Data Mining and Information Security in Big Data Using HACE Theorem

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

[1]PG Student, Department of Computer Engineering
Sinhgad College of Engineering, Pune -41, India
[2]Associate Professor, Department of Computer Engineering
Sinhgad College of Engineering, Pune -41, India

---***---

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 Nave 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.*

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

1. INTRODUCTION

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 more than 2.5 quintillion bytes of data are generated and most of them were created 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 discussed as follows: In Section II, we discuss literature review, Section III discusses the sources of big data, Section IV proposes the HACE theorem to model Big Data characteristics, section V explains the methodology and section VI explains the results of proposed system.

2. LITERATURE REVIEW

In this paper [1] presents a HACE theorem that explain the features of the Big Data revolution, and proposes a Big Data processing model, from the data mining. This paper explains the data-driven model which contains the aggregation of information sources and privacy considerations. Here it provides the challenging issues in the data-driven model and also in the Big Data revolution. It is not more secure, so that for security reason proposed system uses the AES algorithm.

In this paper [3] explains that a number of online photos and videos are now at the scale of tens of billions. To organize, index, and retrieve this large scale rich-media data, a system must employ scalable data management and mining algorithms. The research community always trusts on solving big amount of data instead of solving many small amount of dataset. In this paper presents parallel algorithms for tackling such challenges and introduces key challenges in large-scale rich-media data mining. Here it presents parallel implementation of FP-

Growth, Spectral Clustering, and Support Vector Machines. It is suitable only for single source knowledge discovery methods, so that not suitable for multisource knowledge discovery.

In this paper [4] explains that for paralleling a series of data mining and machine learning problems it uses non-trivial strategy, including 1-class and 2-class support vector machines, non-negative least square problems, and 1 regularized regression (LASSO) problems. This paper strategy fortunately leads to extremely simple multiplicative algorithms which can be directly implemented in parallel computational environments, such as MapReduce, or CUDA. Here it provides difficult analysis of the correctness and convergence of the algorithm. This paper demonstrated the scalability and accuracy of our algorithms in comparison with other current leading algorithms. But it is not suitable for suitable for medium scale and also, it does not provide any security.

In this paper [5] presents the Combined mining concept introduced that it will integrate the many mining algorithm and association and classification rule was done also. The integration of classification and mining done by a special subset of association rules that is CARs. Classifier built is more accurate but simple classification system is not accurate with complex, huge, and heterogeneous data. Association rules are generally extracted from transactional data with a single set.

In this paper [6] introduces a privacy preserving approach that can be applied to decision tree learning, without associated loss of accuracy. It describes an approach to the preservation of the privacy of collected data samples in cases where information from the sample database has been partially lost. Here this approach converts the original sample data sets into a group of unreal data sets from which the original samples cannot be reconstructed without the whole group of unreal data sets. But there are also some problems as data distribution in Centralized format, Storage Complexity and Privacy loss also. Data mining is using the various types of techniques. All those techniques are not more secure. Some of the techniques are not suitable for the large amount of data.

In this paper [7] privacy and security concerns often prevent the sharing of user's data or even of the knowledge gained from it, thus deterring valuable information from being utilized. Privacy preserving knowledge discovery, if done correctly, can alleviate this problem. Here it is using the classification technique, for that it uses the Naïve Bayes algorithm. Here it considers the model where a single provider has centralized access to a dataset and would like to release a classifier while protecting privacy to the best extent possible. But here the classification accuracy is not so good and also it is centralized, which is overcome in the proposed system.

In this paper [16] it discusses different algorithms, and comparing various algorithms and techniques used for cluster analysis using weka tools. This paper presents the comparison of 9 clustering algorithms in terms of their execution time, number of iterations, sum of squared error and log likelihood. Then on that basis it defines that k-means algorithm is easy and simple to use as well as k – means algorithm is suitable for the huge amount of data, that is why in proposed system k-means is used.

Techniques and drawbacks are categorized in table 1.

Table -1: Comparative Study

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.
6.	Naïve Bayesian [7]	Adding noise to classifier's parameters.	Classification accuracy, centralized.

3. TYPES OF BIG DATA AND SOURCES

There are mainly two types of big data that is structured and unstructured.

