Big Data and Big Data Mining: Study of Approaches, Issues and Future scope

Nibedita Chakraborty[#1], Sandeep Gonnade[*2]

[#]*M.Tech. Scholar, Department of Computer Science & Engineering, Mats University*

[*]*Asst. Professor, Department of Computer Science & Engineering, Mats University*

Raipur, India

***Abstract*— Big Data is a latest term introduced to define large and complex datasets. Due to their size and complexity, it is not possible to manage them with our conventional techniques or data mining tools. Using Big Data mining organizations can extract useful information from these large pool of data or streams of data. By analysis of this datasets useful statistics can be extracted. In spite of the usability of Big Data, there are several challenges related to it. This challenge is becoming most evolutionary area of research for the coming years. The paper presents an overview of the topic, methodologies and forecast to the future.**

***Keywords*— Big data, Big data mining, Map Reduce, A-Priori algorithm, PCY algorithm.**

I. INTRODUCTION

Data is the key factor today. It includes personal, professional, social data and more. Digitalization and interconnectivity lead to an unexpected growth of data. The increased use of media and physical networking through sensor networks for business & private purposes generates an enormous amount of data. This in response changes business processes & open up new opportunities worldwide. The internet is a key driver for data growth. The worldwide generated data already exceed the available storage.

Since 2011 interest in an area known as big data has increased exponentially [4]. Unlike the vast majority of computer science research, big data has received significant public and media interest. The era of "big data" has opened several doors of opportunities to upgrade science, boost health care services, improve economic growth, reconstruct our educational system, and prepare new types of social interaction and entertainment services. The area of big data is fast-evolving, and is likely to be subject to improvements and amendments in the future. The paper is all about the study of evaluation of Big data, its mining and issues related to it.

II. BIG DATA

Big data usually includes data sets with sizes beyond the ability of commonly used software tools to capture, curate, manage, and process data within a tolerable elapsed time.[5] Big Data concern large-volume, complex, growing data sets with multiple, autonomous sources. "With the fast development of networking, data storage, and the data collection capacity, Big Data are now rapidly expanding in all science and engineering domains, including physical. biological and biomedical sciences" [1]. Zhu, Wu and Ding propose HACE theorem which explains Heterogeneous, Autonomous, Complex and Evolutionary data sets. In 2012, Gartner has given its definition as follows: "Big data is high volume, high velocity, and/or high variety information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization."[6]. These 3 dimensions of Big data can be understood as: **Volume** which is continuously increasing. It could be in Gigabytes, Terabytes, Petabytes, Exabytes and more. **Variety** means different types of data. They could be in the form of text, audio, sensor data, video, social network data etc. and **Velocity** is continuous arrival of stream of data from which useful information is to be extracted. They can be understood in the form of batch, periodic and real time data. The Boston Consulting Group estimates total growth of 2.5 Exabyte, which equals 2.5 billion gigabytes, per day.

The large volume of such data needs Scalability and parallelism in which today's DBMS systems are not capable. To handle the problem several attempts have been made on developing large parallel processing architectures. The very first attempt was made by Google. Google proposed a programming model named MapReduce. It was composed of two procedures Map() and Reduced(), the previous one performs filtering and sorting while second one generates summaries. It allows large data entries in parallel. The model was coupled with the GFS (Google File System), a distributed file system where the data is partitioned over thousands of nodes within a cluster. In 2005 Yahoo created an Apache open- source version of Google's MapReduce framework, called Hadoop MapReduce. It was an open source software framework for distributed storage and processing of big data on clusters. The Apache Hadoop framework consists of four modules: Hadoop Common, Hadoop Distributed File System(HDFS), Hadoop YARN, Hadoop MapReduced. First one contains library and utilities. The second HDFS is an open source version of the Google's GFS which partitions the files in to large blocks and distributes them among clusters. YARN stands for Yet Another Resource Negotiator, responsible for managing cluster resources and finally Hadoop MapReduce process this large set of data.

