Algorithm and Approaches to Handle Big Data

UzmaShafaque
M.E (1st yr)
Department of Computer Engineering,
JCOET, Yavatmal

Parag D. Thakare
Assistant Professor
Department of Computer Engineering,
JCOET, Yavatmal

Mangesh M. Ghonge
Assistant Professor
Department of Computer Engineering
JCOET, Yavatmal

Milindkumar V. Sarode, Ph.D
Head of Department,
Department of computer engineering.
JCOET, Yavatmal

ABSTRACT

Environment of Big Data produces a large amount of data, in which it need to be analyzed and patterns have to be extracted, to gain knowledge. In this era of big data, with boom of data both structured data and unstructured data, in different field such as Engineering, Genomics, Biology, Meteorology, Environmental research and many more, it has become difficult to manage, process and analyze patterns using architectures and databases that are traditional. So, we should understand a proper architecture to gain knowledge about the Big Data. In this paper a review of various algorithms necessary for handling such large data set is given. These algorithms give us various methods implemented to handle Big Data.

Keywords

Big data, data mining, map-reduce, crowdsourcing, algorithm.

1. INTRODUCTION

Data Mining is the technology to extract the knowledge from the pre-existing databases. It is used to explore and analyse the same. The data which is to be mined varies from a small data-set to a large data-set i.e. **Big Data**.Big data is so large that it does not fit in the main memory of a single machine, and the it need to process big data by efficient algorithms. Modern computing has entered the era of Big Data. The massive amounts of information available on the Internet enable computer scientists, physicists, economists, mathematicians, political scientists, bio-informaticists, sociologists, and many others to discover interesting properties about people, things, and their interactions. Analysing information from Twitter, Google, Facebook, Wikipedia, or the Human Genome Project requires the development of scalable platforms that can quickly process massive-scale data. Such frameworks often utilize large numbers of machines in a cluster or in the cloud to process data in a parallel manner.

Flickr is public picture sharing site, which received 1.8 million photos per day, on average, from February to March 2012 [1]. Assuming the size of each photo is 2 megabytes (MB), this requires 3.6 terabytes (TB) storage every single day. Indeed, as an old saying states: "a picture is worth a thousand words," the billions of pictures on Flicker are a treasure tank for us to explore the human society, social events, public affairs, disasters, and so on, only if we have the power to harness the enormous amount of data. This is excellent example for Big Data processing, as the information comes from multiple, heterogeneous, autonomous sources with complex and evolving relationships, and keeps growing. Along with the above example, the era of Big Data has arrived, Every day, 2.5 quintillion bytes of data are created and 90 percent of the data in the world today were produced within the past two years [2]. Our capability for data generation has never been so powerful and enormous ever since the invention of the information technology in the early 19th century.

Big data is complex data set that has the following main characteristics: Volume, Variety, Velocity and Value [3] [4] [5][6]. These make it difficult to use the existing tools to manage and manipulate [7]. Big Data are the large amount of data being processed by the Data Mining environment. In other words, it is the collection of data sets large and complex that it becomes difficult to process using on hand database management tools or traditional data processing applications, so data mining tools were used. Big Data are about turning unstructured, invaluable, imperfect, complex data into usable information. [8].

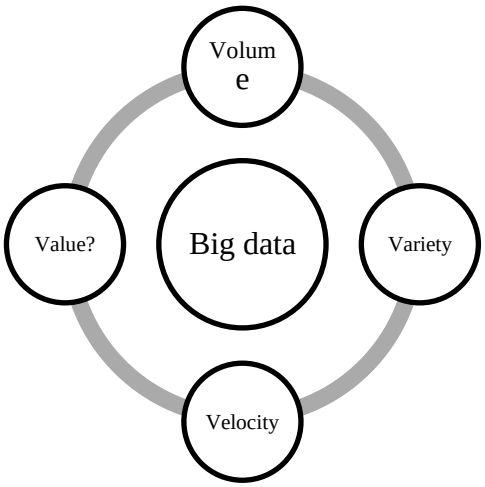

Figure 1: 4V's Big-Data

Browsing is difficult through a large data set and time consuming also, we have to follow certain rules/protocols, proper algorithms and methods is needed to classify the data, find a suitable pattern among them. The data analysis methods such as exploratory, clustering, factorial, analysis need to be extended to get the information and extract new knowledge.

