ISSN:0975-9646

A Survey on Big Data, Data Mining and Overlay Based Parallel Data Mining

Pushpanjali[1], Jyothi S Nayak[2]

[1]*M.Tech (CSE), BMSCE, Bangalore, India.*

[2]*Associate Prof.Dept.of CSE, BMSCE, Bangalore, India*

Abstract: **- The main goal of the data mining process is to extract useful information from Big Data set and transform it into an understandable form for further use. It was not possible to extract useful information from the large datasets or data streams. Now this can be achieved by the capability of Big Data Mining. The overlay- based parallel data mining architecture executes processing by employing the overlay network and fully distributed data management, which can achieve high scalability and service availability. In case of the physical network disruption, an overlay-based parallel mining architecture is not capable of providing data mining services if router/communication line breakdowns, because of this, numerous nodes are removed from the overlay network. An overlay network construction scheme based on physical network structure, including nodes location and a distributed task allocation scheme using overlay network technology is done to overcome with this issue. In this survey paper, a review of Overlay based parallel data mining and its different technologies are studied.**

Keywords **— Data Mining, Big Data, Overlay-based, service availability, physical network disruption.**

I. Introduction

Data Mining and Knowledge Discovery are usually defined as the extraction of patterns or models from observed data, usually the ability to explore much richer and more expressive models, providing new and interesting domains for the application of learning algorithms. The wide availability of large-scale data sets from different domains has created a demand to automate the process of extracting valuable information from them. For example, consider Facebook application, we upload various types of information such as text, images and video. The process of effective mining of such data is known as big data mining [1]. Today is the era of Google, the thing which is unknown is searched in Google and within fractions of seconds; we get the number of links as a result. This would be the best example for the processing of Big Data. This Big Data is not any different thing than out regular term data. The Big Data is nothing but a data in an extreme large amount available of heterogeneous, autonomous sources, which get updated in fractions of seconds [2], [3]. Conventional parallel data mining architectures with centralized management schemes lack scalability, which causes bottleneck in the entire system and this leads to decrease in performance of the system as the number of nodes increases [4]. As a remedy for improving scalability, an overlay-based parallel data mining architecture has been proposed. Since all the nodes execute both management and processing functions by using overlay network, this architecture can balance the management load. Furthermore, this architecture also achieves higher service availability against the breakdown of master node because it keeps providing the data mining until the overlay network is disrupted [5], [6].

However, distributed networks are intolerant of (i.e., not resilient to) network failures due to the lack of a centralized infrastructure to manage the frequent joining and departure of an enormous number of nodes. There has been a great discussion on this critical issue. While the distributed network management has been studied to solve this issue in general [7], [8] and [9]. In order to construct distributed networks, robust enough against network-failures without a management system the present study was undertaken. Categorically, we fixate on the degree distribution of the distributed networks and propose a method to construct networks predicated on the degree distribution robust against network failures. This survey is limited to the Big Data Mining Applications and Overlay Based Parallel data mining, the rest of this paper is organized as follows: In Section 2, we introduce the basic terminology on data mining, big data and Overlay Based Data Mining. In Section 3, we describe previous work on parallelization of most well-known data mining and Big Data algorithms in the research community. With each algorithm we present our envisioned overlay-based parallel data mining architecture are discussed, as well as the experimental works. The conclusion is presented in Section 4.

II. Basic Concepts And Terminology

In this competitive world, top level management needs to take right decisions at the right time for giving better service to customers and to provide a better organizational image. Decisions based on better analysis results in increasing profit and decreasing loss. Management is dependent on better analytical and data mining services for this purpose. Fig.1 describes the relationship between each step in the process, the technologies that can be used to complete each step in Microsoft SQL Server. The Data mining offers a wide range of algorithms used for analysis, pattern discovery and prediction. This includes techniques such as association rule mining, decision trees, regression, support vector machines and much more. In the last twenty years, a lot of research has been done on improvising performance of data mining techniques.

From past to present, three different trends have been observed in the data mining process. Most of the sequential algorithms were part of the centralized approach where all data is needed to be stored on a central node, which was the first trend. The second trend was observed in terms of parallelizing centralized algorithms. For parallelization, the two main approaches were used: Task parallelism and Data parallelism. Parallel computing techniques took a boost with the advent of multi core CPUs and cheaper GPUs. A combination of both GPU and CPU resulted in multifold performance benefit.

