ISSN: 2321-7782 (Online)
Volume 3, Issue 8, August 2015
International Journal of Advance Research in Computer Science and Management Studies
Research Article / Survey Paper / Case Study
Available online at: www.ijarcsms.com

An Overview of Big Data, Issues and Challenges

Pallvi V. Dongardive[1]
M.E 2nd Sem Department Computer Science & Engg.
Anuradha Engg. College, Chikhli
Maharashtra, INDIA

Prof. N. K. Bhil[2]
Department of Computer Science & Engg.
Anuradha Engg. College, Chikhli
Maharashtra, INDIA

Abstract: Innovations in technology, development of networking, greater affordability of digital devices and more use of internet have presided over today's Age of Big Data, big data refers to the explosion in terms of quantity, variety and diversity with high velocity digital data. Turning this big data collected from various sources like mobile-banking transactions, Call logs, online user-generated content such as tweets and blogs, satellite images etc. into an actionable form that can be for future sense refers as big data mining. But extracting knowledge from big data is not possible as huge form of complex data is generating. This paper gives overview of big data, its features, applications and address broad challenging issues while dealing with big data mining.

I. INTRODUCTION

Data is the key factor today that includes professional, personal, social data and more types of data. As this is the digital universe as more of the world goes online, including the physical world and the use of different types of media this interconnectivity and the digitalization lead to an unexpected growth of data. this enormous form of data that is created from Social networking sites, search and retrieval engines, media sharing sites, stock trading sites, and news sources continuously increases the data day- by-day. This collection of large volumes of data refers to as Big Data .Big data nothing but available at heterogeneous, autonomous sources, in extremely huge amount, which get updated in fractions of seconds of time . consider an example of famous social networking site facebook as most of us regularly use it, we upload various information and photos that is stored at the data warehouses at the server of the facebook this is nothing but the big data, a recent study estimated that every minute Facebook has more than 800 million updates per day, YouTube has more than 4 billion views per day. Twitter has more than 250 million tweets per day, some sources are not obvious consider huge amount of data gathered from meteorological and climate systems, or data of patients from patient monitoring systems in the hospital, control systems and acquisition of data from cars, airplanes, health care industry is Inundated with data from patient records alone cell towers, and power plants, all collect

Unending streams of data, Insurance companies collect data after every claim; Health care industry is inundated with data from patient records alone. The data produced now a days is estimated in the order of zettabytes, and it is growing around 40% every year[6].as Consider another example this is an era of google ,thing which we want to know we google it, within a fractions of seconds as the result of query we get number of links this is an processing big data. Big Data concern large-volume, complex, growing datasets with multiple data sources With the fast development of networking, data storage and data collection capacity, big data are now expanding in all science and engineering domains, including physical, biological and biomedical sciences .Data mining uses this data to uncover the hidden valuable information.

II. BIG DATA

Big Data [4],[1] concerns large-volume, complex, growing data sets with multiple, autonomous sources. Generally big data refers to a collection of large volumes of data and these data are generated from various sources such as internet, social media, business organizations etc., consider an example of flickr a public photos consider the size of each photo is 2 megabytes (MB),

which resulted in 3.6 terabytes (TB) storage every single day. As "a picture is worth a thousand words", the billions of photos on Flicker are a treasure tank for us to explore the human society, social events, public affairs, disasters etc., only if we have the power to harness the enormous amount of data. sharing site ,which received, on average 1.8 million pictures per day, this gives an example of big data which increases data collection which grows tremendously and is beyond the ability of commonly used software tools to capture, manage, and process within a "tolerable elapsed time". Different types of data such as Social data – Customer feedback forms for Customer Relationship Management (CRM) in Social media sites such as Twitter, Face book, LinkedIn etc. Machine-generated data in sensor readings, satellite communication, Traditional enterprise data such as and ledger information, Employee information, customer information etc are referred as big data[8].

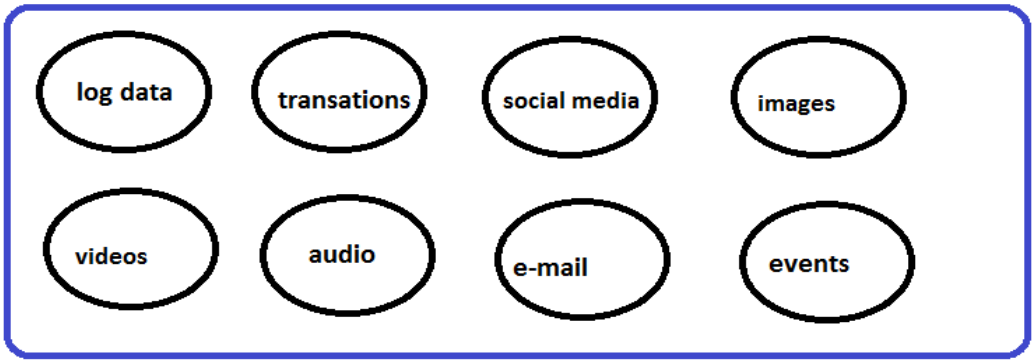

FIG 1 :Sources of big data

Characteristics of Big Data include 4 Vs. It contains Volume, Velocity, Variety and Veracity.

