ISSN 2395-1621

Survey on Data Mining and Information Security in Big Data Using HACE Theorem

[#1]Rajneeshkumar Pandey, [#2]Prof. Uday A. Mande

[1]prajneesh22@gmail.com
[2]uamande.scoe@sinhgad.edu

[#1]Department of Computer Engineering,
[#2]Prof. Department of Computer Engineering,
Sinhgad College of Engineering, Vadgaon(BK), Pune

ABSTRACT

Big Data is a popular term used to describe the exponential growth and availability of data, both structured and unstructured. Big Data might be petabytes(1024 terabytes) or exabytes(1024 petabytes) of data consisting of billions or trillions of records. Big Data are now rapidly expanding in all science and engineering domains, including biological, physical and biomedical sciences. Here presented the HACE theorem that characterizes the features of the Big Data revolution and proposes a Big Data processing model from the data mining perspective. Search in Big Data is cumbersome practice due to the large size and complexity of Big Data. The Big Data challenges are broad in the case of accessing, storing, searching, sharing, and transfer. Managing Big Data is not easy by using traditional relational database management systems; it requires instead parallel computing of dataset. Big data mining and analysis is parallel computing method which uses MapReduce framework of Hadoop and uses the k-means or Naïve Bayes algorithm for mine the data. This paper represents the use of MapReduce function of Hadoop and demand driven aggregation of big data which reduces computational cost. This paper also focuses on security and privacy issues in big data mining. Here it gives the privacy to data with AES algorithm.

Keywords: **Big Data, Data Mining, heterogeneity, autonomous sources, complex and evolving associations, AES algorithm.**

ARTICLE INFO

Article History

Received :30th December 2015

Received in revised form :

31st December 2015

Accepted : 1st January , 2016

Published online :

2nd January 2016

I. INTRODUCTION

The term Big Data appeared for first time in 1998 in a Silicon Graphics slide deck by John Mashey with the title of Big Data. The data produced these days is estimated in the order of zettabytes, and it is growing around 40 percent every year. A new large source of data is going to be generated from mobile devices and big companies like Facebook, Apple, Yahoo and Google. Every day 2.5 quintillion bytes of data are generated and 90 percent of the data in the world today were produced within two years [2]. Our ability for data generation has never been so powerful and massive since the origination of the information technology (IT) in the 19th century. As another example on 4 October 2012 the first presidential debate between Governor Mitt Romney and President Barack Obama triggered more than 10 million tweets within 2 hours. Among all these tweets, the specific moments that generated the most debates exposed the public interests, such as the thoughts about Medicare and vouchers. Such online discussions provide a new means to sense the public interests and generate feedback in real-time, and are mostly appealing compared to generic media, such as radio or TV broadcasting. The above examples show the rise of Big Data applications where data collection has grown extremely and is beyond the capability of commonly used software tools to capture, process and manage within a tolerable elapsed time. The biggest challenge for Big Data applications is explore the large amount of data and take out useful information from system and knowledge for future actions. In many situations the knowledge extraction process has to be very efficient and close to real time because storing all observed data is nearly infeasible. As a result the unmatched data volumes require an effective data analysis and also prediction platform to achieve fast response and real-time classification for such Big Data. The remainder of the paper is structured as follows: In Section II, we discuss literature review, Section III propose the HACE theorem to model Big Data characteristics and section IV, explains the methodology.

II. LITERATURE REVIEW

Data mining is uses the various types of techniques. All those techniques are not more secure. Some of the techniques are not suitable for the huge amount of data. Techniques and drawbacks are categorized as follows:

Sr no	Technique	Description	Drawback
1.	HACE theorem [1]	Uses distributed parallel computing with help of Hadoop. Used three tier framework 1.Big Data Mining Platform 2. Big Data Semantics and Application Knowledge 3. Big Data Mining Algorithm	Not more secure
2.	parallelization strategy Used MapReduce [3]	Used SVM algorithm, NNLS algorithm, LASSO algorithm, converting the problems into matrix-vector Multiplication	It's not provide any kind of security, Suitable for medium scale data
3.	Parallel Algorithms for Mining Large-scale Rich-media Data [4]	Used Spectral, Clustering, FP-Growth, Support Vector Machines	suitable for single source knowledge discovery methods, Not suitable for multisource knowledge discovery
4.	Combined Mining [5]	Multiple data sets, multiple features, multiple methods on demand, Pair pattern, Cluster pattern	Not able to handle large data.
5.	Decision Tree Learning [6]	Converts the original sample data sets into a group of unreal data sets, from which the original samples cannot be reconstructed without the entire group of unreal data sets.	Centralized, Storage Complexity, Privacy loss.

