© 2017 IJSRST | Volume 3 | Issue 7 | Print ISSN: 2395-6011 | Online ISSN: 2395-602X
Themed Section: Science and Technology

Big Data - Applications and Challenges

Rai Singh
Village Gindran, PO Ghoranwali, Rania, Sirsa, Haryana, India

ABSTRACT

These days big data plays a vital role in everyone life in almost every professional or non professional domains. In this paper we will discuss various issues and challenges related Big Data handling. Big Data has numbers of application in different domains. Big Data is now rapidly expanding in all science and engineering domains, including physical, biological and bio-medical sciences, health and medicine, finance, manufacturing, education, transport etc. As the applications, types of data and the rate at which data growing has created a demand to invent new systems to meet requirements of this voluminous data. This paper spread light on various issues to be addressed to handle big data.

Keywords: Temporal, Bid Data, DSS, spatial, veracity, mining Big Data, MIS.

I. INTRODUCTION

The term big data can be described as a complex and large amount of data called voluminous data which may be structured, unstructured or semi-structured in nature and require special attention to handle it on network or cloud. These days as the internet access is becoming cheaper and easy, data sets are growing rapidly at a very high rate. Not only amount, but also the variety and speed or velocity at which it's increasing is a main concern. The name Big Data derived from the amount of data produced everywhere, anywhere and by anyone (means by various fields/domains including data produced by human being and software tools) at any time. Here we are more concern with digital data rather than printed or physical data. This information explosion takes place mostly in last two to three years as the internet and mobile base transaction along with m-commerce increased. Driving force behind this information explosion is easier, wider and cheaper availability of internet and electronic devices and awareness. Apart from little bit handling challenges, this big data is taken as advantageous in decision support systems and strategic planning. Big data is very helpful in improve professional offers, quality of service improvement, improved client support systems and many others business activities. Data mining with Big Data is a main concern which is helpful in all these business and corporate activities. There are various other issues which concern with Big Data like inconsistencies in data, authenticity of data, copyrights etc. In decision support system (DSS) and MIS (management information system), Big Data is collected, analysed, integrated and then selected data is taken as basis for decision making or planning and also to create discoveries. Big data has a diverse scope which covers almost all domains like education, health, medicine, physical science, manufacturing, retail, geographical services, engineering and technology, transport services, banking services, insurance and finance services, these fields can be greatly influence for betterment. In this paper we will review various data technologies to adopt big data management and computing infrastructure requirements. There are various challenges in dealing with Big Data, but main challenge to focus upon is data mining, which is process of extracting useful and core information or data from an unorganised, heterogeneous, from autonomous sources, unstructured and voluminous data base which is to be used in DSS and MIS. In decision making and long term strategic planning data mining process must be very efficient and concurrent because it is not good practice to and also not feasible to store all or large amount of acquired data.

II. CHARACTERSTICS OF BIG DATA

Knowledge base or data base for Big Data contains data in unstructured from, heterogeneous data in nature; unknown sources which difficult to prove authenticity, non centralised control and complex relationship among data and importantly voluminous data. The popular three V's plays important role in understanding the concept of Big Data. These V's refers to Data Volume, Data Velocity and Data Variety.

Volume concern with the amount or quantity of data produced and acquired. The potential and value of data is estimated by the volume of data on the basis of which we can decide whether it is Big Data or Not. Size of data may vary from MB to GB, TB (terabytes) and even PB (petabytes).

Velocity is the rate at which data is being produced in domain and processed to use in DSS by MIS. Historically data was generated in Batch or Periodic basis but these days data is generated near to Real Time or actually real time.

Variety means different dimensions of data in which data is growing. Empirically data was categorised in two basic category, text and multimedia data. But these days, several new categories are required to classify different types of data being produced i.e. database, web, photo/picture, audio, video, unstructured mobile produced data. Various new formats are invented and are in use like gif format is becoming very popular these days.

Apart from these three V's, some other V's are also taken in account. These are

Veracity The dictionary meaning of veracity is quality of being correct or accurate. The extraction of valuable and core data from Big Data is dependent on accuracy or correctness of acquired data to a large extent.