The remaining paper has been described as follows, Section II: deals with the architecture of the big data. Section III: describes the various algorithms used to process Big Data. Section IV: describes potential applications of Big data. Section V: deals with the various algorithms specifying ways to handle big data. Section VI: deals with the various security issues related to the big data.

2. ARCHITECTURE

Big Data are the collection of large amounts of unstructured, heterogeneous data. Big Data means enormous amounts of data, such large that it is difficult to collect, store, manage, analyze, predict, visualize, and model the data. Big Data architecture

typically consists of three segments: storage system, handling and analysis. Big Data typically differ from data warehouse in architecture; it follows a distributed approach whereas a data warehouse follows a centralized one.

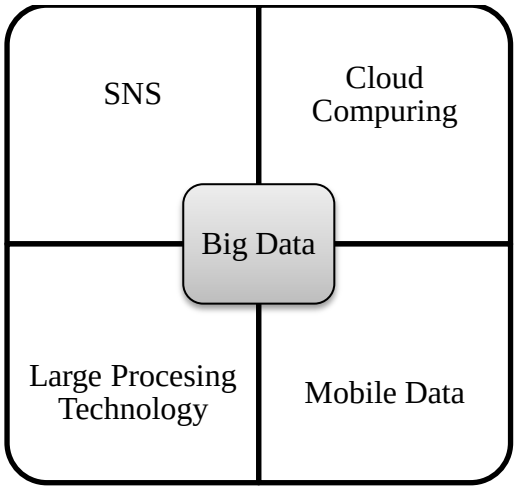

Figure 2: Big-Data m*o*d*e*s

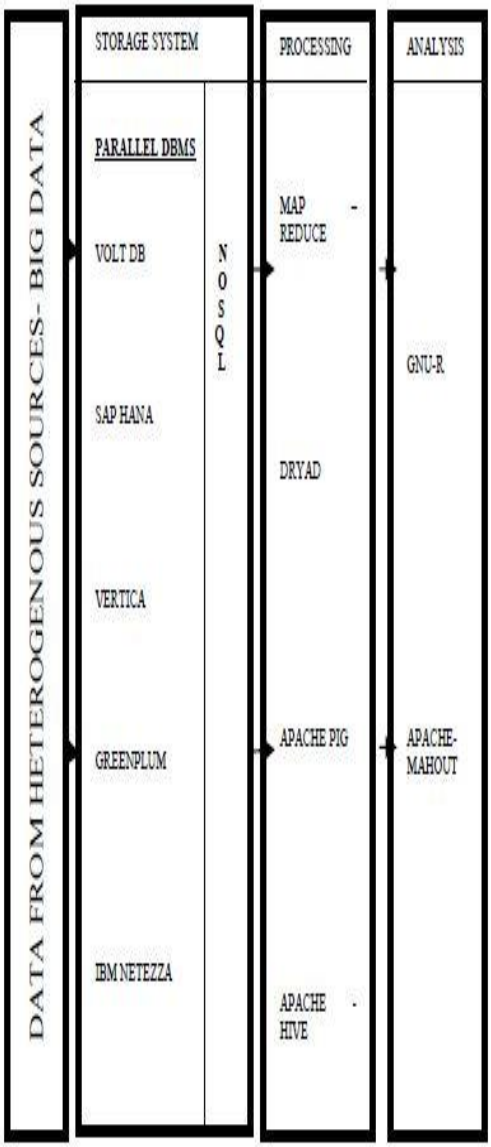

Figure 3: Architecture of Big Data

On a paper An Efficient Technique on Cluster Based Master Slave Architecture Design, the hybrid approach was formed which consists of both top down and bottom up approach. This hybrid approach when compared with the clustering and Apriori algorithm, takes less time in transaction than them. [10]

