Big Data Analytics – A Review

Ajay Rastogi

Assistant Professor, College of Computer Science and Information Technology,
Teerthanker Mahaveer University, India

Abstract:

In recent years data is growing rapidly. Data is generated by social sites, scientific instruments, mobile devices, sensor technology and network devices etc. Big data comes in existence from a decade. Big data refers to the data which is huge, to store and process and to create some sort of knowledge effectively. The current requirement of the real time processing of big data so that one can get the results in real time, which may be used to take an appropriate decision, required advertisement, suggestions to the consumers so that one can create the benefit in term of money, business, social values and status etc. The question is how to improve the performance of big data analysis. The data is very much huge in terms of verity, volume and velocity to store, process and outputting the result in real time.

Key words: Big data, data analysis

INTRODUCTION

As the technology is spreading the ease and access of internet is growing, data creation speed is much faster than before. According to the Lyman and Varian [1] 92% of new information is stored in hard disk however the new data is more than five Exabyte. Laney [3] presented 3V: Volume, Velocity and Variety for the “big” data. It means the data is large enough and created rapidly and having different types of attributes. According to Rijmenam [4] the 3V is not the complete explanation of “big data”. He added veracity, validity, value, variability, venue, vocabulary and vagueness to its explanation.

In order to analyze we know the clustering [2] method used in data mining by partitioning method, sampling method, density based method, grid based and model based method. The

performance of these methods will be degraded while processing high dimensionality data. To process big data having such a volume, that could not be stored at one machine. How to process this data at a run time? It also means that the how data could be processed while not on the single machine because most data mining methods could not be directly implied on this kind of distributed data.

In last few years investment in the field of big data is drastically increasing. Some big IT Companies like Microsoft, Google, facebook, amazon investing in hiring talent skilled people and in AI hardware. The IDC reports - the world wide revenue for big data and business analytics will grow from \$130.1 billion in 2016 to more than \$203 billion In 2020.

In the next section we focus upon the issue of the real time data processing. Time to time these studies evaluate the performance and scalability of such type of methods. We can discuss some data mining analytics and then we will discuss big data analytics.

Data analysis:

In 1996 Fayyad [6] proposed the process of knowledge discovery in database (KDD). In KDD there are many steps like Data selection, Data processing, Transformation, Data Mining, Patterns and interpretation of knowledge. KDD is an iterative process. It can contain loops between any two steps.

The data preprocessing, data cleaning integration and transformation plays a vital role to process it with the any mining methods. This is called raw data, that can be used in mining technique and get the appropriate knowledge from it. Now we will discuss some data mining techniques.

Classification and Prediction- Classification is supervised learning method. Supervised learning can be used if the data has labels for example relational data is having labels. Data will be associated using decision tree technique. Decision tree produces or discovers some hidden patterns. According to these discovered patterns we can take some sort of decision. We can predict the situation on some given parameters. For classification we have to train machines by some training data using some models like neural network, Random forest etc. This training data discovered some patterns using any of the above model. After this training we can test the model by testing data and try to find the accuracy. There are different learning classification models are given. We can test the accuracy of any one of the models and select appropriate one. Once all the training and testing is completed we can easily predict for the upcoming situation.

Clustering is unsupervised learning method. Here we are having data which is not having any labels. By clustering we have to identify some regions in which data could be collected. This is a most challenging task. Data could have different clusters according to its properties. Once clustering will be done. Decision models could be formed. Different clusters have different data objects. Clustering is used in share market research, image processing, and pattern recognition and data analysis. Clustering can partition large data set into groups. These groups are sometimes called segmentation. Different segments have different data objects.

Some of the following issues while analyzing big data

-the biggest challenge in big data analytics is volume. The input is so big to paralyze the data analytics. According to Baraniuk [8] the sensors creating more data to upload into the upper layer of the system. The uploading creates bottleneck problem.

-the another issue is the velocity. When a large data comes in the form of stream of bits into the real time the data analytics may be unable to process such huge data due to input is changing drastically.

- the next issue is the variety. When large data comes in the different types for example images, video and text or have incomplete data. For the handling of such types of data is a very tedious task.

Data analysis responsibility is to find the hidden data or patterns. Many researchers use the term data mining. Apart from data mining new researches in the field of machine learning such as deep learning with the advent of AI is also playing an important role for the data analysis in real time.

Some of the following issues with Big Data Analysis are-

The features of big data is high dimensionality, varied, multifaceted, formless, incomplete and incorrect may change the data analysis approach. That results incorrect information.

