A Comparative Study On Bigdata With Dta Mining And Its Applications

Ganga Patur
Assistant Professor, Department of CSE, KGRCET

Abstract: Presently in real world where there is a massive amount of data that increasing every day. And mostly data is generating from shred networks like face book, twitter and etc, and this data generated in the form of unstructured and if we would be not very intense, if we did not find ways to control and use this data so that we could extract patterns and information from it. One of the ways to extract patterns and use data in an intelligent way is through the use of data mining. But data mining can process on structured data and limited amount of data but whereas big data when the data is too large to process in petabyte so this data can't handle using data mining tool so in this paper we comparing various algorithms of data mining and big data.

Keywords: Big data, big data architecture, Data mining issues

INTRODUCTION:

The concept big data came into being through explosion of data from the Internet, cloud, data center, mobile, Internet of things, sensors and domains that possess and process huge datasets. Volume, velocity and variety are the main features of big data (1). These features make traditional computing models ineffective. A premise of tremendous value in the huge datasets is the motive for big data exploration and exploitation. Big data has been identified with potential to revolutionize many aspects of life (2); applications of big data in some domains have practically changed their practices (3). Data has penetrated each industry and all business functions; it is now considered a major factor in production (4). Big data would inspire a lot of innovative models, companies, products and services. This is due to the strategic insight it behold for the IT industry and businesses. The IT industry would create new products and target untapped markets. Businesses would optimize existing businesses and have new business models. lots of educational analysis is being conducted within the field of massive data; starting from applications, tools, techniques and design. The analysis is knowledge domain and customarily known as knowledge science. seeable of the divergent interest in huge knowledge from distinct domains, a transparent and innate appreciation of its definition, advancement, constituent technologies and challenges becomes predominate (5). during this regard, this paper would provides a literature review on some aspectsof huge data; there area unit definitions, features, opportunities, challenges, platforms, techniques and architectures. The review is exclusive from existing ones thanks to its wide coverage of massive knowledge design offerings from firms and also the unified reference design. These architectures area unit the foremost necessary blueprint to assist navigate the large knowledge parcel. the remainder of this paper is made public as follows; section 2 discuss {the huge|the large|the massive} knowledge context; section 3 examines huge knowledge platforms whereas section four analyze huge knowledge techniques; section five elucidate big knowledge architectures. Section six is the conclusion of the literature review.

Data mining Definition:

Data mining involves the use of sophisticated data analysis tools to discover previously unknown, valid patterns and relationships in large data sets. These tools can include statistical models, mathematical algorithms, and machine learning methods such as neural networks or decision trees. Consequently, data mining consists of more than collecting and managing data, it also includes analysis and prediction. The objective of data mining is to identify valid, novel, potentially useful, and understandable correlations and patterns in existing data. Finding useful patterns in data is known by different names (e.g., knowledge extraction, information discovery, information harvesting, data archeology, and data pattern processing).

The development of Information Technology has generated large amount of databases and huge data in various areas. The research in databases and information technology has given rise to an approach to store and manipulate this precious data for further decision making. Data mining is a process of extraction of useful information and patterns from huge data. It is also called as knowledge discovery process, knowledge mining from data, knowledge extraction or data /pattern analysis.

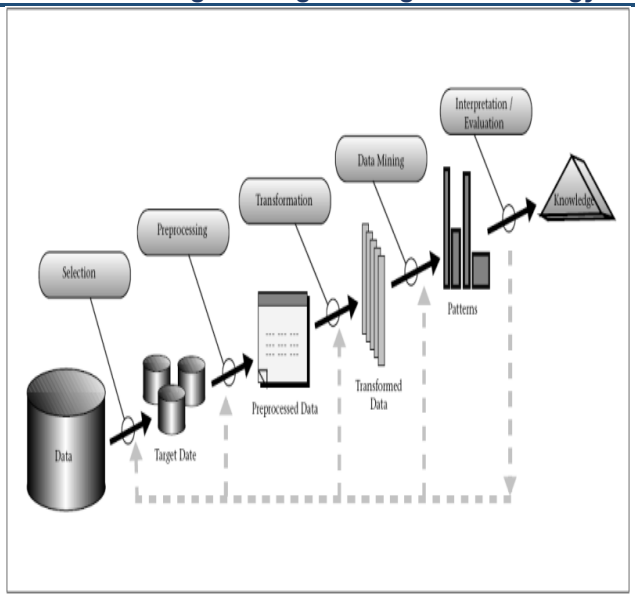

Fig1: Data mining step by step process

Data mining is a logical process that is used to search through large amount of data in order to find useful data. The goal of this technique is to find patterns that were previously unknown. Once these patterns are found they can further be used to make certain decisions for development of their businesses.