2.1 Big Data Characteristics: Hace Theorem

Big Data starts with large-volume, heterogeneous, autonomous sources with distributed and decentralized control, and seeks to explore complex and evolving relationships among data. These characteristics make it an extreme challenge for discovering useful knowledge from the Big Data. In a naı¨ve sense, we can imagine that a number of blind men are trying to size up a giant elephant (see Fig. 1), which will be the Big Data in this context. The goal of each blind man is to draw a picture (or conclusion) of the elephant according to the part of information he collects during the process. Because each person's view is limited to his local region, it is not surprising that the blind men will each conclude independently that the elephant "feels" like a rope, a hose, or a wall, depending on the region each of them is limited to. To make the problem even more complicated, let us assume that 1) the elephant is growing rapidly and its pose changes constantly, and 2) each blind man may have his own (possible

unreliable and inaccurate) information sources that tell him about biased knowledge about the elephant (e.g., one blind man may exchange his feeling about the elephant with another blind man, where the exchanged knowledge is inherently biased).

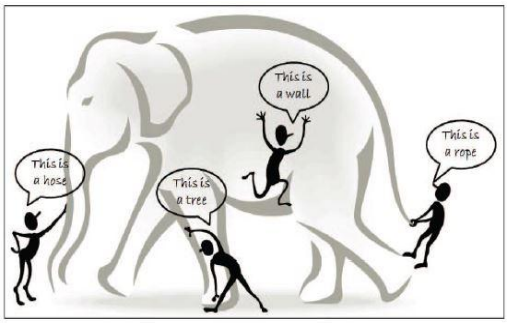

Figure 4: The blind men and the giant elephant: thelocalized(limited) view of each blind man leads to a biased conclusion.

Exploring the Big Data in this scenario is equivalent to aggregating heterogeneous information from different sources (blind men) to help draw a best possible picture to reveal the genuine gesture of the elephant in a real-time fashion. Indeed, this task is not as simple as asking each blind man to describe his feelings about the elephant and then getting an expert to draw one single picture with a combined view, concerning that each individual may speak a different language (heterogeneous and diverse information sources) and they may even have privacy concerns about the messages they deliberate in the information exchange process.[23]

3. ALGORITHM

Many applications in the real world are moving from being computationally-bound to being *data-bound.* We are seeing a wide variety of large datasets. There are billions of emails and search queries, and millions of tweets and photos posted every day, in addition to our every action being tracked online (via cookies) and in the physical world (e.g., through video cameras).

This paper will provide an introduction to algorithm on such large datasets. There are many types of classification algorithms such as tree-based algorithms (C4.5 decision tree, bagging and boosting decision tree, decision stump, boosted stump, and random forest), neural-network, Support Vector Machine (SVM), rule-based algorithms(conjunctive rule, RIPPER, PART, and PRISM), naive Bayes, logistic regression. Along with these algorithm there are many algorithm like Parallel algorithms which partition computation across many machines, large scale machine learning, streaming algorithms that never store the whole input in memory and crowd-sourcing. These classification algorithms have their own advantages and disadvantages, depending on many factors such as the characteristics of the data and results [11,12].

Many algorithms were defined earlier in the analysis of large data set. We will go through the different work done to handle Big Data. In the beginning different Decision Tree Learning was used earlier to analyze the big data. In work done by Hall. et al. [10], there is defined an approach for forming learning the rules of the large set of training data. The approach is to have a single decision system generated from a large and independent n subset of data. Whereas Patil et al, uses a hybrid approach combining both genetic algorithm and decision tree to create an optimized decision tree thus improving efficiency and performance of computation. [13].

Then clustering techniques came into existence. Different clustering techniques were being used to analyze the data sets. A new algorithm called GLC++ was developed for large mixed data set unlike algorithm which deals with large similar type of dataset. This method could be used with any kind of distance, or symmetric similarity function. [14]

Decision trees are simple yet effective classification algorithms. One of their main advantages is that they provide human-readable rules of classification. Decision trees have several drawbacks, especially when trained on large data, where the need to sort all numerical attributes becomes costly in terms of both running time and memory storage. The sorting is needed in order to decide where to split a node.

