Survey on Big Data Analytics and its Applications

S. Sangeetha
Student
Department of EEE, Karpagam College of Engineering
Coimbatore, India

S. Kannimuthu, PhD
Associate Professor
Department of CSE, Karpagam College of Engineering
Coimbatore, India

P. D. Mahendhiran
Assistant Professor
Department of IT, Karpagam College of Engineering
Coimbatore, India

ABSTRACT
Big data are large set or collection of data which cannot be processed by traditional methods such as data processing. The main problems that big data faces are storing, capturing, transferring, data curing (organization and integration of data that are collected from various resources in order to improve the reusability of the data and preservation of the data for a long period of time), querying etc. Analyzing big data has its significance in the field of social networks, spot business trends, internet, medicine, science, finance, business informatics and even in government. Analyzing data would help in great decision making, which may result in improvement in efficiency, reduction in cost and failure risks. Big data analysis becomes a great thirst for the developing organizations since it becomes difficult for those organizations to process thousands of tera bytes of data. Big data analysis even find its application in understanding the reason for natural or man-made disasters by collecting big data in order to recover from the disaster and to develop the communication since communication is the main challenge that the people face while facing disasters.

Keywords
Data Mining, Big Data Analytics, Business Analytics

1. INTRODUCTION
Wang et al [4] evidently said that, by 2025, the Internet will exceed the brain capacity of everyone living on the entire planet. Big data is buzzword that it is most viral and newly originated word, which means enormous collection of both structured and unstructured data which could be handled by traditional data processing techniques and softwares. For a small scale industries or super markets it is very easy to process data which will be small and will not affect the efficiency. But in case of large scale industries, social Medias, government, and those organizations that processes millions of data, storing data and analyzing it faces the main challenge in analyzing the data. Capturing of data, formulating, manipulating, storing and analyzing plays an important role in analyzing the sales or development scenario of every company. When data is in storage disk, applications would starts analyzing the data, which makes the drives very slow and would also reduce the efficiency of the operation. Many organizations including retailers, telecommunications companies, and intellectuals need a speed operation, inorder to seize the opportunities and to respond quickly.

Big data has its greatest impact in artificial intelligence and intelligent system. Have you think about how a robot personifies a human, is responding very accurately to a non-robotic character, there is the place where big data prove itself in artificial intelligence. There is always a question that whether big data is a volume or technology. That isn't the case; the word is likely to be originated from the web search companies who were on the urge to design a query from a very large aggregated loosely-structured data (incomplete or inaccessible data). A best example of big data might be petabytes (1,024 terabytes) or exabyes (1,024 petabytes) of data consisting of billion to trillions of records of very huge number of people all from various sources. Compute intensive data mining algorithms are been widely used by big data analytics inorder to improve the efficiency and high performance to produce timely results. Cloud computing is nothing but, computing the cloud which consists of terrestrial servers across the internet where they could collect, stores, and process the data. Big data plays a very vital role in cloud computing. Since processing and analyzing such a huge data becomes possible by cloud computing.

Big data has been characterized in many ways, which is recently extended from "3V" to the latest "4V".

The major 3V's are:

- Velocity.
- Volume.
- Variety.

The fourth V's are:

- Variablility
- Value
- Virtual

2. DATA MINING
Data mining is one of the types of analyzing tools that is used to extract specific information from massive number of data collected from various resources and summarizing it into some useful information which is used mainly to increase revenues, cut costs and also to develop the productivity of a company and also to increase the number of users in case of internets. Data mining is collection of data from various resources, categorizing it based on the data, establishing relationship between the data, and summarizing it based on the relationship established. While relationship is established data mining involves in finding the patterns or common relationship between the data which were subjected to summarization. There was a situation where there is a need to change hard disks periodically in the areas where data is to stored in a large quantity. Most common, to analyze the local buying pattern Oracle software's data mining capacity is used in many grocery areas.

Data mining provides a link between large-scale information technology and analytical system. Data mining uses open-ended user query in order to analyze patterns and relationship. The types of relationship being analyzed are: class, cluster, association and sequential pattern. As such relationships are classified; levels of analysis are also classified into various

types: artificial neural network, genetic algorithm, decision tree, nearest neighbor method, rule induction and data visualization method. A proper analysis of data results in the reduction of time and cost and improves the efficiency of data mining method. The important parameters of data mining are:

Association

Pattern that the connection between two data**.**

Sequence or Path Analysis

Pattern in which one data leads another

Classification

Pattern that is different and innovative or unique

Clustering

Binding up of many data together which is new and not known priory

Fore Casting

Patterns that shows something about the future scope of a data

Data is the input that is given to data mining and patterns are the output obtained from it. Patterns are of three types:

- Classification rule
- Association rule
- Summaries

Classification Rule

Classification rule defines the output of some process that was happen over a period of time or the credits of a task being held.