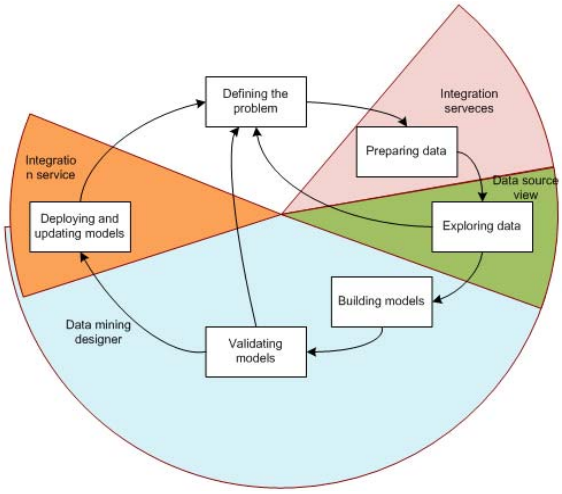

Fig. 1 Data Mining Process

Last trend is distributed data mining where data mining techniques were applied in different distributed computing paradigms like peer to peer, clusters, grids and cloud environment. The evaluation was done on how eager the people were to recognize Big Data today, and know its importance, how to deal with and to know its outcomes and benefits. Apart from rampant speculation, what other companies are actually doing and planning about Big Data today? Are they looking for a need to adopt new technologies? Which direction they are going in? What are their major concerns?

The three major results of the survey are

- An overwhelming majority of organizations view their Big Data processing as mission critical.
- The companies that are handling Big Data, there is a need for both significance and growth of real-time functionality. The survey indicated that there is increasing readiness to use streaming solutions to deal with the challenges of Big Data and speed up Big Data processing.
- Most of the companies have been planning to move their Big Data to the Cloud.
- It has been observed that only 20% of the IT professionals have surveyed and indicated that their company had no Plans to move their Big Data to the cloud.
- The first goal of our survey was to assess how important Big Data is for the companies, today. Nearly 80% of responders indicated that Big Data was important to their business, with 43% indicating it is mission critical.

Due to largeness in size, decentralized control and different data sources with different types, the Big Data becomes much complex and harder to handle. They cannot be managed by the local tools those we use for managing the regular data in real time. For major Big Data-related applications, such as Google, Flicker, Facebook, a large number of server farms are deployed all over the world to ensure nonstop services and quick responses for local markets. And when Big Data is divided into a number of subsets, and apply the mining algorithms on them, the results of these mining algorithms will not always point us to the actual result as we want when we collect the results together.

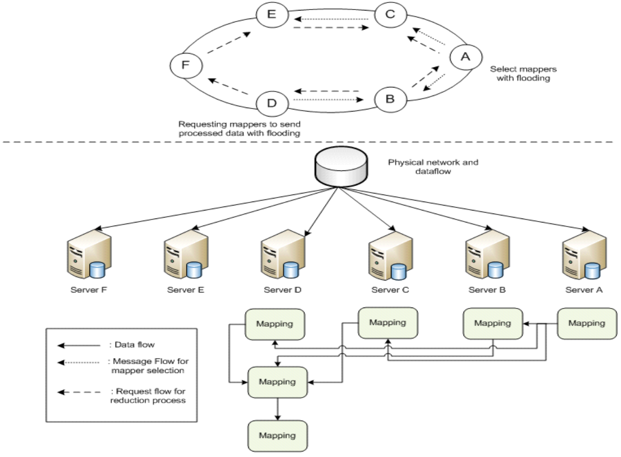
E
C
A
F
D
B
Select mappers with flooding
Requesting mappers to send processed data with flooding
Physical network and dataflow
Server F
Server E
Server D
Server C
Server B
Server A
: Data flow
: Message Flow for mapper selection
: Request flow for reduction process
Mapping
Mapping
Mapping
Mapping
Mapping
Mapping

Fig. 2 Mapping and reduction processes in overlay-based parallel data mining architecture.

Overlay-based parallel data mining is one of the architectures that improve the service availability against server break- downs. In this architecture, all the servers execute both management and processing functions. The overlay network is constructed by all servers and utilized to and processing nodes, which are similar to the master nodes in the conventional architecture. This overlay architecture can keep providing the service even if some nodes are removed from the overlay network. Fig. 2 shows an example of mapping and reduction processes in the overlay-based parallel data mining architecture. When a data processing request is injected, a node that received the request (node A in the Fig. 2) executes a reception function using the overlay network [10]. In other words, the mappers are found by the node using flooding message, where mappers are randomly selected (nodes B, C, and D in the Fig. 2). Then, a mapper that initially finished the mapping process (node D in the Fig 2) becomes a reducer, and it requests to other mappers to transmit the processed data to itself, where the request message can be forwarded by using flooding scheme. When the processed data are received from mappers, the reducer executes the reduction process and outputs the analysed result [24].

