Volume 2, Issue 12, December-2015, pp. 1147-1152 ISSN (O): 2349-7084

International Journal of Computer Engineering In Research Trends

Available online at: www.ijcert.org

Review on Data Mining Techniques with Big Data

[1]TEMPALLI NARENDRA BABU, [2]R.MADHURI DEVI

[1]M.Tech (CSE), Priyadrshini Institute of Technology & Management

[2]AssociateProfessor (Dept.of CSE), Priyadrshini Institute of Technology & Management

Abstract: The term Big Data comprises large- volume, complex, growing data sets with multiple, autonomous sources. With the fast development of networking, data storage, and the data collection capacity, Big Data are now rapidly expanding in all science and engineering domains, including physical, biological and biomedical sciences. This paper presents a HACE theorem that characterizes the features of the Big Data revolution, and proposes a Big Data processing model, from the data mining perspective. This data-driven model involves demand- driven aggregation of information sources, mining and analysis, user interest modeling, and security and privacy considerations. We analyze the challenging issues in the data-driven model and also in the Big Data revolution.

Keywords: HACE,Hadoop,Big Data, heterogeneity, autonomous sources, complex and evolving associations

1. INTRODUCTION

According to the sources, every day, 2.5 quintillion bytes of data are created and 90 percent of the data in the world today were produced within the past two years [16]. Our capability for data generation has never been so powerful and enormous ever since the invention of the information technology in the early 19th century. Let's take an example; Dr. Yan Mo won the 2012 Nobel Prize in Literature. This is probably the most controversial Nobel Prize of this category. Searching on Google with "Yan Mo Nobel Prize," resulted in 1,050,000 web pointers on the Internet (as of 3 January 2013). "For all praises as well as criticisms," said Mo recently, "I am grateful." What types of praises and criticisms has Mo actually received over his 31-year writing career? As comments keep coming on the Internet and in various news media, can we summarize all types of opinions in different media in a real-time fashion, including updated, cross-referenced discussions by critics? This type of summarization program is an excellent example for Big Data processing, as the information comes from multiple, heterogeneous, autonomous sources with complex and evolving relationships, and keeps growing. As another example, on 4 October 2012, the first presidential debate between President Barrack Obama and Governor Mitt Romney triggered more than 10 million tweets within 2 hours [1]. Among all these tweets, the specific moments that generated the most discussions actually revealed the public interests, such as the discussions about Medicare and vouchers. Such online discussions provide a new means to sense the public interests and generate feedback in real time, and are mostly appealing compared to generic media, such as radio or TV broadcasting. Another example is Flickr, a public picture sharing site, which received 1.8 million photos per day, on average, from February to March 2012 [2]. Assuming the size of each photo is 2 megabytes (MB), this requires 3.6 terabytes (TB) storage every single day. Indeed, as an old saying states: "a picture is worth a thousand words," the billions of pictures on Flicker are a treasure tank for

us to explore the human society, social events, public affairs, disasters, and so on, only if we have the power to harness the enormous amount of data.

The above examples demonstrate the rise of Big Data applications where data collection has grown tremendously and is beyond the ability of commonly used software tools to capture, manage, and process within a "tolerable elapsed time." The most fundamental challenge for Big Data applications is to explore the large volumes of data and extract useful information or knowledge for future actions [3]. In many situations, the knowledge extraction process has to be very efficient and close to real time because storing all observed data is nearly infeasible. For example, the square kilometer array (SKA) [4] in radio astronomy consists of 1,000 to 1,500 15-meter dishes in a central 5-km area. It provides 100 times more sensitive vision than any existing radio telescopes, answering fundamental questions about the Universe. However, with a 40 gigabytes (GB)/second data volume, the data generated from the SKA are exceptionally large. Although researchers have confirmed that interesting patterns, such as transient radio anomalies [5] can be discovered from the SKA data, existing methods can only work in an offline fashion and are incapable of handling this Big Data scenario in real time. As a result, the unprecedented data volumes require an effective data analysis and prediction platform to achieve fast response and real-time classification for such Big Data.

