International Journal of Emerging Trends & Technology in Computer Science (IJETTCS)
Web Site: www.ijettcs.org Email: editor@ijettcs.org
Volume 3, Issue 4 July-August 2014 **ISSN 2278-6856**

Issues and Challenges in the Era of Big Data Mining

B R Prakash[1*], Dr. M. Hanumanthappa [2]
[1]Assistant Professor, Department of MCA,
Sri Siddhartha Institute of Technology, Tumkur

[2]Professor, Department of Computer Science & Applications,
Bangalore University, Bangalore

Abstract
The amount of data being generated and stored is growing exponentially, owed in part to the continuing advances in computer technology. While "big data" has become a highlighted buzzword since last two year, "big data mining", i.e., mining from big data, has almost immediately followed up as an emerging, interrelated research area. This paper provides an overview of big data mining and discusses the related challenges and the new opportunities. The discussion includes a review of state-of-the-art frameworks and platforms for processing and managing big data as well as the efforts expected on big data mining. We address broad issues related to big data and/or big data mining, and point out opportunities and research topics as they shall duly flesh out..

Keywords:-data mining, big data, big data mining, knowledge discovery, data-intensive computation.

1. INTRODUCTION

Enormous amounts of data are generated every minute. Some sources of data, such as those found on the Internet are obvious. Social networking sites, search and retrieval engines, media sharing sites, stock trading sites, and news sources continually store enormous amounts of new data throughout the day. For example, a recent study estimated that every minute, Google receives over 2 million queries, e-mail users send over 200 million messages, YouTube users upload 48 hours of video, Facebook users share over 680,000 pieces of content, and Twitter users generate 100,000 tweets. Some sources of data are not as obvious. Consider the vast quantity data collected from sensors in meteorological and climate systems, or patient monitoring systems in hospitals. Data acquisition and control systems, such as those found in cars, airplanes, cell towers, and power plants, all collect unending streams of data. The health care industry is inundated with data from patient records alone. Insurance companies collect data for every claim submitted, fervently working to catch increasing quantities of fraudulent claims. Regardless of the source of the data, contained within them are nuggets of knowledge that can potentially improve our understanding of the world around us. The challenge before us lies in the development of systems and methods that can extract these nuggets. The topic of "data" is no longer a subject for only the computer science major. As educators, we need to encourage and engage our students in all STEM fields in a way that raises their awareness of the challenges and opportunities in being able to work with large datasets. We need to work hard to provide our students with the knowledge required to work competently in the emerging field of Big Data. Scalability is at the core of the expected new technologies to meet the challenges coming along with big data. The simultaneously emerging and fast maturing cloud computing technology delivers the most promising platforms to realize the needed Scalability with demonstrated elasticity and parallelism capacities. Numerous notable attempts have been initiated to exploit massive parallel processing architectures[1]. Google's novel programming model, MapReduce [2], and its distributed file system, GFS (Google File System) [3], represent the early groundbreaking efforts made in this line. From the data mining perspective, mining big data has opened many new challenges and opportunities. Even though big data bears greater value (i.e., hidden knowledge and more valuable insights), it brings tremendous challenges to extract these hidden knowledge and insights from big data since the established process of knowledge discovering and data mining from conventional datasets was not designed to and will not work well with big data. The cons of current data mining techniques when applied to big data are centered on their inadequate Scalability and parallelism. In general, existing data mining techniques encounter great difficulties when they are required to handle the unprecedented heterogeneity, volume, speed, privacy, accuracy, and trust coming along with big data and big data mining. Improving existing techniques by applying massive parallel processing architectures and novel distributed storage systems, and designing innovative mining techniques based on new frame works/platforms with the potential to successfully overcome the aforementioned challenges will change and reshape the future of the data mining technology.

