International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-9 Issue-2, December, 2019

A Bigdata Process for Practical Privacy-Preserving Utilizing k-Means Clustering

Praveen S. Banasode, Sunita Padamannavar

Abstract: *Now a day's privacy preservation is the big issue on growing big data in various field such as medical, engineering and physical with the fast growing network. One of the most important challenges in handling big data is security issues. To overcome such security issues cryptographic concepts have been used in this paper to provide high security of big data's with the low consumption of time for both encryption and decryption process. In this paper the proposed method is Indexed RSA (IRSA) which is developed with modified scheme. We offered a method to index the keyword before encrypting the file and based on the indexed keyword the search has been done. Finally the security analysis was carried out and the analysis showed that our modified scheme can meet the security requirement against brute force attack and SQL injection attack.*

Keywords: IRSA, Encryption, Decryption, Big Data

I. INTRODUCTION

Big data has large set of data which are complex, heterogeneous, unstructured and disorganized with the multiple independent sources. These days, the concept of big data is rapidly growing in all fields such as engineering, science, medical, social domains including biological, physical and biomedical sciences with the fast development of the networking, data collection and data storage capacity. Typically the big data are heterogeneous, i.e., all object in the big data is multi model. It includes several interconnected kinds of objects include audios, texts, and images. Due to the structured and unstructured data it results in high heterogeneity. While they are interrelated with each other it carries different information [Zhang et al., 2016]. Now a days it is a big issue for maintaining such big data's though every day more and more attacks to grab the important information's from internet and thus it takes more efforts for such data security. These efforts includes the designing the new algorithms for encryption and security systems [Mehta B. B & Rao, U.P. 2016]. Cryptography is the art and technique of preserving privacy and security of information / data. Cryptosystems are of two types such secret-key system and public key system. In secret key the encryption and decryption are done with same key where as in public key it is done by different keys. Secret key system is also called symmetric cryptosystem and public key is known as asymmetric cryptosystem. Secret key system is easy to implement but suffers from key distribution, authentication and non repudiation problems. The public key resolves the issues of secret key system. RSA cryptosystem is the most popular approach in the public key cryptosystem. The RSA cryptosystem was developed in 1977 by Ronald L. Rivest, Adi Shamir, and Leonard Adleman at MIT and first published in 1978 [Rivest et al., 1978]. The most commonly used algorithm for the data encryption and digital signature application is RSA algorithm. Chaotic systems in encryption is one of the most significant methods used in the maintenance of data security. It is a matter of debate that the science of the chaos and cryptology are closely coupled, due to the sensitive dependence on the initial conditions and control parameters of the chaotic systems.

Revised Manuscript Received on December 30, 2019.

* Correspondence Author
Praveen S. Banasode, Jain College of Engineering, Belagavi affiliated to Visvesvaraya Technological University, Belagavi Karnataka, India. Email: praveenb.jce@gmail.com
Dr. Sunita Padamannavar, KLS Gogte Institute of Technology, Belagavi Affiliated to Visvesvaraya Technological University, Belagavi. Karnataka, India. Email:sunitapdm@gmail.com

II. RELATED WORK

There are numerous literatures have been studied based on the RSA encryption and other privacy preserving algorithms. Yuan and Tian (2017) presented K means clustering scheme based on Map Reduce in cloud computing. Even though the scheme is perfect in getting accuracy in clustering and speed that are comparable to the K-means clustering, it fails in protecting privacy of the data. They have implemented the system with 5 million data objects. The implementation demonstrates that the scheme is scalable, efficient and accurate for K-means clustering for large-scale dataset. Zhang et al., (2017) made a high-order PCM scheme for heterogeneous clustering of data. They have used cloud servers to improve the efficiency of clustering big data by presenting a distributed HOPCM scheme based on Map Reduce. Further they developed Privacy Preserving Higher Order PCM using BGV encryption to preserve privacy of the big data. The implementation results show that PPHOPCM can cluster big data by using the cloud computing technology without disclosing privacy. Manjusha and Hari Kumar (2016) made a comparative study on the performance. This is carried out for these spectral values. They made the classification using both KNN and K means clustering. The experiment proved that k-means performs better than KNN. Li et al., (2017) devised a high-order neuro fuzzy c means algorithm to cluster heterogeneous data in cloud. This is developed as an advance work of PPHOPCM, for the encryption of the data they have used BGV encryption. The method was enhanced by employing the cloud computing. Experiment simplified that PPHOFCM outperforms PPHOPCM for clustering heterogeneous data in IOT. But the scheme is not feasible for noisy data though it is highly scalable.

Retrieval Number: B4085129219/2019©BEIESP
DOI: 10.35940/ijeat.B4085.129219
Journal Website: www.ijeat.org

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

Lou et al., (2014) presented DPQR algorithm. This is based on data perturbation and query restriction. This base is achieved through multi-parameters perturbation.