Table 1: Different Decision tree Algorithm

AUTHOR'S NAME	TECHNIQUE	CHARACTERISTIC	SEARCH TIME
N. Beckmann, H. -P. Kriegal, R. Schneider, B. Seeger [8]	R-Tree R*-Tree	Have performance bottleneck	O (3^D)
S. Arya, D. Mount, N. Netanyahu, R. Silverman, A. Wu [9]	Nearest Neighbor Search	Expensive when searching object is in High Dimensional space	Grows exponentially with the size of the searching space. O(dn log n)
Lawrence O. Hall, NiteshChawla, Kevin W. Bowyer [10]	Decision Tree Learning	Reasonably fast and accurate	Less time consuming
Zhiwei Fu, Fannie Mae [11]	Decision Tree C4.5	Practice local greedy search throughout dataset	Less time consuming

D. V. Patil, R. S. Bichkar [12]	GA Tree (Decision Tree + Genetic Algorithm)	Improvement in classification, performance and reduction in size of tree, with no loss in classification accuracy	Improved performance- Problems like slow memory , execution can be reduced
Yen-Ling Lu, Chin-ShyurngFahn [17]	Hierarchical Neural Network	High accuracy rate of recognizing data; have high classification accuracy	Less time consuming - improved performance

Whereas Koyuturk et al. Defined a new technique PROXIMUS for compression of transaction sets, accelerates the association mining rule, and an efficient technique for clustering and the discovery of patterns in a large data set. [15].With the growing knowledge in the field of big data, the various techniques for data analysis- structural coding, frequencies, co-occurrence and graph theory, data reduction techniques, hierarchical clustering techniques.Multidimensional scaling was defined in Data Reduction Techniques for Large Qualitative Data Sets. It described that the need for the particular approach arise with the type of dataset and the way the pattern are to be analyzed. [16] The earlier techniques were inconvenient in real time handling of large amount of data so in Streaming Hierarchical Clustering for Concept Mining, defined a novel algorithm for extracting semantic content from large dataset. The algorithm was designed to be implemented in hardware, to handle data at high ingestion rates. [17].

Then in Hierarchical Artificial Neural Networks for Recognizing High Similar Large Data Sets., described the techniques of SOM (self-organizing feature map) network and learning vector quantization (LVQ) networks. SOM takes input in an unsupervised manner whereas LVQ was used supervised learning. It categorizes large data set into smaller thus improving the overall computation time needed to process the large data set. [18]. Then improvement in the approach for mining online data come from archana et al.Where online mining association rules were defined to mine the data, to remove the redundant rules. The outcome was shown through graph that the number of nodes in this graph is less as compared with the lattice. [19]. Then after the techniques of the decision tree and clustering, there came a technique in Reshef et al. In which dependence was found between the pair of variables. And on the basis of dependence association was found. The term maximal information coefficient (MIC) was defined, which is maximal dependence between the pair of variables. It was also suitable for uncovering various non-linear relationships. It was compared with other approaches was found more efficient in detecting the dependence and association. It had a drawback –it has low power and thus because of it does not satisfy the property of equitability for very large data set. [20]. Then in 2012 wang, uses the concept of Physical Science, the Data field to generate interaction between among objects and then grouping them into clusters. This algorithm was compared with K-Means, CURE, BIRCH, and CHAMELEON and was found to be much more efficient than them. [21]. Then, a way was described in "Analyzing large biological datasets with association network" to transform numerical and nominal data collected in tables, survey forms, questionnaires or type-value annotation records into networks of associations (ANets) and then generating Association rules (A Rules). Then any visualization or clustering algorithm can be applied to them. It suffered from the drawback that the format of the dataset should be syntactically and semantically correct to get the result. [22]

4. CONCLUSION

Because of Increase in the amount of data in the field of genomics, meteorology, biology, environmental research, it becomes difficult to handle the data, to find Associations, patterns and to analyze the large data sets.

As an organization collects more data at this scale, formalizing the process of big data analysis will become paramount.The paper describes methods for different algorithms used to handle such large data sets. And it gives an overview of architecture and algorithms used in large data sets.

5. REFERENCES

[1] F. Michel, "How Many Photos Are Uploaded to Flickr Every Day and Month?"http://www.flickr.com/photos/franckmichel/6855169886/, 2012.

[2] "IBM What Is Big Data: Bring Big Data to the Enterprise," http://www-01.ibm.com/software/data/bigdata/, IBM, 2012.

[3] Feifei Li, Suman Nath "Scalable data summarization on big data", Distributed and Parallel DatabasesAn International Journal,15 February 2014.

