ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

Data Mining with Big Data

M.Rajitha
M.Tech Student,
Dept of CSE,
KITS for Women's, Kodad, T.S, India.

Mrs.P. Sravanthi
Associate Professor,
Dept of CSE,
KITS for Women's, Kodad, T.S, India.

Abstract:

Big Data is a new term used to identify the datasets that due to their large size and complexity. Big Data are now rapidly expanding in all science and engineering domains, including physical, biological and biomedical sciences. Big Data mining is the capability of extracting useful information from these large datasets or streams of data, that due to its volume, variability, and velocity, it was not possible before to do it. The Big Data challenge is becoming one of the most exciting opportunities for the next years. This study paper includes the information about what is big data, Data mining, Data mining with big data, Challenging issues and its related work. Keywords — Big Data, Data mining, Challenging issues, Datasets, Data Mining Algorithms.

2. INTRODUCTION:

Today is the era of Google. The thing which is unknown for us, we Google it. And in fractions of seconds we get the number of links as a result. This would be the better example for the processing of Big Data. This Big Data is not any different thing than out regular term data. Just big is a keyword used with the data to identify the collected datasets due to their large size and complexity? We cannot manage them with our current methodologies or data mining software tools. Another example, the first strike of Anna Hajare triggered number of tweets within 2 hours. Among all these tweets, the special comments that generated the most discussions actually revealed the public interests. 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. This example demonstrates the rise of Big Data applications.

The data collection has grown tremendously and is beyond the ability of commonly used software tools to capture, manage, and process within a tolerable time. 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 [40].

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) [17 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 [41] 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. The remainder of the paper is structured as follows: In Section 2, we propose a HACE theorem to model Big Data characteristics.

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

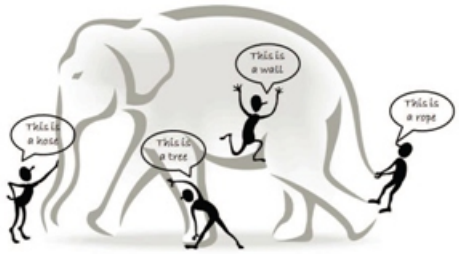


Fig. 1. The blind men and the giant elephant: the localized (limited) view of each blind man leads to a biased conclusion.

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.2.1 Sources of Big data

3. HACE Theorem:

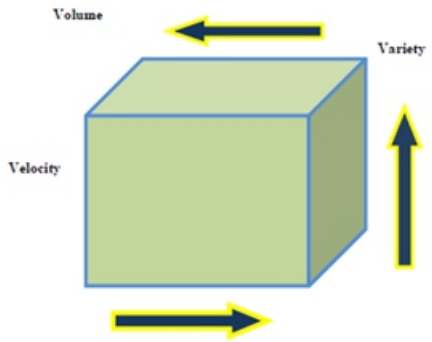
Big Data starts with large-volume, heterogeneous, autonomous sources with distributed and decentralized control, and seeks to explore complex and evolving relationships among data.These characteristics make it an extreme challenge for discovering useful knowledge from the Big Data. In a naïve sense, 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). 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.

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

4. Three V's in Big Data Volume:



Doug Laney was the first one talking about 3V"s in Big Data Management Volume: The amount of data. Perhaps the characteristic most associated with big data, volume refers to the mass quantities of data that organizations are trying to harness to improve decision-making across the enterprise. Data volumes continue to increase at an unprecedented rate.Variety: Different types of data and data sources. Variety is about managing the complexity of multiple data types, including structured, semi-structured and unstructured data. Organizations need to integrate and analyze data from a complex array of both traditional and non-traditional information sources, from within and outside the enterprise. Withshe explosion of sensors, smart devices and social collaboration technologies, data is being generated in countless forms, including: text, web data, tweets, audio, video, log files and more. Velocity: Data in motion. The speed at which data is created, processed and analyzed continues to accelerate. Nowadays there are two more V"s Variability:- There are changes in the structure of the data and how users want to interpret that data. Value:- Business value that gives organization a compelling advantage, due to the ability of making decisions based in answering questions that were previously considered beyond reach.

5. 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 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 AdaBoost. 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 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 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.

