Big Data Mining: Problems and Prospects

Nayani Sateesh
CVR College of Engineering/ Information Technology, Hyderabad, India
Email: nayanisateesh@gmail.com

***Abstract* - Nowadays decision making is going towards data centric and the data is becoming crucial in understanding the user needs and behavior so as to enhance the services. As the communication is going over the internet, lots and lots of data is being generated like in social networks, blogs etc., which is to be managed and analyzed properly. Big Data is becoming more prominent nowadays in the data management and processing. Big data deals with huge amount of data which is large in size, heterogeneous in nature, complex to process. In this paper I would like to review various problems in mining the Big Data and its applications in various sectors etc.**

***Index Terms*— Big Data, Data Mining, KDD.**

I. INTRODUCTION

From the recent past, we are hearing the most prominent topic or area is the big data in the information and communication technology (ICT) world. Because of data centric computing in understanding the user's behavior and their trend in the usage of services, Big Data is playing a vital role to gain the competitive edge in the market and to reach and increase the global customer base. In Big Data Analysis understanding the nature, significance of data and data visualization [1, 2] is more important. "Big Data" describes data sets so large and complex they are impractical to manage with traditional software tools. It relates to data generation, data storage, data retrieval and analysis of data that is remarkable in terms of size, type, and rate in which data being generated or stored.

II. BIG DATA: TREND AND CHARACTERISTICS

As per NESSI forum big data is defined [3] as "Big Data" is a term encompassing the use of techniques to capture, process, analyze and visualize potentially large datasets in a reasonable timeframe not accessible to standard IT technologies. By extension, the platform, tools and software used for this purpose are collectively called "Big Data technologies".

A. Data Trend

As per Oracle's review, data is being growing at 40% annual rate from the last few years and reaching by 45 ZB by 2020. The trend shows that the volume of business data significantly grows every year which emphasis the need for analysis of the data being generated to know the insights of the data to get the competitive edge in the market to showcase their products and service in an effective way to the users and hence to stand in the top place in the competitive global market.

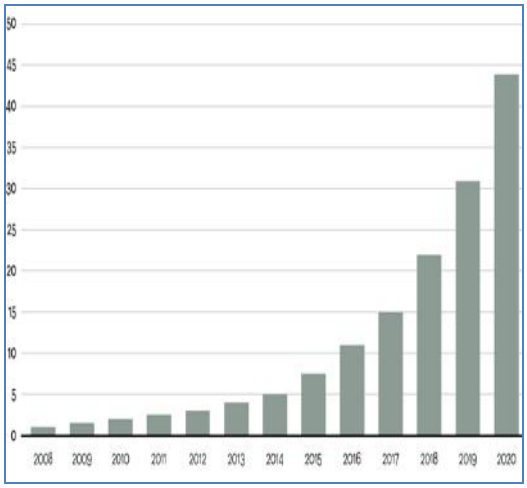


Figure 1: Data Growth (2008-2020)
Source : Oracle 2012 (Data in Zetta Bytes)

A. Characteristics

Characteristics of BIG Data can be usually called as 5 V's [4, 5]. They can be classified as Primary and Secondary characteristics based on their significance in Data Analytics.

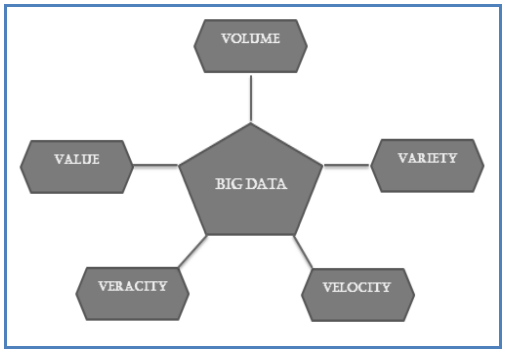


Figure 2: Characeristics of Big Data (5V's)

Primary Characteristics – 3V's (Volume, Variety, Velocity)
Secondary Characteristics – 2V's (Veracity, Value)
Volume: It reference to the size of the data being stored. The data size can be terabytes or petabytes or even more.

