Mining Big Data: scope and challenges

Arti S. Bhoir [#1], Archana Gulati [*2]

[#] *Computer Department , University of Mumbai*
S. S. Jondhale College of Engineering,
Dombivli, INDIA.
[1] artibhoir03@gmail.com

[*] *Datta Meghe College of Engineering,*
Airoli, Navi Mumbai, INDIA.
[2] gulatiarchana81@gmail.com

Abstract—**Nowadays, Big Data is considered as a driving force for business & technology innovations as well as economic growth. Big data is the term for collection of data sets, which are large & complex. With Big Data databases, enterprises can save money, generate more revenue, and achieve many other business goals. Data Mining is the process of analysing data from different perspectives & summarizing it into useful information. This paper focuses on various features of Data mining and challenges in Big Data. It further provides solution to the challenges faced by mining in Big Data.**

Keywords— **Data Mining, Big Data, Data mining techniques, Data mining Algorithms.**

I. INTRODUCTION

Big Data is a new term used to identify the datasets that are large in size and complexity. Big Data mining is the capability of extracting useful information from these large datasets or streams of data.

The most fundamental challenge for big data Application is to explore the large volumes of data & Extract useful information or Knowledge for future action [1]. In Knowledge extracting process, it is infeasible for storing all observed data. HACE theorem [2] suggests that the key characteristics of the big data are

- Huge with Heterogeneous & different Data.
- Decentralized control
- Complex Data & Knowledge associations

There are many data mining challenges [3] with Big Data with respect to data processing

1. Computing effective platform
2. Data Privacy
3. The Big data mining algorithm designs in tackling the difficulties rose by the Big Data volumes, distributed data distributions, and by complex and dynamic data characteristics.

To tackle the problems of Big Data processing, we have studied the solution for challenges of big data mining are as follows

1] MapReduce –MapReduce is a parallel programming technique, which is proposed by Google for large scale data processing in the distributed computing environment

2] There are four privacy preserving Methods such as K-anonymity, L-Diversity-T-closeness for maintaining Privacy in big Data.

3] Multi-Database mining technique that selects relevant database & searches only the set of all relevant database, this overcomes the problems of existing data mining techniques or tools. This approach is effective in reducing search cost for the application. The Database selection approach is application-dependent.Organization of this paper is as follows. The section 2 focuses Big Data and data mining. Section 3 describes the data mining challenges in Big Data. Section 4 describes solutions to the challenges in mining Big Data and finally Section 5 concludes the paper.

II. BIG DATA AND DATA MINING

There are two types of Big Data:

1. Structured [4]
2. Unstructured [4]

Structured data are numbers and words that can be easily categorized and analyzed. These data are generated by things like network sensors embedded in electronic devices, smartphone and global positioning system (GPS) devices. Structured data also include things like sales figures, account balances and truncation.

Unstructured data include more complex information such as customer reviews from commercial websites, photos and other multimedia, and comments on social networking sites. These data cannot easily be separated into categories or analyzed numerically [5].

2.1 Objectives for BIG DATA

1. Provides security to the data in Cloud using Cryptographic schemes.
2. Prevent unauthorized access to the data stored in cloud.
3. Provides Independent and Concurrent access to the data in Cloud.
4. Provide Role-Based Access, i.e. provides access only if the user has access permission.

2.2 Data mining for big data

Generally, Data mining is the process of analyzing data from different perspectives and summarizing it into useful information –information that can be used to increase revenue, cost cutting or both. Technically, Data Mining is the process of finding correlations or pattern among dozens of fields in large relational database. Applying existing data mining algorithms and techniques to real-world problems has been recently running into many challenges due to the inadequate scalability

(and other limitations) of these algorithms and techniques that do not match the three V's of the emerging Big Data. Not only the scale of data generated today is unprecedented, the produced data is often continuously generated in the form of streams that require being processed as well as mined in real time.