Example: result=pass, marks>90% credit -> first class

Association Rule

Association rule defines the possible result that would comes when two or more correlated thing happens simultaneously.
Example: {uniform, socks} -> shoes

Summary

Summary tells the brief of a data. Summary helps to decide the progress or current position of any organization based on the performance of its employee. Consider a class of 100 students. The result of all the students of a class for a total mark of 500 can be summarized as below:

Table 1. Result of all the students of a class

Marks	No of students	Percentage
Above 400	30	30%
Above 300	50	50%
Below 300	20	20%

Interestingness is the best treatment which is a wide concept in emphasizing the concepts of conciseness, reliability, peculiarity, coverage, diversity, novelty, surprisingness, utility and actionability.

Conciseness

It defines the property of being concise. A patter is concise when it has less number of attributes and a set of patterns is said to be concise when it has less number of patterns. Conciseness helps in best understanding and remembering of particular pattern by the users.

Coverage

The pattern must include all the important information of a data or it must contain the sub set of a large data set. The subset which contains the fraction of a data must belong to the dataset. Coverage is an important property that decides the order in which a pattern must be designed. Coverage designs a pattern in such way that it may act as a map in guiding the mining of data.

Reliability

The pattern must be accurate and confident; the relationship established by the pattern must contain the high percentage of applicable case [1]. A pattern must be in such a way that is must in consideration when there is need for comparison at time of arrival of new pattern.

Peculiarity

A pattern must be peculiar from other pattern that where established for different data in order to show to distinguishing behavior of one from another. Peculiarity of a pattern avoid the confusion among data and also for data with even minute difference, pattern must be generated uniquely.

Diversity

Many numbers of patterns can be generated for a same set of data, but all the patterns generated must be different from one another. Diversity is an important factor for summary. Diversity among summaries creates interestingness among the users.

Novelty

No data mining system can represent what and user know and did not know, hence novelty could not measure explicitly with reference to the user's knowledge and user's ignorance respectively [1]. Hence in order to find the novelty the users must be involved in testing the pattern explicitly or else the previously discovered pattern must be bring to their knowledge. The pattern is not opposed by any of the patterns already known to the user.

Surprisingness

The pattern is said to be surprised when it satisfies the expectations of the users and also beyond the expectation. The pattern that resists the contents of the patterns known already to the users gives much pleasure.

Utility

The utility of pattern is decided based upon the success rate that a pattern can provide a user to achieve his goal. Different users using the patterns have different goal. The same pattern may be used by many for different purposes. A pattern which provides good results in various fields where it is used shows a great utility. But utilizing a pattern successfully depends on the users.

Actionability

The pattern used for a particular application must involve in making the decision of future action of that application. Association rule plays an important role in actionability. Since predicting the future status of the domain where pattern is supposed to use has a largest impact on the development of the domain.

3. CRUNCHING BIG DATA BY IN-MEMORY ANALYTICS

Garber [2] proposed an idea on in-memory analytics on investigating or exploring data in RAM instead of using data

in the disk that would increase the accessibly and produce accurate results even for a large volume of data. Using RAM would increase the cost even though a massive collection of data could be analyzed instantly. In-memory analytics could also perform parallel analytics of massive data by operating multiple systems as a single system. In-memory analytics approaches data in such a way that it divides data into multiple set of tasks and involve them in parallel operation by transferring them across blade servers. The benefits are saves time, efficient, optimize the data hence reduce the volume that could be handled easily and precise. On the other hand it is less efficient when worked with old and large quantity of data and also data with duplicate and unnecessary materials. In-memory analytics could easily handle data in RAM, but the main challenge is fitting large set of data in RAM is not much easy. Multitenancy is not very easy hence cloud computing is not possible.

4. BUSINESS ANALYTICS

The volume of data is been explosive in social media, e-commerce and raising interest in business. However enterprises don't know how to automate the big data resulting in a good decision for the development of that stream. Sunil el at [3] has proposed new ideas for better understanding of big data and making best use of it by the enterprises and also points out the health care field for controlling the cost, improvement of techniques for better care and efficiency and improvement of quality. The paper has also explained about the cross-organizational supply chain process and optimization-from upstream R&D .The content also alerts about the hype, confusion and fears in case of big data and also explains about the vendors who were hijacking their own commercial benefit. In order to leverage the potentiality of big data and business analytics completely the strategies and capability of an organization must be completely synchronized. The organizations must develop governance process and additional systems so as to manage the internal data and also coherently integrate data from both internal and external data sources.