1. **Structured data** are words and numbers, and easily can be categorized in tabular format and easy to analysed. These data are generated many sources as global positioning system (GPS) devices, smart phones and network sensors embedded in electronic devices.

2. **Unstructured data** are not easy to analysed and it will not maintained in the tabular format. It contains more complex information such as commenting on any websites, customer review of any commercial websites. Most of the time unstructured data is not easily readable.

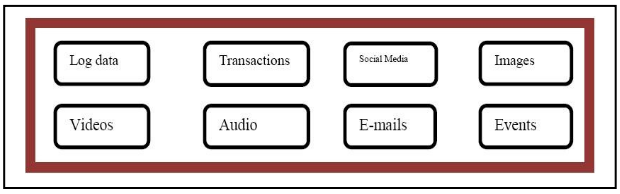

Fig -1: Sources of Big Data

4. 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, so now we consider a very big animal like Camel, so if any blind man is trying to size up a massive Camel, which will be consider as 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 thinks about the animal or whatever information he collects from his source. It is human nature that every person thinks differently, so that's why blind men will conclude that the camel feels like a wall, rope and hose. 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 (example as one blind man may share his feeling about the camel with another blind man where the exchanged knowledge is fundamentally biased). According to getting information from the sources it again make changes in the thinking in the blind man. 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 just asking to any blind man to explains his feeling about the Camel and integrating all the information getting from the different-different sources, and by using only those information to draw the one single picture combined view by focusing on the each individual sources (diverse information sources and heterogeneous) getting to him. The term Big Data literally concerns about data volumes, HACE theorem suggests that the key characteristics of the Big Data are:

1) 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.

2) 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.

3) 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 [8]. Multi-structure, multisource data is complex data, Examples of complex data types are bills of materials, word processing documents, maps, time-series, video, and images. Such combined characteristics suggest that Big Data require a big mind to consolidate data for maximum values.

5. PROPOSED SYSTEM

5.1 Problem statement

To implement data mining with big data by using HACE theorem that characterizes the features of the big data revolution, for mining uses the k-means and Naïve Bayes algorithm. Also provide the security by using AES algorithm.

5.2 Methodology

The data mining in big data mainly divided into three-tier structure, those are shown in figure 2.

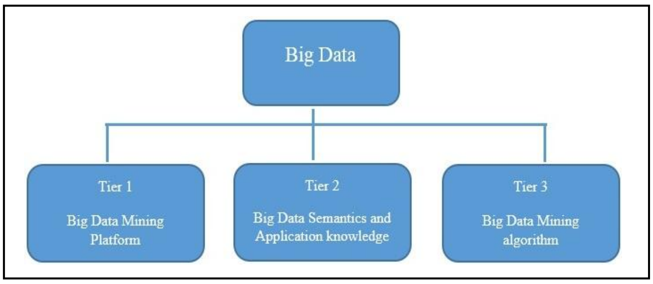
Big Data
Tier 1
Big Data Mining Platform
Tier 2
Big Data Semantics and Application knowledge
Tier 3
Big Data Mining algorithm

Fig -2: 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, hence effective computing platform will have to take distributed large-scale data storage and also do arithmetic operation on them. So that for doing such type of operation common solutions are depend on parallel computing [9]. For Big Data mining, because data scale is in very huge quantity, which is not easily handle by the single personal computer. So that for handling this kind of information distribution of data concept is used. For Big Data processing framework parallel programming tools used such as MapReduce [10, 11].

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 [12].

At Tier III, the data mining challenges concentrate on algorithm designs in tackling the difficulties raised by the Big Data volumes, complex and dynamic data characteristics and distributed data distributions. The circle at Tier III contains three stages. First uncertain and sparse, heterogeneous, and multisource data are preprocessed. Second dynamic and complex data are mined after preprocessing operation. Third the global knowledge get by local learning and relevant information is feedback to the preprocessing stage. Then the model and parameters are adjusted according to the feedback.

5.3 System Architecture

Overall working of the system is shown in figure 3. It shows all the phases of the data mining concept from the big data.