A. *volume*

Volume is the term used for vast amount of data generated in fractions of seconds. This volume data is in the form of rest state. An example of volume data is machine generated data. Now days volume of data enlarging exponentially.

B. *velocity or speed*

Velocity is the speed at which data generated. The streaming data may not be massive and its state is in motion. It should have high speed data. Social networking site is the best example of high velocity or speed data.

C. *Variety*

It is an important factor of big data. Various different formats ,types structure of data are referred as variety of data .it contains variety of Text, numerical, images, audio, video, sequences, time series, social media data, multi-dim arrays, etc ,also includes static and streaming data..

D. *Veracity*

Veracity means data which contain doubt. Unclear, uncertain data is found due to inconsistency and data in incomplete state. Abbreviation, colloquial speech etc may result in data veracity

FIG: 2 Vs (volume, variety, velocity, veracity)

III. BIG DATA AND DATA MINING

Data mining is the term used to describe the process of extracting value from a database, data mining also known as data or knowledge discovery is the process which analyses data from different perspectives and discover useful information from it. With these data some useful information can be extracted with the help of data mining.It extracts interesting patterns and relationships hidden in a large volume of raw data, and the results tapped out may help make valuable predictions or future observations in the real world[6]. Big data includes structured, semi-structured, and unstructured data.

This unstructured or semi-structured data contains useful and valuable information that is hidden in it. Data mining is a technique for discovering interesting patterns as well as descriptive, understandable models from large scale data[2]. The goals of big data mining techniques go beyond fetching the requested information or even uncovering.

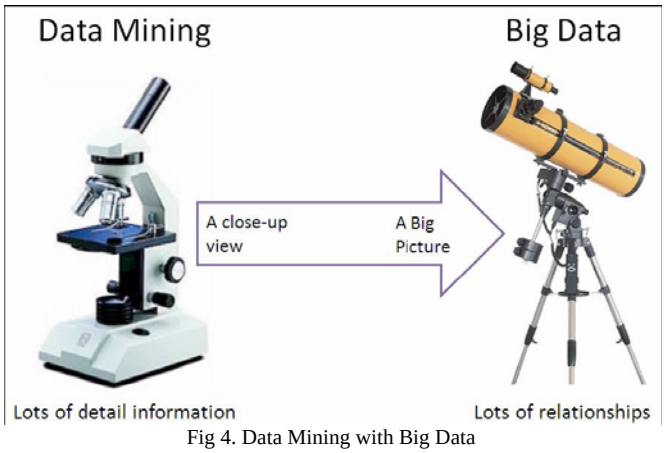

Fig 4. Data Mining with Big Data

The above figure 4 [2]relationship between big data with data mining where big data gives lots of relationship and data mining gives lots of information with closed view[2].

IV. KEY FEATURE OF BIG DATA HACK 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[1].this characteristics of big data makes it as extreme challenge for useful knowledge discovery. For above consider the scenario we can imagine that a number of blind men are trying to size up a giant elephant, where blind people are asked to draw the picture of an of the elephant according to the part of information each collected during the process. As each persons view is limited to his region each can think that trunk of elephant as a wall, leg as a tree, body as a wall and tail as a rope[2]. Problem can be make more complicated by assuming that

size of elephant growing rapidly and the pose also changing continually .Also blind men are exchanging information and learn on their respective feelings.

This information may be biased[1].So the results of each blind person's prediction is something different than actually what it is.

A. Vast data with heterogeneous and diverse sources

One of the fundamental characteristics of the Big Data is the huge volume of data represented by heterogeneous and diverse dimensionalities. Heterogeneity means same person having different representation and diverse means single information having variety of features Consider example in the biomedical filed a single human being is represented by various entities name, age, gender, family history etc., For X-ray and CT scan images and videos are used[2].

B. Autonomous with distributed and de-centralized control

This is the important characteristics of big data**.** Autonomous means information generated automatically without any central control. Like World Wide Web (WWW) as server provides information without depending on any other servers[1].

