Challenges and Future Scope

[1] B.V.Hemalatha, [2] R.Vijayalatha
[1] Assistant professor, Theni Kammavar Sangam College of Arts and Science,Theni, Tamilnadu, India
[2] Research scholar, Manonmaniam Sundaranar University, Tirunelveli, Tamilnadu, India

Abstract: **- Big Data has gained much attention from the academia and the IT industry. In the digital and computing world, information is generated and collected at a rate that rapidly exceeds the boundary range. The term, 'Big Data' has been coined to refer to the gargantuan bulk of data that cannot be dealt with by traditional data-handling techniques. A huge repository of terabytes of data is generated each day from modern information systems and digital technologies such as Internet of Things and cloud computing. Analysis of these massive data requires a lot of efforts at multiple levels to extract knowledge for decision making. Therefore, big data analysis is a current area of research and development. The utilization of Big Data Analytics after integrating it with digital capabilities to secure business growth and its visualization to make it comprehensible to the technically apprenticed business analyzers has been discussed in depth. Aside this, the incorporation of Big Data in order to improve population health, for the betterment of finance, telecom industry, food industry and for fraud detection and sentiment analysis have been delineated. The challenges that are hindering the growth of Big Data Analytics are accounted for in depth in the paper.**

Keywords: **Big Data, Data Visualization, Integration.**

I. 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 petabytes and beyond

II. II. DEFINITIONS AND CHARACTERISTICS

Big Data can be defined by five V's: Volume, Velocity, Variety, Veracity, and Value. Figure 1 depicts the overall five characteristics (Five V's) of big data. Each of the characteristics of big data (shown in figure 1 is briefly described below.

A) Volume: It refers to the size of data which is larger than terabytes and petabytes or even exabytes.
B) Variety: In big data, data is collected from different sources and usually has three types: structured, semi-structured, and unstruc as, text, image, audio, video, emails, sensors data, online transactions, etc.
C) Velocity: It refers to the high motion of data, that means, how fast data is being produced and how fast data is being processed. This is especially needed for time-limited or time-critical or real-time processing.
D) Veracity: It refers to the genuineness of the data.
E) Value: It refers to the quality of data for generating the intended result.

Big Data is characterized by three aspects: (a) the data are numerous, (b) the data cannot be categorized into regular relational databases, and (c) data are generated, captured, and processed very quickly. Big Data can be simply defined by explaining the 3V's – volume, velocity and variety which are the driving dimensions of Big Data quantification. Gartner analyst, Doug Laney introduced the famous 3 V's concept in his 2001 Metagroup publication, '3D data management: Controlling Data Volume, Variety and Velocity'.

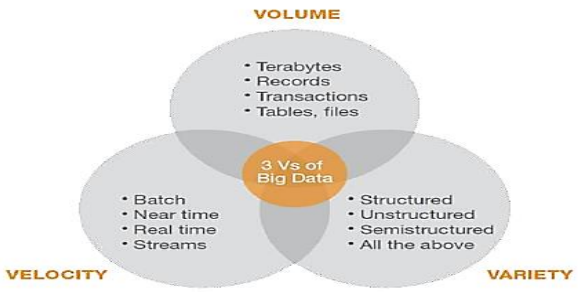

Fig. 1 Schematic representation of the 3V's

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 3, March 2018

Volume refers to the huge amount of data that are being generated everyday whereas velocity is the rate of growth and how fast the data are gathered for being analysis. Variety provides information about the types of data such as structured, unstructured, semi-structured etc.,

III. TECHNOLOGIES

Big Data provides a new method to traditional data analysis, which has a variety of technologies, including Hadoop and MapReduce, cloud computing, grid computing and so on. This paper sorts out the following technologies

A) Hadoop and MapReduce

In the related technologies, more representative one is Hadoop, which is represented by non-relational data

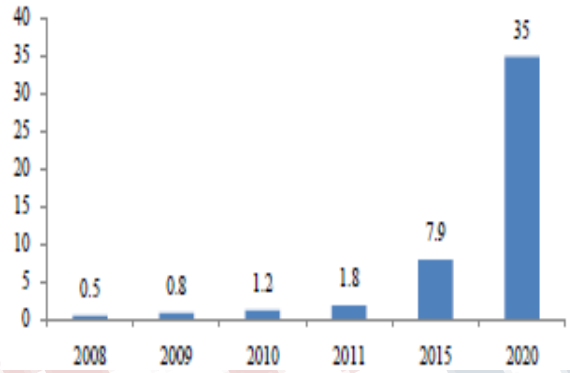

Fig. 2. The forecast of global data growth (unit: ZB).

