IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

AN INTEGRATED METHODOLOGIES FOR PRIVACY PRESERVING IN CLOUD OVER BIG DATA USING HADOOP FRAMEWORK

Ms. A. Kanimozhi [*1] , Dr. N. Vimala[2]

[*1]Ph.D (Part time) Scholar ,

Department of Computer Science ,LRG Govt Arts College for Women ,Tirupur.

[*2]Assistant Professor, Department of Computer Science ,LRG Govt Arts College for Women , Tirupur.

Abstract:

One of the most important aspects of big data research is data security. The majority of cloud systems deal with perceptive in order, such as personal, business, or health-related information. The rise of cloud storage systems poses a threat to this type of data. Traditional security methods, on the other hand, fail to protect large data transmissions. An effective privacy-preserving methodology to handle data generation and protection issues. Initially, cloud-based big data is clustered with multiple security procedures, which are balanced with the help of a Hadoop-based map-reduce mechanism.

Furthermore, the data and information are encrypted using a variety of data security measures, and the convolution process is carried out over the data encryption process. The deep neural network establishes the evaluation process (DNN). The data will be passed through the convolution process again if the encrypted data is not adequately encrypted. The proposed framework combines experimental discoveries with current methodology, and the big data-based privacy-preserving scheme outperforms existing strategies when data size is changed.

Keywords: Big data, encryption, cloud storage, classification, clustering, optimization.

1. Introduction

Big Data is a big volume of data that continues to rise exponentially over time. It's a data set so massive and complicated that no typical data management technologies can effectively store or process it. Big data is similar to regular data, except it is much larger. Medical researchers and clinicians use big data to uncover disease indicators and risk factors, as well as diagnose illnesses and medical problems in patients. Furthermore, data from electronic health records, social media sites, the internet, and other sources is combined to provide healthcare organisations and government agencies with up-to-date information on infectious disease threats and outbreaks.

Data security is the combination of confidentiality, which prevents unauthorised disclosure of information, integrity, which prevents unauthorised amendment or deletion of information, availability, which prevents unauthorised withholding of information, and privacy, which is the ability of an individual or group to reveal information about themselves selectively.

For the future generation of IT applications, cloud computing is a promising and developing technology. The fast expansion of cloud computing has raised concerns about data security and privacy. Researchers have offered a number of strategies for data protection and achieving the maximum level of data security in the cloud.

Hadoop is an Apache v2 licensed open-source software platform that facilitates data-intensive distributed applications. Simple programming models abstract and simplify the storage and processing of vast and/or fast growing data collections, as well as structured and unstructured data. It provides high scalability and availability by utilizing commodity hardware with minimal redundancy and fault tolerance. Hadoop prioritizes compute over data. The cluster's main nodes house the majority of the system's computational power and storage.

The efficiency of privacy-preserving bigdata security in the cloud via Hadoop has been improved utilizing several ways in this research. The first method encapsulates certain mechanisms with the MapReduce programming paradigm to explore and realize privacy preserving knowledge discovery from big data, as well as the framework known as IDEA with Hadoop framework. Optimal Support Vector Machine is the second method. The Hadoop framework was utilized to reduce repetitious encryption and decryption procedures for public and hybrid files using a classification method for privacy preservation in the cloud. The categorizing strategy shows a significant improvement.

The remainder of the article is organized as follows: the overview of the related work discussed in section 2 and the main objective of the proposed research summarized in section 3 and Design and Development of Enhanced Privacy Preserving Methods detailed in section 4 and performance evaluation elucidated in section 5 and finally the paper concludes with discussion and summary of future enhancement discussed in section 6.

2. Related Works

Zhao et al. [11] have introduced a safe authentication mechanism for users across many servers based on passwords. For user authentication, Elliptic Curve Cryptography (ECC) is presented, which handles two security attacks namely Offline Password Guessing and Impersonation Attack. Experimentations were carried out to examine the performance in terms of efficiency and security. The proposed plan utilizes four stages, the First stage is the Initialization stage where the RC focuses on the initialization of the framework as indicated by a security parameter. Second is the Registration Phase. In this stage, every system user Ui registers with the enrolment centre RC (the Registration Center) to get a smart card containing a secret key, as a qualification of Ui to demonstrate his/her realness to service provider (SP). Each SP Sj registers with RC to acquire a secret key, as a qualification of Sj to demonstrate its legitimateness to the system user. Next is the Authentication Phase. By running a confirmation strategy between a system user Ui and a service provider Sj, they can check the legitimacy of one another and set up a safe channel. That is, Sj guarantees that Ui is an enlisted client, and Ui accepts that the administration gave by Sj is lawful. The last one is the Password Update Phase, then a client Ui needs to refresh his/her unique secret key PWi. The suggested method is also suitable for real-time applications. ECC is an asymmetric encryption technique that creates smaller key pairs (public and private). This is one of the disadvantages, as well as the fact that the processing and communication costs are higher than with hashing approaches.

