Big Data Mining: A Literature Review

Imran Rashid Banday
Department of Computer Sciences University of Kashmir, Srinagar, J&K, India

Majid Zaman Scientist,
Directorate of Information Technology & Support Systems, University of Kashmir, Srinagar, J&K, India

S. M. K. Quadri
Professor, Department of Computer Science, Jamia Millia Islamia, India

Muheet Ahmed Butt
Scientist ,PG Department of Computer Sciences, University of Kashmir, Srinagar, J&K, India

Abstract— Big Data concern large-volume, Structured or Unstructured , complex, growing data sets with multiple, independent heterogeneous sources. Data has grown tremendously and increasing every day. The velocity of its generation and growth is increasing, driven in part by the number of internet connected devices. Moreover, the variety of data being generated is also expanding, and organization's capability to capture and process this data is limited. Current technology, architecture, management and analysis approaches are unable to cope with the flood of data, and organizations will need to change the way they think about, plan, govern, manage, process and report on data to realize the potential of big data. Exploring the big data and extracting useful information and knowledge from large volume of data is a challenging one.

Along with the increase in the applications of Big Data there has been parallel increase in problems allied with this technology. In order to gain a higher productivity and efficient data utilization, the analysis of Big Data at enterprise level is must.correct.

Key words —Big Data, Data Mining, Heterogeneous sources, 3Vs, 6Vs , HACE theorem

I. INTRODUCTION

IN organizations data is growing enormously and doubles in size every two years. With the growth in size it is hard for organizations to manage, consume, process, transform and analyze these massive data sets. The size of these data sets is beyond the ability of traditional software technologies to capture, store, manage and process it in "tolerable elapsed time". For more than 20 years many organizations have already being using enterprise data warehouses (EDW) for providing information to support decision makers and to serve as central data system for reporting. These Data warehouses are now becoming overloaded with data and are almost unable to handle all the data. Considering an example of Google search engine, it has to process approximate 20 Petabytes of data every day. Certain survey states that in 2010, there were 1.2 Zetta bytes of data present as digital data and in the same way in 2011 there were 300 Quadrillion of file present as unstructured data. Thus, considering such huge data collection and its efficient retrieval, a concept called Big Data came into existence. The concept of present trends of Big Data is different from the traditional Business Intelligence approach. There have been a huge applications of Big Data in present day technical scenario, such as applications for financial sector, retails, manufacturing, healthcare, mobile, social media and government agencies or applications and in education sector. Actually the era of Big Data has begun. Each day 2.5 quintillion bytes of data is created.

In last two years the 90% of the data which the world is having today has been created[1]. Almost every one whether physicist, computer scientists, physicists, mathematicians, bio-informaticists, political scientists, sociologists, and many others are yelling for access to the massive quantities of information created by and about things, people and their interactions. Big data are now expanding in all engineering, science fields. A report by the forum, "Big Data, Big Impact," declared data a new class of economic asset, like currency or gold. Sam Madden from Massachusetts Institute of Technology (MIT) defines Big data as " data that's too big, too fast, or too hard for existing tools to process.

II. BIG DATA SOURCES AND TYPES

Big data is a term for large data set with multiple autonomous and heterogeneous sources. Data from different sources varies in its format. This data comes from databases, posts to social media sites , weblogs, sensors used to gather climate information, purchase transaction records , cell phone, GPS signals, digital pictures and videos to name a few. Primarily data is obtained from following types of sources

1) Internal Sources : Internal sources are those data sources that provide generally organized data that originates from within the enterprise. Data from Internal sources (like student Life Cycle, Customer Management, Resource Management etc) is generally operational data. This type of data is used by enterprises for their daily operations like OLTP.