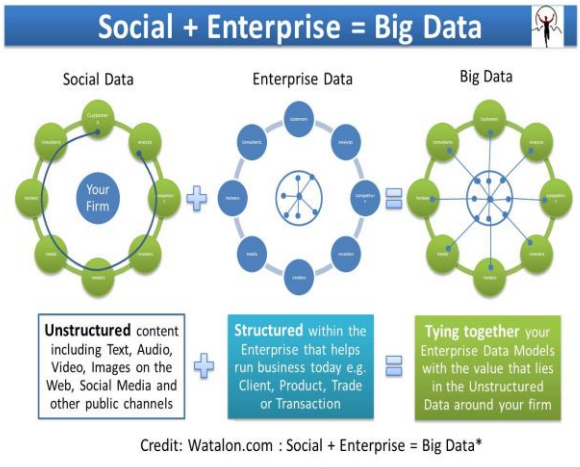

Fig 1. Theme of Big data

II. TYPES OF BIG DATA AND SOURCES:

There are two types of big data: structured and unstructured. Structured data are numbers and words that can be easily categorized and analyzed. These data are generated by things like network sensors embedded in electronic devices, smartphones, and global positioning system (GPS) devices. Structured data also include things like sales figures, account balances, and transaction data. Unstructured data include more complex information, such as customer reviews from commercial websites, photos and other multimedia, and comments on social networking sites. These data can not easily be separated into categories or analyzed numerically. "Unstructured big data is the things that humans are saying," says big data consulting firm vice president Tony Jewitt of Plano,Texas. "It uses natural language."Analysis of unstructured data relies on keywords, which allow users to filter the data based on searchable terms. The explosive growth of the Internet in recent years means that the variety and amount of big data continue to grow. Much of that growth comes from unstructured data.

Fig 2. Big data Sources

III. TECHNIQUES AND TECHNOLOGY

For the purpose of processing the large amount of data, the big data requires exceptional technologies. The various techniques and technologies have been

introduced for manipulating, analyzing and visualizing the big data [20]. There are many solutions to handle the Big Data, but the Hadoop is one of the most widely used technologies.

A. Hadoop Hadoop is a Programming framework used to support the processing of large data sets in a distributed computing environment. Hadoop was developed by Google's MapReduce that is a software framework where an application break down into various parts. The Current Appache Hadoop ecosystem consists of the Hadoop Kernel, MapReduce, HDFS and numbers of various components like Apache Hive, Base and Zookeeper [17]. MapReduce is a programming framework for distributed computing which is created by the Google in which divide and conquer method is used to break the large complex data into small units and process them. Map Reduce have two stages which are [18]: Map ():- The master node takes the input, divide into smaller subparts and distribute into worker nodes. A worker node further do this again that leads to the multi-level tree structure. The worker node process the m=smaller problem and passes the answer back to the master Node. Reduce ():- The, Master node collects the answers from all the sub problems and combines them together to form the output

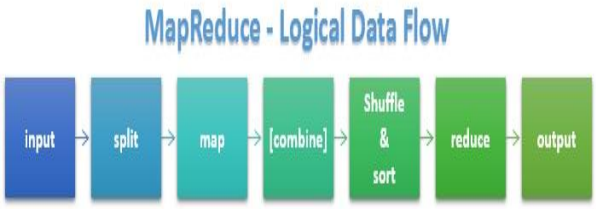

Fig 3. MapReduce logical dataflow

IV. HACE Theorem.

Big Data starts with large-volume, heterogeneous, autonomous sources with distributed and decentralized control, and seeks to explore complex and evolving relationships among data. These characteristics make it an extreme challenge for discovering useful knowledge from the Big Data. In a naïve sense, we can imagine that a number of blind men are trying to size up a giant Camel, which will be the Big Data in this context. The goal of each blind man is to draw a picture (or conclusion) of the Camel according to the part of information he collects during the process. Because each person" s view is limited to his local region, it is not surprising that the blind men will each conclude independently that the camel "feels" like a rope, a hose, or a wall, depending on the region each of them is limited to. To make the problem even more complicated, let us assume that the camel is growing rapidly and its pose changes constantly, and each blind man may have his own (possible unreliable and inaccurate) information sources that tell him about biased knowledge about the camel (e.g., one blind man may exchange his feeling about the camel with another blind man, where the exchanged knowledge is inherently biased).