- The flow of input in big data is the big issue. How to control this flow?
- This large input flow is typical to handle in real time.
- This large input flow may have different types of data or incomplete data or streaming data.

Big data input issue could be solved by compression, sampling and feature selection and so on. For the faster compression FPGA technique discussed by Jun et al. data compression, sampling and handling fast data input is still is an open research issue.

Big Data Analysis frameworks and platform-

There are several platform/ solution available for big data analysis.

- Processing –hadoop [9]
- Nvidia CUDA [10]
- Twitter Storm [11]
- Storage: Titan or HDFS [12]
- Analytics: MLPACK [13]
- Mahout [14]

We can find tools available for data analysis from well known organization. Cloud computing technology is widely used in big data analysis to satisfy the large demand of computing power and storage. Searching of the useful information is equal to finding a needle into haystack.

Huai [15] proposed the replacement of old computer systems with new and larger computer system, in traditional parallel processing model for increasing the processing performance. Huai also presented a matrix model with three matrix DOT. D-Data Set, O- concurrent processing operations, T- transformation

GLADE [16] having two types of nodes i.e. coordinator and worker. In terms of execution times GLADE is better than hadoop.

Another study of Essa et al. [17] shows that hadoop is a single master and once it goes down entire system will down. To avoid this problem MRAM-Map Reduce Agent Mobility framework is device. The architecture of MRAM is now ready to work on distributed environment. Herodotou [18] considered the issue of user need and system workload based on hadoop. User must get the system according to his need so that the performance of the system could be increased. The study [18] presents BDAF- Big data architecture framework that include security,

life cycle management, data structure, data analytics and infrastructure. Some studies [19] proves that R is better in data analysis services called CBDMASP. In [20], Laurila et al. focused on the security and privacy of data analytics.

In [21], Talia classified three services using the cloud platform – data analytics infrastructure as a service, Data analytics platform as a service, data analytics software as a service. In [22], Cuzzocrea et al. discussed uncorrelated data filtering, data source heterogeneity and computational emergency. In [23], Zhang and Huang talked about 5Ws- what kind of data, what is the requirement of these data, from where the data comes, when data comes, who is user of the data and what is the data transfer technique. He explain the method and framework to be used in big data analysis using his 5Ws approach. A later study [24] used the features like complexity, user, workload etc to compare the Drill, storm and hadoop and found that the hadoop is having high latency.

Other technologies like machine learning and Artificial intelligence are also have been used for big data analysis. For different mining and analysis problems machine learning is already used. For example one of the machine learning algorithm GA- genetic analysis is not only solve the clustering problem but also solve frequent mining problem. It enhance the performance of the other part of KDD. A recent study[25] shown that some traditional tools are still used in the big data analysis. Machine learning will be necessarily become the important part of big data analysis. Some methods of machine learning is based on centralized computing however some algorithm can be used for parallel processing which is the most essential for big data analysis.

CONCLUSION

We have discussed about traditional data analysis, issues regarding the big data analysis-that is comes from its 3Vs. issue of how to process such a huge input. Suggestions are to use a number of connected computing nodes or use the existing cloud based architecture. Use of compression is also take an important role to gather that huge in a limited space. Sampling and feature

selection is discussed as an open research issue. Then we discussed several available tools available for big data analysis. Like hadoop, twitter, mahout, etc. somewhere studies shows that hadoop is better when load balancing is done according to the user requirements. Somewhere study shows hadoop is having high latency time. Different criteria is used for the comparisons. Machine learning is also plays an important role in data mining/analysis. Because its most of the traditional algorithms best suits the central computing environments. But in future ML will be a key player in big data analysis. Some of its algorithms can do well with parallel computing environment and some of the algorithms are trying to implement on parallel environment. This implementation is also opens a new research topic.