Variability is different issue than variety in the sense that variety refers to different types of same thing or object whereas variability refer to the variation in similar things. It deals with inconsistencies in data.

Big Data is a multidimensional vast knowledge base in which data is collected in unstructured and unorganised manner because it just observes what is happening and keeps tracks of these happenings. Further, different user may have different interpretation of same set of data, depending upon the part of context they are interested or context which is available or accessible to them. Localised view to data leads to different conclusions on same data set from same knowledge base.

III. EXTRACTION AND MINING PROCESS

Data mining with Big Data and knowledge/information extraction are used interchangeably. Data mining is process of discovering useful pattern from a large data set. Extraction is method of retrieving data from an unstructured database. In mining with Big Data, data analysis is done on voluminous data and then data processing is performed to add Meta data and some integration to present mined data in acceptable form to use for DSS. The data mining techniques evolved from combination of statistics, database management and artificial intelligence. There are mainly two objectives for which Data Mining with Big data is carried out, classification means grouping data and prediction. Various data mining techniques are in use according to nature and type of Big Data base. Some of these are discussed below.

Hierarchical mining a broad category is further divided in dependent sub categories based on measurement or analysis of one or more pattern. This type of classification forms a tree like structure with vertices and arc or links.

Logistic Regression is a technique based on statistics data analysis techniques which deal with classification. To predict the possibilities of independent variables as a function, a formula is formulated.

Techniques based on neural network based neural network basically a software algorithm. As we know a neural network is consist of input, output nodes and some internal layers. Some weight is assigned to each unit in network. Network is expertise by training. Input is faded up to input nodes, on the basis of training, the algorithm adjust the weights to met a certain condition, then the traffic that passes by meeting conditions or patterns reaches to output nodes.

Clustering techniques like K-nearest neighbours it is technique identifies similar types of rows from table or data record from database and group them. This process is similar to k- nearest neighbours. In K-nearest neighbour technique, the distances between the record and points in the historical (training) data is computed. Then these records are assigned to nearest neighbour class in database. These diverse natures of big data are the causes of challenges for extracting useful pattern or data from large data set. In the next section of paper we will discuss issues related to these challenges in Big Data mining process.

IV. CHALLENGES IN MINING WITH BIG DATA

Various obstacles and challenges in mining with Big Data can be broadly categorised in two categories. One is issues related to data acquired and other is issues related to mining tools.

Issues related to Data Acquired

Mining Complex and Dynamic Data

The nature or characteristics of data is a major concern while dealing with Big Data. The swift production of complex data and their changes in volumes and in nature like Documents posted on WWW servers, Internet backbones, social networking sites, other communication networks, and transportation networks etc. are all featured with complex data are driving force behind evolution of Big Data. This complex nature of data presented various difficulties for our mining and learning systems. For example, some researchers have successfully used a well-known social networking site, Twitter to detect events like earthquakes and various other social activities, with almost real time speed and with very high accuracy. Dealing with such complex data set is a big challenge for Big Data applications. The number of active users on social networking sites like facebook or twitter already reached to trillions, and these active users are interested to each other and with different types of connections, such connections are quadratic with respect to numbers of users. Analyzing data from such enormous networking sites is a big issue.

Various data mining tools have been invented to meet above challenges and extract useful patterns from complex data collection. For example, identifying communities and tracking their dynamically evolving relationships are required for understanding and handling complex systems.

Extraction of data from Incomplete Data Set, Sparse and Uncertain collection

In a data collection where each data field is not deterministic but is subject to some random/error distributions are a special type of data reality is Uncertain data. Uncertain data is mainly related to domain specific applications with inaccurate data readings and collections. For example, data acquired through from GPS equipment like mobile phone is probably uncertain, mainly because the technology barrier of the device limits the precision of the data to certain levels. Another example, one individual may not be able to tell us what is his actual exact income, but he may tell us rough idea or a range. The defining features for Big Data applications are Incomplete Data Set, Sparse and Uncertain collection. Meaning of sparse here is the number of data points is very small for drawing reliable conclusions. In a situation where data in a high dimensional space does not conclude a real image or trends or distributions, generally it is considered as a complication of the data dimensionality issues. High dimensional spare data significantly deteriorate the difficulty and the reliability of the models derived from the data in most of machine learning and data mining techniques. In such situations we need to employ a mechanism to minimize these dimensions. Other method to deal with it is to core feature selection, in which to decrease data additional samples can be included. There might be a situation where data field value may not be present in sample taken, it refers to incomplete data. Reasons for this incomplete data can be various realities like a systematically policy to skip some vale intentionally or it is also possible that a data capturing machine like a sensor device not performing well and accurate. In data mining algorithms various solutions are evolved to handle missing data like an algorithm may ignore missing vale, or data imputation is an established research field which seeks to impute missing values in order to produce improved models.