Big data analytics is also been employed in business analytics and evidence based medicines [EBM]. EBM are conscientious, judicious, explicit usage of present evidence of various analytics in medical field in order to improve the personal medical care. Evidence from the external clinics would help in developing the quality of an individual clinic. Omar El-Gayar el at [5] has explained about the opportunities on big data analytics in evidence based medicines. Kolodziej [6] has demonstrated how EBM is beneficial in three different ways 1. Analysis could produce best result on the best therapy for a particular disease that has high success rate 2. The cheap and best medicine among many medicines of same composition could be determined by analyzing data of exploit volume 3. EBM also ranks out the best hospital with good quality and care but analyzing the data extracted from various physicians. The paper also reviewed about the seven steps involved in EBM process initially it involves in identifying the patient's condition. Once the condition is identified the patient's are questioned with formulated EBM questions. From the report the evidence are gathered and evaluated. Later the evidences are converted into consumable unit. Then the evidences are presented and used. Finally the evidence is brought upto practice.

Improvement in the areas of Business Intelligence (BI) and data mining technology had result in the development of business operation methods. Internet environment, pathway for communication, has provided a platform for collection of enormous amount of data from various communicating resources [10]. Data obtained from business sectors that undergo rescheduling and re-planning may reduce the risk of failure and also improves the efficiency and profitability of the operation.] F. Zhang *et al.*, [11] has demonstrated the use of big data distributed in clouds to create a dynamic workload scheduling problem. Zhang *et al.* [12] in his another paper has also analyzed about the issue of cost minimization when the data moves around a geographically dispersed data. W. Dou el at [13] has developed a service optimization model concerning the data privacy to handle data stored in clouds. If data is not ready to be provided to privacy issues by the user the service quality may be compromised. Service quality can be improved by service optimization model which can be verified by simulation study.

5. GREEN SMART GRID

Smart grid promises to bring green revolution by integrating renewable resources come into the energy of mainstream. The author describes that processing big data is the main energy black hole in green revolution. Zakia Asad et al [14] also proposed the need for transforming data enterprises into energy efficient enterprises. The paper explains that when data gathered from transmission systems, PMUs and AMI, weather reports, customer billing and behavioral extracts, distributed generations, electricity market and clearing house transferred into a single processing system will help in enhancement in energy trading, economic dispatch, load planning, day ahead forecast, trading and selling megawatts, load flow analysis, micro grid analytics etc. When big data is transferred from small grid to the data centers the outcome will be the analytics about the big data. The components of cross plane green orchestrator are as follows: green lessor, pre-execution analyzer, network state predictor, server state predictor, network traffic optimizer, VMizer, Pizer, server-dynamic power manager, network-dynamic power manager, post execution analyzer.

6. PRIVACY-PRESERVING COMPUTING

Lu et al [15] has proposed the architecture of big data analytics. The architecture composed of three main parts including multi-source big data collecting, distributed big data storing, intra/ inter big data processing. Multi-sourcing of data is characterized by high volume, high velocity and high variety. This method improves the efficiency, speed, and accuracy of data by categorizing it into structured and unstructured data to maximize the value of big data. Storing the massive volume of data in a centralized data storing data will increase the risk of storing hence data is distribute into various data centers which reduce the risk of data losing. Processing all the distributed data in parallel will help in quick accessing due to which many new knowledge and innovations nevertheless the data belongs to the same organization or not. When all the data belongs to the same organization are processed in parallel they are known as intra big data processing. When data belongs to different organization is processed in parallel they are known as inter big data processing. The most challenging between the two is inter-big data processing. The paper also proposed the operations on encrypted data in order to protect individual privacy in big data analytics. The issues of flexility, efficiency or risks in de-identifications are the main challenges in preserving the privacy in big data analytics. Developing privacy-preserving algorithms might help in reducing the risk of re-identification in privacy-preserving of big data analytics.

7. INTERNET OF THINGS

Sun et al [16] promotes the concept of Smart and Connected Communities (SCC). The characteristics of SCC are livability, preservation, revitalization, sustainability. The paper has also discussed the challenges and opportunities of IoT in SCC. Some of them are: community or participatory sensing collection of social and public sensings. A new community sensing paradigm called mobile crowd sensing is emerging due to the development in the smart phones with various sensors such as camera, audio, accelerometer, GPS etc., the interconnecting layer is provided in order to transfer data among different domains and devices. The data layer is provided in order to store massive, trivial, and heterogeneous data in the sensing layer. The service is provided in order to provide various services for communities. Mobile crowdsensing and cyber-physical cloud computing are the main opportunities of big data analytics in IoT. Big data analytics also faces major challenges in SCC they are cyber security and privacy, data heterogeneity, decision making under uncertainity and resource limitations. Big data analytics in SCC finds its applications in health care, smart interconnected automobile and trucks and smart building.