C. Complex and Evolving relationships

As size of data increases complexity and relationships of data also increases exponentially. Whenever there is small amount of data complexity is less but increasing data for example data from social sites and other networking sources complexity and evaluating relational knowledge is difficult [1].

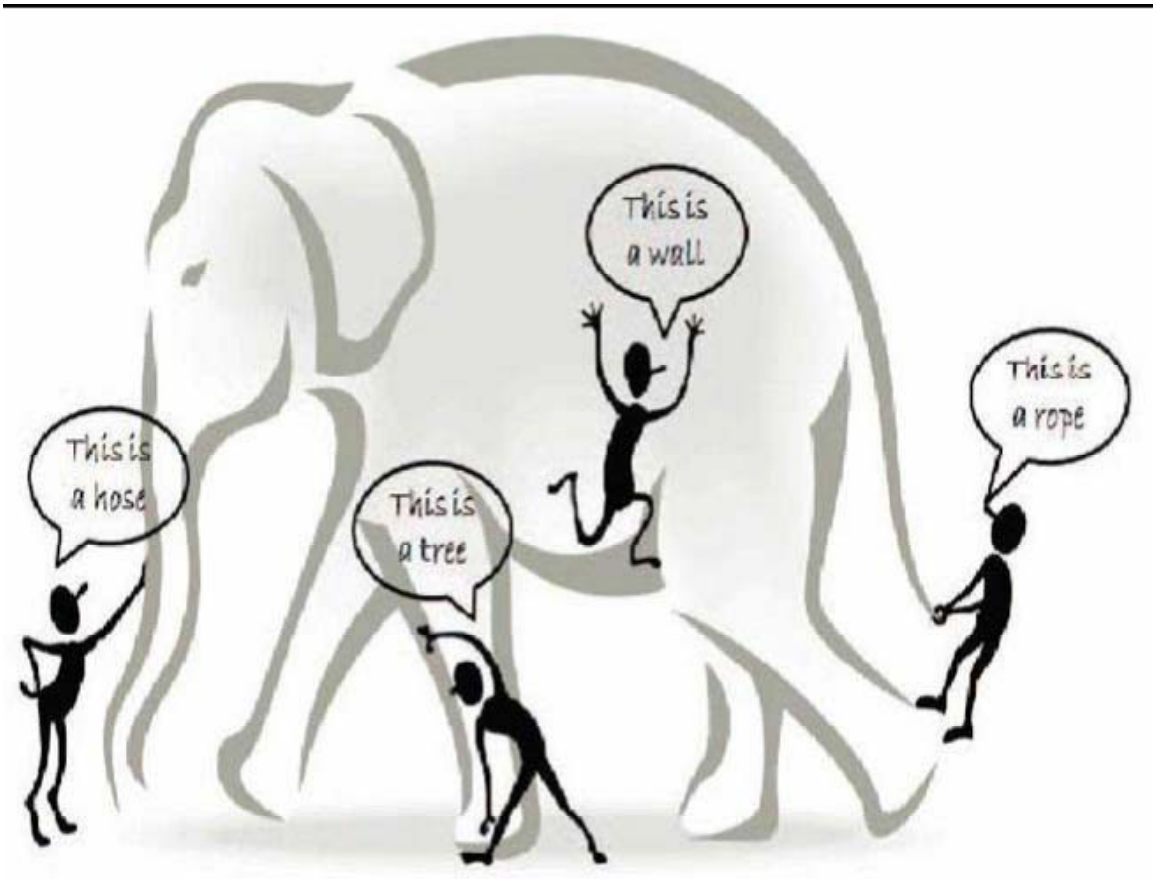

Fig 5. Blind men and the giant elephant

V. DATA MINING FOR BIG DATA

Data mining contains several algorithms which falls into four categories.

a) Association: is used to search relationship between variables. It is applied in searching for most frequent visited items establish relationship among objects. **Name of the Algorithm used for association is** Apriori, Partition, FP growth, ECLAT[2].

b) Clustering: it discovers groups i.e. data belongs to which group and structures in the data. Classification deals with associating an unknown structure to a known structure. Clustering uses K-Means, Expectation Maximization, DBSCAN, fuzzy C Means algorithms, clustering has attracted a significant amount of research attention in past decades.

c) Classification uses Decision Trees, C4.5, KNN, Naïve Bayes, SVM. Algorithms

d) Regression finds a function to model the data. Name of the Algorithm For regression is Multivariate Linear regression [2].

TABLE I
Differences between big data and data mining[3].

