ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 7, July 2017

Analysis of Big Data in the Data Mining

S.Sathiyakala Jayakanthan

Assistant Professor, Department of Computer Applications, Sri Akilandeswari Women's College – Vandavasi, India

ABSTRACT: Big data include huge volume with multiple evolutions of data sets, complexity on the independent sources. In the field of science and engineering, biological and biomedical sciences the big data are rapidly expanded. The capability of extracting useful information from large set of data is referred as big data which is not possible in earlier days. Data Mining is the term of exploring and analyzing the large quantity of data to locate different molds for big data. Analysis of big data is a troublesome one. It often involves the collection and storage of mixed data depends on different patterns and rules. The big data becomes one of the most stimulating concepts for the subsequent years. This paper consists of big data with data mining, issues on big data with data mining and related works on the big data.

I. INTRODUCTION

The advanced technology gave the rapid growth in storing of data which is implied as digitization. The amount of storing data can be structured, semi structured and unstructured data so called Big data. The word 'big' here means the data with larger size and with collected set that makes complexity. 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. The big data has become expandable in all science and engineering domains. Data Mining focuses on relevant and unique information from large set of data (Big Data). The most important target of this paper is to offer the readers with the information about data mining with big data, challenging issues on it and its related works. The big data are nowadays plays a challenging role in social medias.

For example: In twitter any public or general topic messages are tweeted by millions of people within an hour and the best comments are to be top listed and feedbacks are to be generated in real time.

II. BIG DATA

Big data is a term of massive collection of information or huge volume of data with different types from distributed sources that pursues very complex data. It also provides the various relationships among data. This pays the way to determine the knowledge from the big data.

Big data are classified into two types. They are:

1. Structured data sets.
2. Unstructured data sets.

STRUCTURED DATA SETS

The structured data sets are which the collection of data that forms a relationships among them. The information can be represented in rows and columns. The relationship implies that these types of data are grouped using one similarity. The best example is information of an organization.

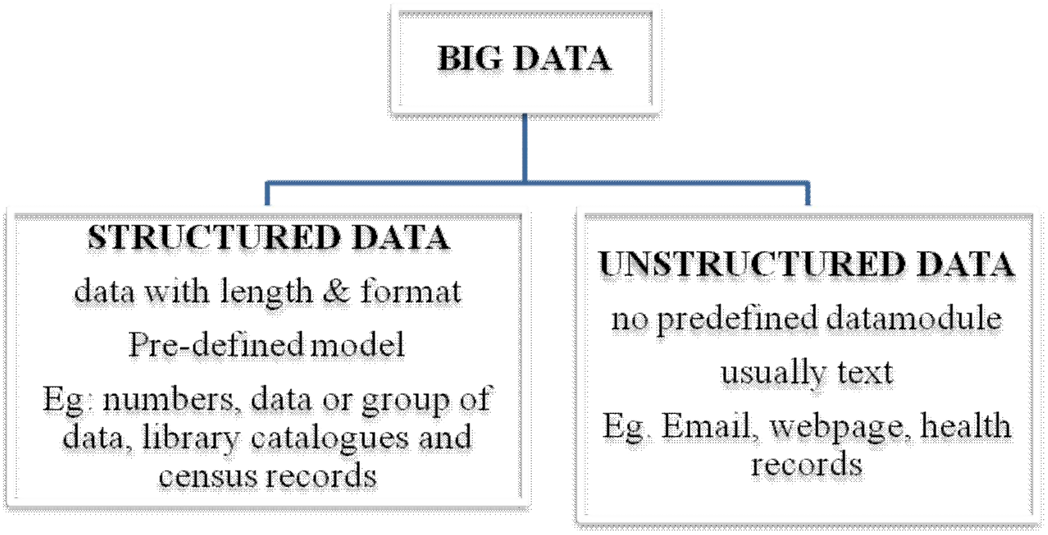
SEMI STRUCTURED DATA

The semi structured data does not provide a fixed schema. The information provided here are heterogeneous.

UNSTRUCTURED DATA

Unstructured data includes the very complex information. This type of data cannot be indexed easily for analyzing or querying. The unstructured data can be on the format image files, audio files, video files and complex health records so on.

DOI: 10.15680/IJIRCCE.2017. 0507021


BIG DATA
STRUCTURED DATA
data with length & format
Pre-defined model
Eg: numbers, data or group of data, library catalogues and census records
UNSTRUCTURED DATA
no predefined datamodule
usually text
Eg. Email, webpage, health records

III. THREE V'S ON BIG DATA

According to Doug Laney, the big data management is defined into three v' s. They are

1. volume
2. variety
3. velocity

VOLUME:

The storing of data are in the huge manner. It refers to the amount of data that are stored. Volume is the most important aspect in dealing with big data. It is an expandable one. The data stored can be extended then and there required.

VELOCITY:

The velocity refers to the speed on which the data are circulated. It is just like that of status on whatsapp, timeline on facebook, comments on messages and various updates. It implies the data movement on real time and their updations in fractions of seconds.

VARIETY:

The variety in name itself implies different, same that data are with multiple or different data types: The data can be of structured or unstructured type. Unstructured data can be short message service(SMS), audio files, media files, conversation in social media, portable document format and social sensor data.