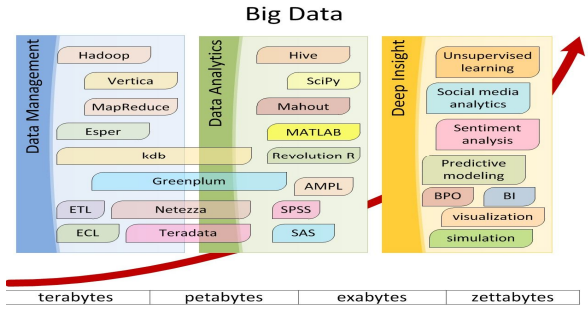

Figure 1: Big Data Management, Data Analytics, Deep insight

2. DATA MINING

Knowledge discovery (KDD) is a process of unveiling hidden knowledge and insights from a large volume of data [4], which involves data mining as its core and the most challenging and interesting step (while other steps are also indispensable) . Typically, data mining uncovers interesting patterns and relationships hidden in a large volume of raw data, and the results tapped out may help make valuable predictions or future observations in the real world. Data mining has been used by a wide range of applications such as business, medicine, science and engineering. It has led to numerous beneficial services to many walks of real businesses – both the providers and ultimately the consumers of services. 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 Vs 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 and mined in (nearly) real time. Delayed discovery of even highly valuable knowledge invalidates the usefulness of the discovered knowledge. Big data not only brings new challenges, but also brings opportunities – the interconnected big data with complex and heterogeneous contents bear new sources of knowledge and insights. Big data would become a useless monster if we don't have the right tools to harness its "wildness". Current data mining techniques and algorithms are not ready to meet the new challenges of big data. Mining big data demands highly scalable strategies and algorithms, more effective pre processing steps such as data filtering and integration, advanced parallel computing environments (e.g., cloud Paas and IaaS), and intelligent and effective user interaction. Next we examine the concept and big data and related issues, including emerging challenges and the (foregoing and ongoing) attempts initiated on dealing with big data.

3. BIG DATA

We are sure living in an interesting era – the era of big data and cloud computing, full of challenges and opportunities. Organizations have already started to deal with petabyte-scale collections of data; and they are about to face the exabyte scale of big data and the accompanying benefits and challenges. Big data is believed to play a critical role in the future in all walks of our lives and our societies. For example, governments have now started mining the contents of social media networks and blogs, and online-transactions and other sources of information to identify the need for government facilities, to recognize the suspicious organizational groups, and to predict future events (threats or promises). Additionally, service providers start to track their customers' purchases made through online, instore, and on-phone, and customers' behaviors through recorded streams of online clicks, as well as product reviews and ranking, for improving their marketing efforts, predicting new growth points of profits, and increasing customer satisfaction. The mismatch between the demands of the big data management and the capabilities that current DBMSs can provide has reached the historically high peak. The three Vs (volume, variety, and velocity) of big data each implies one distinct aspect of critical deficiencies of today's DBMSs.

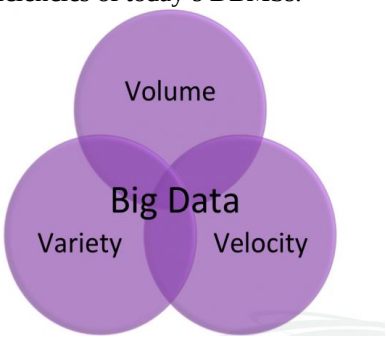

Figure 2: The three Vs (volume, variety, and velocity)

Gigantic volume requires equally great scalability and massive parallelism that are beyond the capability of today's DBMSs; the great variety of data types of big data particularly unfits the restriction of the closed processing architecture of current database systems [5]; the speed/velocity request of big data (especially stream data) processing asks for commensurate real-time efficiency which again is far beyond where current DBMSs could reach. The limited availability of current DBMSs defeats the velocity request of big data from yet another angle (Current DBMSs typically require to first import/load data into their storage systems that enforces a uniform format before any access/processing is allowed. Confronted with the huge volume of big data, the importing/loading stage could take hours, days, or even months. This causes substantially delayed/reduced availability of the DBMSs). To overcome this scalability challenge of big data, 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 [5] that was coupled with (and facilitated by) the GFS (Google File System [6]), 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 framework allows users to define two functions, map and reduce, to process a large number data entries in parallel [7]. More specifically, in MapReduce, the input is divided into a large set of key-value pairs first; then the map function is called and forked into many instances concurrently processing on the large key-value pairs. After all data entries are processed, a new set of key-value pairs are produced, and then the reduce function is called to group/merge the produced values based on common keys.