Variety: it refers to the structure or type of the data being stored. They include unstructured, semi-structured and structured. Examples of such variety are audio, video, xml, sensor data, text files etc.
Velocity: it refers to data rate at which data is being generated and stored into the databases.
Veracity: It refers to trustworthiness of the data being stored
Value: It refers to insight of the data being extracted from the stored data is useful and which will extend the business in terms of returns and market share [6].

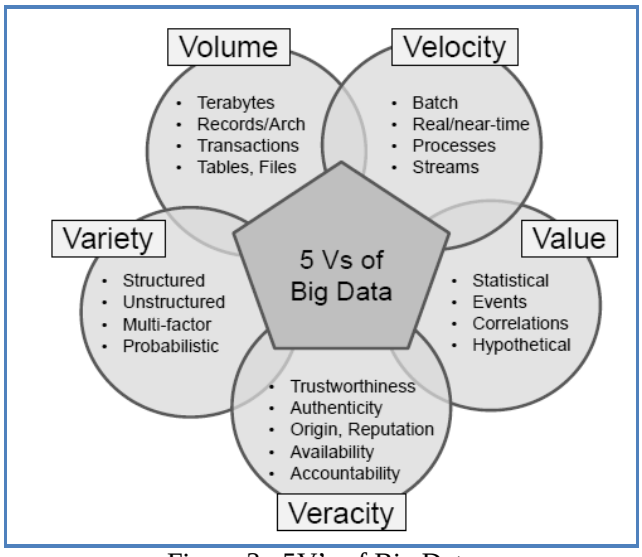


Figure 3: 5V's of Big Data

III – DATA MINING vs. BIG DATA MINING

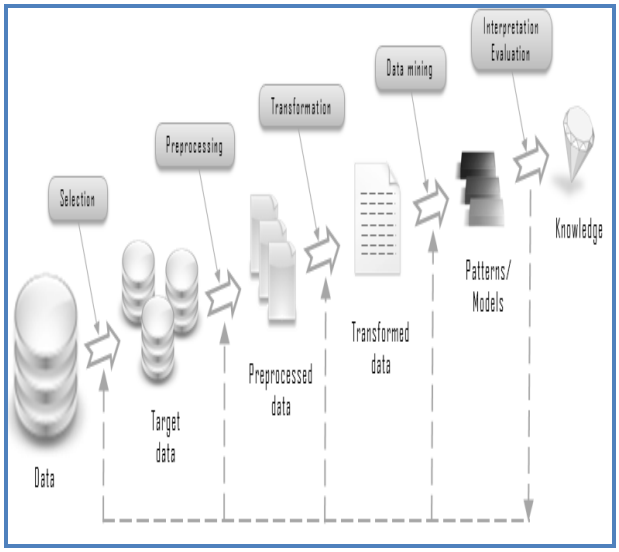
*Big Data for the Enterprise:*With Big Data databases, in any vertical enterprises can save money, achieve many other business objectives, grow revenue.

Using Big Data organizations can do the following:

Build new applications: Helps the organization to optimize the real data and build new applications to analyze and reuse the data
Improve effectiveness and lower the cost of existing applications: Many big data technologies are open source technology based; they can be implemented at low cost than proprietary technologies.
Realize competitive advantage: Big data can help businesses to act quickly to adapt to changes faster than their competitors.
Increase customer loyalty: Increasing the amount of data shared within the organization – and the speed with which it is updated – allows businesses and other organizations to more rapidly and accurately respond to customer demand.

A. *Data Mining*

According to Simoudis (1996) - Data mining can be defined as process of extracting previously unknown, comprehensible and actionable information from large databases and using it to make crucial business decisions. Data from various sources is integrated and then data is transformed to standard format after preprocessing. Using Data mining Algorithms interesting patterns and rules are extracted. The patterns and rules that are extracted are interpreted into the useful information which increases the knowledge of the end users.



