Volume 4, Issue 5, May 2014 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Data Mining for Big Data: A Review

Bharti Thakur, Manish Mann
Computer Science Department
LRIET, Solan (H.P), India

Abstract :- ***Big data is the term for a collection of data sets which are large and complex, it contain structured and unstructured both type of data. Data comes from everywhere, sensors used to gather climate information, posts to social media sites, digital pictures and videos etc This data is known as big data. Useful data can be extracted from this big data with the help of data mining. Data mining is a technique for discovering interesting patterns as well as descriptive, understandable models from large scale data. In this paper we overviewed types of big data and challenges in big data for future.***

Keywords:- Big data, Data mining, Hace theorem,3V's,Privacy

I. INTRODUCTION

The term 'Big Data' appeared for first time in 1998 in a Silicon Graphics (SGI) slide deck by John Mashey with the title of "Big Data and the NextWave of InfraStress". Big Data mining was very relevant from the beginning, as the first book mentioning 'Big Data' is a data mining book that appeared also in 1998 by Weiss and Indrukya . However,the first academic paper with the words 'Big Data' in the title appeared a bit later in 2000 in a paper by Diebold .The origin of the term 'Big Data' is due to the fact that we are creating a huge amount of data every day. Usama Fayyad in his invited talk at the KDD BigMine' 12Workshop presented amazing data numbers about internet usage, among them the following: each day Google has more than 1 billion queries per day, Twitter has more than 250 milion tweets per day, Facebook has more than 800 million updates per day, and YouTube has more than 4 billion views per day. The data produced nowadays is estimated in the order of zettabytes, and it is growing around 40% every year.A new large source of data is going to be generated from mobile devices and big companies as Google,Apple,Facebook,Yahoo are starting to look carefully to this data to find useful patterns to improve user experience. "Big data" is pervasive, and yet still the notion engenders confusion. Big data has been used to convey all sorts of concepts, including: huge quantities of data,social media analytics, next generation data management capabilities, real-time data, and much more. Whatever the label, organizations are starting to understand and explore how to process and analyze a vast array of information in new ways. In doing so, a small, but growing group of pioneers is achieving breakthrough business outcomes. In industries throughout the world, executives recognize the need to learn more about how to exploit big data. But despite what seems like unrelenting media attention, it can be hard to find in-depth information on what organizations are really doing. So, we sought to better understand how organizations view big data – and to what extent they are currently using it to benefit their businesses.

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.

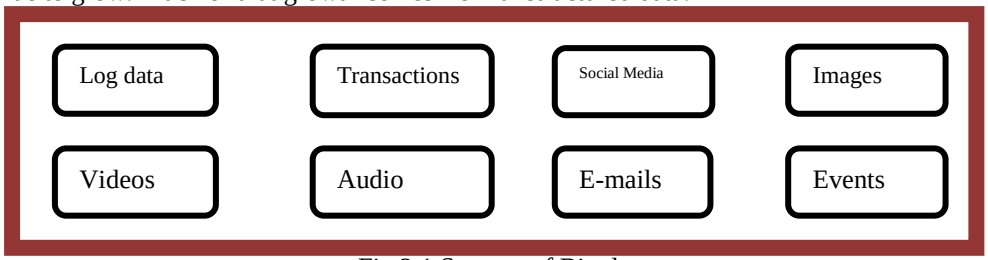

Fig.2.1 Sources of Big data

III. 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.

IV. Three V's in Big Data

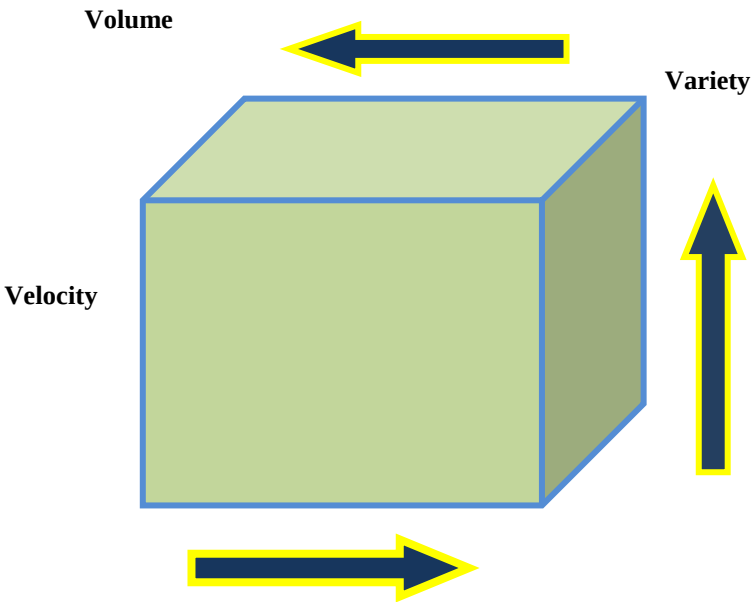

Fig 4.1 3 V's in Big Data Management

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. With

the 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.

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

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. P**rediction***

It's a statement about the way things will happen in the future , often but not always based on experience or knowledge. P*rediction* 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 1
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" varies 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

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.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.

VII. 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

VIII. CONCLUSION

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. Alex Berson and Stephen J.Smith Data Warehousing,Data Mining and OLAP edition 2010.
2. Department of Finance and Deregulation Australian Government Big Data Strategy-Issue Paper March 2013
3. NASSCOM Big Data Report 2012
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.