analysis techniques. By the virtue of processing for non-structural, massively parallel processing, easy using and other advantages, Hadoop becomes a mainstream technology. Map Reduce is a model proposed for parallel processing and generating big data by Google in 2004. which is a linear, scalable programming model. Hadoop is an open source realization of Map Reduce. With its open source and easy using, Hadoop has become the first choice for big data processing. It not only creates targeted marketing applications, make full use of transaction data, but also improve accuracy and timeliness of fraud detection. Many Internet companies, including Facebook, Google, eBay and Yahoo, have developed a large scale applications based on Hadoop. Map Reduce and Hadoop can significantly improve the efficiency of big data processing.

B) Big Data Acquisition Engine

In addition to the requirements of efficiency and speed, big data collection also requires security. A general data acquisition engine which combines rule engine and finite state automaton together, helps to verify the security and correctness of the big data acquisition flow. When adding a new collection node, the rule engine will automatically make the whole system more flexible and scalable. At the same time, it ensures the state transition, and improves safety and clear logic. Big data acquisition, integrated with JESS rule engine, not only can control the state transitions and match, but also to monitor the unusual status and location errors. Rules engine can clearly show the errors and details which are matching wrong, ensure the safety and accuracy of the data acquisition.

C) MFA (Mean Field Analysis)

Big data processing system requires some related components to use in parallel multiple instances of the same task, so as to achieve the desired level of performance applications. In order to enable administrators and developers to maintain the growth rate of the data, these systems' reliability assessment is critical. A set of methods for approximate inference of probabilistic models, based on MFA, can solve the performance evaluation system problem of big data. Through behavioral modeling to assess the performance of data structure, MFA can calculate the related basic performance in a limited time. In addition, MFA can set up and evaluate in a shorter time, because it does not depend on the number of instances. In the process of assessing the performance of big data, MFA technology is very effective.

D) Other Technologies

In addition to the above-mentioned techniques, M2M (Machine To Machine) technology is an important one. M2M platform can expand the number of data producers and data consumers flexibility, accomplish new services in a very short period of time, re-use and combine data from different sources. Existing studies have shown that automatically creating M2M decision support system has much room for development. There are also grid computing, cloud computing and other technologies in big data analysis and processing. Big data technology is not a single technology, but mix with a variety of other techniques, so as to play the biggest role in the storage and analysis.

IV.REAL-TIME APPLICATIONS

Real-time applications differ from regular applications in "time attribute". That is, real-time applications guarantee

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 3, March 2018

the response or action or decision within a very short time or within a specific time line. The applications are called real-time because of the real-time processing of its response or required decision. But if the processing takes over of specific time line, it becomes useless. So, real time applications require all the needed resources available while processing to guarantee the fast response or timely response. There are various domains where real-time big data analytics applications are needed. In this section, examples of such different real-time big data analytics applications are discussed

A) Transportation

Transportation system is one of the major area where real-time data analytics is very much needed because of the required processing of data within a very short time for various purposes or services. For example, real-time data analytics of current traffic conditions could provide very useful information to the end user within a very short time for making a efficient decision, such as

—route selection for the destination

—estimate time to reach to the destination

—changing route because of any kind of sudden incidents like accident, or roadblocks.

—quick delivery of orders for any kind of goods, like pizza delivery, or emergency postal delivery.

—dynamic time calculation for emergency vehicles like ambulance, fire service car, police van for the quick arrival to the destination.

Road Sensor Technologies	Inductive-Loop Detectors
	Monitoring Cameras
	Capacitance Mats
	Road Tube Axie
	Inductive Loop
	Piezoelectric Axie
Vehicle Sensor Technology	On-board Cameras
	Proximity Sensors
	GPS Systems
	Speedometers
Communication Technologies	Satellite
	GSM
	WiFi
	Bluetooth

Table 1. Sensor and Communication Technologies

In recent time sensor technology is developed a lot which could be used for monitoring traffic condition associated with the communication technology. Table 1 lists some sensor technologies categorized into road sensor and vehicle sensor and of course the list of communication mechanisms for reliable and time efficient communication.

