Big data analytics vs Data Mining analytics

Vinti Parmar,
[1]Department of Computer Science,
Indira Gandhi University, Meerpur, Rewari
Haryana, INDIA

Itisha Gupta
Department of Computer Science,
Bright Riders School, Abu Dhabi, UAE
Itishagupta07@gmail.com

Abstract

Big data is collection of complex, large varieties of data set – structured, unstructured and semi structured data that is arriving continuously at high velocity and have capability for giving or revealing meaningful and valuable information. Unlimited and unknown number of sources likes sensors, cameras, internet, social media sites, satellite, email etc. are generating large volume of diversified data but it become a challenging opportunity to gain insight from data flood. Big data have the potential in untapping the resources for delivering insight that transform strategic and operations initiatives of organisation. This tsunami of big data is testing the ability of traditional data analysis tool that are not able to cope up with analysis of big data and extracting valuable insight cost effectively. This paper provides a discussion on why the traditional data mining tools cannot be used for big data mining and how the big data is different from data mining. In this paper we address various issues and challenges related to big data mining.

Keywords: big data, data mining, 3V's, data analytics, challenges.

1. Introduction

In 1998 first time name of big data appeared in silicon graphics slide (SGI) by John Mashey. Big data name appeared first time in title of an academic paper in 2000 by Diebold [1]. In 1999 Berkeley researchers also estimated production of 1.5 billion gigabytes of data and that amount got doubled in next 3 years. Origin of big data is due to fact that data comes from everywhere like internet, sensors, mobile devices, facebook, google etc. and is getting bigger day by day as large volume of data is generating every day. Even if you put your mobile phone in pocket, data is generated by network regarding location. Most of data comes from internet as internet usage is in boom now days. Google receive more than one billion queries per day. There are 800 million updates every day for facebook and 4 billion views per day for You Tube [1]. There is unprecedented growth in data size form petabytes to zettabytes. "Big data refers to high volume, high velocity and variable data that needed advanced tools and technologies for capturing, storing, managing and analysing data as it's beyond the capability of current data mining tools for capturing, storing and managing big data [2]. The volume, velocity and variety as mentioned in above definition are called 3V's of big data that characterize the big data.

- **Volume:** specifies quantity or amount of big data in petabyte, Exabyte, zettabyte and that volume is growing enormously every day. Petabyte era has almost gone and now we have to confront era of Exabyte. Google estimated that in 2010 data generated by world at every two days is same as sum of data generated in 2003 so it means data is growing at unprecedented scale.
- **Velocity:** specifies speed of data creation and processing and is accelerating continuously. Time is one of the important dimensions of big data that led to another feature of big data called velocity. Real time processing is must otherwise delayed information extracted from big data have no value but existing data mining tools are not suitable for real time processing of such huge amount of data.
- **Variety:** specifies that big data is collection of varieties of data like structured data, unstructured data and semi structured data from unlimited sources. Due to amalgamation of varieties of data more associations, correlations and patterns can be find out for decision making from big data.

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 [6].

So we conclude that big data is not all about volume or size of data even though size is foremost feature of big data. Also with each of V serious challenges are unfolded like heterogeneity, scalability, privacy, integration etc. that cannot be addresses by current data mining tools and techniques as they are designed to work with structured data only and not in real time but today's generation of unstructured data is more than structured one so existing data mining techniques must be incorporated with fast, novel distributed storage system and massive parallel processing architecture to deal with big data.

Types and Forms of Big data

There are three types of big data:

- **Structured data:** Structured data is in proper format and can be categorized and analysed easily. Sources of structured data generation are sensors, cameras, GPS devices .Financial data and transaction data also comes under that category.
- **Unstructured data:** It is not in format form so it's difficult to analyse and categorize them easily. Unstructured data have more complex information and provides better insight than structured one and it is the unstructured data that necessities the fast and novel tools and techniques as such data cannot be handled by existing data mining tools and techniques. Much of the big data comes from unstructured part as generation of unstructured data is more than structured data. Audio, video, pdf, email etc. are unstructured data.