A. *Mining Big Data – Problems*

Whatever the knowledge discovery process that we discussed above is also applicable for the Big Data also. Here in this section we will look on the possible problems [4, 7, 8, 9, 10, 11] that we face during the Big Data mining process with reference to knowledge discovery process.

a. *Data integration:* Since we are integrating the data from various sources, there exists a possibility of integrating unstructured, semi-structured and structured data. Data to be mined in big data will be huge in terms of terabytes, petabytes etc. This is one problem for the integration tools to integrate high volume and variety of data. Reliable and High quality data should be integrated and transformed to standard form.
b. *Preprocessing:* Since large data to be pre-processed due to Velocity of Big Data, the pre-processing tools or applications should be able to process the data in minimal time and to convert into standard form. Data scaling should be handled properly. Since large data is being stored and taken for the pre-processing, reliable and accurate data only to be taken.
c. *Data mining:* data mining algorithms should be able to process the huge amount of data and Variety of the data should be handled appropriately to extract the interesting patterns in minimal time. Here

10.32377/cvrjst0603

sophisticated data mining algorithms are needed which will process the data at faster rate. Also people working on this domain should be highly skilled in processing and extracting interesting patterns.

d. *Evaluation and Interpretation:* The patterns should be evaluated effectively and should be presented visually which will increase the end user knowledge. Presenting the data in visual way in an effective manner and the people involved in the processing and presentation of the visualized data is an important factor in mining big data. Appropriately visualized data will provide the value to the data being mined and helps in increasing the business market. Visualization tools are more important in presenting the data patterns. Visualization tools that we are using should be appropriate and capable of handling and processing high loads of data.

B. Mining Big Data – Prospects & Applications

With the invent of internet, services nowadays are provided over the internet and with the social networking activities on day to day , large amount of data is being generated and stored in in the data bases in various formats. Integrating such data to understand the insights of the data and hence to understand the needs of the user and the trends in usage, Big data mining is playing a vital role.

Nowadays various organizations investing huge amounts in storing and processing of historical and real-time data as the decision making process is going towards the data centric and accurate decisions need to be taken based on the historical and real time data which is available. Lot of research work is going on towards data integration and effective visualization of qualitative patterns to be extracted.

Since Big data mining is an interdisciplinary, it has significance in various sectors in decision making process. It is generating huge amount of business returns [12].

The following figure illustrate the trend in Big Data market forecast

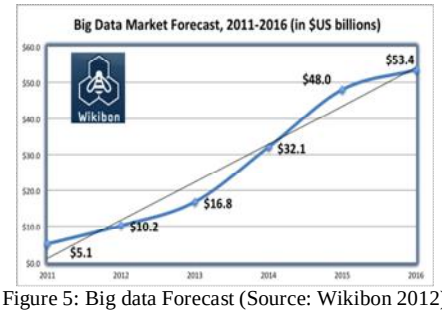


Figure 5: Big data Forecast (Source: Wikibon 2012)

Some of the applications [7, 8, 13, 14] of big data mining are:

1. *Social networking* – Based on user access patterns and data being used and generated, User Profiles can be analyzed effectively and the web pages can be personalized.
2. *Manufacturing*–Product modeling, quality and performance of the products can be analyzed and hence to customize the product to meet the customer needs.
3. *Weather forecasting* – Time series data can be analyzed effectively which helps in weather forecasting so as to alert the people and governments in case of natural disasters and other calamities.
4. *Marketing Research* – Product Reviews by the customers can be analyzed which helps in Sales Promotion, pricing and increasing the sales. It also, helps in understanding user buying behavior and sentiment analysis
5. *Advertising & Brand Promotion* – Designing the advertisements based on perspective of product performance and promoting the brands to retain the customers and also to create new customer base.
6. *BioInformatics*–Biological and genetic data like DNA sequence, protein sequence etc. can be analyzed in classification of species.
7. *Financial Data Analysis* – To detect the uncertainty and fraud in the financial data
8. *Healthcare* – To understand various vital statistics to classify the patients and disease level. Also helps in understanding the effectiveness of the drugs being given to the patients in curing the disease.
9. *Government & Political* - Evaluate the policies being implemented and the trend of the resources and budget being utilized can be analyzed.
10. *Retailing* – Managing the customer relationships and understating the user's needs and pricing modeling can be done.
11. *Energy* – Operation modeling can be done.
12. *Media & Telecommunications* – Network Optimization and fraud detection can be done.

CONCLUSIONS

During the next few years also Big Data is going to grow and continue since the need of the historical data and the management of real-time data which is generating in huge volume, which helps to understand the insights of the data and the user requirements to enhance the products and services so as to retain their space by the organizations in the global market since decision making is becoming data centric. Efficient Analytical Architecture is needed to handle the various problems that we discussed in this paper and the proper management of the variety of data being stored and the efficient visualization tools are also needed to improve

the visual appeal of the data insights which will improve the knowledge being mined. Big Data is becoming the new Final Frontier for scientific data research and for business applications and contributing to data science research.

FUTURE DIRECTIONS

Business Analytics and Big Data Tools need to be reviewed to get further insight into the research areas of Mining and visualizing Big Data.

REFERENCES

[1] Big Data Visualization: Turning Big Data Into Big Insights - A White Paper, Intel IT Center, Intel, March 2013

[2] Dennis McCafferty,"Data Visualization: Making Sense of Big Data", Baseline Magazine, April 2014.

[3] Big Data: A New World of Opportunities, NESSI White Paper, December 2012.

[4] Mrs. Deepali Kishor Jadhav, "Big Data: The New Challenges in Data Mining", International Journal of Innovative Research in Computer Science & Technology, Vol 1, Issue 2, Sept 2013, pp. 39-42.

[5] Yuri Demchenko, "Defining the Big Data Architecture Framework", University of Amsterdam, July 2013.

[6] Zhong Li, "Harness Big Data Value and Empower Customer Experience Transformation", Infosys Labs Briefings, Vol. 11, 2013, pp. 27-34.

[7] Katina Michael, Keith W. Miller , "Big Data: New Opportunities and New Challenges", IEEE Computer Society, Vol. 46, Issue No. 6, June 2013, pp. 22-24.

[8] Alexandros Labrinidis, H. V. Jagadish," Challenges and Opportunities with Big Data", Proceedings of the VLDB Endowment, Vol. 5, No. 12, pp. 2032- 2033.

[9] Big Data Strategy — Issues Paper, © Commonwealth of Australia, 2013.

[10] Divyakant Agrawal, et.al "Challenges and Opportunities with Big Data" – A community white paper, United States, March 2012.
http://cra.org/ccc/docs/init/bigdatawhitepaper.pdf

[11] E. Letouzé ,"Big Data for Development: Opportunities & Challenges", May 2011.

[12] The Emerging Big Returns on Big Data - A TCS Global Trend Study, 2013.

[13] Wei Fan, Albert Bifet, "Mining Big Data: Current Status, and Forecast to the Future", SIGKDD Explorations, Volume 14, Issue 2, pp. 1-5.

[14] J. Manyika, M. Chui, B. Brown, J. Bughin, R.Dobbs, C. Roxburgh, and A. H. Byers. "Big data: The next frontier for innovation, competition, and productivity",McKinsey Global Institute,May 2011.

[15] Chris Snijders, Uwe Matzat, Ulf-Dietrich Reips, "Big Data: Big Gaps of Knowledge in the Field of Internet Science", International Journal of Internet Science, Vol. 7, 2012, pp. 1-5.

10.32377/cvrjst0603