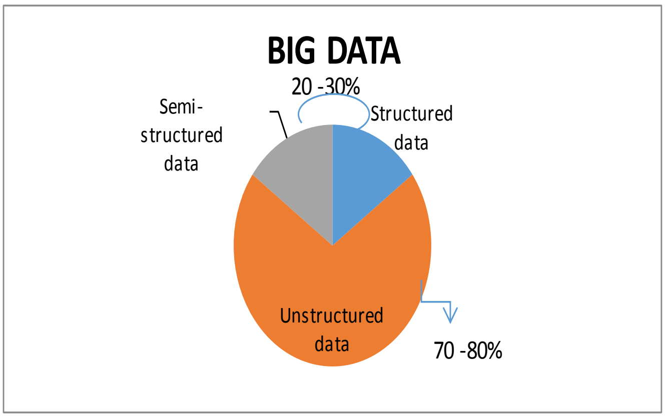
2)External Sources: External Sources are those data sources that provide unorganized data that originates from the external environment of an enterprise. Data from these sources (like Internet, industry partners, Government, weblogs etc) is analysed to understand the external entities (students, customers, Competitors).

On the bases of sources the big data comprises following types of data

a) Structured data: structured data can be defined as the data that has predefined format and usually stored in tabular format. The sources for this type of data is Flat files(delimiter separated values),relational databases

b) Unstructured data: Unstructured data is the data that may or may not have and predefined format or repeating patterns. Unstructured data consists of data having different formats like text, images, audio, video, emails etc. Sources for this type of data is documents, logs, survey results, feedbacks, social networking platforms, mobile data.

c) Semi-structured data: Semi-structured data can be defined as the data that does not follow the proper structure of the data models as in relation databases. This type of data contains labels or mark-up components in order to separate elements and generate hierarchies of records and fields in the given data.



The above figure illustrates the types of data that big data comprises.

Big Data Characteristics: 3Vs by Doug Laney in Big Data Management are used to characterize different aspects of big data.

Volume:-The amount of data. Perhaps the characteristic most associated with big data, volume refers to the mass quantities of data that organizations are trying to harness to improve decision making across the organization. Data volumes continue to increase at an unprecedented rate.

Variety:- Different types of data and data sources. Variety is about managing the complexity of multiple data types, including structured, semi-structured and unstructured data. Organizations need to integrate and analyze data from a complex array of both traditional and non-traditional information sources, from within and outside the enterprise. With the explosion of sensors, smart devices and social collaboration technologies, data is being generated in countless forms, including: text, web data, tweets, audio, video, log files and more.

Velocity:- Data in motion. The speed at which data is produced, processed and analyzed continues to accelerate.

Nowadays from a business viewpoint there are three more V's added to the Big data characteristics.

Variability:- Refers to number of inconsistencies in the data and multitude of data dimensions resulting from multiple disparate data types and sources. It depends on the users how they want to interpret data as there are changes in the structure of Data.

Veracity:-It refers to the uncertainty of data which include noises, biases and abnormality in the data. Lot of efforts are required in cleansing data so that only valuable data is stored and it is suitable for analysis.

Value: This is the most important among all other characteristics of Big Data. Big data Mining is meaningless if we don't gain business value from the data.

III. RELATED WORK

The era of Big Data has begun [2], [3], [4]. According to IDC (technology research firm) there is huge data, all the time and growing at 50 percent every year or doubles in every two years. The 'Big Data' term came into existence because huge date is being created every day. Big Data is the data sets having large volume, complexity and multiple autonomous heterogeneous sources. Big Data is now rapidly growing in all science and engineering domains. The term big data was for first time sighted in as slilcon Graphics (SGI) slide deck by John Mashey in 1998 with title " Big Data and the Next Wave of InfraStress"[5]. The first book that mentioned Big Data is a book on data mining that appeared in 1998 by Weiss and Indrukya[6]. But the first academic paper was in 2000 by Diebold in which the words 'Big Data' appeared in its title[7]. At the KDD BigMine'12Workshop, Usama Fayyad [8] in his invited talk, gave amazing data figures about internet usage ,like there are 250 million tweets per day on twitter, ever day Google has to process more than 1 billion queries, on the Facebook there are more than 800 million updates per day, and likewise there are billion views per day on YouTube. New Algorithms and tools are required as the data is growing around 40% every year. Estimation of data that we create these days is made in order of Terabytes , Petabytes, Exabytes, Zettabytes.

a) Wei Fan and Albert Bifet in their paper titled " Mining Big Data : Current Status and Forecast to the Future" say that Big Data is the data sets having large size and complexity due to its volume, variability, and velocity cannot be managed with current data mining tools or methodologies. This Paper shows that Big data is becoming new area of research and we are at the beginning of this era and with Big data mining we can discover that knowledge that has not been discovered before by any one.