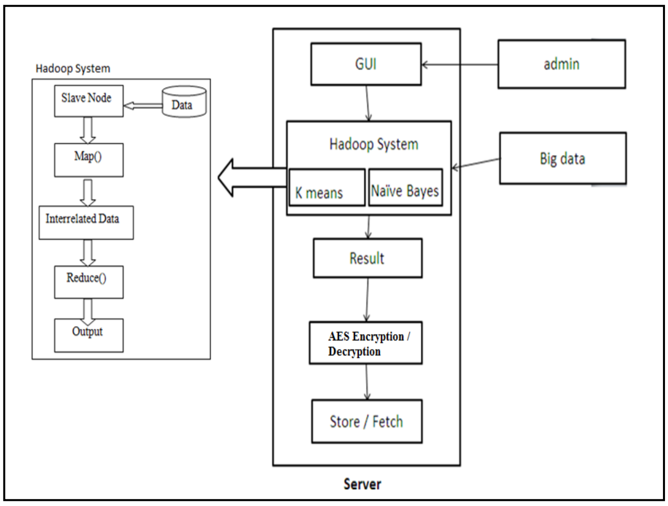

Fig -3: System Architecture

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.

C. K-means algorithm

- Partition of a dataset into given k non-empty set.
- Identification of cluster mean point called centroids for the current partition.
- Assign each point to a specific cluster.
- The minimum distance from each point to centroid is computed, and then points are allotted to the cluster
- Computation of distance between each point and allocation of minimum distanced points from mean point to cluster.
- Repeat the above steps for re-allotted points and find the mean point for the new cluster.

D. Naïve Bayes Algorithm

Based on Bayes theorem and frequency table. It gives the Estimation, Classification, and Prediction. It is used when large data set. It is very easy to construct. Not using complicated iterative parameter estimations. It solves the zero frequency problems. It uses the following formula as:

$$P(Y|X_1,\ldots,X_n) = \frac{\overset{\text{Likelihood}}{P(X_1,\ldots,X_n|Y)}\,\overset{\text{Prior}}{P(Y)}}{\underset{\text{Normalization Constant}}{P(X_1,\ldots,X_n)}}$$

E. 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 [13, 14].

F. Flow Diagram

In this section flow of overall proposed system is explained.

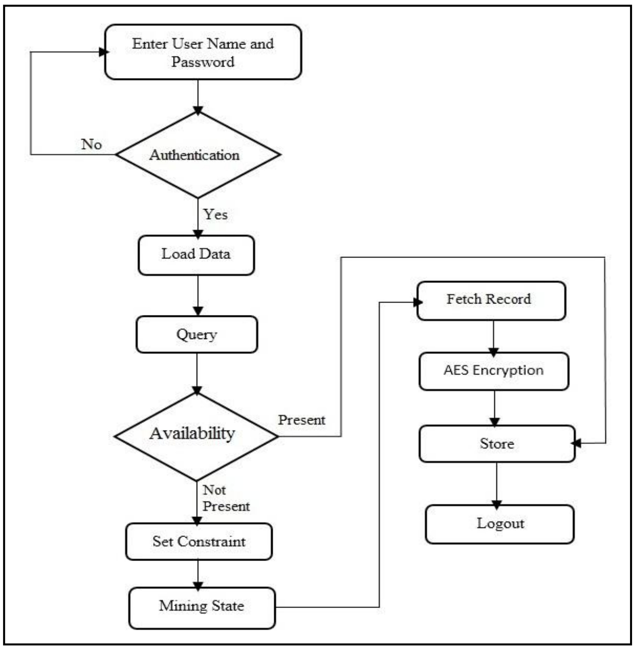

Fig -4: Flow Diagram

6. RESULTS AND DISCUSSIONS:

In this section presented the evaluation of proposed system. After describing experimental setup, then we have measured require execution time for algorithm and also about security. Execution time may be change in different systems.

6.1 Experimental Setup

Here we are using medical dataset for proposed system. Initially we have to load our dataset in the hadoop system then in Mapreduce functionality is applied on it . Data mining is done with two ways that is by using k-means algorithm or Naïve Bayes algorithm. After naïve Bayes algorithm AES algorithm is applied for encryption.