III. DATA MINING CHALLENGES IN BIG DATA

To handle Big Data, the essential key is to scale up to the exceptionally large volume of data and provide treatments for the characteristics featured by the aforementioned HACE theorem. There are some challenges in handling a massive dataset which is a Big Data. These challenges are due to the large and complex data.

3.1 Computing Big Data mining platform

Basically, the Big Data is stored at different places and also the data volumes may get increased as the data keeps on increasing continuously. So, to collect all the data stored at different places is that much expensive. Suppose, if we use these typical data mining methods (those methods which are used for mining the small scale data in our personal computer systems) for mining of Big Data, and then it would become an obstacle for it. Because the typical methods are required data to be loaded in main memory, this is a clear technical barrier for big data because moving data across different location is expensive.

3.2 Data Privacy

Data privacy has been always an issue even from the beginning when data mining was applied to real-world data. The concern has become extremely serious with big data mining that often requires personal information in order to produce relevant/accurate results such as location-based and personalized services, e.g., targeted and individualized advertisements [6]. Also, with the huge volume of big data such as social media that contains tremendous amount of highly interconnected personal information, every piece of information about everybody can be mined out, and when all pieces of the information about a person are dug out and put together, any privacy about that individual instantly disappears.

3.3 Design BIG DATA algorithms

While designing such algorithms, we face various challenges. When we divide the Big Data into number of subsets, and apply the mining algorithms on those subsets, the results of those mining algorithms will not always point us to the actual result as we want when we collect the results together [7].

IV. Solutions to the challenges

The large datasets are new challenges for data mining algorithms to efficiently process within given conditions such as memory and execution time. To overcome the challenges, data mining algorithms can be implemented with Map-Reduce, which break the large datasets into small chunks and process them in parallel on multiple cluster nodes and scales easily to mine hundreds of terabytes data by adding inexpensive commodity computers.

4.1 Solution for computing Big Data mining

For Big Data mining, as data scale is far beyond the capacity that a single personal computer (PC) can handle, a typical Big Data processing framework will rely on cluster computers with a high-performance computing platform, with a data mining task being deployed by running some parallel programming tools, such as MapReduce [8], several attempts have been made on exploiting massive parallel processing architectures. The first such attempt was made by Google. Google created a programming model named MapReduce that was coupled with (and facilitated by) the GFS (Google File System) [9], a distributed file system where the data can be easily partitioned over thousands of nodes in a cluster.

Later, Yahoo and other big companies created an Apache open- source version of Google's MapReduce framework, called Hadoop MapReduce. It uses the Hadoop Distributed File System (HDFS) – an open source version of the Google's GFS. The MapReduce system manages by marshaling the distributed servers, running various tasks in parallel, managing all communications and data transfers between the various parts of the system. MapReduce stable inputs and outputs are usually stored in a distributed file system. The transient data is usually stored on local disk and fetched remotely by the reducers. The mapping operation can be performed in parallel, when each mapping operation is independent of the other. The parallelism helps to overcome failure, when one node fails the work is assigned to other nodes as the input data is still available. The data structure (key, value) pair is used to define both map and reduce methods.

Map step: The master node divides the problems into smaller pieces and assigns them to work node.

Map (key1, value1) → list (key2, value2)

The Map function is applied in parallel to every pair in the input dataset. This produces a list of pairs for each call. The Reduce function is then applied in parallel to each group.

Reduce step: The master node collects the solution from work node and combines them with some procedure

Reduce (k2, list (v2)) → list (v3)

Another way to look at MapReduce is as a 5 step parallel and distributed computation:

Prepare the Map () input –the "MapReduce system" designates Map processors, assigns the K1 input key value each processor would work on, and provides

1. That processor with all the input data associated with that key value.
2. Run the user provided Map () code–Map () is run exactly once for each K1 key value, generating output organized by key values K2.
3. "Shuffle" the Map output to the Reduce processors– the MapReduce system designates Reduce processors, assigns the K2 key value each processor should work on, and provides that processor with all the Map-generated data associated with that key value.