In MapReduce method, the input is firstly divided into a large set of key-value pairs; then the map function is called. After all data entries are processed, a new set of key-value

pairs are produced, and then the reduce function is called to group or merge the produced values based on common keys. To support the MapReduce computing model, Google developed the BigTable – "a distributed storage system designed for managing structured data. BigTable can scale well to a very large size: petabytes of data across thousands of commodity severs" [2].

III. MINING OF BIG DATA

A. Data Mining

There are many conventional data mining algorithms available like Market basket analysis, A-priori algorithm, hash based improvement techniques and multistage algorithm etc. The market basket analysis finds the frequent item sets from the database which appears at least n times. The data is stored in a file basket-by-basket. The A-priori algorithm is used which searches in the database to find frequent itemsets where k-itemsets are used to generate k+1- itemsets. It has a two pass approach which reduces the need of main memory. It finds the pairs of items that appear at least s times together. Data is stored in a file one basket at a time. The limitation is the time complexity of the algorithm. PCY algorithm is the hash based improvement of to A-priori. During pass 1 of A-priori most memory is idle. PCY uses that memory to keep count of buckets in to which pairs of items are hashed. In the second pass it gives extra condition that candidate pairs must satisfy. The idea of multistage algorithm is After Pass 1 of PCY, rehash only those pairs that qualify for Pass 2 of PCY. On middle pass, fewer pairs contribute to buckets, so fewer *false drops* buckets that have count *s* , yet no pair that hashes to that bucket has count *s* .[8]

B. Big Data Mining

The traditional algorithms explained above are not capable of mining such large distributed data. The above mentioned data mining algorithms extract from and analyze the historical datasets for decision making. The purpose of Big data mining is to go beyond the usual request-response processing, market basket analysis or uncovering some hidden relationships and patterns between numerical parameters of data but to design and implement very large scale parallel data mining algorithm. Comparing with the results derived from mining the conventional datasets, unveiling the huge volume of interconnected heterogeneous big data has the potential to maximize our knowledge in the target domain. However, this brings a series of new challenges to the research community. Overcoming the challenges will reshape the future of the data mining technology, resulting in a spectrum of groundbreaking data and mining techniques and algorithms.

One feasible approach is to improve existing techniques and algorithms by exploiting massively parallel computing architectures. Big data mining must deal with heterogeneity, extreme scale, velocity, privacy, accuracy, trust, and interactivity that existing mining techniques and algorithms are incapable of.

The operations related to Big data value chain can be divided in to following 3 processes: [11]

1) Big data collection: It focuses on data collection and data control over data access. They contain large datasets, collect and append new data and provide access to those data on request. Companies related to it sell or license data access & data sets. One of the examples is twitter that sales data feeds to Gnip.

2) Big data Aggregation & Integration: The operation focuses on building technical infrastructure for data aggregation, management and backup of Big data. They provide software tools to manage and restore useful data from Big data pool and integrate them for decision making. Companies like Oracle sale technical infrastructure service & consulting services. Oracle big data provide appliance as an integrated Big data solution.

3) Big data Analytics: The operation focuses on goal oriented analysis of big data sets. The analysis tools extract meaningful dataset from large pool of data and analyze them according to user perspective and for users benefit. Organizations like kaggle, SAP and others sale data analysis and visualization services.

C. Big Data Application

Big data can be useful in both commercial & non-commercial purpose. It can be listed as follows:

1) Targeted customer approach
2) Identification & solution of customer issues in real time.
3) Easier information search & knowledge building.
4) Competence building by R&D.
5) Improved analysis capabilities etc.

These applications will help people and organizations to have better services & better customer experience.

IV. OPEN SOURCE TOOLS

The Big data phenomenon brought with itself the open source software revolution. Many big companies like Facebook, Yahoo!, Twitter, LinkedIn, Amazon etc. are working collaboratively or independently for open source projects. The infrastructure generally deals with Apache Hadoop, a software for data intensive distributed application. With it applications can be written which rapidly process large data. Apache Pig, Apache Hive, Apache HBase, Cascading, Scribe are the projects related to Apache Hadoop. Apache Mahout, MOA are Big data mining open source tools while GraphLab is graph mining tools built open source.