- **Semi structured data:** It is between structured and unstructured data

This combination of varieties of data presents a lot of challenges for extracting knowledge form them. Big data comes in two forms:

- Big data at rest: It means analysis of data set will be done after its collection that is end of day, month etc. Until then data is stored in warehouse so it means no real time processing. [2]
- Big data in motion: It means processing of big data in real time just after its collection means time is a constraint over here. That is velocity matters a lot as timely information is valuable only so processing must be finished with in real time only. For example: in case of delayed information in stock market prediction, earthquake prediction result can be disastrous.

2. Related Work

Provides an overview of big data mining and discusses the related challenges and the new opportunities. The discussion includes a review of state-of-the-art frameworks and platforms for processing and managing big data as well as the efforts expected on big data mining. We address broad issues related to big data and/or big data mining, and point out opportunities and research topics as they shall duly flesh out [1]. “Big data” appears to have become a buzzword overnight. The term describes innovative techniques and technologies to capture, store, distribute, manage and analyse petabyte- or larger-sized datasets with high-velocity and diverse structures that conventional data management methods are incapable of handling [2]. Provides an overview of types of big data and challenges in big data for future. 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 [3]. Describes the journey of big data starting from data mining to web mining to big data. It discusses each of this method in brief and also provides their applications. It states the importance of mining big data today using fast and novel approaches [4]. The enormous growth in the amount of data that the global economy now generates has been well documented, but the magnitude of its potential impact to drive competitive advantage has not. It is my hope that this briefing urges all stakeholders—executives who must fund analytics initiatives, IT teams that support them and data scientists, who uncover and communicate meaningful insight—to go boldly in the direction of “Big Analytics.” [5]. Big Data is a new term used to identify the datasets that due to their large size and complexity, we cannot manage them with our current methodologies or data mining software tools. Provide a broad overview of the topic, its current status, controversy, and forecast to the future [6].

3. Big data analytics versus Data mining analytics

We are living in big data era which is full of opportunities and a lot of challenges like scalability, heterogeneity, privacy etc. Revolution in technology makes it possible to produce enormous heterogeneous stream data from diversified sources like sensors, digital devices, internet etc. but current data mining techniques are not ready yet to meet scalability, privacy and other challenges as they have inadequate scalability and are not designed to go through the 3V’s of big data . This becomes more clear through below points:

Large volume of big data demands high scalability and massive parallelism which is beyond competency of existing data mining techniques. For existing tool and techniques of data mining it is found too hard to handle the large and heterogeneous big data.

Also generated big data is in stream form so its meaning is required in real time but real time processing is not possible with existing data mining tool that cause delay in valuable information extraction and that delayed information will be of no use.

Current data mining techniques does not fulfil velocity requirement of big data as in current technique data must be loaded into uniform format into storage system before processing but loading huge amount of big data requires a lot of time that delay processing.[1]

Conventional data mining techniques works on data set which is structured and homogeneous in nature but big data is heterogeneous in nature captured from diversified sources and unstructured data cannot fit well in conventional techniques.

Data mining techniques can not reveal maximum associations, relations and insight into data.

Table 1: Showing difference between Traditional data analytics and big data analytics

Traditional data analytics	Big data analytics
a) • Analysis suitable for structured data only. • Usual size of data is megabyte/gigabyte.	a) • Analysis suitable for structured data ,semi structured data and unstructured data • Usual size of data is terabyte/ petabyte
b) Analysis of data captured from limited and known sources	b) Analysis of data captured from unlimited and unknown sources.
c) • SQL approach to data • Relational database (data to function model) • No open source • Batch processing (offline) of “historical,” static data	c) • Massively parallel processing and NoSQL approach to data, but almost SQL compliant • Hadoop framework (function to data model) • Open source • Stream processing (online) of (near) real time, live data
d) Individual manufacturer/ developer can work independently	d) Nobody works alone; all related parties must work together

4. Difference between Data Mining and Big Data

Big data and data mining are two different things. Both of them relate to the use of large data sets to handle the collection or reporting of data that serves businesses or other recipients. However, the two terms are used for two different elements of this kind of operation [3].

