DATA MINING FOR BIG DATA

Mr. DUKARE TUKARAM.S[1], Mr. Chavan Dharmaraj .D[2]
Lecturer , Computer Department, M. S. Polytechnic Beed, Maharashtra, India.

ABSTRACT

big data is term used for collection of data set which are large and complex which contain structured semi-structured and unstructured and complex type of data. data comes from everywhere from difernt department ,social media sites digital video pictures etc. this data is known as big data and this data we have to extracted from data mining tool. Data mining is technic for discovering intersecting pattern as well as descriptive understandable models from large scale data. big data is term used to identify the dataset that due to their large size and complexity we cannot manage them with our current methodologies or data mining tools.in this paper the overview of big data and future scope of the data mining and big data.

Keyword: *big -data, data mining tool, Hace theorem, privacy, 3vs*

1. Introduction

In todays era the drastic changes in increasing the collecting data from various sensors, devices, in different formats, from independent or connected applications. One of research found that web pages which is indexed by he google which is around 1million in 1998 bu reached 1 billion in 2000 and to dyas it may crossed 1 trillion this rapid changes in he field of data and the expansion of the drastic increase in the accepting the social networking application such as twitter, facebook etc. that allow user to view amplify ythe content of that to free actually web volume is the term of big data which appeared every day for various transaction like google which has more than1 billion queries on a per day.twitter which has 250million w queries per day and facebook has 800 million updates per day. Youtube have 4 million views per day. The data generated now is very large and produce every day very lagre format like zettabytes and this data is generating day by day.growing data in the form of big companies like apple facebook twitter yahoo etc.and this different format of data we have o carefully mainted and improve their performance to extracting this data by different pattern and anlysze it.

1.1TYPES OF BIG DATA AND RESOURCES

Thereare two types of big data: structured and unstructured, semi-strucured. Structured data are numbers and words that can be easily categorized and analyzed. s. Structured data conrast with unstructured and semi-strucred and transaction data. Unstructured data being the least formatted .and structured data have most formatted data .. These data can not easily be separated into categories or analyzed numerically .analysis of umstrured data mainly depend upon the keyword that is used for users to filter required data depending on the searchable data.. 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.

1.2

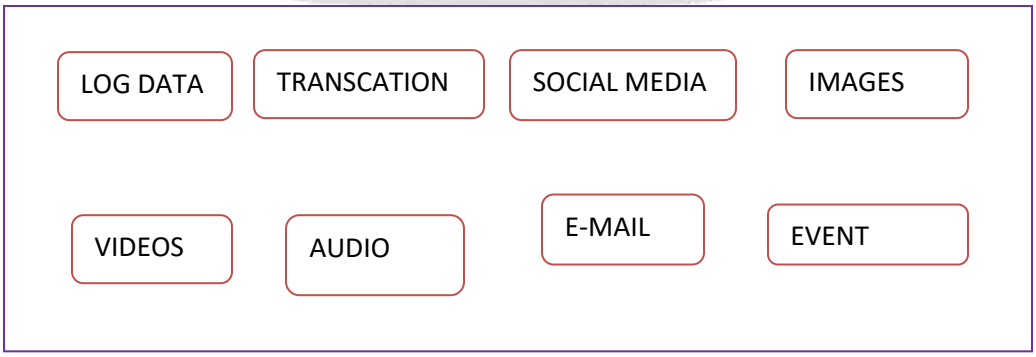

Figure 1.2

1 HACE THEREM

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 types of data for exporing requires more challengef rom big data while retrieving the Big Data. In a naïve sense, actually we can imagine that a number of blind men who are trying size up a giant Camel,. that the Big Data in this context. The goal of every blind man have to draw a picture of the Camel according to the part of information he collects during the process.. To make the problem even more complicated, let us assume that the camel is growing rapidly and its pose changes constantly, that blind man have hos own kowldge abut the came and he should his information so that he can umderstand.means hat one blind person can understand another blind persons information.acually his he big data term which is to be scenario which can implement o to aggregate the hetrogenus information different sources of data to help a drwing a best picture o understand the gesture of camel on real-time fashion.so that its not easy like blind man can share their views abut the camel .so that big data is most important term that where the volume of data also we have consider. HACE theorem suggests that the key characteristics of the Big Data are A. the main big daya issue is he herogenus data which is diverse and complex data which nis present in large volume the big data which is comes from different sources like twitter, orkut ,linklidin myspace etc.

B. Decentralized control:-

The decentralized control sysem which is controlling the autonomus data sources which distributed like www so that they can acces like web browser which can generate and collect information wihou depending in cengtralized control.

C. Complex data and knowledge associations

While he complex data and association the data od multi structure and multi source is complex type of dataand that complex type of data used in bills of materials ,maps,images,video, etc. so that such type of complex data should be required.

Big Data Management

Volume:

the concept of dara volume means that the amount of data which also reffered as big data.voulme refers to mass quaniy of data that organization want to for improving the decision makong in large enterprise.this type of data coninusively increasing.

Variety:

The varitey of different ypes of data and data sources which is multiple types which include structured data,semi-structured, unstructured data organization need to analyze ,integrate he data from complex array

.