In this architecture, since the connectivity of overlay network dramatically affects the service availability of data mining, there are numerous works, which tackled the connectivity issue from the various points of view, example, context- cognizant, graph theory predicated, and intricate network theory predicated overlay network construction schemes [11]. These works make overlay networks, tolerant to minute-scale server breakdowns, but do not consider the sizable voluminous- scale server breakdowns, i.e., a physical network disruption. Therefore, this paper develops an overlay-based parallel data mining architecture that is tolerant to physical network disruption so that data mining is available at all times and at any place.

III. Challenging issues in data mining with big data.

There are three sectors at which the challenges of Big Data arrive. They are:

- Mining platform.
- Privacy.
- Security
- Design of mining algorithms.

Basically, the Big Data is stored in different places and the data volumes may get increased as the data keeps on increasing continuously. So, to collect all the data stored in 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. Though we have super large main memory, the typical methods are required to load the data in main memory [12]. Variety, Volume, Velocity and Accuracy are essential characteristics of big data. Variety, data from multiple sources inherently possesses many types and different forms like structural, semi structured and unstructured data. Scalability, large volume of big data requires high scalability of its data management and mining tools. Speed of data mining depends on the data access time and efficiency [13], [21].

To maintain the privacy is one of the main aim of data mining algorithms. Presently, to mine information from Big Data, parallel computing based algorithms such as Map Reduce are used [14]. The large data sets are divided into a number of subsets and then, mining algorithms are applied to those subsets, in such algorithms. Finally, summation algorithms are applied to the results of mining algorithms to meet the goal of the Big Data mining. During this whole procedure, the privacy statements obviously break as we divide the single Big Data into a number of smaller datasets [1], [14], [13].

An emerging topic in data mining is privacy preserving data mining, the basic idea of privacy preserving data mining is performing data mining algorithms effectively without compromising the security of sensitive information contained in the data [15]. Fuzzy fingerprint is one of the data mining technique that enhances data privacy during data leak detection operation which is based on sensitive data. The main goal of privacy preservation is protecting private data while processing or releasing sensitive information [16], [17]. S. Moncrieff et.al [18] proposes a solution which is based on environmental context to dynamically alter the privacy levels in the smart house. Fabio Borges [19] proposes a privacy preserving protocol for smart metering systems to ensure customers' privacy and security in the network data. The security concerns have become a major barrier to the widespread growth of cloud computing. Distributed architecture is used to eliminate the risks during data mining based attacks [20].

From the above study, we found that, Data mining process is not easy and the algorithm used for mining is very complicated. The data needs to be integrated from the various heterogeneous data sources as is not available at one place. Data mining derives its name from the similarities between searching for valuable business information in a large database. For example, finding the linked products in gigabytes of store scanner data and mining a mountain for a vein of valuable one [22],[23]. Both processes require either shifting through an immense amount of material, or reasonably probing it to find exactly where the value resides. The databases of sufficient size, quality and the data mining technology can generate new business opportunities by providing some capabilities.

IV. Techniques in data mining process

A. *Decision trees*

Tree- shaped structures that represent sets of decisions. These decisions generate the rules for classification of a dataset. Classification and Regression Trees (CART) and Chi Square Automatic Interaction Detection (CHAID) are included under specific decision tree methods.

B. Genetic algorithms

Genetic combination, mutation, and natural selection are the process used in Optimization techniques for design based on the concepts of evolution.

C. Nearest neighbor method

The technique that classifies each record in a dataset based on a combination of the classes of the records(s) most similar to it in a historical Sometimes called the k-nearest neighbor technique, where k is the number of neighbors.

D. Artificial neural networks

Non-linear predictive models, which are based on biological neural systems.

At present, on the level of the mining platform sector, parallel programming models like Map Reduce are being used for the purpose of analysis and mining of data. Map Reduce is a batch-oriented parallel computing model [1], [11], [13], [14]. There is still a certain gap in performance with relational databases. Improving the performance of Map Reduce and enhancing the real-time nature of large-scale data processing have received a significant amount of attention with Map Reduce parallel programming being applied to many machine learning and data mining algorithms. These data mining algorithms usually need to scan through the training data for obtaining the statistics to solve or optimize the model.

V. Conclusion

In this paper, we briefly reviewed the sundry data mining applications and this survey is based on various issues of data mining. Conventional parallel data mining architecture will not provide data mining services in case of network disruption. In the future, we will review the sundry relegation algorithms and consequentiality of evolutionary computing (genetic programming) approach in designing of efficient relegation algorithms for data mining and withal by utilizing overlay-predicated data mining architecture can potentially provide scalable data mining in immensely colossal-scale network. Big Data is becoming the new Final Frontier for scientific data research and for business applications. We are at the commencement of an incipient era where Big Data mining will avail us to discover cognizance that no one has discovered afore.