Three steps involved are

- Exploration
- Pattern identification
- Deployment

Big Data Definition

There are a lot of definitions of big data due to its wide usage in numerous fields. There are the product oriented, process oriented, cognition oriented and social movement perspective (6). The product oriented perspective looks at features of the data; especially its volume, velocity and variety. Big data imply contemporary technologies and architectures that have been designed to efficiently deduce benefit from huge and variety of datasets (7). Another definition states big data involves huge volume, heterogeneous, localized control and finds intricate and dynamic correlations between data (8). A third definition states as a result of exponential increase in global data, big data signifies huge datasets. Relative to conventional data, it includes huge unstructured data that need real time analysis. It also comes with new opportunities for dissecting value and deeper understanding (9). The second perspective of big data leverages the processes involved in its operation.

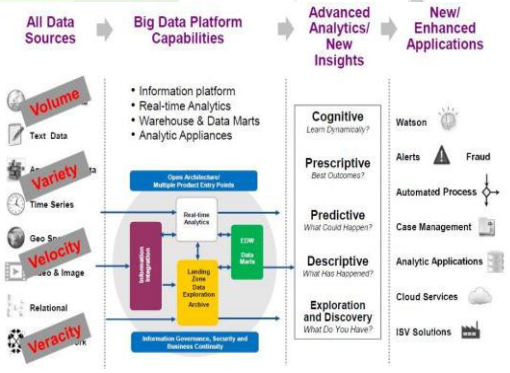

Fig2: big data step by step process

The process oriented perspective looks at processing activities that are involved or essential in dealing with big data. These processes are searching, aggregating, storage and analysis of data. These processes in the context of big data are novel in terms of architecture, tools and techniques. This perspective to big data further highlights the unique technological infrastructure; most importantly tools and programming methodologies required in creating big data. A definition developed at University of California Berkeley resonates within the process oriented perspective. It states big data is when the traditional computing technology cannot meet timely, cost effective and quality answers that are demanded by data driven questions requested by users (10). The next perspective to big data definition involves the cognition dimension.

The third school of thought is the cognition oriented perspective. Big data is defined in relation to the limited ability of the human mind to understand, hence a need for technological infrastructure, numerous techniques from diverse disciplines and visualization techniques in conveying meaning of data.

Big Data Opportunities and Challenges

Big Data opportunities have been well articulated in the McKinsey study (4). There are increased productivity and cost savings for numerous organizations. Innovative business models such as the booming sharing economy and on-demand services would rely on big data. New companies would keep on springing up and provide cutting edge services for consumers. Big data has also been premised to refine the scientific method (7). Big data has also opened the door for data scientists; these are skilled people that effectively and creatively deduce big data value for their organizations and clients.There are however a number of challenges to be surmounted to harness big data value.

Some big data challenges are privacy, security, storage and processing. Privacy is the most challenging aspect of big data; this is due to the age old sensitive nature of peoples' data. When publicly available data of an individual is subjected to inference techniques, insightful informationof an individual can be deduced. Legal, technical and policy strategies can be put in place to safeguard the privacy of individuals' data. Laws could guide data collection, processing, usage and transfer to third party. Technical mechanisms include encrypting data at rest and on transmission. Organization policy should ensure workers clearly know their responsibilities, limitations and applicable sanctions (1) (2) (3). Security is challenging to big data due to its pervasive nature. Secure computation need to be incorporated for distributed programming frameworks such as Map Reduce (19). Storage and processing are other challenges of big data; the amount of effort required on networks, servers to store and process data is enormous. The processing of data would warrant effort to distribute the stored data for processing. Another processing challenge is the size of data for machine learning, this requires several scan of the data which is costly because learning objectives and assessment measures are non-linear, non-smooth, non-convex and non-decomposable over samples (5). The viable solution is to bring processing to the storage point or taking only the relevant data for processing through reduction techniques. Cloud computing is an overall viable option to the storage and processing dilemma, this is due to its prospect of scalable storage and processing capacity (7).