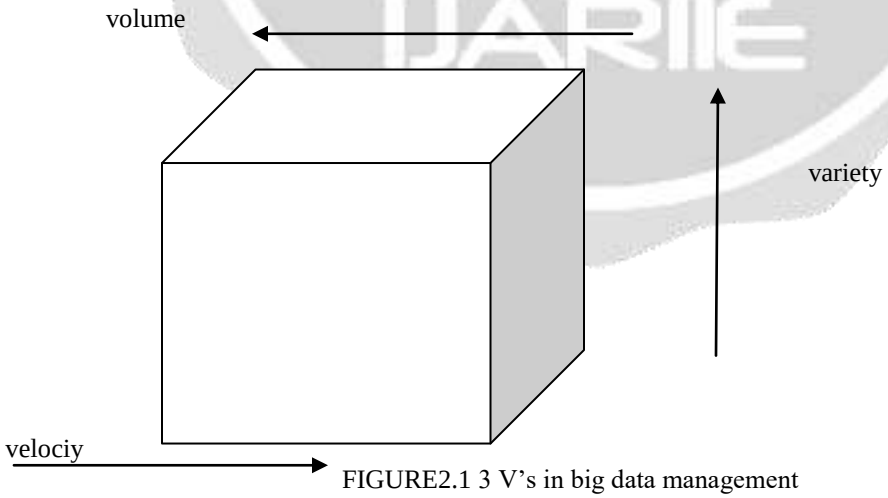

FIGURE2.1 3 V's in big data management

With the explosion of sensors, smart devices and social collaboration technologies, Velocity: Data in motion. The speed at which data is created, processed and analyzed continues to accelerate .Nowadays there are two more Variability:- There are changes in the structure of the data and how users want to interpret that data

.

2 DATA MINING FOR BIF DATA

Generally, data mining is the term or process for analysis of data from different perspectives and summary of that data it into useful information - information that can be used to increase revenue, cuts costs, or both. The term data mining is the process for finding he co-relations or patterns beween dozens of fields in large relational database management system .Data mining is term can b e classified ino six or tasks as follows:

1. Classification of data mining
2. Estimation of data mining
3. Prediction of data mining
4. Association rules for data mining
5. Clustering while mining
6. Description abut data mining

A. Classification

Classification is a process of which to generalize the data according to different instances .according classification the alogorithms which can be useful while data mining is decision tree,, bayes algorihm etc.

B. Estimation

Estimation which can be continuously valued outcomes. While some of the input data values we use estimation value for which is unknown variable used such as income, height or credit card balance.

C. Prediction

Its a statement about the way things will happen in the future , often but not always based on experience or knowledge. rediction may be a statement in which some outcome is expected.

D .Association Rules

An association used for depending upon different types of relationship to established between object in that database.

E .Clustering

Clistering is very important concept which learning principle of unsupervised; in that clustering to find the structure ofor un-labelled data .

Comparision beween big data and data mining

TABLE-1

Big data	Data mining
Big data is a term for large type of data set	Data mining refers to the activity of going through big data type of set to look for relevant information while mining.
Big data is the asset	Data mining is to used for handler for the purpose of beneficial result.
Big data which to be depend upon the capacity of that organization sets, and applications that are traditionally used to process and analyze data	Data mining refers to the operation that involve relatively sophisticated search operation

4. CHALLENGES IN BIG DATA AND APPLICAION

In todays era the big data is most challenges and difficult. The volume of data is already enormous and increasing very day. In todays generation the use of internet connecting device increased largely . Furthermore, the variety of data being generated is also expanding, and organization‟s aablity to capture whatever data required and process that data.and the technology should match with the architecture management and analysis of required data a, and organizations will need to change the way they think about, plan, govern ,manage, process and report on data to ealize the potential of big data.

A. Privacy, security and trust.

For the purpose of security and privacy some government have implement their privacy and security laws.

Privacy Act (through the passing of the Privacy Amendment (Enhancing Privacy Protection) Bill 2012)to implement privacy and protection laws for personal information which is collected by he government agencies., when

collecting or managing citizens data, are subject to arange of legislative controls, and must omply with the a number of acts and regulations such the Freedom of Information Act (1982), the Archives Act(1983), the Telecommunications Act (1997) ,the electronic ranscion act in 1999 and he intellgence services act 2001 act which to be legislate or designed for the public confidence and heir data will be confendential and secured. maintain public confidence in the government as an effective and secure repository and steward of citizen information. So that government agency should not change the data of customers . 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 relese of readable format data or any sensible data should be secured as government responsibility to care all abut he sensible information... The potential value of big data is afunction of the number of relevant, disparate datasets that can be linked and analysed to reveal new patterns, trends and insights. public have to.trust to government before sharing information because the govenmen cal link also heir information to oher department of respective government so that privacy and shold checked.

B. Data management and sharing

Accessible information is the lifeblood of a robust democracy and a productive economy.2 Government agencies realize that for data to have any value it needs to be discoverable, accessible and usable, and the significance of thee quirements 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, utilized 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 viasta ndardised APIs, format sand metadata. If he he data qualiy will improve that willn affect overa all performance of database and tah will easy to used for business analysis decision making cost saving and improving he producviy of compannies
seen a focus on making data sets available to the public, however these „the open data available also useful for data sharing and standerrised data between agencies so thag communicate iner governmental use and collaborate and privacy also apply their.

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 he any government agency have he business analyst which can analysis the analysis of data ,processing he data and archiving the data.so that the government agency can manage their requirement .

5. CONCLUSIONS

Big data is the term for a collection of complex data sets, Data mining is an analytic process designed to explore data big data which o be continue growing during the next years so that data scientist have manage the blarge amoun of data every year. and that type of of data have large diverse and faster. To support Big data mining, high-performance computing platforms are required,. 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. big datais becoming the new final frontier for scientific data researchand data application.

6. 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 .