Win et al. [12] have conducted a security analytics service that runs on a cloud of virtualized equipment and stores data in HDFS A two-step machine learning approach was described in this study, which included logistic regression (to calculate the conditional probabilities of assaults through characteristics) and belief propagation (to estimate the belief in the presence of an attack).The utilization of logistic regression empowers the quick count of assault's contingent probabilities. All the more significantly, calculated relapse likewise empowers the retraining of the individual logistic regression classifiers utilizing the new assault includes as they are obtained from assault identification. The utilization of belief propagation ascertains the total conviction of an assault nearness by considering the contingent probabilities as for individual properties, which consequently accomplishes a comprehensive perspective on the visitor VM's behaviour.

Jun et al. [13] has proposed a big data investigation technique for smart grid systems based on security situational insights. To present a security investigation in smart grid, Reinforcement Learning, Game Theory, and a Fuzzy Cluster based Analytical System are merged. As input variables, real security values are given into the neural network. The smart grid can benefit from the extraction of network security scenario variables, network situational evaluation, and security situational forecasting. Security-related Big data, as well as security situational element data, are gathered from an electric power company's telecommunications network. This game has been played by both legitimate users and insider attackers, according to the game theory method. The use of game theory and deep learning methodologies adds complexity.

The Advance Encryption Standard (AES) is used to encode and decode data in a fully working engine. It has the ability to send data at a rapid rate in both encryption and decryption operations [14]. The information is encrypted three times using the Triple Data Encryption Standard (DES), which encrypts it three times [15]. Data is encrypted and decoded using the RC2 encryption technique [16]. In a large data context, the Hadoop security architecture is used to safeguard data. The installation of proper security procedures over vast volumes of data, as well as the clustering and classification processes, are all faults in these systems.

3. Objectives

The objective of the Research is to improve the performance of big data security and privacy preserving in cloud over hadoop frame work. The proposed methods have been developed for

- Decreasing data uploading time
- Improving the performance of Transmission.
- Reduce Error rate
- Enlarge data storage and increase data processing time
- Avoiding Unnecessary Transmission
- Increase data classification time
- High level data accuracy and throughput

4. Design and Development of Enhanced Privacy Preserving Methods

I. Privacy Preserving using Map Reduce based IDEA Algorithm and weighted Auto Encoder KNN classifier. This section discusses about the proposed methodology where information generated from the cloud source is clustered with the assistance of Incremental density based K-means clustering (IDK-means) algorithm. The time taken to process, encrypt and decrypt is comparatively minimum in proposed methodology.

II. Optimal Support Vector Machine Classification method for privacy preserving using density peak-based weighted FCM algorithm over enhanced word auto key encryption classifier. In this methodology deals with the evaluation of findings demonstrate the applicability and utility of the proposed integrated methodology for securing confidential information while being transmitted over several cloud nodes. The encrypted data is classified using deep learning approach.

III. An Intelligent Classification Method using K-Modes based Fast Mutation Artificial Bee Colony clustering technique over Advanced super encryption classifier.

In the proposed method the discusses the suggested strategy Advanced SET with Hadoop framework, and the KDD dataset is utilised to apply the methodology. The suggested technique of efficient convolution for privacy-preserving across vast data using map reduction notion in a cloud computing environment was tested using the census-income KDD data set.

5.Performance Evaluation

The comparison of results obtained by the proposed method with the existing approach are shown below. From the results the following are the summary ofthe findings in this research work.

(A) Decreasing data uploading time

Data uploading is one the main task of data transferring from one system to another memory through online.Table1 and Fig1 shows the proposed method takes to upload the minimum time of the given quantity of data. In the existing WAEK classifier with map reduce the data uploading time is 11.8 secs and decryption time is 17.4 (secs) and encryption time is 19(secs) and this method reduce the time delay for advanced super encryption technique with K-Modes based Fast Mutation Artificial Bee Colony clustering , time delay is reduced to (9.6 secs).

(B) **Improving Performance of transmission**

In this categorization technique to demonstrates a considerable enhancement and efficient in performance of big data security technique is assessed. The performance analysis shown in table1.