Challenges in Data Mining

Data mining algorithms embody techniques that have typically existed for several years, however have solely latterly been applied as reliable and climbable tools that point and once more beat older classical applied mathematics ways. Whereas data processing remains in its infancy, it's changing into a trend and omnipresent. Before data processing develops into a standard, mature and sure discipline, several still unfinished problems need to be addressed. a number of these problems are addressed below. Note that these problems don't seem to be exclusive and don't seem to be ordered in any means. Security and social issues: Security is a vital issue with any information assortment that's shared and/or is meant to be used for strategic decision-making. Additionally, once information is collected for client identification, user behavior understanding, correlating personal information with alternative info, etc., massive amounts of sensitive and personal info regarding people or corporations is gathered and keeps. This becomes disputable given the confidential nature of a number of this information and therefore the potential misappropriated access to the data. Moreover, data processing might disclose new implicit information regarding people or teams that might be against privacy policies, particularly if there's potential dissemination of discovered info. Another issue that arises from this concern is that the applicable use of information mining. owing to the worth of information, databases of all kinds of content are frequently sold , and since of the competitive advantage that may be earned from implicit information discovered, some necessary info might be withheld, whereas alternative info might be cosmopolitan and used while not management. Programmer issues: The information discovered by data processing tools is beneficial as long because it is fascinating, and particularly apprehensible by the user. smart information visualization eases the interpretation of information mining results, also as helps users higher perceive their desires. several information exploratory Analysis tasks ar considerably expedited by the flexibility to examine information in an applicable visual presentation. There ar several visualization ideas and proposals for effective information graphical presentation. However, there's still a lot of analysis to accomplish so as to get smart visualization tools for giant datasets that might be accustomed show and manipulate well-mined information. The major issues related to user interfaces and visualization are "screen real-estate", information rendering, and interaction. Interactivity with the data and data mining results is crucial since it provides means for the user to focus and refine the mining tasks, as well as to picture the discovered knowledge from different angles and at different conceptual levels. Mining methodology issues: These issues pertain to the data mining approaches applied and their limitations. Topics such as versatility of the mining approaches, the diversity of data available, the dimensionality of the domain, the broad analysis needs (when known), the assessment of the knowledge discovered, the exploitation of background knowledge and metadata, the control and handling of noise in data, etc. are all examples that can dictate mining methodology choices. For instance, it is often desirable to have different data mining methods available since different approaches may perform differently depending upon the data at hand. Moreover, different approaches may suit and solve user's needs differently. Most algorithms assume the data to be noise-free. This is of course a strong assumption. Most datasets contain exceptions, invalid or incomplete information, etc., which may complicate, if not obscure, the analysis process and in many cases compromise the accuracy of the results. As a consequence, data preprocessing (data cleaning and transformation) becomes vital. It is often seen as lost time, but data cleaning, as time-consuming and frustrating as it may be, is one of the most important phases in the knowledge discovery process. Data mining techniques should be able to handle noise in data or incomplete information. More than the size of data, the size of the search space is even more decisive for data mining techniques. The size of the search house is usually relying upon the amount of dimensions within the domain house. The search house sometimes grows exponentially once the amount of dimensions will increase. this is often referred to as the curse of spatial property. This "curse" affects thus badly the performance of some data processing approaches that it's turning into one amongst the foremost imperative problems to unravel. Performance issues: several computer science and applied mathematics strategies exist for knowledge analysis and interpretation. However, these strategies were typically not designed for the terribly giant knowledge sets data processing is addressing these days. T sizes area unit common. This raises the problems of quantifiability and potency of the knowledge the info the information mining strategies once process significantly giant data. Algorithms with