The processing of data is done different ways so as to preserve the privacy using various parameters. The algorithm reduces the time complexity getting by the adjacency matrix recursive relations and matrix block method The experiment and results show that the execution efficiency of DPQR is highly improved over the MASK algorithm. But this DPQR handles only Boolean data and not found suitable for numerical and other types of data.

Zhu et al., (2016) proposed a practical privacy preserving frame work called eDiag. This is with the help of nonlinear kernel SVM and the data used is medical data. The model is introduced on an improved expression for the nonlinear support vector machine. They have incorporated the techniques of lightweight multi-party random masking and polynomial aggregation techniques. For the unregistered user, the query is answered directly at the service provider without decryption and for the registered user the result can be decrypted, meanwhile the query result is consistent with that of un-privacy-preserving scheme. Lin et.al (2017) constructed an efficient CL-PKE (certificate less public key encryption) scheme from RSA which is the defeat internet standard. The security is based on a variant of RSA known as Kilian–Petrank's RSA assumption. The proposed CL-PKE algorithm is based on the nature BDHP difficulty assumption and therefore avoids paring computation on elliptic curves, which is the most expensive operation in the encryption algorithm. In CL-PKE algorithm, the selective-ID model is applied instead of the random oracle model. Even RSA outperforms in giving privacy and security, it is rarely used in big data's because of computation time. It is mainly used in various fields such digital signatures. Some components, such as multi-leveled equations, tables and graphics are not prescribed, although the various table text styles are provided. To tackle above problems this paper proposes an IRSA encryption algorithm for big data privacy preservation. In our proposed method it reduce the computational time and the proposed method creates indexes that help the user to get the required documents in a secure environment.

III. PROBLEM DEFINITION

Privacy has turn out to be a substantial problem as there are numerous applications of big data are intensely growing. Big Data has enormous volumes of diverse data which is created at excessive speeds. As data is generated at higher speeds and in huge sizes we are in need of a fresh set of tools and application to execute and succeed the data. The existing system is time consuming and lengthy. Besides they are not validated on larger actual data sets . The complexity of encryption and decryption of the cryptographic techniques has to be reduced which in turn does not reduce the security level of the encrypted data To the Sensitive Data, the data must be encrypted before subcontracted to guard user's and company's privacy, but the encrypted data can't be examined over old search method. Moreover the security algorithms have been applied to prevent from common attacks against IRSA algorithm.

IV. METHODOLOGY

We have implemented a new method for preserving privacy of data. This is to index the keyword before encrypting the file and based on the indexed keyword the search will be done. If a fresh search is made then the index is explored and based on the index. The confidentiality of the novel scheme will be preserved and computation complexity of the novel scheme will also be decreased. In our method indexes are used instead of actual values of public and private keys at the time of communication between sender and receiver. The proposed system consists of five modules

1. Indexing the data/file
2. Key Generation
3. Encryption
4. Search the file based on index
5. Decryption.

Indexing

Our algorithm creates index based on keyword. The following details are contained in the key word:

a) ID of the file that contains particular keyword
b) Frequency of the keyword in the particular file
c) The length of file
d) Score of relevance
e) Total number of files which contains number of key word.

Key Generation

a) Let a and b are two prime number such that a not = b
b) Find k = a*b
c) Find ø (k) = (a -1) * (b-1).
d) Choose integer p with gcd (Ø (k), e) = 1.
e) Find q.
f) Create Public key PU = {e, k}.

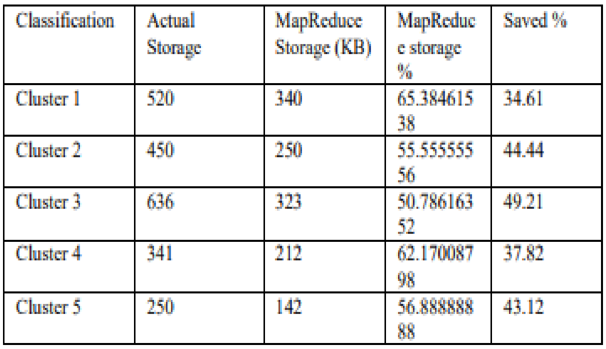
Encryption The proposed system encrypt the file with random bits sizes of RSA like 1024, 2048, 4096 bits. As the collections are encrypted with random bits sizes, it will be difficult to decrypt without a key. The encrypted file is indexed with the keywords that are in the files so when the search is made the files need not decrypt every time. It will check the index and decrypt the selected files. If the key is not provided then the decryption will be aborted. With this proposed method secure communication has been provided. Search the file Based on Index as the encrypted data is indexed with the keywords, we can search the data/file according to the index and also decrypt the selected file/data. The algorithm will check the index and decrypt only those files which are needed.