Complex relationship networks in data

The relationships between individuals exist in big data context. The web pages on internet can be considered as individuals and these pages are linked with each other through hyper links. These linked pages construct a complex network on web. For example, on twitter, a big complex social network is formed. New computer architectures for real-time data-intensive processing has spawned as result of rise of Big Data, such as the open source project Apache Hadoop which runs on high-performance clusters. Real-time processing for complex data is a very challenging task in the context of Big Data.

Complex intrinsic semantic associations in data

Look at a scenario where users on facebook posting videos and pictures, same on flickers and YouTube, and news on blogs may discuss some events like conferences or seminars etc. It is clear that there is strong semantic relationship between these data. Mining of complex associations from text, videos and pictures or images data will surely helpful in improving performance of

application systems such as prediction systems, recommendations and decision support systems. But it is a big challenge in Big Data to describe these semantic associations in data and to build a model to make necessary arrangements to fill gap in association if any.

Inconsistencies in Big Data

Inconsistencies in acquired in big data are non avoidable factor in human behaviours and decision support processes. It is common thing in acquired, processed or represented data. Inconsistency or conflict in Big Data base is harmful because it can cause adverse effect on quality of result of mining or extraction process which is to be used in DSS by MIS. In Big Data inconsistency can be easily captured at various stages and dimensions like in data, Meta data, information or knowledge. So inconsistency creates a big challenge in Big Data mining or analysis. Various tasks are involved in mining or analysis which aimed to support in decision making, prediction, classification of facts, regression and association analysis among data etc. there are different types of inconsistency which has impact on these above mentioned various ambition. So the type of inconsistency is also need to be described.

Inconsistencies Due To Functional Dependency

Functionally dependency plays an important role in database where database is managed by relational model of database management systems. If violation of these functional dependencies occurs, inconsistencies in Big Database arise.

Temporal Inconsistencies

Temporal inconsistency may occur in a database where dataset holds data items with temporal characteristics. These temporal data item may overlap or coincide in time. In IBM temporal inconsistencies have been utilized as problem solving heuristics. Inconsistencies concerning with time may occur as partial temporal of fully temporal inconsistency.

Text Inconsistencies

Text producing sources like emails, blog spots, social networking sites and sms etc might be the origin of text inconsistencies. The integrity of the Big data can be highly disrupted by these text inconsistencies. Co-reference is said to be occur when same thing or event are referred by two or more texts.

Spatial Inconsistencies

This type of inconsistency may arise in data set where data items with geometric dimensions present in Database. Information about objects in space is represented with spatial properties like geometric location, direction, shape, distance between objects etc. When objects are aggregated or composited with different representations for same object different sour resulting in violation of the constraint that objects must have unique geometric representation, spatial inconsistency may arise.

Issues related to Mining Tools

Selection of suitable Language for Big Data

A specific language for a specific field plays a vital role in boosting up in that field. Same in the case of Big Data concept is. It's an important issue to find and choose suitable language for mining when dealing with Big Data. There is also a need of inventing an algebraic notation to better analysis of Big Data

Reliability of Data to Work On

When dealing with Big Data, it's a big challenge to separate the Signal of Data and Valuable information. Unfortunately, in this matter various organisations feeling difficulties in identifying the correct and reliable data and determining how to best use it. Spam data is a serious issue when identifying quality data is crucial. To overcome this problem, organisations must think out of box to invent models different from traditional models.

Technical Infrastructure

The technology landscape in the electronic field is advancing at a very fast speed. Next version is already available when existing version comes into fully fictional mode. Mining and extracting useful information from big data means collaboration with a strong and innovative technology partner that can help create the correct technical infrastructure which should be able adapt to changes in the landscape in an efficiently.