Performance Analysis	AES	Triple DES	RC2	Waek classifier with Map reduce	Enhanced Word key encryption with DPWFCM algorithm	Advanced Super Encryption Algorithm KK-Modes based FMABC
Data uploading time	40 (seconds)	35 (seconds)	38 (seconds)	11.8 (seconds)	10.3 (seconds)	9.6 (seconds)
Encryption time	39 (seconds)	8.337 (seconds)	9.4 (seconds)	19 (seconds)	20.1 (seconds)	24.1 (seconds)
Decryption time	0.12 (seconds)	0.23 (seconds)	0.20 (seconds)	17.4 (seconds)	18.6 (seconds)	23.7 (seconds)

Table 1: Overall comparison of existing and the proposed methods of data uploading time

(C) **Reduce Error rate**

In this proposed method to reduce error rate and increase data accuracy by uploading of data processing in the convolution process. Advanced super encryption technique used to upgrade the level of data security and minimize the error rate.

(D)Enlarge data storage and increase data processing time

In the existing enhanced word auto key encryption classifier with density peak-based weighted FCM algorithm using the data is 2000 mb of data processing time (434 secs) and this method to increase for advanced super encryption technique with K-Modes based Fast Mutation Artificial Bee Colony clustering , data processing time is increased to (486 secs). In the proposed method to take maximum duration of processing time and accurate result compare than existing methodology to show the table 2.

File size (MB)	100	300	500	1000	2000
GMPLS/MPLS networks	58	76	100	135	190
IDEA with hadoop framework	17.04	51.12	85.2	170.4	340.8
Enhanced WAKE	18.6	53	86	272	434
Advanced SET	21.7	61.4	93.3	311	486

Table 2: overall comparison of data processing time

(E)Avoid unnecessary Transmissions:

In this classification of approach is identified by the unnecessary transmission of the Encrypted data through convolution process and classification after decrypt the data properly.

(F)High data classification time

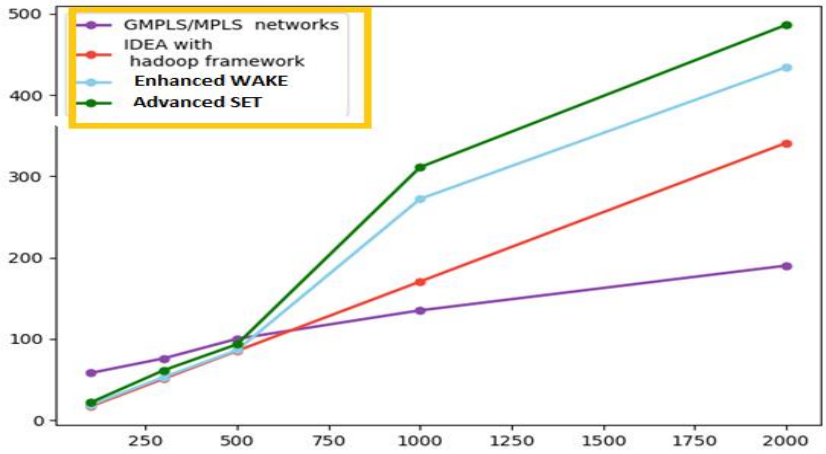
Data classification time is identified that the time consumption of the proposed methodology is higher than when compared with existing approaches AES, Triple AES is given in table 3. Because of display the accurate results of data accuracy and data security performance is much very high.

(G)High level data accuracy and Throughput

The effectiveness of the proposed classification approach is identified by the accuracy of the classification. To increase the data accuracy and throughput of efficient data encryption and decryption process performed very efficiently compare than other

existing techniques. In this proposed methodology to process change volume of data transmission to take maximum time and show the accuracy of results display in table 4.

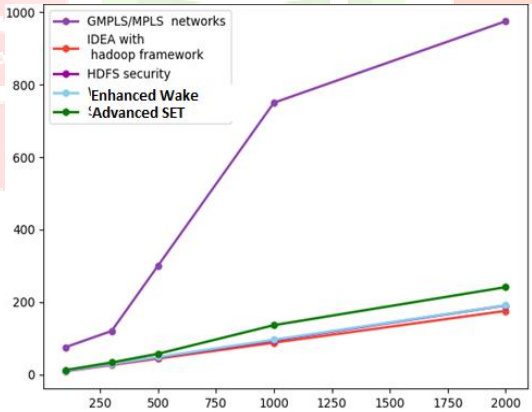
Comparison of Data Processing Time



Comparison of Data Processing Time

File size(MB)	100	300	500	1000	2000
GMPLS/MPLS networks	75	120	300	750	975
IDEA with hadoop framework	8.77	26.31	43.85	87.7	175.4
HDFS security	9.4	28.2	47	94	190.5
Enhanced WAKE	10	29	48	96	191
Advanced SET	12	33	57	136	241