6.2 Compare Algorithms

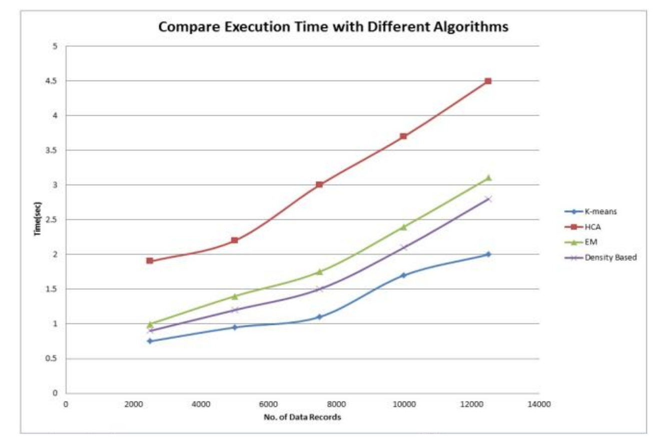

Fig -5: Compare Algorithms

In the above figure it shows that k means is better than the other algorithms, that's why k-means algorithm is used for doing the data mining, in the given figure it shows that for 10k data records it required about 3.7s but k-means can do it just 1.7s which is lowest value as compare to other algorithms.

6.3 Execution time

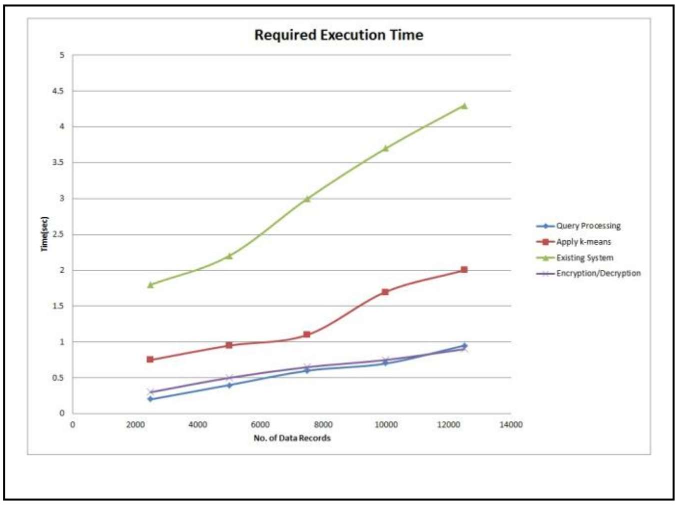

Fig -6: Required Execution Time

Here we use the analysis on the basis of parameter time. K-means algorithm is used for doing query processing and it required very less time for it, it takes only 0.7s for query processing and for clustering all the 10k records it takes 1.7s only.

6.4 Clustering of data records

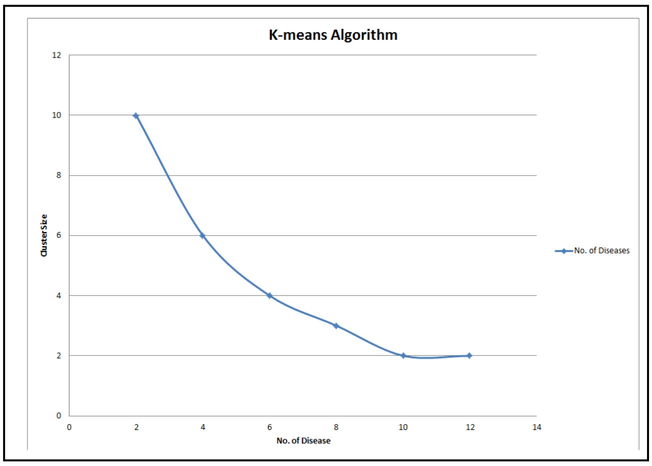

Fig -7: Clustering of data records

Here we use the k-means algorithm for clustering the data on the attribute of 'Disease', here we uses 10k data records. So in the dataset total 21 diseases are available so that if we take cluster size as 2 then near about 50% of disease in one cluster, similarly if we take cluster size 10 then up to 12 to 18 percent disease in each cluster, which is not possible in other algorithms as explained by the Prakash Singh and Aarohi Surya [16].