TABLE I
Difference between Big data and Data mining

Big data	Data mining
Big data is a term for large data set.	Data mining refers to the activity of going through big data set to look for relevant information
Big data is the asset	Data mining is the handler which provide beneficial result.
Big data' value depending on the capabilities of the organization managing the set, and on the capabilities of the applications that are traditionally used to process and analyze the data.	Data mining refers to the operation that involve relatively sophisticated search operation

6. 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.

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

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.2 Government agencies realise 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 standardised 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 standardised 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.

7. RESEARCH INITIATIVES AND PROJECTS:

To tackle the Big Data challenges and "seize the opportunities afforded by the new, data driven resolution,"the US National Science Foundation (NSF), underPresident Obama Administration's Big Data initiative,announced the BIGDATA solicitation in 2012. Such afederal initiative has resulted in a number of winning projects to investigate the foundations for Big Data management (led by the University of Washington), analytical approaches for genomics-based massive datacomputation (led by Brown University), large scale machine learning techniques for high-dimensional data sets that may be as large as 500,000

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

dimensions (led by Carnegie Mellon University), social analytics for largescale scientific literatures (led by Rutgers University), and several others. These projects seek to develop methods, algorithms, frameworks, and research infrastructures that allow us to bring the massive amounts of data down to a human manageable and interpretable scale. Other countries such as the National Natural Science Foundation of China (NSFC) are also catching up with national grants onBig Data research. Meanwhile, since 2009, the authors have taken the lead in the following national projects that all involve Big Data components: . Integrating and mining biodata from multiple sources in biological networks, sponsored by the US National Science Foundation, Medium Grant No. CCF-0905337, 1 October 2009 - 30 September 2013. Issues and significance.

We have integrated and mined biodata from multiple sources to decipher and utilize the structure of biological networks to shed new insights on the functions of biological systems. We address the theoretical underpinnings and current and future enabling technologies for integrating and mining biological networks. We have expanded and integrated the techniques and methods in information acquisition, transmission, and processing for information networks. We have developed methods for semantic-based data integration, automated hypothesis generation from mined data, and automated scalable analytical tools to evaluate simulation results and refine models. .

Big Data Fast Response. Real-time classification of Big Data Stream, sponsored by the Australian Research Council (ARC), Grant No. DP130102748, 1 January2013 - 31 Dec. 2015. Issues and significance. We propose to build a stream-based Big Data analytic framework for fast response and real-time decision making. The key challenges and research issues include: - designing Big Data sampling mechanisms to reduce Big Data volumes to a manageable sizefor processing; - building prediction models from Big Data streams.

8. FORECAST TO THE FUTURE:

There are many future important challenges in Big Data management and analytics, that arise from the nature of data: large, diverse, and evolving. These are some of the challenges that researchers and practitioners will have to deal during the next years:

A. Analytics Architecture:

It is not clear yet how an optimal architecture of an analytics systems should be to deal with historic data and with real-time data at the same time. An interesting proposal is the Lambda architecture of Nathan Marz.The Lambda Architecture solves the problem of computing arbitrary functions on arbitrary data in real time by decomposing the problem into three layers: the batch layer, the serving layer, and the speed layer. It combines in the same system Hadoop for the batch layer, and Storm for the speed layer. The properties of the system are: robust and fault tolerant, scalable, general, extensible, allows ad hoc queries, minimal maintenance, and debuggable.

B. Statistical significance:

It is important to achieve significant statistical results, and not be fooled by randomness. AsEfron explains in his book about Large Scale Inference it is easy to go wrong with huge data sets and thousands of questions to answer at once.

C. Distributed mining:

Many data mining techniques are not trivial to paralyze. To have distributed versions of some methods, a lot of research is needed with practical and theoretical analysis to provide new methods.

D. Hidden Big Data.:

Large quantities of useful data are getting lost since new data is largely untagged file based and unstructured data. The 2012 IDC study on Big Data explains that in 2012, 23% (643 exabytes) of the digital universe would be useful for Big Data if tagged and analyzed. However, currently only 3% of the potentially useful data is tagged, and even less is analyzed.