IV EXTENSION OF V' S IN BIG DATA

Apart from the 3 v's given by Doug Laney, nowadays it is extended into some more

1. Value
2. Veracity
3. Variability

VALUE:

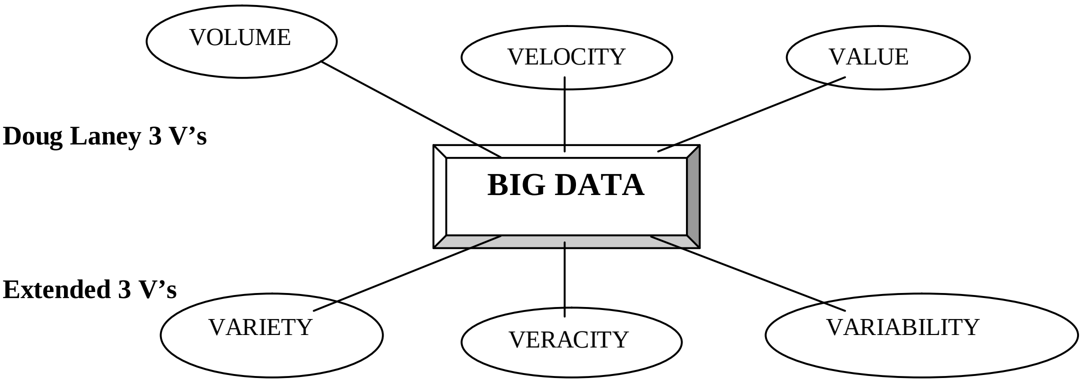
This includes the result of implementation of big data in an organization. simply the profit gained by an organization after the invest of big data.

VERACITY:

Many large volume of data, lack in their accuracy and quality. But veracity is available in big data. It gives truefull information and repetitions of informations are cleaned.

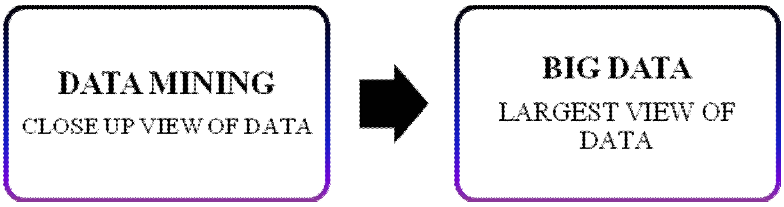
VARIABILITY OR COMPLEXITY:

Variability refers variations in the data flow rates. It imposes a critical challenge, the need to connect, match, cleanse and transform data received from different sources.



V. DATA MINING WITH BIG DATA

Data Mining is the gaining information or knowledge from big data. Big data is the collection of large volume of complex data in which focusing on appropriate data is meant by data mining.



Mining refers to finding exact information or knowledge gained from the collected data.

VI. KEY FEATURES OF BIG DATA

The salient features of big data are given below.

- Contains large volume of data size.
- Data sources are from various phases.
- Time to time data get changed
- It is free from control of anyone.
- Difficult to handle data and more complex in nature.

DOI: 10.15680/IJIRCCE.2017. 0507021

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 7, July 2017

VII. CHALLENGING ISSUES IN DATA MINING WITH BIG DATA

- Big data stored are scattered in different places and so on data volumes are get increased continuously.
- It is important to achieve unique statistical results and should not have any duplicate information.
- Time complexity is more on the big data.
- Storing of large volume of data required huge memory capacity.

VIII. CONCLUSION

Increased usage of huge data then big data may get increased for forth coming years. It provides the final borders for researchers and business applications. The important challenge on big data structure is to concentrate more on complicated interaction between data sources and developing changes by time to time on the models. High performance computing platforms are necessary to manage big data.

REFERENCES

1. Labrinidis and H. Jagadish, “Challenges and Opportunities with Big Data,” Proc. VLDB Endowment, vol. 5, no. 12, 2032-2033, 2012.
2. Y. Lindell and B. Pinkas, “Privacy Preserving Data Mining,” J. Cryptology, vol. 15, no. 3, pp. 177-206, 2002.
3. Alex Berson and Stephen J.Smith Data Warehousing,Data Mining and OLAP edition 2010.
4. Wei Fan and Albert Bifet “ Mining Big Data:Current Status and Forecast to the Future”,Vol 14,Issue 2,2013
5. E.Y. Chang, H. Bai, and K. Zhu, Parallel Algorithms for Mining Large-Scale Rich-Media Data,Proc. 17th ACM Int’l Conf. Multi-media, (MM ’09,) pp. 917-918, 2009.
6. Rajaraman and Ullman, 2011, A. Rajaraman and J. Ullman, Mining of Massive Datasets, Cambridge University Press, 2011.
7. Banerjee and Agarwal 2012, Soumya Banerjee, Nitin Agarwal, Analyzing collective behavior from blogs using swarm intelligence, Knowledge and Information Systems, December 2012, Volume 33, Issue 3, pp 523-547.
8. Silva et al. 2012, Alzennyr da Silva, Raja Chiky, Georges Hébrail, A clustering approach for sampling data streams in sensor networks, Knowledge and Information Systems, July 2012, Volume 32, Issue 1, pp 1-23.