4. Run the user-provided Reduce () code–Reduce () is run exactly once for each K2 key value produced by the Map step.
5. Produce the final output–the MapReduce system collects all the Reduce output, and sorts it by K2 to produce the final outcome.

Pros of MapReduce

1. Easy to use for programmers with no experience in distributed systems
2. Hides details of parallelization, load balancing, fault tolerance, task management from the user
3. Massively scalable
4. Provides status monitoring tools

4.2 Solution to design of Big Data Algorithm

When we divide the Big Data into number of subsets, and apply the mining algorithms on those subsets, the results of those mining algorithms will not always point us to the actual result as we want when we collect the results together

Design Big data mining algorithm is having challenges from tackling the difficulties raised by the Big data volume ,Distributed data distribution and by complex and dynamic data characteristic but new process are introduced for Multi–database Mining algorithm.

For Example interstate (or international) company consists of multiple branches. The National Bank of Australia, for example, has many branches in different locations. Each branch has its own database, and the bank data is widely distributed and thus becomes a multi-database [12] problem. This is responsible for the development and decision-making, an interstate company consists of n branches at different places its own database multiple databases places.

Many organizations have a pressing need to manipulate all the data from their different branches rapidly and reliably. This need is very difficult to satisfy when the data is stored in many independent databases, and the data is all of importance to an organization. Formulating and implementing queries requires data from more than one database. It requires knowledge of where all the data is stored, mastery of all the necessary interfaces and the ability to correctly combine partial results from individual queries into a single result.

The computing environment is becoming increasingly widespread through the use of Internet and other computer communication networks. In this environment, it has become more critical to develop methods for building multi-database systems that combine relevant data from many sources and present the data in a form that is comprehensible for users.

One possible way for multi-database mining is to integrate all the data from these databases to amass a huge dataset for discovery by mono-database mining techniques.

While collecting all data together from different branches might produce a huge database amylose some important patterns for purpose of centralized processing, forwarding the local patterns (rather than the original raw data) to central company headquarters provides a feasible means of dealing with multiple database problems. The patterns forwarded from branches are called *local patterns*. However, the number of forwarded local patterns may be so large that browsing the pattern set and finding interesting patterns can be rather difficult for central company headquarters. Therefore, it can be difficult to identify which of the forwarded patterns (including different and identical ones) are really useful at the central company level.

To mine multi-databases, the first method (mono-database mining technique) is to put all the data together from multiple databases to create a huge mono-dataset. There are various problems with this approach. In order to confront the size of datasets, we have studied an alternative multi-database mining technique that selects relevant databases and searches only the set of all relevant databases. Their work has focused on the first step in multi-database mining, which is the identification of databases that are most relevant to an application. A relevance measure was thus proposed to identify relevant databases for mining with an objective to find patterns or regularity within certain attributes. This can overcome the drawbacks that are the result of forcedly joining all databases into a single huge database upon which existing data mining techniques or tools are applied. The approach is effective in reducing search costs for a given application.

There are limitations in existing multi-database mining techniques. For these reasons, we have studied a high-performance prototype system for multi-database mining (MDM). Below we introduce our MDM design through defining a new process of multi-database mining and describing its functions.

A. Three Steps in MDM

The MDM process focuses on local pattern analysis as follows. Given n databases within a large organization, MDM performs three steps:

1. Searching for a good classification of these databases.
2. Identifying two types of new patterns from local patterns: high-vote patterns and exceptional patterns.
3. Synthesizing local patterns by weighting.

(a) Searching for a good classification:- In a multi-database environment, a pattern has attributes such as the name of the pattern, the rate voted for by branches, and supports (and confidences for a rule) in branches that vote
for the pattern. In other words, a pattern is a super-point of the form; we have studied a local pattern analysis strategy by using the techniques.

The key problem to be solved is how to analyse the diverse projections of patterns in a multi-dimension space consisting of local patterns within a company.