B) Stock Market

A stock market is the aggregation of buyers and sellers where they buy or sale shares or stocks of listed companies. In stock market huge number of data are generated in every working day. These data is not only big in volume but also very dynamic. By analyzing these data in real-time both buyers and sellers could be benefited and it also helps to detect fraud and illegal activities which certainly improves the performance of the stock market. Below is listed some points which can be achieved by real-time data analytics of stock market.

—Prediction of share prices before actual changes occur in share prices. So that timely selling or buying of shares can be done for higher profit margin.

—Earlier decision making ability for buying or selling shares.

—Financial threads detection in quick time.

—Detection of illegal activities in market which helps to improve market performance.

—Automated trading of shares and threads detection system, which could increase number of buyer and seller in the market.

All these merits of stock market can be achieved if real-time data analytics of stock market is possible. Otherwise it will take longer time if it is done manually. As a result it will neither help the buyer nor the seller to earn higher profit or not even the market itself to detect threads to improve market performance.

C) Clinical Care

Clinical research in real-time big data refers making correct prediction in real-time so that physicians can provides better treatment and fast accurate decisions to their patients by analyzing patients data in a timely and reliable manner. The amount of data produced within medical area or in clinic has grown to be huge in volume, where analysis of of those data and generating time response can improve the quality of clinical care for the patients. By considering the importance of real-time big data analytic in clinical care, many research are being conducted in this scope. Zhang et al. develop a clinical support system to facilitate real-time prognosis and diagnosis as quick as possible. Thommandram et al. designed a system called Artemis to detect cardiorespiratory spell in real-time. But the stream processing of their Artemis system is done by the InfoSphere Streams (a middleware system developed by IBM). But the research in this area is still fairly young and more research and development is needed here to generate results in real-time by reliably analyzing

medical data. However, there are a number of challenges that arise when dealing with these huge quantities of data such as, how to analyze this data in a reliable manner and generating real-time result to offer right treatment to the patient in real-time. By making it possible, the risk of human life can be minimized. The main purpose of clinical care base application is to provide real-time health care to the end users (i.e. patients) by collecting the real world medical data from all levels of human existence and analyzing them in a reliable manner in real-time.

D) Defense

In defense sector or in intelligent service real-time data analytics can have an important impact for making right decision in time for saving human lives. For example, winning in war, power or strength is not only the concern but also making the right decision in time. And for that lot of information needed to analyze like information about different vehicles used in war, opposition strengths or any movement, current situation or historical information related to the war, number of soldiers, and other related information or resources needed to collect and analyzed in time to take right movement or decision in the war. Also in war data generation is very big in volume and very dynamic. These data should be collected and analyzed dynamically along with other static information to plan for the next step or to make decision on the fly in the war. This sort of various action or decision are needed to make in defense or military or national security center for the sake of life of human being. Hence, real-time data analytics application or system of such kind of big data is very important to deal with these sort of situation arise in security sector.

E) Events/Festivals

Nowadays large events are taking place in a regular basis all over the world. Such events include open-air concerts, various types of sports games, New Year celebrations or various types of religious or traditional programs/festivals where big numbers of crowds are get together. In this circumstances, controlling crowds is very much important in terms of security or the smoothness completion of events. Control Center needs to take some real-time decisions based on the crowd movement like increasing parking lots, traffic control, medical supports, or number of security force presence in some areas of the events becomes important. This monitoring /controlling can be done using GPS, satellite or location tracking technology, traffic or vehicle sensing technologies (see table 1 for Sensor and Communication Technologies). After collecting necessary information, Control Center can use these information to provide their services in real-time over the over-crowded areas. As these sort of large events can have huge number of crowds, the amount of data collected would be huge in volume. Hence, these huge number of data has to be organized and analyzed in real-time to make immediate decisions or emergency response which might help to save crowds lives in case of any accidents or any disasters. But in reality, to manage large amount data in real-time is very challenging tasks because these data is not only big in volume but also they are variety in characteristics and velocity is high on their motion.