Data Access

A serious obstacle may meet while Big Data mining is Data access and connectivity. According to a survey, lot of data is still not connected. To manage and access data, Organisation dealing with Big Data mining does not have suitable platform.

Time Variable Integration

An important dimension in Big Data is time. Data over internet is increasing at a very rapid rate and almost real time. It is a big challenge in big data mining tools

development. To analyse causalities in long term and real time data is also big issue. In terms of storage, there are also some issues to address as the volume of data produced may exceed than the capacity of storage. In such situation, extra care is to be taken in selection of data for storage.

Data Embedding and Complex Data

As the data increasing in various dimensions like volume and diverse in nature of data, new and reinvention of Big Data tools is current requirement in Big Data mining. Earlier developed system and tools which were suppose to simple data like tables, sheets or graphs are unsuitable when dealing with modern Big data like audio, video and complex images etc.

Security

Last but not the least, is issue related to security. As we have discussed, Big Data means a Voluminous and complex data collection. To keep secure this Big data set is itself a big issue. One approach may be to push some limit on data access based on some parameters. But this scheme needs utilisation of an efficient authentication mechanism for each term. It will also require lot of encryption methods to adopt to avoid various problems.

V. CONCLUSIONS AND DISCUSSION

In this paper, we discussed about what the Big Data is and its various applications. We also focus our attention on one of the challenges in Data Mining with Big Data like inconsistencies in Big Data and their impact. We analyse two types of challenges i.e. infrastructural and data related issues in Big Data. We also examine different types of inconsistencies in Big Data. From the literature review we can conclude that a new system should be invented so that voluminous and unstructured data can be connected through their complex relationships and useful patterns and information can be extracted which can be used as an aid in prediction or future panning. We regard Big Data as an emerging trend and the requirement of Big Data mining has arisen in various domains.

VI. ACKNOWLEDEMENTS

The author expresses his gratitude to his parents and teachers for their continuous support and help.

VII. REFERENCES

[1] Alam et al. 2012, Md. Hijbul Alam, JongWoo Ha, SangKeun Lee, Novel approaches to crawling important pages early, *Knowledge and Information Systems*, December 2012, Volume 33, Issue 3, pp 707-734

[2] http://www.arpnjournals.com/jeas/research_papers/rp_2015/jeas_0515_1931.pdfwww.arpnjournals.com

[3] Machanavajjhala and Reiter 2012, Ashwin Machanavajjhala, Jerome P. Reiter: Big privacy: protecting confidentiality in big data. *ACM Crossroads*, 19(1): 20-23, 2012.

[4] Suriya Begum, Dr. Prashanth C.S.R, "Investigational Study of 7 Effective Schemes of Load Balancing in cloud Computing", *International Journal of Computer Science* Issues, ISSN (Print): 1694-0814 | ISSN (Online): 1694-0784, Vol. 10, Issue 6, No. 1, November-2013, pg. 276 - 287.

[5] 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

[6] http://www.slideshare.net/minujoseph/inconsistencies-in-bigdata

[7] Borgatti S., Mehra A., Brass D., and Labianca G. 2009, Network analysis in the social sciences, *Science*, vol. 323, pp.892-895.

[8] Rajaraman and Ullman, 2011, A. Rajaraman and J. Ullman, *Mining of Massive Datasets*, Cambridge University Press, 2011.

[9] IBM 2012, What is big data: Bring big data to the enterprise, http://www-01.ibm.com/software/data/bigdata/, IBM.

[10] *Ghoting et al., 2009*, Ghoting A., Pednault E., Hadoop-ML: An infrastructure for the rapid implementation of parallel reusable analytics, In: *Proceedinds of the Large-Scale Machine Learning: Parallelism and Massive Datasets Workshop (NIPS-2009)*.

[11] 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

[12] Reed C., Thompson D., Majid W., and Wagstaff K. 2011, Real time machine learning to find fast transient radio anomalies: A semi-supervised approach combining detection and RFI excision, *Int'l Astronomical Union Sym. on Time Domain Astronomy*, UK. Sept. 2011