Big data is a term for a large data set. Big data sets are those that outgrow the simple kind of database and data handling architectures that were used in earlier times, when big data was more expensive and less feasible. For example, sets of data that are too large to be easily handled in a Microsoft Excel spreadsheet could be referred to as big datasets [7].

Data mining refers to the activity of going through big data sets to look for relevant or pertinent information. This type of activity is really a good example of the old axiom "looking for a needle in a haystack." The idea is that businesses collect massive sets of data that may be homogeneous or automatically collected. Decision-makers need access to smaller, more specific pieces of data from those large sets.

Data mining can involve the use of different kinds of software packages such as analytics tools. It can be automated, or it can be largely labour intensive, where individual workers send specific queries for information to an archive or database. Generally, data mining refers to operations that involve relatively sophisticated search operations that return targeted and specific results [4]. For example, a data mining tool may look through dozens of years of accounting information to find a specific column of expenses or accounts receivable for a specific operating year. In short, big data is the asset and data mining is the "handler" that is used to provide beneficial results [3].

Table 2: showing difference between big data and data mining

BIG DATA	DATA MINING
a)Big Data is defined as large set of data that is very unstructured and disorganized.	a)Data mining is the analysis of data for relationships that have not previously been discovered.
b)Big data deals with lots of relations in large data set.	b)Data Mining deals with lots of details in large data set.
c)Big Data cannot be handled by standard database management systems like DBMS, RDBMS or ORDBMS".	c)It involves database and data management aspects, data pre-processing, model and inference considerations, interestingness metrics, complexity considerations, post-processing of discovered structures, visualization, and online updating.
d)The challenges include analysis, capture, curation, search, sharing, storage, transfer, visualization, and privacy violations.	d)The challenge for data mining is tackling the problem of mining richly structured, heterogeneous datasets.

5. Conclusion

Big data is a term used for complex and large heterogeneous data set that is still growing continuously at unprecedented scale and is a hot topic now a days as it has lot of potential for revealing meaningful information and valuable insight. It provides a way for organisation to remain in competition by wealth of information provided by big data analysis. Though big data is at its early stage but in future it will revolutionize our society and life. For example government already started mining social media data, on line transaction for predicting pattern, determining need of government facilities etc. In this paper we have a discussion on difference between data mining and big data and also disclosed challenges or difficulties faced by conventional data mining techniques on big data. So we conclude that there is need of revolution in existing data mining technique that is existing techniques must be supplemented with novel and fast distributed storage system, massive parallel processing architecture and new innovative technique must be designed on new platform to overcome challenges like heterogeneity, scalability, privacy etc.

6. References

[1] From Big Data to Big Data Mining: Challenges, Issues, and Opportunities Dunren Che1 , Mejdl Safran1 , and Zhiyong Peng2 1 Department of Computer Science, Southern Illinois University Carbondale, Illinois 62901, USA {mejdl.safran@, dche@cs.}siu.edu 2 Computer School, Wuhan University, Wuhan, 430072, China peng@whu.edu.cn

[2] Big Data, Bigger Opportunities – Data gov's roles: Promote, lead, contribute, and collaborate in the era of big data Jean Yan April 9, 2013

[3] Data Mining for Big Data: A Review Bharti Thakur, Manish Mann Computer Science Department LRIET, Solan (H.P), India

[4] Journey from Data Mining to Web Mining to Big Data Richa Gupta Department of Computer Science University of Delhi

[5] The Rise of Big Data Spurs a Revolution in Big Analytics by Norman H. Nie, CEO Revolution Analytics

[6] Mining Big Data: Current Status, and Forecast to the Future Wei Fan Huawei Noah's Ark Lab Hong Kong Science Park Shatin, Hong Kong david.fanwei@huawei.com Albert Bifet
Yahoo! Research Barcelona Av. Diagonal 177 Barcelona, Catalonia, Spain abifet@yahooinc.
com