V. CHALLENGES RELATED TO BIG DATA MINING

During the study following issues are found related to Big Data Mining:

A. Variety of Data

There are unlimited information sources that generate or create big data. This leads to huge variety of big data. Mining useful information from such heterogeneous environment is great challenge.

B. Scalability of Data

The undefined volume of big data requires high scalability of its data management & mining tools. The scalability

gradually increases because more data generates more knowledge. Cloud computing with parallelism can deal with the situations.

C. Security

Big data uses large volumes of data that may be held in the cloud and it may involve distributed processing across several servers. It has been suggested that the growth of big data increases the threats to the security of information. The European Union Agency for Network and Information Security (ENISA) has identified [10] a number of emerging threats arising from the potential misuse of big data. ENISA consider that the 'threat trend' is increasing in this area. The ENISA report says that "uncontrolled collection, usage and dissemination of user and systems data are the perfect playground for malicious activities". This implies that a key issue is how far the growth of big data is "uncontrolled".

D. Reliability

In past data mining systems were relatively accurate & reliable because the data resources were well known & countable. With emerging trend of big data, the problem generated because not all sources are well known, verifiable & also the number of sources are limitless. Therefore reliability has become a big issue.

E. Mining & cleaning of unused data

In the large environment of big data presence of unused data is a big issue. The unused data captures most of the useful space of memory but to have a durable & sustainable Big data mining system, mining & cleaning of unused garbage data is very essential & recommended.

VI. CONCLUSION

In the present time Big data is a buzzword from news article to media, from tweets to YouTube, from blog discussions to science projects. The area is, not surprisingly, computer science; but one can notice other disciplines that investigate the topic such as engineering, mathematics, business and also social and decision sciences. But there is still much to do in developing a professional approach to big data mining. The paper focuses on several issues related to big data and improving those issues will result in better environment for Big Data. In our proposed work we are going to work on the security issue related to Big Data. The proposed work is to develop a security system to protect the Big data server from unauthorized access.

ACKNOWLEDGMENT

We are grateful to Amulya Prasad for helpful discussion.

REFERENCES

[1] Xingquan Zhu, Gong-Qing Wu, Wei Ding "Data Mining with Big Data", Knowledge and DataEngineering, IEEE Transactions on vol: 26, issue 1, June 2013.

[2] DunrenChe, MejdlSafran and ZhiyongPeng "From Big Data to Big Data Mining: Challenges, Issues, and Opportunities" Springer: Database System for Advanced Application, volume 7827, 2013

[3] Jonathan Stuart Ward and Adam Barker "Undefined By Data: A Survey of Big Data Definitions" arXiv: 1309, 20 Sep 2013.

[4] Google. Google Trends 2013

[5] Chris Snijders, UweMatzat, Ulf-Dietrich Reips "Big data: Big gaps of Knowledge in the field of internet science" International Journal of Internet Science, vol 7(1), pp 1-5, 2012.

[6] Laney, Douglas "The importance of 'Big Data': A definition, Gartner, Retrieved 21 June 2012.

[7] Dean J. Ghemawat S. "Map Reduce a flexible data processing tool" In communications of the ACM vol 53 no. 1 pp 72-77,Jan 2010.

[8] Aggarwal Shruti, Ranveer Kaur "Comparative Study of Various Improved Versions of Apriori Algorithm" International Journal of Engineering Trends and Technology (IJETT), Volume4Issue4- April 2013.

[9] By the SMW "Security and Privacy in the Era of Big Data", A RENCI/ National Consortium for Data Science WHITE PAPER.

[10] By Big Data Working Group "Big Data Analytics for Security Intelligence", CLOUD SECURITY ALLIANCE Big Data Analytics for Security Intelligence September 2013.

[11] By Arnold Picot "From open Data to Big Data Opportunities and Challenges from a Business Perspective", 2nd International Open Data Dialog, Berlin, November 18, 2013.