6.5 Security with Naïve Bayes algorithm

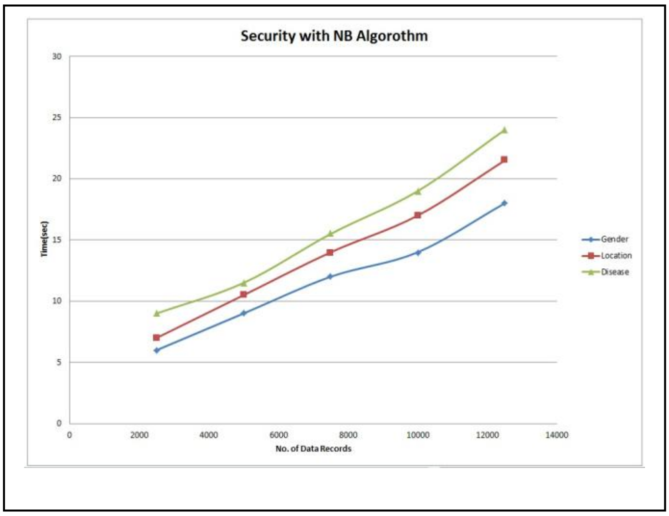

Fig -8: Security with NB Algorithm

Here data is initially classified with the help of the naïve Bayes algorithm then on that classified records or we can say mined data AES algorithm is applied. Both the functionality is run after 1st is completed. After apply AES encryption algorithm we get encrypted file.

7. CONCLUSION AND FUTURE WORK

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.

In the given system security approach may further improved by using key aggregation cryptosystem which is more secure and reliable [18]. As well as analysed data can be used for business purpose. Also different data mining algorithm may be used for data mining for better result.

ACKNOWLEDGEMENT

I 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.

REFERENCES

[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, "Parallelization 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, "Combined Mining: Analyzing Object and Pattern Relations for Discovering Actionable Complex Patterns", *sponsored by Australian Research Council discovery Grants*, 2012.

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

[7] J. Vaidya, B. Shafiq, A. Basu, and Y. Hong, "Differentially private Naïve Bayes classification," in *Proc. IEEE/WIC/ACM Int. Joint Conf. Web Intell. (WI) Intell. Agent Technol. (IAT)*, vol. 1. Nov. 2013, pp. 571_576.

[8] 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.

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

[10] C. Ranger, R. Raghuraman, A. Penmetsa, G. Bradski, and C. Kozyrakis, "Evaluating MapReduce for Multi-Core and Multiprocessor Systems," *Proc. IEEE 13th Intl Symp. High Performance Computer Architecture (HPCA 07)*, pp. 13-24, 2007.

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

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

[13] 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.

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

[15] A. Rajaraman and J. Ullman, "Mining of Massive Data Sets," *Cambridge Univ. Press*, 2011.

[16] Prakash Singh, Aarohi Surya, "Performance Analysis of clustering algorithms in data mining in Weka", *IJAET*, Jan 2015.

[17] Rajneeshkumar Pandey, Prof. Uday A. Mande, "Survey on data mining and Information security in Big Data using HACE theorem," *International Engineering Research Journal (IERJ)*, vol. 1, issue 11 , 2016, ISSN 2395-1621.

[18] Cheng-Kang Chu, Sherman S. M. Chow, Wen-Guey Tzeng, Jianying Zhou, and Robert H. Deng, "Key-Aggregate Cryptosystem or Scalable Data Sharing in Cloud Storage," *IEEE Transactions On Parallel And Distributed System*, Vol 25, No. 2 February 2014.

BIOGRAPHIES

Rajneeshkumar Pandey (prajneesh22@gmail.com) received bachelor's degree in Information Technology from University of Pune and currently pursuing Master of Engineering in Computer Networks from Sinhgad College Of Engineering, Pune. Also doing internship in Avaya India pvt. ltd.

Prof. Uday A Mande (uamande.scoe@sinhgad.edu) received his Bachelor's degree in Computer from Marathwada University, Aurangabad and Master's degree in CSE-IT from Pune University. He is Associate Professor, Department of Computer Engineering, Sinhgad College of Engineering, Pune.