F) Natural Disasters

There are significant number of natural disasters the world faced so far, which costs huge number of human lives, health, economies and various resources. Example of such natural disaster includes earthquakes, floods, tsunami, cyclones, volcanoes, etc. Early predicting and warnings of such natural disasters can save the lives of thousands of people and resources. The early warnings can include information about shelter, to-do, and required action, emergency support, etc. But this early warning system for natural disaster involves real-time processing of huge amount of distributed data that also collected in real-time. For this type of data collection various sensors and GPS or Satellite technology can be used. However the main challenging task is to analyze those data in timely fashion to provide early warnings. Example of such early warnings systems include the Indian Tsunami Early Warning System to detect tsunamis, the global early warning system for wild land fire. and so on. The response of these type of application will need to be accurate as it involves thousands of human lives. Though it is a very challenging tasks however, having a early natural disaster system which is capable of dealing with huge distributed data and time critical processing of those information, it is possible to deliver a accurate or almost accurate (at least) early warnings of natural disasters to people. And this warning can save thousands of human lives, economies, and resources and help them to take required action/initiative at a time.

G) Daily Resources

In this paper, daily resources refer to electricity, gas, water, etc. which are required in daily life for a human. The production of these resources is based on the daily usage/consumption and the future usages. However, on a particular day or on a specific time period, the usage of these resources could increase or decrease. Besides, on of a sudden the demand of any of these resources could increase. So, a real-time monitoring about the usage of these resources could handle such type of situation. As a result, required amount of production of these resources can be managed efficiently on a timely fashion. Hence,

consumer always has their required resources for usages. In addition, this type of real-time data analytic application can be used to predict future usages rate and prepared/plan for that in advance. Also, it allows efficient allocation of resources quite efficiently on time to time. Moreover, corresponding authority of these resources can manage the resources in real-time in a efficient way by using this type of data analytic application.

IV. FUTURE SCOPE AND DEVELOPMENT

Today, Big Data is influencing IT industry like few technologies have done before. The massive data generated from sensor-enabled machines, mobile devices, cloud computing, social media, satellites help different organizations improve their decision making and take their business to another level. Recently it was announced that, Indian Prime Minister‘s office is using Big Data analytics to understand Indian citizen‘s sentiments and ideas through crowd sourcing platform www.mygov.in and social media to get a picture of common people‘s thought and opinion on government actions. Google is launching the Google Cloud Platform, which provides developers to develop a range of products from simple websites to complex applications. It enables users to launch virtual machines, store huge amount of data online, and plenty of other things. Basically, it will be an one stop platform for cloud based applications, online gaming, mobile applications, etc. All these required huge amount of data processing where Big Data plays an immense role in data processing.

V. SUGGESTIONS FOR FUTURE WORK

Additionally, machine learning concepts and tools are gaining popularity among researchers to facilitate meaningful results from these concepts. Research in the area of machine learning for big data has focused on data processing, algorithm implementation, and optimization. Many of the machine learning tools for big data are started recently needs drastic change to adopt it. We argue that while each of the tools has their advantages and limitations, more efficient tools can be developed for dealing with problems inherent to big data. The efficient tools to be developed must have provision to handle noisy and imbalance data, uncertainty and inconsistency, and missing values.

VI. 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. This literature survey discusses Big Data from its infancy until its current state. It elaborates on the concepts of big data followed by the applications and the challenges faced by it. Finally we have discussed the future opportunities that could be harnessed in this field. Big Data is an evolving field, where much of the research is yet to be done. Big data at present is handled by the software named Hadoop. However, the proliferating amounts of data is making Hadoop insufficient. We belive that in future researchers will pay more attention to these techniques to solve problems of big data effectively and efficiently.

REFERENCES

1. Samiddha Mukherjee, Ravi Shaw, “Big Data – Concepts, Applications, Challenges and Future Scope”, IJARCCE, vol 5, issue 2, feb 2016.

2. D.P.Acharjya, Kauser Ahmed P, “A Survey on Big Data Analytics: Challenges, Open Research Issues and Tools”, IJACSA, vol 7, no. 2, 2016.

3. Akinul Islam Jony, “Applications of Real-Time Big Data Analytics”, International Journal of Computer Applications, vol 144 – no.5, june 2016.

4. Zan Mo, Yanfei Li, “Research of Big Data Based on the Views of Technology and application”, American Journal of Industrial and Business Management, 5, 192-197