ACKNOWLEDGMENT

The work reported in this paper is supported by the college through the TECHNICAL EDUCATION QUALITY IMPROVEMENT PROGRAMME [TEQIP-II] of the MHRD, Government of India.

REFERENCES

[1] *Ubiquitous Analytics: Interacting with Big Data Anywhere, Anytime*, vol.46, Issue 4, IEEE, 2013.

[2] *Data Mining with Big Data*, vol. 26, no. 1, IEEE, 2014.

[3] Rohit Pitre and Vijay Kolekar, “A Survey Paper on Data Mining With Big Data”, *International Journal of Innovative Research in Advanced Engineering (IJIRAE)*, vol. 1, Issue 1, April 2014.

[4] "GFS: Evolution on Fast-Forward”, *Communications of the ACM*, March 2010, vol. 53, no. 3.

[5] Jing Zhang, Gongqing Wu, Xuegang Hu and Xindong Wu, "A Distributed Cache for Hadoop Distributed File System in Real-time Cloud Services" 2012 ACM/IEEE 13th International Conference on Grid Computing, DOI 10.1109, Grid.2012.17.

[6] *Dynamic Cloud Deployment of a Map Reduce Architecture*, IEEE Std 1089-7801, 2012.

[7] *Distributed Data Mining in Peer-to-Peer Networks*, IEEE Std. 1089-7801, 2006.

[8] S. V. S. Ganga Devi, "A Survey on Distributed Data Mining and it’s Trends", *IMPACT: IJRET*, vol. 2, Issue 3, Mar 2014.

[9] Rekha Sunny and Sabu M. Thampi, “Survey on Distributed Data Mining in P2P Networks”.

[10] Wenjun Xiao, Mingxin He and Huomin Liang, “Cayley CCC: A Robust P2P Overlay Network with Simple Routing and Small-World Features”, *Journal of Networks*, vol. 6, Issue 9,September 2011

[11] Katsuya Suto, Hiroki Nishiyama, Xuemin Shen and Nei Kato1, “Designing P2P Networks Tolerant to Attacks and Faults Based on Bimodal Degree Distribution”, *Journal of Communications*, vol 7, Issue 8, August 2012.

[12] Nikita Jain and Vishal Srivastava “Data Mining Techniques: A Survey Paper”, *International Journal of Research in Engineering and Technology*, vol. 2, Issue 11, November 2013.

[13] Dunren Che, Mejdl Safran, and Zhiyong Peng, *From Big Data to Big Data Mining: Challenges, Issues, and Opportunities*, B. Hong et al. (Eds.): DASFAA Workshops 2013, Springer-Verlag Berlin Heidelberg 2013

[14] Michael Cardosa, Aameek Singh, Himabindu Pucha and Abhishek Chandra, "Exploiting Spatio-Temporal Tradeoffs for Energy-aware MapReduce in the Cloud", *IEEE 4th International Conference on Cloud Computing*, 2011.

[15] *Information Security in Big Data: Privacy and Data Mining*, IEEE Std., vol 2, 2014.

[16] Privacy-Preserving Detection of Sensitive Data Exposure, IEEE Std. 1556-6013, 2015.

[17] *Privacy Preserving Data Analytics for Smart Homes*, IEEE Security and Privacy Workshops, 2013.

[18] *Dynamic Privacy in a Smart House Environment*, IEEE Std. 1-4244-1017-7, 2007.

[19] *EPPP4SMS: Efficient Privacy-Preserving Protocol for Smart Metering Systems and Its Simulation Using Real-World Data*", IEEE, vol. 5, no. 6, November 2014.

[20] An Approach to Protect the Privacy of Cloud Data from Data Mining Based Attacks", IEEE Std. 978-0-7695-4956-9, 2012.

[21] *Big Data Processing in Cloud Computing Environments*", IEEE, 1087-4089, 2012.

[22] Amandeep Kaur Mann and Navneet Kaur, “Survey Paper on Clustering Techniques”, *International Journal of Science Engineering and Technology Research (IJSETR)*, vol. 2, Issue 4, 2013.

[23] Anirban Mukhopadhyay, Ujjwal Maulik, Sanghamitra Bandyopadhyay and Carlos A. Coello Coello, "A Survey of Multi-Objective Evolutionary Algorithms for Data Mining: Part-II", *IEEE Trans. Evol. Comput.*

[24] *An Overlay-Based Data Mining Architecture Tolerant to Physical Network Disruptions*, IEEE Std.10.1109, 2014.