Volume 6, Issue 6, June 2016 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Issues and Challenges in the Era of Big Data Mining

Abhinav Kathuria
Department of Computer Science and Application, Panjab University,
Chandigarh, India

Abstract:-Data is an important part of every industry, organization, economy and individual. The term Big data is used to define huge amount of data. The data is so huge that the traditional database management systems are unable to store, process and analyze. It introduces many challenges, opportunities and research topics which require new tools and techniques for processing and analyzing of Big Data. This paper presents review on the term Big Data Mining and presents various challenges on analyzing Big Data.

Keywords:-Data mining, big data, big data mining, knowledge discovery, Hadoop, MapReduce.

I. INTRODUCTION

The era of petabyte has come and gone, leaving us to angrily face/stand up to the exabytes time in history now. Technology revolution has been helping millions of people by creating huge/extreme data via ever-increased use of digital devices and especially remote sensors that create continuous streams of digital data, resulting in what is known as "big data". It has been a confirmed important thing/big event that huge amounts of data have been being constantly created at never-before-seen and ever increasing scales. For example according to a survey, Google receives over 2 million queries, YouTube users upload 72 hours of video, Facebook users share over 2 million pieces of content etc. The main challenge before us is collecting useful information from this huge amount of data. Various technologies are coming up to cope up with these requirements like Cloud computing, Google's model i.e. MapReduce etc. From data mining point of view mining of data from Big Data is a major challenge before us.

The extracted information can be useful for making various business decisions and for predicting the future trends. Organizations can make knowledge driven decisions. Various data mining techniques are available for discovery of knowledge from databases and these techniques are often applied with parallel processing architectures and distributed storage systems to improve the performance.

II. DATA MINING

It is a process of finding hidden knowledge and insights from huge amount of data. It finds various patterns and relationships hidden in this huge amount of data. Various data mining algorithms and techniques are available, but these algorithms are not scalable and techniques are not able to match the volume, velocity and variety of emerging data. However these techniques are not able to work in real time, so new algorithms and techniques are required to work with huge amount of data otherwise value of this information will be useless. So new tools are required that can work in parallel, that are scalable and that can work in real time having interaction with users.

III. BIG DATA MINING

The desires of massive facts mining strategies go beyond fetching the asked information or maybe uncovering some hidden relationships and styles between numeral parameters. Studying fast and big move statistics may additionally result in new valuable insights and theoretical concepts [1]. Evaluating with the results derived from mining the conventional datasets, unveiling the big quantity of interconnected heterogeneous huge data has the capability to maximize our knowledge and insights in the target area. but, this brings a sequence of latest demanding situations to the research community. Overcoming the challenges will reshape the destiny of the statistics mining era, ensuing in a spectrum of groundbreaking information and mining techniques and algorithms. One possible approach

is to improve present strategies and algorithms by exploiting vastly parallel computing architectures (cloud platforms in our thoughts). Big facts mining need to deal with heterogeneity, extreme scale, speed, privacy, accuracy, believe, and interactiveness that current mining strategies and algorithms are incapable of. The need for designing and implementing very-massive-scale parallel machine learning and facts mining algorithms (ML-DM) has remarkably improved, which accompanies the emergence of effective parallel and very-large-scale records processing structures, e.g., Hadoop MapReduce. NIMBLE[2] is a transportable infrastructure that has been mainly designed to permit fast implementation of parallel MLDM algorithms, jogging on top of Hadoop. Apache's Mahout[3] is a library of system studying and information mining implementations. The library is also implemented on top of Hadoop using the MapReduce programming model. some vital components of the library can run stand-alone. the principle drawbacks of Mahout are

that its mastering cycle is simply too lengthy and its loss of person-pleasant interaction aid. besides, it does not put in force all of the wanted records mining and system getting to know algorithms. BC-PDM (big Cloud-Parallel data Mining)[4], as a cloud-primarily based statistics mining platform, additionally based on Hadoop, gives access to huge telecom facts and commercial enterprise answers for telecom operators; it supports parallel ETL technique (extract, remodel, and cargo), data mining, social

community analysis, and text mining. BC-PDM attempted to conquer the problem of unmarried feature of different strategies and to be more relevant for enterprise Intelligence. PEGASUS (Peta-scale Graph Mining system)[5] and Giraph[6] each implement graph mining algorithms the use of parallel computing and they each run on pinnacle of Hadoop. GraphLab[7] is a graph-based totally, scalable framework, on which sever all graph-primarily based device learning and records mining algorithms are carried out.