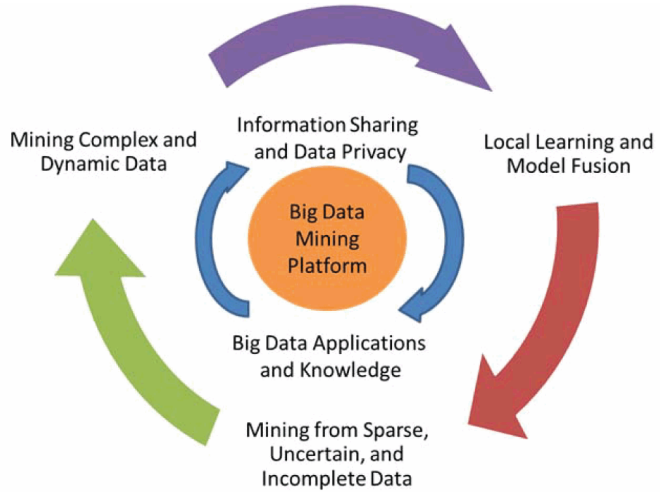

Fig. 4. A Big Data processing framework:

Exploring the Big Data in this scenario is equivalent to aggregating heterogeneous information from different sources (blind men) to help draw a best possible picture to reveal the genuine gesture of the camel in a real-time fashion. Indeed, this task is not as simple as asking each blind man to describe his feelings about the camel and then getting an expert to draw one single picture with a combined view, concerning that each individual may speak a different language (heterogeneous and diverse information sources) and they may even have privacy concerns about the messages they deliberate in the information exchange process. The term Big Data literally concerns about data volumes, HACE

theorem suggests that the key characteristics of the Big Data are A. Huge with heterogeneous and diverse data sources:-One of the fundamental characteristics of the Big Data is the huge volume of data represented by heterogeneous and diverse dimensionalities. This huge volume of data comes from various sites like Twitter,Myspace,Orkut and LinkedIn etc. B. Decentralized control:- Autonomous data sources with distributed and decentralized controls are a main characteristic of Big Data applications. Being autonomous, each data source is able to generate and collect information without involving (or relying on) any centralized control. This is similar to the World Wide Web (WWW) setting where each web server provides a certain amount of information and each server is able to fully function without necessarily relying on other servers C. Complex data and knowledge associations:-Multistructure, multisource data is complex data, Examples of complex data types are bills of materials, word processing documents, maps, time-series, images and video. Such combined characteristics suggest that Big Data require a "big mind" to consolidate data for maximum values.