REFERENCES

- I. Lyman P, Varian H. How much information 2003? Tech. Rep, 2004. [Online]. Available:[http://www2.sims.berkeley.edu/research/projects/how-much-info-2003 /printable_report.pdf](http://www2.sims.berkeley.edu/research/projects/how-much-info-2003/printable_report.pdf).
- II. Xu R, Wunsch D. Clustering. Hoboken: Wiley-IEEE Press; 2009
- III. Laney D. 3D data management: controlling data volume, velocity, and variety, META Group, Tech. Rep. 2001. [Online]. Available: <http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Control-ling-Data-Volume-Velocity-and-Variety.pdf>.
- IV. van Rijmenam M. Why the 3v's are not sufficient to describe big data, BigData Startups,Tech. Rep. 2013. [Online]. Available: <http://www.bigdata-startups.com/3vs-sufficient-describe-big-data/>.
- V. Gil Press, 6 Predictions For The \$203 Billion Big Data Analytics Market, Forbes, Tech. Rep. 2017[online] <https://www.forbes.com/sites/gilpress/2017/01/20/6-predictions-for-the-203-billion-big-data-analytics-market/#1964f5632083>
- VI. Fayyad UM, Piatetsky-Shapiro G, Smyth P. From data mining to knowledge discovery in databases. AI Mag. 1996;17(3):37–54.
- VII. Han J. Data mining: concepts and techniques. San Francisco: Morgan Kaufmann Publishers Inc.; 2005.
- VIII. Baraniuk RG. More is less: signal processing and the data deluge. Science.2011;331(6018):717–9.

- IX. Pospiech M, Felden C. Big data—a state of the art. In: Proceedings of the Americas Conference on Information Systems, 2012, pp 1–23. [Online]. Available: <http://aisel.aisnet.org/amcis2012/proceedings/DecisionSupport/22>
- X. Apache Hadoop, February 2, 2015. [Online]. Available: <http://hadoop.apache.org>
- XI. Cuda, February 2, 2015. [Online]. Available: URL: http://www.nvidia.com/object/cuda_home_new.html
- XII. Apache Storm, February 2, 2015. [Online]. Available: URL: <http://storm.apache.org/>
- XIII. Curtin RR, Cline JR, Slagle NP, March WB, Ram P, Mehta NA, Gray AG. MLPACK: a scalable C machine learning library. J Mach Learn Res. 2013;14:801–5.
- XIV. Apache Mahout, February 2, 2015. [Online]. Available: <http://mahout.apache.org/>
- XV. Huai Y, Lee R, Zhang S, Xia CH, Zhang X. DOT: a matrix model for analyzing, optimizing and deploying software for big data analytics in distributed systems. In: Proceedings of the ACM Symposium on Cloud Computing, 2011. Pp 4:1–4:14.
- XVI. Rusu F, Dobra A. GLADE: a scalable framework for efficient analytics. In: Proceedings of LADIS Workshop held in conjunction with VLDB, 2012. pp 1–6.
- XVII. Essa YM, Attiya G, El Sayed A. Mobile agent based new framework for improving big data analysis. In: Proceedings of the International Conference on Cloud Computing and Big Data. 2013, pp 381–386.
- XVIII. Demchenko Y, de Laat C, Membrey P. Defining architecture components of the big data ecosystem. In: Proceedings of the International Conference on Collaboration Technologies and Systems, 2014. pp 104–112.
- XIX. Ye F, Wang ZJ, Zhou FC, Wang YP, Zhou YC. Cloud based big data mining and analyzing services platform integrating . In: Proceedings of the International Conference on Advanced Cloud and Big Data, 2013. pp 147–151.
- XX. Laurila JK, Gatica Perez D, Aad I, Blom J, Bornet O, Do T, Dousse O, Eberle J, Miettinen M. The mobile data challenge: big data for mobile computing research. In: Proceedings of the Mobile Data Challenge by Nokia Workshop, 2012. pp 1–8.

- XXI. Talia D. Clouds for scalable big data analytics. *Computer*. 2013;46(5):98–101 ,100.
Cuzzocrea A, Song IY, Davis KC. Analytics over large scale multidimensional data: The big data revolution!. In: *Proceedings of the ACM International Workshop on Data Warehousing and OLAP*, 2011. pp 101–104.
- XXII. Cuzzocrea A, Song IY, Davis KC. Analytics over large scale multidimensional data: The big data revolution!. In: *Proceedings of the ACM International Workshop on Data Warehousing and OLAP*, 2011. pp 101–104.
- XXIII. Zhang J, Huang ML. 5Ws model for big data analysis and visualization. In: *Proceedings of the International Conference on Computational Science and Engineering*, 2013. pp 1021–1028.
- XXIV. Chandarana P, Vijayalakshmi M. Big data analytics frameworks. In: *Proceedings of the International Conference on Circuits, Systems, Communication and Information Technology Applications*, 2014. pp 430–434.
- XXV. Ma C, Zhang HH, Wang X. Machine learning for big data analytics in plants. *Trends Plant Sci*. 2014;19(12):798–808.