exponential and even medium-order polynomial quality can't be of sensible use for data processing. Linear algorithms area unit sometimes the norm. In same theme, sampling may be used for mining rather than the total dataset. However, issues like completeness and selection of samples could arise. different topics within the issue of performance area unit progressive change, and parallel programming. There's little doubt that similarity will facilitate solve the dimensions drawback if the dataset may be divided and therefore the results may be unified later. progressive change is vital for merging results from parallel mining, or change data processing results once new knowledge becomes on the market while not having to re-analyze the entire dataset. knowledge supply problems: There area unit several issues associated with the info sources, some area unit sensible like the variety of information sorts, whereas others area unit philosophical just like the knowledge glut drawback. we have a tendency to definitely have associate a lot of than way over knowledge since we have a tendency to have already got more knowledge than we can we will we area unit able to} handle and that we are still aggregation knowledge at a good higher rate. If the unfold of management systems has helped increase the gathering of knowledge, the arrival of knowledge| information mining is definitely encouraging additional data harvest. This follow is to gather the maximum amount knowledge as potential currently and method it, or try and method it, later. The priority is whether or not we have a tendency to area unit aggregation the proper knowledge at the suitable quantity, whether or not we all know what we would like to try and do with it, and whether or not we have a tendency to distinguish between what knowledge is vital and what knowledge is insignificant. relating to the sensible problems associated with knowledge sources, there's the topic of heterogeneous knowledge bases and therefore the target numerous complicated data sorts. We have a tendency to area unit storing differing kinds of information during a sort of repositories. it's tough to expect an information mining system to effectively and expeditiously come through good mining results on all types of information and sources. The size of the search house is usually relying upon the amount of dimensions within the domain house. The search house sometimes grows exponentially once the amount of dimensions will increase. this is often referred to as the curse of spatial property. This "curse" affects thus badly the performance of some data processing approaches that it's turning into one amongst the foremost imperative problems to unravel. Performance issues: several computer science and applied mathematics strategies exist for knowledge analysis and interpretation. However, these strategies were typically not designed for the terribly giant knowledge sets data processing is addressing these days. T sizes area unit common. This raises the problems of quantifiability and potency of the knowledge the info the information mining strategies once process significantly giant data. Algorithms with exponential and even medium-order polynomial quality can't be of sensible use for data processing. Linear algorithms area unit sometimes the norm. In same theme, sampling may be used for mining rather than the total dataset. However, issues like completeness and selection of samples could arise. Different topics within the issue of performance area unit progressive change, and parallel programming. there's little doubt that similarity will facilitate solve the dimensions drawback if the dataset may be divided and therefore the results may be unified later. Progressive change is vital for merging results from parallel mining, or change data processing results once new knowledge becomes on the market while not having to re-analyze the entire dataset. Knowledge supply problems: There are a unit several issues associated with the info sources, some area unit sensible like the variety of information sorts, whereas others area unit philosophical just like the knowledge glut drawback. We have a tendency to definitely have associated a lot of than way over knowledge since we have a tendency to have already got more knowledge than we can we will we area unit able to handle and that we are still aggregation knowledge at a good higher rate. If the unfold of management systems has helped increase the gathering of knowledge, the arrival of knowledge of knowledge of information mining is definitely encouraging additional data harvest. This follow is to gather the maximum amount knowledge as potential currently and method it, or try and method it, later. The priority is whether or not we have a tendency to area unit aggregation the proper knowledge at the suitable quantity, whether or not we all know what we would like to try and do with it, and whether or not we have a tendency to distinguish between what knowledge is vital and what knowledge is insignificant. Relating to the sensible problems associated with knowledge sources, there's the topic of heterogeneous knowledge bases and therefore the target numerous complicated data sorts. we have a tendency to area unit storing differing kinds of information during a sort of repositories. It's tough to expect an information mining system to effectively and expeditiously come troughs good mining results on all types of information and sources. Different kinds of data and sources may require distinct algorithms and methodologies. Currently, there is a focus on relational databases and data warehouses, but other approaches need to be pioneered for other specific complex data types. A versatile data mining tool, for all sorts of data, may not be realistic. Moreover, the proliferation of heterogeneous data sources, at structural and semantic levels, poses important challenges not only to the database community but also to the data mining community.

Conclusion

Data mining has significance concerning finding the patterns, forecasting, discovery of knowledge etc., in unlike business domains. Data mining has extensive application domain nearly in each industry where the data is produce that's why data mining is measured one of the most important frontiers in database and information systems and one of the most promising interdisciplinary developments in Information Technology.

Big data platforms and tools are the cornerstone of the data revolt taking place. Volume, velocity and variety are the major skin texture that makes big data exceptional from conventional computing. Big data has been premised to make thoughtful insights in all domains. Original models, products, services and huge cost savings are some of the opportunities big data present. Big data due to its huge potential and user base has multiple definitions reflecting wide perspective of its stakeholders. Platforms for big data have varied purposes with regard to speed at which data needs to be fashioned and progression. The platforms are batch giving out, stream dispensation and interactive analytics.

REFERENCES:

[1] Laney D. 3D Data Management: Controlling Data Volume, Velocity and Variety. Application Delivery Strategies. 2001.

[2] Philip Chen CL, Zhang CY. Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. Inf Sci. 2014; 275: 314–347.

[3] Huang T, Lan L, Fang X, An P, Min J, Wang F. Promises and Challenges of Big Data Computing in Health Sciences. Big Data Res. 2015; 2(1): 2–11.

[4] James M, Michael C, Brad B, Jacques B, Richard D, Charles R. Big data: The next frontier for innovation, competition, and productivity. McKinsey Glob Inst. 2011.

[5] Hu H, Wen Y, Chua T-S, Li X. Toward Scalable Systems for Big Data Analytics: A Technology Tutorial. IEEE Access. 2014; 2: 652–687.

[6] McKerlich R, Ives C, McGreal R. Measuring use and creation of open educational resources in higher education. Int Rev Res Open Distance Learn. 2013; 14(4): 90–103.

[7] Gantz BJ, Reinsel D. Extracting Value from Chaos State of the Universe : An Executive Summary. IDC iView. 2011: 1–12.

[8] Jiawei Han and Micheline Kamber (2006), Data Mining Concepts and Techniques, published by Morgan Kauffman, 2nd ed.

[9]. Dr. Gary Parker, vol 7, 2004, Data Mining: Modules in emerging fields, CD-ROM.

[10] Crisp-DM 1.0 Step by step Data Mining guide from http://www.crisp-dm.org/CRISPWP-0800.pdf.