Table 1. Comparative Study

III. HACE THEOREM

Big Data starts with heterogeneous, large-volume, autonomous sources with distributed and decentralized control, and complex and evolving relationships among data. These characteristics make very important for determining useful information from the big data. In immature sense, we can imagine that a Number of blind men are trying to imagine the size up a massive Camel, which will be the Big Data in this context. The purpose of each blind man is to draw a picture of the Camel according to the part of information he collects during the process. Because each person's view is limited to according to his local region, it is not shocking that the blind men will each conclude independently that the camel feels like a wall, rope, hose and depending on the region each of them is limited to make the problem even more complex. Now consider that the camel is growing quickly and also its pose changes constantly and each blind man may have his own (inaccurate and possible unreliable) information sources that tell him about different-different knowledge about the camel. Exploring the Big Data in this scenario is alike to merging or integrating heterogeneous information from different sources (example as blind men) to help draw a best possible diagram of the camel. Definitely this task is very difficult that means it is not as simple as asking each blind man to explains his feelings about the camel and then getting an expert by using all information to draw one single picture combined view, focusing on that each individual person may speak a different language and they may even have privacy concerns about the messages they deliberate in the information exchange process. The term Big Data literally concerns about data volumes, HACE theorem suggests that the key characteristics of the Big Data are:

A. Huge With Heterogeneous and Diverse Data Sources

Big data is heterogeneous because different data collectors use their own big data protocols and schemata. For example, data stored by DNA scanning, CT scan and X-ray are in the different form depends on its use. It may be videos, images or series of images. Big challenging issue in data aggregation is to collect data from heterogeneous and diverse dimensionality resources. This huge volume of data comes from different sites like Orkut, MySpace, LinkedIn and Twitter etc.

B. Autonomous Sources and Decentralized Control

Autonomous data sources with distributed and decentralized controls are a main characteristic of Big Data applications. Being autonomous, each data source is able to collect and generate the information without including any centralized control. This is totally alike to the World Wide Web (WWW) setting where each web server Author and Title details provides a proper amount of information and also each server can function fully without necessarily depend on other servers.

C. Complex Data and Knowledge Associations

Complexity and relationships among data grow the increase in a volume of data day by day. The relationship between individual such as in Facebook friends or twitter represents complex relationship because everyday friends are added and to maintain the relationship among them is big challenge for developers. Such a complexity is becoming the challenging issue with consideration of changes in data in every day [7]. Multi-structure, multisource data is complex data.

IV. METHODOLOGY

The data mining in big data mainly divided into three-tier structure, those are as follows:

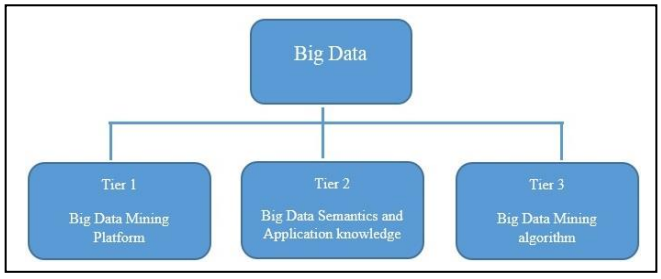

Fig 1. Three-tier Structure

The challenges at Tier I concentration on data accessing and doing arithmetic operation on them. Big data are not able to store in single place so that it is stored on diverse locations and day by day it constantly increasing. So that tackling from such types of challenges, common solutions are

depend on parallel computing [8]. For Big Data mining, a high-performance computing platform with a data mining task is also running some parallel programming tools such as MapReduce [9].

The challenges at Tier II focus on semantics and domain knowledge for different Big Data applications. Such information can provide benefits to the mining process and add technical barriers Tier I and data mining algorithms that is in Tier III. For example, rely on various domain applications, the information sharing and data privacy mechanisms between data producers and data consumers can be significantly different [10].

At Tier III, contains three stages. First uncertain and sparse, heterogeneous, and multisource data are pre-processed. Second dynamic and complex data are mined after pre-processing operation. Third the global knowledge get by local learning and relevant information is feedback to the pre-processing stage. Then the model and parameters are adjusted according to the feedback.