9 CONCLUSIONS:

Driven by real-world applications and key industrial stakeholders and initialized by national funding agencies, managing and mining Big Data have shown to be a challenging yet very compelling task.While the term Big Data literally concerns about data volumes, our HACE theorem suggests that the key characteristics of the Big Data are

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

1) huge with heterogeneous and diverse data sources, 2) autonomous with distributed and decentralized control, and 3) complex and evolving in data and knowledge associations. Such combined characteristics suggest that Big Data require a "big mind" to consolidate data for maximum values [27].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. At the data level, the autonomous information sources and the variety of the data collection environments, often result in data with complicated conditions, such as missing/uncertain values. In other situations, privacy concerns, noise, and errors can be introduced into the data, to produce altered data copies. Developing a safe and sound information sharing protocol is a major challenge. At the model level, the key challenge is to generate global models by combining locally discovered patterns to form a unifying view. This requires carefully designed algorithms to analyze model correlations between distributed sites, and fuse decisions from multiple sources to gain a best model out of the Big Data. At the system level, the essential challenge is that a Big Data mining framework needs to consider complex relationships between samples,models, and data sources, along with their evolving changes with time and other possible factors. A system needs to be carefully deigned so that unstructured data can be linked through their complex relationships to form useful patterns,and the growth of data volumes and item relationships should help form legitimate patterns to predict the trend and future. We regard Big Data as an emerging trend and the need for Big Data mining is arising in all science and engineering domains. With Big Data technologies, we will hopefully beable to provide most relevant and most accurate social sensing feedback to better understand our society at realtime. We can further stimulate the participation of the public audiences in the data production circle for societal and economical events. The era of Big Data has arrived.

9.ACKNOWLEDGMENTS:

I am M.RAJITHA and would like to thank the publishers, researchers for making their resources material available. I am greatly thankful to Associate Prof MRS.P. SRAVANTHI for their guidance. We also thank the college authorities, PG coordinator and Principal for providing the required infrastructure and support. Finally, we would like to extend a heartfelt gratitude to friends and family members.

11.REFERENCES:

[1] R. Ahmed and G. Karypis, "Algorithms for Mining the Evolution of Conserved Relational States in Dynamic Networks," Knowledgeand Information Systems, vol. 33, no. 3, pp. 603-630, Dec. 2012.
[2] M.H. Alam, J.W. Ha, and S.K. Lee, "Novel Approaches to Crawling Important Pages Early," Knowledge and InformationSystems, vol. 33, no. 3, pp 707-734, Dec. 2012.
[3] S. Aral and D. Walker, "Identifying Influential and Susceptible Members of Social Networks," Science, vol. 337, pp. 337-341, 2012.
[4] A. Machanavajjhala and J.P. Reiter, "Big Privacy: Protecting Confidentiality in Big Data," ACM Crossroads, vol. 19, no. 1, pp. 20-23, 2012.
[5] S. Banerjee and N. Agarwal, "Analyzing Collective Behavior from Blogs Using Swarm Intelligence," Knowledge and Information Systems, vol. 33, no. 3, pp. 523-547, Dec. 2012.
[6] E. Birney, "The Making of ENCODE: Lessons for Big-Data Projects," Nature, vol. 489, pp. 49-51, 2012.
[7] J. Bollen, H. Mao, and X. Zeng, "Twitter Mood Predicts the Stock Market," J. Computational Science, vol. 2, no. 1, pp. 1-8, 2011.
[8] S. Borgatti, A. Mehra, D. Brass, and G. Labianca, "Network Analysis in the Social Sciences," Science, vol. 323, pp. 892-895, 2009.
[9] J. Bughin, M. Chui, and J. Manyika, Clouds, Big Data, and Smart Assets: Ten Tech-Enabled Business Trends to Watch. McKinSey Quarterly, 2010.
[10] D. Centola, "The Spread of Behavior in an Online Social Network Experiment," Science, vol. 329, pp. 1194-1197, 2010.
[11] E.Y. Chang, H. Bai, and K. Zhu, "Parallel Algorithms for Mining Large-Scale Rich-Media Data," Proc. 17th ACM Int'l Conf. Multimedia, (MM '09,) pp. 917-918, 2009.

Author's Details:

Ms.M.Rajitha. MTech student, in M.Tech Student, Dept of CSE in KITS for women's,kodad, T.S, India

Mrs.P. Sravanthi working as a Associate at CSE in KITS for women's,kodad, T.S, IndiaJNTUH Hyderabad. He has 2 years of UG/PG Teaching Experience