8. CONCLUSION

This paper presents the basic concepts pertaining to the big data analytics and data mining. This paper also investigates the applications of big data analytics in various domains like business intelligence, internest of things and smart grid. Issues in privacy preserving data analytics is discussed in clear way.

9. REFERENCES

[1] Liqiang Geng & Howard J. Hamilton, "Interestingness Measures for Data Mining: A Survey", ACM Computing Surveys, Vol. 38, No. 6, pp. 1-5, 2006

[2] Lee Garber, "Using In-Memory Analytics to Quickly Crunch Big Data".

[3] Sunil Mithas, University of Maryland,Maria R. Lee, Shih Chien University, Taiwan,Seth Earley, Earley & Associates,San Murugesan, BRITE Professional Services, Australia,Reza Djavanshir, Johns Hopkins University, "leveraging big data and business analytics",Vol. No. pp

[4] Junbo Wang, Member, IEEE, Yilang Wu, Student Member, IEEE, Neil Yen, Member, IEEE, Song,Guo, Senior Member, IEEE, and Zixue Cheng, Member, IEE: "Big Data Analytics for Emergency Communication Networks: A Survey".

[5] Omar El-Gayar,Dakota State University, USA, Prem Timsina, Dakota State University, USA, "Opportunities for Business Intelligence and Big Data Analytics In Evidence Based Medicine".

[6] http://www.cancernetwork.com/practice/content/article/10165/1821731, accessed Aug 30, 2013,

[7] http://searchsqlserver.techtarget.com/definition/data-mining.

[8] https://www.google.co.in/search?q=data+mining&aq=f&oq=data+mining&aqs=chrome.0.57j0l2j5.11931&sourceid=chrome&ie=UTF-8.

[9] Tsan-Ming Choi, Member, IEEE, Hing Kai Chan, Senior Member, IEEE, and Xiaohang Yue, "Recent Development in Big Data Analytics for Business Operations and Risk Management", Vol.

[10] S. Chaudhuri, U. Dayal, and V. Narasayya, "An overview of business intelligence technology," *Commun. ACM*, vol. 54, no. 8, pp. 88–98, 2011.

[11] F. Zhang *et al.*, "Evolutionary scheduling of dynamic multitasking workloads for big-data analytics in elastic cloud," *IEEE Trans. Emerg. Topics Comput.*, vol. 2, no. 3, pp. 338–351, Sep. 2014.

[12] W. Dou, X. Zhang, J. Liu, and J. Chen, "HireSome-II: Towards privacyaware cross-cloud service composition for big data applications," *IEEE Trans. Parallel Distrib. Syst.*, vol. 26, no. 2, pp. 455–466, Feb. 2015.

[13] W. Dou, X. Zhang, J. Liu, and J. Chen, "HireSome-II: Towards privacyaware cross-cloud service composition for big data applications," *IEEE Trans. Parallel Distrib. Syst.*, vol. 26, no. 2, pp. 455–466, Feb. 2015.

[14] Zakia Asad, *Student Member, IEEE*, and Mohammad Asad Rehman Chaudhry, *Member, IEEE "*A Two-Way Street: Green Big Data Processingfor a Greener Smart Grid*"*

[15] Rongxing Lu, Hui Zhu, Ximeng Liu, Joseph K. Liu, and Jun Shao," Toward Efficient and Privacy-Preserving Computing in Big Data Era".

[16] YUNCHUAN SUN1, (Member, IEEE), HOUBING SONG2, (Senior Member, IEEE), ANTONIO J. JARA3, (Member, IEEE), AND RONGFANG BIE4, (Member, IEEE), "Internet of Things and Big Data Analytics for Smart and Connected Communities".

[17] Y. Sun *et al.*, ``Organizing and querying the big sensing data with event- linked network in the Internet of Things," *Int. J. Distrib. Sensor Netw.*, vol. 2014, 2014, Art. No. 218521.

[18] Y. Sun and A. J. Jara, ``An extensible and active semantic model of information organizing for the Internet of Things," *Pers. Ubiquitous Comput.*, vol. 18, no. 8, pp. 1821_1833, 2014.

[19] Y. Sun, H. Yan, C. Lu, R. Bie, and Z. Zhou, Constructing the Web of events from raw data in theWeb of things," *Mobile Inf. Syst.*, vol. 10, no. 1, pp. 105_125, 2014.

[20] Y. Sun, C. Lu, R. Bie, and J. Zhang, ``Semantic relation computing theory and its application," *J. Netw. Comput. Appl.*, vol. 59, pp. 219_229, Jan. 2016