Overall working of the system is shown in the follows system architecture

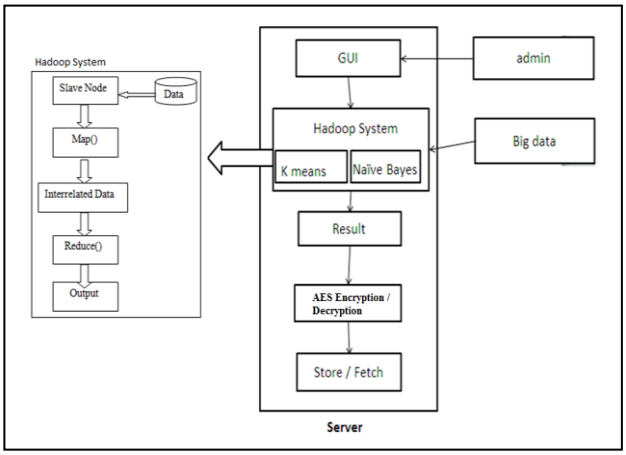

Fig 2. System Architecture

A. Admin

Admin is responsible for the fired the queries according to his need. When admin has fired the query, means admin is interacting with the GUI of the system. After that the Hadoop System is responsible for the processing the mining further.

B. Hadoop System

Hadoop uses MapReduce programming model to mine data. This MapReduce program is used to separate datasets which are sent as input into independent subsets. Those are process parallel map task. Map() procedure that performs filtering and sorting. Reduce() procedure that performs a summary operation. After doing the MapReduce operation then whatever the output system created it given to the K-means or Naive Bayes algorithm for doing clustering and classification.

D. AES Encryption Algorithm

AES is symmetric key based encryption algorithm that means the same key is used for both encrypting and decrypting the data. It has mainly three types as AES-128, AES-192 and AES-256. Each cipher encrypts and decrypts data in blocks. There are 10 rounds for 128-bit keys, 12 rounds for 192-bit keys, and 14 rounds for 256-bit keys a round consists of several processing steps that include substitution, transposition and mixing of the input plaintext and transform it into the final output of cipher text [11, 12].

V. Conclusion

Big data is the term for a collection of complex data sets, Data mining is an analytic process designed to explore data in search of consistent patterns and then to validate the findings by applying the detected patterns to new subsets of data. To support Big data mining, high-performance computing platforms are required. In this paper proposed the HACE theorem for data mining with Big data. For data mining uses the k-means and Naive Bayes algorithm. This system is providing security by using AES algorithm, hence it is more secure than traditional systems.

ACKNOWLEDGEMENT

I am Rajneeshkumar Pandey and would like to thank the publishers, researchers for making their resources material available. I am greatly thankful to Associate Prof. Uday A. Mande for their guidance. I also thank the college authorities, PG coordinator and Principal for providing the required infrastructure and support. Finally, I would like to extend a heartfelt gratitude to friends and family members.

REFRENCES

[1] X. Wu, X. Zhu, G-Q. Wu, and W. Ding, "Data Mining with Big Data," IEEE Trans. On Knowledge and data engineering, vol. 26, no. 1, pp. 97-107, Jan. 2014.

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

[3] D. Luo, C. Ding, and H. Huang, with multiplicative Algorithms for Big Data Mining," Proc. IEEE 12th Intl Conf. Data Mining, pp. 489-498, 2012.

[4] Lo, B. P. L., and S. A. Velastin. "Parallel Algorithms for Mining Large-Scale Rich-Media Data," Proc. 17th ACM Intl Conf. Multimedia, (MM 09,), pp. 917-918, 2009.

[5] Longbing Cao (2012), Combined Mining: Analyzing Object and Pattern Relations for Discovering Actionable Complex Patterns, sponsored by Australian Research Council discovery Grants.

[6] P. K. Fong and J. H. Weber,"Privacy preserving decision tree learning using unrealized data sets," IEEE Trans. Knowl. Data Eng., vol. 24, no. 2, pp. 353_364, Feb. 2012.

[7] Y.-C. Chen, W.-C. Peng, and S.-Y. Lee, "Efficient Algorithms for Influence Maximization in Social Networks," Knowledge and Information Systems, vol. 33, no. 3, pp. 577-601, Dec. 2012.

[8] J. Shafer, R. Agrawal, and M. Mehta, SPRINT: A Scalable Parallel Classifier for Data Mining, Proc. 22nd VLDB Conf., 1996.

[9] D. Gillick, A. Faria, and J. DeNero, MapReduce: Distributed Computing for Machine Learning, Berkley, Dec. 2006.

[10] E. Schadt, "The Changing Privacy Landscape in the Era of Big Data," Molecular Systems, vol. 8, article 612, 2012.

[11] C. Wang, S.S.M. Chow, Q. Wang, K. Ren, and W. Lou, "Privacy- Preserving Public Auditing for Secure Cloud Storage" IEEE Trans. Computers, vol. 62, no. 2, pp. 362-375, Feb. 2013.

[12] "Advance Encryption Algorithm", https://en.wikipedia.org/wiki/Advanced_Encryption_Standard.