b) Jimmy Lin and Dmitriy Ryaboy (Twitter,Inc.) "Scaling Big Data Mining Infrastructure: The Twitter Experience" , Over the past few years at Twitter the analytics platform has shown tremendous growth in terms of complexity, number of users and size . This paper presents that to perform analytics is difficult with existing data mining tools. Preparatory work for applying data mining process is time consuming and significant amount of effort is required to get robust solutions from preliminary models.

c) Yizhou and Sun Jiawei Han in their paper "Mining Heterogeneous Information Networks: A Structural Analysis Approach" this paper presents that real world data & objects are mostly interconnected that results in information networks that are heterogeneous, complex and often semi structured. These information networks which are heterogeneous can surprisingly discover rich knowledge from data. Mining these information networks is new and promising frontier in big data research.

d) U Kang and Christos Faloutsos in their paper "Big Graph Mining: Algorithms and discoveries " presented an overview of mining big graphs, focusing in the use of Pegasus tool and showing some findings in the Twitter social network and Web Graph. The paper gives future research directions for mining big graphs.

e) Xavier Amatriain in his paper titled as "Mining Large Streams of User Data for Personalized Recommendations " presented some lessons learned with the Netix Prize, and discuss the recommender and personalization techniques used in Netix. It also discusses current important problems and future research directions.

f) Danah boyd & Kate Crawford (2011) presented that the Big Data era has begun. Almost every one like Computer scientists, economists, physicists, political scientists ,mathematicians , bio-informaticists, sociologists, and other scholars are clamoring for access to the gigantic quantities of information produced by and about people, things, and their interactions. They also gave six provocations for Big Data.

g) Mukherjee et al. (2012) "Shared disk big data analytics with Apache Hadoop" Big data analytics is a process of analyzing large amount of data sets to discover hidden patters and other useful information. Google developed Mapreduce Framework which is referred by Big data analytics and Apache Hadoop is an open source platform used for the purpose of implementation of Google's Mapreduce Model . In this the performance of HDFS is compared with SF-CFS using the SWIM. SWIM by the facebook contains the workloads of job traces with complex data arrival and computation patterns.

h) Hsinchun Chen et al (2012) said ,"Now, in this era of Big Data, even while BI&A 2.0 is still maturing, we find ourselves poised at the brink of BI&A 3.0, with all the attendant uncertainty that new and potentially revolutionary technologies bring". In a paper submitted in April 2013, Renu Kanwar, Prakriti Trivedi & Kuldeep Singh claimed that NoSQL is the solution for use cases where ACID is not the major concern and uses BASE instead which works up on eventual consistency.

i) Vibha Shukla and Pawan Kumar Dubey in their research paper said that " we have entered an era of Big Data. The amount of information and knowledge that can be extracted from the digital universe is continuing to expand as users come up with new ways to massage and process data. Big data is not formally and structurally defined. The big data technology is still in its infancy. The analysis of big data is confronted with many challenges, but the current research is early stage. This paper is a collaborative research effort to begin examining the traditional view of data analytics and big data analytics, introduces new techniques and technologies to handle big data. We have identified some major challenges regarding big data which big data users specifically face. We must support and encourage fundamental research towards addressing these technical challenges if we are to achieve the promised benefits of Big Data. Our future research will concentrate on developing a more complete understanding of challenges associated with big data. "

j) Bharti Thakur, Manish Mann in their paper "Data Mining for Big Data: A Review" This paper presents an overview of mining big data. Big data is a collection of large, complex data sets containing both structured and unstructured data and data mining is a technique to discover knowledge from large amount of data. High performance computing platforms are required for big data mining. This paper also presents challenges in Big Data.

k) Xindong Wu, Xingquan Zhu, Gong-Qing Wu, and Wei Ding in their research paper titled as "Data Mining with Big Data" has presented a HACE theorem that characterizes the features of the Big Data revolution, and proposes a Big Data processing model, from the data mining perspective. This data-driven model involves demand-driven aggregation of information sources, mining and analysis, user interest modeling, and security and privacy considerations.