Decryption since RSA is a asymmetric cryptographic method, the party should use private key for the decrypting the data. The algorithm generates public and private key. It will be stored either in memory or in a cryptographic key container. Let A and B are two parties and A sends public key to B, using that B encrypts the data and send back to A. Using private key which B used for encryption A can decrypt the data

V. EXPERIMENTAL RESULT

A system called K-means clustering is proposed that uses Map Reduce technique over Large-scale Dataset [4]. First the trained data sets are initialized for every different cluster which is related to Reuter's collection Information .After, the clustering algorithm divide file into number of chunks and for every chucks hash code is generated for the security purpose [2].

Before storing into HDFS System, classification algorithm classifies that file belong to which cluster category [1]. Advantages of proposed system are: Hadoop is a distributed file system and provides fast storage of files. Security is more because of Hash code generation Speed of Transmission is high because of duplication concept is used while uploading file to the HDFS storage



Classification	Actual Storage	MapReduce Storage (KB)	MapReduce storage %	Saved %
Cluster 1	520	340	65.38461538	34.61
Cluster 2	450	250	55.55555556	44.44
Cluster 3	636	323	50.78616352	49.21
Cluster 4	341	212	62.17008798	37.82
Cluster 5	250	142	56.88888888	43.12

Table : Performance of MapReduce

VI. CONCLUSION

In this paper, we proposed the privacy preserving algorithm for big data with the modification in the RSA algorithm by indexing the keywords called IRSA. Based on the indexed keyword the search has been done. The fresh search was made then the index was explored and based on the index. The results showed that the novel scheme preserved the big data security and also computational time and complexity was also decreased. Finally security analysis was done to know the attack resistance against SQL and brute force attacks. Thus the security analysis showed that our modified scheme met the security requirement.

REFERENCES

1. Bertino, E. (2015). Big data - security and privacy. 2015 IEEE International Congress on Big Data. doi:10.1109/bigdatacongress.2015.126
2. Irshad Hussain, N., Choudhury, B., & Rakshit, S. (2014). A novel method for preserving privacy in bigdata mining. International Journal of Computer Applications, 103(16), 21–25. doi:10.5120/181599378.
3. Li, P., Chen, Z., Yang, L. T., Zhao, L., & Zhang, Q. (2017). A privacy-preserving high-order neuro-fuzzy cmeans algorithm with cloud computing. Neurocomputing, 256, 82–89. https://doi.org/10.1016/j.neucom.2016.08.135 [4] Lin, X. J., Sun, L., & Qu, H. (2018). An efficient RSAbased certificateless public key encryption scheme. Discrete Applied Mathematics, 241, 39–47. https://doi.org/10.1016/j.dam.2017.02.019
4. Lou, H., Ma, Y., Zhang, F., Liu, M., & Shen, W. (2014). Data mining for privacy preserving association rules based on improved MASK algorithm. Proceedings of the 2014 IEEE 18th International Conference on Computer Supported Cooperative Work in Design (CSCWD), 265–270. https://doi.org/10.1109/CSCWD.2014.6846853
5. Manjusha, M., & Harikumar, R. (2016). Performance analysis of KNN classifier and K-means clustering for robust classification of epilepsy from EEG signals. In Proceedings of the 2016 IEEE International Conference on Wireless Communications, Signal Processing and Networking, WiSPNET 2016(pp. 2412– 2416). Presses Polytechniques Et Universitaires Romandes. https://doi.org/10.1109/WiSPNET.2016.7566575
6. Mehta, B. B., & Rao, U. P. (2016). Privacy preserving unstructured big data Analytics: Issues and challenges. Procedia Computer Science, 78, 120–124. doi:10.1016/j.procs.2016.02.020
7. Saraladevi, B., Pazhaniraja, N., Paul, P. V., Basha, M. S. S., & Dhavachelvan, P. (2015). Big data and Hadoop-a study in security perspective. Procedia Computer Science, 50, 596–601. doi:10.1016/j.procs.2015.04.091
8. Yuan, J & Tian Y. (2017). Practical Privacy – Preserving MapReduce Based K-means Clustering over Large-scale Dataset. IEEE Transactions on Big Data, 1–1. https://doi.org/10.1109/TCC.2017.2701816
9. Zhang, Q., Yang, L. T., Chen, Z., & Li, P. (2017). PPHOPCM: Privacy-preserving High-order Possibilistic c-Means Algorithm for Big Data Clustering with Cloud Computing. IEEE Transactions on Big Data, 1–1. https://doi.org/10.1109/TBDATA.2017.2701816
10. Zhu, H., Liu, X., Lu, R., & Li, H. (2017). Efficient and Privacy-Preserving Online Medical Prediagnosis Framework Using Nonlinear SVM. IEEE Journal of Biomedical and Health Informatics, 21(3), 838–850. https://doi.org/10.1109/JBHI.2016.2548248