IV. ISSUES AND CHALLENGES

- Variety and Heterogeneity:- Variety is the characteristic of Big Data. Data is collected from many sources that may generate data on its own or may contribute to it. It means there is variety as well as heterogeneity in the data. These types of data are interconnected, interrelated and inconsistent. Data may be structured which may fit in the database, semi-structured which may partially fit into the database or unstructured may not fit in the database. So mining hidden patterns and knowledge from these heterogeneous data is a challenge before the data scientists.
- Scalability:- Big data requires high scalability of its data management and mining tools. This data may contain knowledge and information which may not be possible to collect from conventional data.
- Speed/Velocity:- The data mining algorithm must be able to finish the processing within a particular time. The data must be accessed and processed quickly otherwise the results obtained from these will become worthless. The factors which affect the speed of data mining depends include data access time and efficiency of mining algorithms. Parallelism may be included to increase the accessing speed.
- Accuracy and Trust:- With an increase in the amount of data, there are many data sources from where data is collected, which may not be verifiable or trustable. Therefore the accuracy and trust of data source becomes an issue, which may propagate to results as well. Therefore data validation and accuracy becomes an important issue for discovery of useful information.
- Privacy:- It is an important issue that data must be kept private and invisible to others. Data mining requires personal information in order to produce results. Social media contains all of the information about an individual and information can be mined from that information and then the privacy disappears. So this is the issue which must be considered by data scientists and tools must be built by taking this consideration into account.
- Interactiveness:- It means feature of the data mining system that allows user interaction by using feedback/guidance. It is an important issue as it allows the users to visualize, evaluate and interpret intermediate and final results.

V. CONCLUSION

In this era data is generated at unprecedented speed. This paper presented the limitations in existing data mining techniques used for data mining. More work is required to be done to cope with the challenges related to it. New techniques must be developed and parallelism must be used for improving the speed of analysis and accessing of data. We are in the beginning of era where Big data mining allows us to discover knowledge and use that knowledge for different applications.

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] 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, pp.334-342, San Diego, California, USA (2011).

[3] Mahout, http://lucene.apache.org/mahout/.

[4] Yu, L., Zheng, J., Shen, W.C., et al: BC-PDM: Data Mining, Social Network Analysis and Text Mining System Based on Cloud Computing. In: 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1496-1499 (2012).

[5] Kang, U., Tsourakakis, C.E., Faloutsos, C.: PEGASUS: A Peta-Scale Graph Mining System Implementation and Observations In: 9th IEEE International Conference on Data Mining, pp. 229-238 (2009).

[6] Apache Giraph Project, http://giraph.apache.org/.

[7] 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, vol. 5, no. 8, pp.71-727 (2012).

[8] Madden, S.: From Databases to big data. In: IEEE Internet Computing, vol. 16, no. 3, pp. 4-6. IEEE Computer Society (2012).

[9] B R Prakash, Dr. M. Hanumanthappa. Issues and Challenges in the Era of Big Data Mining, Volume 3, Issue 4,pp 321-325, IJETTCS(2014).

[10] Dunren Che , Mejdl Safran , and Zhiyong Peng. From Big Data to Big Data Mining: Challenges, Issues and Opportunities,pp.1-15.

[11] Jaseena K.U and Julie M. David, ISSUES, CHALLENGES, AND SOLUTIONS: BIG DATA MINING, Natarajan Meghanathan et al. (Eds) : NeTCoM, CSIT, GRAPH-HOC, SPTM – 2014 pp. 131–140, 2014.

[12] Fayyad, U.M., Gregory, P.S., Padhraic, S.: From Data Mining to Knowledge Discovery: an Overview. In: Advances in Knowledge Discovery and Data Mining, pp. 1-36. AAAI Press, Menlo Park, CA (1996).