(b) Identifying high-vote patterns: - Within a company, each branch, large or small, has a power to vote for patterns for global decision-making. Some patterns can receive votes from most of the branches. These patterns are referred to as high vote patterns. These patterns may be far more important in terms of global decision-making within the company. Because traditional mining techniques cannot identify high vote patterns, these patterns are regarded as novel patterns in multi-

databases. We have studied a mining strategy for identifying high-vote patterns of interest based on a local pattern analysis. The key problem to be solved in this mining strategy is how to post-analyse high-vote patterns.

(c) Finding exceptional patterns: - Like high-vote patterns, exceptional patterns are also regarded as novel patterns in multi-databases. But an exceptional pattern receives votes from only a few branches. While high-vote patterns are useful when a company is making global decisions, headquarters are also interested in viewing exceptional patterns when special decisions are made at only a few of the branches, perhaps for predicting the sales of a new product. Exceptional patterns can capture the individuality of branches. Therefore, these patterns are also very important.

(d) **Synthesizing patterns by weighting**. Although each branch has a power to vote for patterns for making global decisions, branches may be different in importance their company.

For example, if the sale of branch A is 4 times of that of branch B in a company, branch A is more important than branch B in the company. The decisions of the company can be reasonably partial to high-sale branches. Also, local patterns may have different supports in different branches. We will a new strategy for synthesizing local patterns based on an efficient model for synthesizing patterns from local patterns by weight.

IV. CONCLUSION

Big data is the term for a collection of complex data sets, Data mining is an analytic process designed to explore data from different perceptive. There are number of Data mining challenges for handling Big Data. This paper studies the solution for choosing Big Data computing Platform, handling Data privacy and design of Big Data algorithm.

REFERENCES

[1] Wei-Fan &Albert Bifet "Mining Big data: current status & forecast to the future", SIGKDD Explorations volume 14 Issue 2, 2013.

[2] Xindog Wu, Gong-Quins Wu &Wei-dins "Data mining with Big data", IEEE transaction paper knowledge and data engineering Volume 26 issue1, JAN 2014.

[3] Bhoj Raj Sharman, Daljeet Kaur, Manjub, "A review on data mining: its challenges, issues and Applications", IJCET volume 3 issue 2, June 2013.P.P. 695-700.

[4] Bharti Thakur, Manish Mann, "Data mining for Big Data: Review", IJARCSSE Volume 4, Issue 5, May 2014.

[5] Mary-Hall, Robert M , "Rethinking Abstraction for Big Data why where how & what", http://www.cs.utah.edu/~jeffp/papers/BD-white.pdf accessed on Sept 2014.

[6] Rongxing Lu, Hui Zhu, Ximeng Liu, Joseph K. Liu, and Jun Shaoo "Toward Efficient and Privacy-Preserving Computing in Big Data Era". Network, IEEE Volume 28, Issue 4 July-August 2014 P.P. 46-50.

[7] Dunren Che, Mejdi safran, Zhiyong Peng, "From Big Data to Big Data mining: Challenges, issues and opportunities", DASFAA Workshop 2013 P.P. 1-15.

[8] D. Jayalatchumy , p. Thombiduraj, A. Alomela "Parallel processing of big data Using Power Iteration clustering over MapReduce" IEEE Computing and Communication Technologies, World Congress 2014.

[9] Jeffrey dean sanjay Ghemawat, "MapReduce: Simplified Data Processing on Large Clusters", Communications of the ACM - 50th anniversary issue: 1958 – 2008.

[10] Andrew Ton and M. saravanan Erricson research, " Big DataPrivacyPreservation", http://www.ericsson.com/research-blog/data-knowledge/big-data-privacy-preservation accessed on Sept 2014.

[11] Ashwin Machanavajjhala, Jerome P. Reiter, "Big Privacy: protecting confidentially in Big Data" XRDS: Crossroads, The ACM Magazine for Students - Big Data Volume 19 Issue 1, Fall2012 P.P. 20-23.

[12] Shichao Zhang, Xingdong Wu, Chengaizhang, "Multi-database Mining", IEEE Computational Intelligence Bulletin, June 2003.