In order to match/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 [8]. In the same spirit, Amazon created Dynamo [9], which is also a key-value pair storage system. The Apache open-source community acted quickly again, created HBase – an open-source version of Google's BigTable built on top of HDFS and Cassandra – an open-source version of Amazon's Dynamo. Apache Hive [10] is an open source data warehouse system built on top of Hadoop for querying and analyzing files stored in HDFS using a simple query language called HiveQL. Hadoop is not alone; it has other competitor platforms. All these platforms lack many niceties existing in DBMSs. Some of the competitors improved on existing platforms (mostly on Hadoop), and others came up with a fresh system design. However, most of these platforms are still in their infancy. For example, BDAS, the Berkeley Data Analytics Stack [11], is an open-source data analytics stack developed at the UC Berkeley AMPLab for computing and analyzing complex data. It includes the following main components: Spark, Shark,, and Mesos. Spark is a high-speed cluster computing system that performs computations in memory and can outperform Hadoop by up to 100x. Shark is a large-scale data analysis system for Spark that pro vides a unified engine running SQL queries, compatible with Apache Hive. Shark can answer SQL queries up to 100x faster than Hive, and run iterative machine learning algorithms up to 100x faster than Hadoop, and can recover from failed mid-queries within seconds [12]. Mesos is a cluster manager that can run Hadoop, Spark and other frameworks on a dynamically shared pool of compute nodes. ASTERIX [13] is data intensive storage and computing platform. Some notable drawbacks of Hadoop and other similar platforms, e.g., single system performance, difficulties of future maintenance, inefficiency in pulling data up to queries and the unawareness of record boundaries, are properly overcome in ASTERIX by exploring runtime models inspired by parallel database system execution engines . In ASTERIX, the open software stack is layered in a different way that it sets the data records at the bottom layer, facilitating a higher-level language API at the top. While the majority of the big data management and processing platforms have been (or are being) developed to meet business needs, SciDB is an open source data management and analytics (DMAS) software system for data-intensive scientific applications like radio astronomy, earth remote sensing and environment observation and modeling. The difference between SciDB and other platforms is that SciDB is designed based on the concept of array DBMS (i.e., raster data) where big data is represented as arrays of objects in unidimensional or multidimensional spaces. SciDB is designed to support integration with high-level imperative languages, algorithms, and very large scales of data [4].

4 BIG DATA MINING

The goals of big data mining techniques go beyond fetching the requested information or even uncovering some hidden relationships and patterns between numeral parameters. Analyzing fast and massive stream data may lead to new valuable insights and theoretical concepts [2]. 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 and insights 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 (cloud platforms in our mind). Big data mining must deal with heterogeneity, extreme scale, velocity, privacy, accuracy, trust, and interactiveness that existing mining techniques and algorithms are incapable of. The need for designing and implementing very-large-scale parallel machine learning and data mining algorithms (ML-DM) has remarkably increased, which accompanies the emergence of powerful parallel and very-large-scale data processing platforms, e.g., Hadoop MapReduce. NIMBLE [11] is a portable infrastructure that has been specifically designed to enable rapid implementation of parallel ML-DM algorithms, running on top of Hadoop. Apache's Mahout [12] is a library of machine learning and data mining implementations. The library is also implemented on top of Hadoop using the MapReduce programming model. Some important components of the library can run stand-alone. The main drawbacks of Mahout are that its learning cycle is too long and its lack of user-friendly interaction support. Besides, it does not implement all the needed data mining and machine learning algorithms. BC-PDM (Big Cloud-Parallel Data Mining) [13], as a cloud-based data mining platform, also based on Hadoop, provides access to large telecom data and business solutions for telecom operators; it supports parallel ETL process (extract, transform, and load), data mining, social network analysis, and text mining. BC-PDM tried to overcome the problem of single function of other approaches and to be more applicable for Business Intelligence. PEGASUS (Peta-scale Graph Mining System) and Giraph both implement graph mining algorithms using parallel computing and they both run on top of Hadoop. GraphLab is a graph-based, scalable framework, on which sever all graph-based machine learning and data mining algorithms are implemented.

5 EFFORTS AND CHALLENGES OF BIG DATA MINING AND DISCOVERY

Considering big data a collection of complex and large data sets that are difficult to process and mine for patterns and knowledge using traditional database management tools or data processing and mining systems. While presently the term big data literally concerns about data volumes, Wu et al. have introduce HACE theorem that described the key characteristics of the big data as (1) huge with heterogeneous and diverse data sources, (2) autonomous with distributed and decentralized control,

and (3) complex and evolving in data and knowledge associations. Generally, business intelligence applications are using data analytics that are grounded mostly in data mining and statistical methods and techniques. These strategies are usually based on the mature commercial software systems of RDBMS, data warehousing, OLAP, and BPM. Since the late 1980s, various data mining algorithms have been developed mainly within the artificial intelligence, and database communities. The 10 most influential data mining algorithms were identified based on expert nominations, citation counts, and a community survey. In ranked order, these techniques are as follows C4.5, k-means, SVM (support vector machine), Apriori, EM (expectation maximization), PageRank, AdaBoost, kNN (k-nearest neighbors), Naïve Bayes, and CART. These algorithms are for classification, clustering, regression, association rules, and network analysis. Most of these well known data mining algorithms have been implemented and deployed in commercial and open source data mining systems. Hadoop was originally a (distributed) file system approach applying the MapReduce framework that is a software approach introduced by Google in 2004 to support distributed computing on large/big data sets. Recently, Hadoop has been developed and used as a complex ecosystem that includes a wider range of software systems, such as HBase (a distributed table store), Zookeeper (a reliable coordination service), and the Pig and Hive high-level languages that compile down MapReduce components. Therefore in the recent conceptual approaches Hadoop is primarly considered an ecosystem or an infrastructure or a framework and not just the file system alongside MapReduce components. The big data and cloud computing frameworks include the Google MapReduce, Hadoop Reduce, Twister, Hadoop++, Haloop, and Spark etc. which are used to process big data and run computational tasks. The cloud databases are used to store massive structured and semistructured data generated from different types of applications. The most important cloud databases include the BigTable, Hbase, and HadoopDB. In order to implement an efficient big data mining and analysis framework, the data warehouse processing is also important. The most important data warehouse processing technologies include the Pig, Hive etc. Strambei suggests a different conceptual interpretation of the OLAP technology considering the emergence of web services, cloud computing and big data. One of the most important consequences could be widely open access to web analytical technologies. The related approach has evaluated the OLAP Web Services viability in the context of the cloud based architectures. There are also a few reported practical applications of big data mining in the cloud. Patel et al. have explored a practical solution to big data problem using the Hadoop data cluster, Hadoop Distributed File System alongside Map Reduce framework, and a big data prototype application scenarios. The results obtained from various experiments indicate promising results to address big data problem. The challenges for moving beyond existing data mining and knowledge discovery techniques are defined as follows:

1. a solid scientific foundation to be able to select an adequate analytical method and a software design solution
2. new algorithms (and demonstrate the efficiency and scalability, etc.) and machine learning techniques
3. the motivation of using cloud architecture for big data solutions and how to achieve the best performance of implementing data analytics using cloud platform (e.g. big data as a service)
4. dealing with data protection and privacy in the context of exploratory or predictive analysis of big data

6 CONCLUSION

The big data movement has energized the data mining, knowledge discovery in data bases and associated software development communities, and it has introduced complex, interesting questions for researchers and practitioners. Big data discloses the limitations of existing data mining techniques, resulted in a series of new challenges related to big data mining. Big data mining is a promising research area, still in its infancy. In spite of the limited work done on big data mining so far, we believe that much work is required to overcome its challenges related to heterogeneity, Scalability, speed, accuracy, trust, provenance, privacy, and interactiveness. This paper also provides an overview of state-of- the-art frameworks/platforms for processing, managing and mining big data.

References

[1] Berkovich, S., Liao, D.: On Clusterization of big data Streams. In: 3rd International Conference on Computing for Geospatial Research and Applications, article no. 26. ACM Press, New York (2012)

[2] Beyer, M.A., Laney, D.: The Importance of 'Big Data': A Definition. Gartner (2012)

[3] Madden, S.: From Databases to big data. IEEE Internet Computing 16(3), 4–6 (2012)

[4] Shmueli, G., Patel, N.R., Bruce, P.C.: Data Mining for Business Intelligence: Concepts, Techniques, and Applications in Microsoft Office Excel with XLMiner, 2nd edn. Wiley & Sons, Hoboken (2010)

[5] Ghoting, A., Kambadur, P., Pednault, E., Kannan, R.: NIMBLE: a Toolkit for the Implementation of Parallel Data Mining and Machine Learning Algorithms on MapReduce. In: 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Diego, California, USA, pp. 334–342 (2011)

[6] Low, Y., Bickson, D., Gonzalez, J., Guestrin, C., Kyrola, A., Hellerstein, J.M.: Distributed GraphLab: A Framework for Machine Learning and Data Mining in the Cloud. VLDB Endowment 5(8), 71–727 (2012)

[7] Borkar, V.R., Carey, M.J., Li, C.: big data Platforms: What's Next? ACM Crossroads 19(1), 44–49 (2012)

[8] Sun, Y., Han, J., Yan, X., Yu, P.S.: Mining Knowledge from Interconnected Data: A Heterogeneous Information Network Analysis Approach. VLDB Endowment 5(12), 2022–2023 (2012)

[9] Tene, O., Polonetsky, J.: Privacy in the Age of big data: A Time for Big Decisions. Stanford Law Review Online 64, 63–69 (2012)

[10] Gartner Press Release. "Gartner Says Big Data Will Drive $28 Billion of IT Spending in 2012." October 17, 2012

[11] NewVantage Partners: Big Data Executive Survey(2013) http://newvantage.com/wp-content/uploads/2013/02/ NVP-Big- Data-Survey-2013-Summary-Report.pdf

[12] Xin, R.S., Rosen, J., Zaharia, M., Franklin, M., Shenker, S., Stoica, I.: Shark: SQL and Rich Analytics at Scale. In: ACM SIGMOD Conference (accepted, 2013)

[13] Agrawal, D., Bernstein, P., Bertino, E., et al.: Challenges and Opportunities With big dataA Community White Paper Developed by Leading Researchers Across the United States(2012), http://cra.org/ccc/docs/init/bigdatawhitepaper. pdf