IV. BIG DATA ANALYTICS

The real power of big data lies in its analysis. In today's world people want to understand data and its importance and use it in their decision makings. Data Analytics is the process to extract information from data. In other words data analytics is the process of finding patterns, relationships and information among number of fields. What data is to be mined and for what use varies radically from one organization to another, as does the nature and organization of the data, so there can be no such thing as a generic "data mining tool". Following are the domains in which data Analytics is highly useful in

- Customer Retention
- Market Analysis and Management
- Production Control
- Fraud Detection
- Corporate Analysis & Risk Management
- Science Exploration
- Education

- Astrology
- Sports

Data analytics of Big Data is not possible on conventional databases. Actually big data affects the analytical process and technologies that are used for analytics. For big data analytics high performance computing platforms are required.

V. CONCLUSION

Enormous Data will keep developing amid the following years, and every data researcher should oversee substantially more measure of information consistently. This information will be more assorted, bigger, and quicker. Enormous Data is turning into the new Final Frontier for scientific information examine and for business applications. Actually Big data is an emerging trend and its need is arising in almost in very filed of science and engineering. We are toward the start of another time where Big Data mining will help us to find learning that nobody has found some time recently. In this era of data flood big data analytics is of great importance that can provide unforeseen insights and help in making better decision in various areas.

REFERENCES

[1] "IBM What Is Big Data: Bring Big Data to the Enterprise", http://www.ibm.com/software/data /bigdata/, IBM, 2012.

[2] Nature Editorial, "Community Cleverness Required", Nature, vol. 455, no. 7209, p. 1, Sept. 2008.

[3] A. Labrinidis and H. Jagadish, "Challenges and Opportunities with Big Data" Proc. VLDB Endowment , vol. 5, no. 12, 2032-2033,2012.

[4] J. Mervis, "U.S. Science Policy: Agencies Rally to Tackle Big Data" Science, vol. 336, no. 6077, p. 22, 2012.

[5] F. Diebold. On the Origin(s) and Development of the Term "Big Data". Pier working paper archive, Penn Institute for Economic Research, Department of Economics, University of Pennsylvania, 2012.

[6] S. M. Weiss and N. Indurkhya, "Predictive data mining: a practical guide" Morgan Kaufmann Publishers Inc, San Francisco, CA, USA, 1998.

[7] F. Diebold., "Big Data" Dynamic Factor Models for Macroeconomic Measurement and Forecasting. Discussion Read to the Eighth World Congress of the Econo-metric Society, 2000.

[8] U. Fayyad. "Big Data Analytics: Applications and Opportunities in On-line Predictive Modeling", http://big-data-mining.org/keynotes/#fayyad, 2012.

[9] Danah boyd & Kate Crawford , "Six Provocations for Big Data", Oxford Internet Institute's "A Decade in Internet Time: Symposium on the Dynamics of the Internet and Society" on September 21, 2011.

[10] Danah boyd & Kate Crawford , "CRITICAL QUESTIONS FOR BIG DATA",Routledge, Information, Communication & Society Vol. 15, No. 5, June 2012, pp. 662–679 ISSN 1369-118 (2012).

[11] Mukherjee, A.; Datta, J.; Jorapur, R.; Singhvi, R.; Haloi, S.; Akram, W. , "Shared disk big data analytics with Apache Hadoop" , Dec.,2012.

[12] Hsinchun Chen , Roger H. L , Veda C., "BUSINESS INTELLIGENCE AND ANALYTICS: FROM BIG DATA TO BIG IMPACT", MIS Quarterly Vol. 36 No. 4/December 2012,Eller College of Management, University of Arizona, Tucson, AZ 85721 U.S.A.(2012).

[13] Vibha Shukla and Pawan Kumar Dubey , "Big Data: Moving Forward with Emerging Technology and Challenges", International Journal of Advance Research in Computer Science and Management Studies(IJARCSMS), Volume 2, Issue 9, September 2014.

[14] Bharti Thakur, Manish Mann, "Data Mining for Big Data: A Review ", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 4, Issue 5, May 2014.

[15] Xindong Wu, Xingquan Zhu, Gong-Qing Wu, and Wei Ding, "Data Mining with Big Data", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 26, NO. 1, JANUARY 2014.