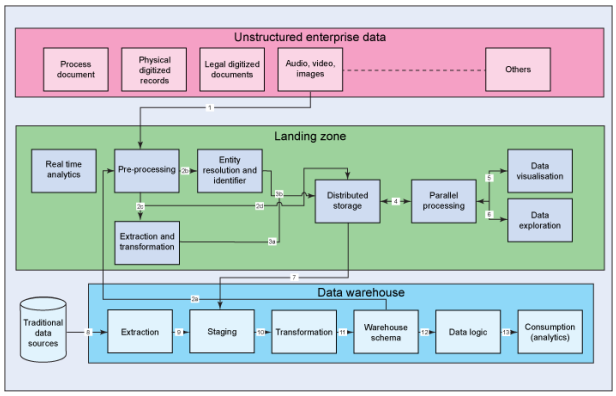

Fig 5.Datamining Process on Unstructured data

V. DATA MINING FOR BIG DATA

Generally, data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cuts costs, or both. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational database. Data mining as a term used for the specific classes of six activities or tasks as follows: 1. Classification 2. Estimation 3. Prediction 4. Association rules 5. Clustering 6. Description A. Classification is a process of generalizing the data according to different instances. Several major kinds of classification algorithms in data mining are Decision tree, k-nearest neighbor classifier, Naive Bayes, Apriori and Gadabouts. Classification consists of examining the features of a newly presented object and assigning to it a predefined class. The classification task is characterized by the well-defined classes, and a training set consisting of reclassified examples. B. Estimation deals with continuously valued outcomes. Given some input data, we use estimation to come up with a value for some unknown continuous variables such as income, height or credit card balance. C. Prediction It‟s a statement about the way things will happen in the future , often but not always based on experience or knowledge. Prediction may be a statement in which some outcome is expected. D.Association Rules An association rule is a rule which implies certain association relationships among a set of objects (such as "occur together" or "one implies the other") in a database. E.Clustering can be considered the most important unsupervised learning problem; so, as every other problem of this kind, it deals with finding a structure in a collection of unlabeled data.

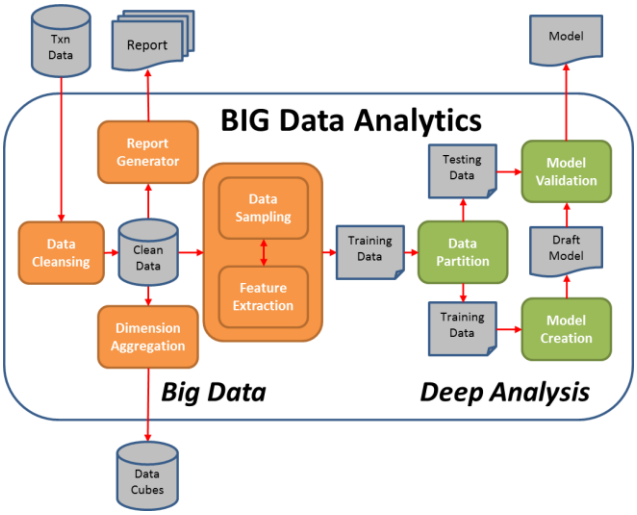

Fig 6. Big data Analytics

V1. CHALLENGES IN BIG DATA

Meeting the challenges presented by big data will be difficult. The volume of data is already enormous and increasing every day. The velocity of its generation and growth is increasing, driven in part by the proliferation of internet connected devices. Furthermore, the variety of data being generated is also expanding, and organization's capability to capture and process this data is limited. Current technology, architecture, management and analysis approaches are unable to cope with the flood of data, and organizations will need to change the way they think about, plan, govern, manage, process and report on data to realize the potential of big data.

A. Privacy, security and trust:

The Australian Government is committed to protecting the privacy rights of its citizens and has recently strengthened the Privacy Act (through the passing of the Privacy Amendment (Enhancing Privacy Protection) Bill 2012) to enhance the protection of and set clearer boundaries for usage of personal information. Government agencies, when collecting or managing citizens data, are subject to a range of legislative controls, and must comply with the a number of acts and regulations such as the Freedom of Information Act (1982), the Archives Act (1983), the Telecommunications Act (1997) ,the Electronic Transactions Act (1999), and the Intelligence Services Act (2001). These legislative instruments are designed to maintain public confidence in the government as an effective and secure repository and steward of citizen information. The use of big data by government agencies will not change this; rather it may add an additional layer of complexity in terms of managing information security risks. Big data sources, the transport and delivery systems within and across agencies, and the end points for this data will all become targets of interest for hackers, both local and international and will need to be protected. The public release of large machine-readable data sets, as part of the open government policy, could potentially provide an opportunity for unfriendly state and non-state actors to glean sensitive information, or create a mosaic of exploitable information from apparently innocuous data. This threat will need to be understood and carefully managed. The potential value of big data is a function of the number of relevant, disparate datasets that can be linked and analysed to reveal new patterns, trends and insights. Public trust in government agencies is required before citizens will be able to understand that such linking and analysis can take place while preserving the privacy rights of individuals.