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
Bigdata 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,	Data mining refers to the operation that involve relatively sophisticated search operation
Finding interesting patterns	Involves large scale storage and processing of large data sets
Data size is smaller	Data size is Larger

VI. APPLICATION OF BIG DATA IN DATA MINING

While talking about applications of Big Data [8] we have to consider that Data mining techniques can be used to find the characteristics of object evaluation or the trend of changes for objects in the database. this information may be useful for making decisions and for planning strategy.Big Data mining offers opportunities to go beyond their relational databases to rely on less structured data such that data of weblogs, data from social media, data from email, sensors, and data from photographs that can be mined to extract useful information. Many business intelligence companies, such Oracle, IBM, Teradata etc., have all featured that their own products and helps customers acquire and organize these diverse data sources and coordinate with customers' existing data to find new insights and capitalize on hidden relationships.

Following are the various applications of Big Data [9]:

- In Astronomy
- In Sensor networks
- In Government data
- In Web logs
- In social networking sites to find for useful patterns/personal data
- In google search
- In Mobile phones
- In Scientific research
- In Natural disaster and resource management
- Health care
- Global personal location data (this is very common gives the rise of
- mobile devices)
- Manufacturing

VII. Challenges of Big Data Mining and Discovery

As the rate at which speed the data is increasing the volume of enormous data also increases now a days. Furthermore variety of data and the veracity of data expanding it is difficult to deal with this data since current architecture, latest technologies, big data analysis are unable to deal with this data.

Following are the issues related to data mining un big data[4].

A. *Variety of Data*

Variety refers to the different form of data. as there is unlimited forms of sources that generate a information and forms big data. This heterogeneous form of data leads to variety of data and extracting useful information from this type of big data is a challenge.

B. *Scalability of Data*

Scalability is at the core of upcoming technologies, to meet an issues come from big data .it require more scalability for undefined huge form of data for its data management and tool management tools. When data mining applies for big data that are centered to parallelism and scalability.

C. Security:

Cloud is used for large amount of data to be held and requires distributed processing across servers. As there is enormous increase of data more risk to the treat to the security of information.

ENISA found that various emerging treat arising for misuse of big data. ENISA says that volume of uncontrolled form of data, use and dissemination of user are good for malicious activities.

D. Data Discovery[10]

This is an important challenge that how to find out high quality knowledge that will be use for future scope which is out there on web.

E. *Reliability*

In traditional data mining systems are efficient and also reliable as there is limited and resources of data are well known. Enormous big data increases, data is not limited, countable and verified, this leads to big issue for reliability.

F. *Mining & cleaning of unused data*

As there is large environment of big data there is another issue that is presence of unused data. The unused data acquires most of the useful space of memory but to have a durable & sustainable Big data mining system, mining & cleaning of unused garbage data is very essential & recommended.

G. *Quality and Relevance*

For a particular issue to determine quality of data sets and relevance is a challenge for handling with the big data

VIII. CONCLUSION

In this paper, WE have reviewed some insights about big data and big data mining. Big Data can be literally explain by the HACE theorem we focuses on several issues related to big data and overcoming those issues will result in better environment for knowledge discovery that no one has discovered it before. Big Data is becoming the new Final Frontier for research related to scientific domain and for other business applications as it is having tremendous wealth of information. Thus we conclude that big data will become an excellent Opportunity in the forth coming years.

References

1. Xindong Wu, Xingquan Zhu, Gong Qing Wu,Wei Ding, „Data mining with Big data", IEEE, Volume 26, Issue 1, January 2014
2. Shobana.V, Maheshwari.S, Savithri.M "Study on Big data with Data Mining" April 2015
3. Rohit Pitre Vijay Kolekar " A survey paper on data mining wiyh big data" nternational Journal of Innovative Research in Advanced Engineering (IJIRAE)Volume 1Issue 1 (April 2014).
4. Nibedita choudhary ," Big Data and Big Data Mining: Study of Approaches, Issues and Future scope, International Journal of Engineering Trends and Technology(IJETT) –Volume 18 Number 5–Dec 2014
5. Bharti Takur ,Manish Mann,"data mining for big data : a review" may 2014
6. B R prakash ,Dr Hanumanthappa," issues and challenges in the era of big data mining" International Journal of Emerging Trends & Technology in Computer Science (IJETTCS) August 2014
7. Qiang Yang, Xindong Wu ,"10 challenging problems in data mining research
8. SMITHA T," Application of Big Data in Data Mining", International Journal of Emerging Technology and Advanced Engineering, July 2013
9. Richa Gupta ," Journey from Data Mining to Web Mining to Big Data "International Journal of Computer Trends and Technology (IJCTT) –volume 10 number 1–April 2014
10. Roberto V. Zicari ," Big Data: Challenges and Opportunities"