[4] United Nations Global Pulse, 2012, Big Data for Development:Challenges& Opportunities, May 2012

[5] Office of Science and Technology Policy | Executive Office of the President, 2012, Fact Sheet: Big Data across the Federal Government, March 29□2012□www.WhiteHouse.gov/OSTP

[6] Office of Science and Technology Policy | Executive Office of the President, 2012, Obama Administration Unveils □Big Data□ Initiative:Announces $200 Million in New R&D Investments, March 29□2012□ www.WhiteHouse.gov/OSTP

[7] McKinsey Global Institute, 2011, Big Data: the Next Frontier for Innovation, Competition, and Productivity, May 2011

[8] Rajaraman A, Ullman J D□Mining of Massive Datasets, Cambridge University Press, 2011

[9] EdmonBegoli, James Horey, "Design Principles for Effective Knowledge Discovery from Big Data", Joint Working Conference on Software Architecture & 6th European Conference on Software Architecture, 2012

[10] IvankaValova, Monique Noirhomme, "Processing Of Large Data Sets: Evolution, Opportunities And Challenges", Proceedings of PCaPAC08

[11] NehaSaxena, NiketBhargava, UrmilaMahor, Nitin Dixit, "An Efficient Technique on Cluster Based Master Slave Architecture Design", Fourth International Conference on Computational Intelligence and Communication Networks, 2012

[12] R. Caruana and A. Niculescu-Mizil, \An Empirical Comparison of Supervised Learning Algorithms," in Proceedings of the 23rd international conference on Machine learning, ICML '06, (New York, NY, USA), pp. 161{168, ACM, 2006.

[13] R. Caruana, N. Karampatziakis, and A. Yessenalina, \An Empirical Evaluation of Supervised Learning in High Dimensions," in Proceedings of the 25th international

[14] Mr. D. V. Patil, Prof. Dr. R. S. Bichkar, "A Hybrid Evolutionary Approach To Construct Optimal Decision Trees with Large Data Sets", IEEE, 2006

[15] Guillermo Sinchez-Diaz , Jose Ruiz-Shulcloper, "A Clustering Method for Very Large Mixed Data Sets", IEEE, 2001

[16] Mehmet Koyuturk, AnanthGrama, and NarenRamakrishnan, "Compression, Clustering, and Pattern Discovery in very High-Dimensional Discrete-Attribute Data Sets", IEEE Transactions On Knowledge And Data Engineering, April 2005, Vol. 17, No. 4

[17] Emily Namey, Greg Guest, Lucy Thairu, Laura Johnson, "Data Reduction Techniques for Large Qualitative Data Sets", 2007

[18] Moshe Looks, Andrew Levine, G. Adam Covington, Ronald P. Loui, John W. Lockwood, Young H. Cho, "Streaming Hierarchical Clustering for Concept Mining", IEEE, 2007

[19] Yen-ling Lu, chin-shyurngfahn, "Hierarchical Artificial Neural Networks For Recognizing High Similar Large Data Sets. ", Proceedings of the Sixth International Conference on Machine Learning and Cybernetics, August 2007

[20] Archana Singh, MeghaChaudhary, Dr (Prof.) Ajay Rana, GauravDubey, "Online Mining of data to Generate Association Rule Mining in Large Databases", International Conference on Recent Trends in Information Systems, 2011

[21] David N. Reshef et al.,"Detecting Novel Associations in Large Data Sets", Science AAAS, 2011, Science 334

[22] Shuliang Wang, WenyanGan, Deyi Li, Deren Li "Data Field For Hierarchical Clustering", International Journal of Data Warehousing and Mining, Dec. 2011

[23] Tatiana V. Karpinets, ByungH.Park, Edward C. Uberbacher, "Analyzing large biological datasets with association network", Nucleic Acids Research, 2012

[24] M. Vijayalakshmi, M. Renuka Devi, "A Survey Of Different Issues Of Different Clustering Algorithms Used In Large Data Sets", International Journal Of Advanced Research In Computer Science And Software Engineering, March 2012

[25] Xindong Wu, XingquanZhu,Gong-Qing Wu, and Wei Ding, "Data Mining with Big Data," IEEE Transactions On Knowledge And Data Engineering, Vol. 26, No. 1, January 2014.