B. Data management and sharing

Accessible information is the lifeblood of a robust democracy and a productive economy to Government agencies realize that for data to have any value it needs to be discoverable, accessible and usable, and the significance of these requirements only increases as the discussion turns towards big data. Government agencies must achieve these requirements whilst still adhering to privacy laws. The processes surrounding the way data is collected, handled, utilised and managed by agencies will need to be aligned with all relevant legislative and regulatory instruments with a focus on making the data available for analysis in a lawful, controlled and meaningful way. Data also needs to be accurate, complete and timely if it is to be used to support complex analysis and decision making. For these reasons, management and governance focus needs to be on making data open and available across government via standardized APIs, formats and metadata. Improved quality of data will produce tangible benefits in terms of business intelligence, decision making, sustainable cost-savings and productivity improvements. The current trend towards open data and open government has seen a focus on making data sets available to the public, however these „open" initiatives need to also put focus on making data open, available and standardized within and between agencies in such a way that allows inter-governmental agency use and collaboration to the extent made possible by the privacy laws.

C.Technology and analytical systems:

The emergence of big data and the potential to undertake complex analysis of very large data sets is, essentially, a consequence of recent advances in the technology that allow this. If big data analytics is

to be adopted by agencies, a large amount of stress may be placed upon current ICT systems and solutions which presently carry the burden of processing, analysing and archiving data. Government agencies will need to manage these new requirements efficiently in order to deliver net benefits through the adoption of new technologies.

VII.CONCLUSION:

To explore Big Data, we have analyzed several challenges at the data, model, and system levels. To support Big Data mining, high-performance computing platforms are required, which impose systematic designs to unleash the full power of the Big Data. Big data is the term for a collection of complex data sets, Data mining is an analytic process designed to explore data(usually large amount of data-typically business or market related-also known as "big data")in search of consistent patterns and then to validate the findings by applying the detected patterns to new subsets of data. To support Big data mining, high-performance computing platforms are required, which impose systematic designs to unleash the full power of the Big Data. We regard Big data as an emerging trend and the need for Big data mining is rising in all science and engineering domains. With Big data technologies, we will hopefully be able to provide most relevant and most accurate social sensing feedback to better understand our society at real time.

REFERENCES:

1. Bakshi, K.,(2012)," Considerations for big data: Architecture and approach"

2. Mukherjee, A.; Datta, J.; Jorapur, R.; Singhvi, R.; Haloi, S.; Akram, W., (18-22 Dec.,2012) , "Shared disk big data analytics with Apache Hadoop"

3. Aditya B. Patel, Manashvi Birla, Ushma Nair ,(6-8 Dec. 2012),"Addressing Big Data Problem Using Hadoop and Map Reduce"

4. Wei Fan and Albert Bifet " Mining Big Data:Current Status and Forecast to the Future",Vol 14,Issue 2,2013

5. Algorithm and approaches to handle large Data-A Survey,IJCSN Vol 2,Issue 3,2013

6. Xindong Wu , Gong-Quing Wu and Wei Ding " Data Mining with Big data ", IEEE Transactions on Knoweledge and Data Enginnering Vol 26 No1 Jan 2014

7. Xu Y etal, balancing reducer workload for skewed data using sampling based partioning 2013.

8. X. Niuniu and L. Yuxun, "Review of Decision Trees," IEEE, 2010 .

9. Decision Trees for Business Intelligence and Data Mining: Using SAS Enterprise Miner "Decision Trees-What Are They?"

10. Weiss, S.H. and Indurkhya, N. (1998), Predictive Data Mining: A Practical Guide, Morgan Kaufmann Publishers, San Francisco, CA