Table 3 Comparison of Data Classification Time



Comparison of Data classification Time

Overall accuracy

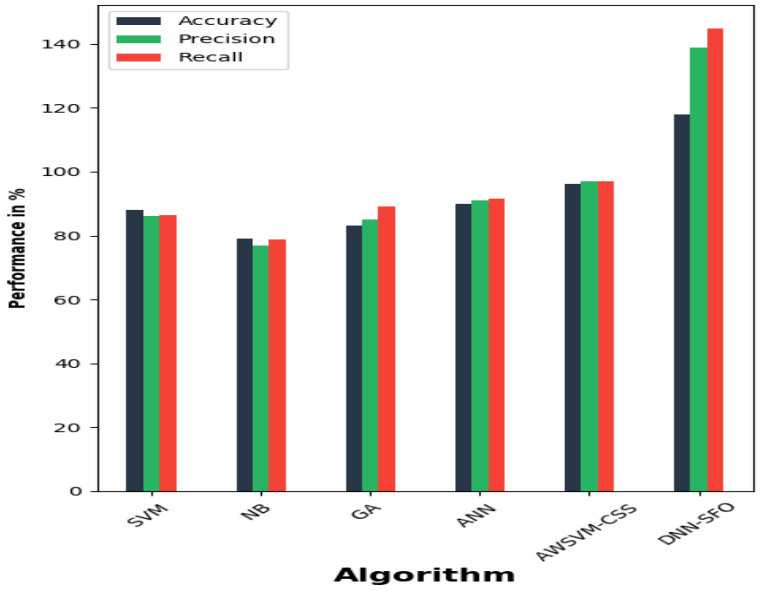


Table 4- Overall accuracy

6. Conclusion and Future Enhancement

This research has presented efficient techniques to improve the privacy preservation of big data security using Hadoop in a cloud environment to boost throughput and leverage clustered data of map reduce algorithms to eliminate time delays. An effective privacy-preserving solution was necessary to deal with the huge amount of data created and its security consequences. If the encrypted data is not sufficiently encrypted, the data will be run through the convolution process again. When the suggested framework's experimental findings are compared to existing techniques, the big data-based privacy-preserving methodology wins. Using the proposed methodology, data generated in the cloud is encrypted and decrypted. The proposed methodology takes a relatively short amount of time to process, encrypt, and decode data. In the future this study has proposed effective methods for Artificial intelligence and machine learning techniques could be added to the approach in the future. These files, however, require special treatment because their nature differs from that of txt, csv, log, xls, and sql files. As a result, new techniques will be required to deal with file attributes for such large data sets.

References

[1]. Roy, S., Shovon, A. R., & Whaiduzzaman, M. (2017, December). Combined approach of tokenization and mining to secure and optimize big data in cloud storage. In 2017 IEEE Region 10 Humanitarian Technology Conference (R1s0-HTC) (pp. 83-88). IEEE.

[2]. Wu, J., Dong, M., Ota, K., Li, J., & Guan, Z. (2018). Big data analysis-based secure cluster management for optimized control plane in software-defined networks. IEEE Transactions on Network and Service Management, 15(1), 27-38.

[3]. Ji, C., Li, Y., Qiu, W., Jin, Y., Xu, Y., Awada, U., .&Qu, W. (2012). Big data processing: Big challenges and opportunities. Journal of Interconnection Networks, 13(03n04), 1250009.

[4]. Chaudhary, R., Aujla, G. S., Kumar, N., & Rodrigues, J. J. (2018). Optimized big data management across multi-cloud data centers: Software-defined-network-based analysis. IEEE Communications Magazine, 56(2), 118-126.

[5]. Li, Z., Xu, W., Shi, H., Zhang, Y., & Yan, Y. (2021). Security and Privacy Risk Assessment of Energy Big Data in Cloud Environment. Computational Intelligence and Neuroscience, 2021.

[6]. Yang, C., Huang, Q., Li, Z., Liu, K., & Hu, F. (2017). Big Data and cloud computing: innovation opportunities and challenges. International Journal of Digital Earth, 10(1), 13-53.

[7]. Gupta, B. B., Yamaguchi, S., & Agrawal, D. P. (2018). Advances in security and privacy of multimedia big data in mobile and cloud computing. Multimedia Tools and Applications, 77(7), 9203-9208.

[8]. Zhang, L., Wu, C., Li, Z., Guo, C., Chen, M., & Lau, F. C. (2013). Moving big data to the cloud: An online cost-minimizing approach. IEEE Journal on Selected Areas in Communications, 31(12), 2710-2721.

[9]. Subramanian, E. K., &Tamilselvan, L. (2020). Elliptic curve Diffie–Hellman cryptosystem in big data cloud security. Cluster Computing, 23(4), 3057-3067.

[10]. Guo, L., &Qiu, J. (2018). Optimization technology in cloud manufacturing. The International Journal of Advanced Manufacturing Technology, 97(1), 1181-1193.

[11]. Zhao, Y., Li, S., & Jiang, L. (2018). Secure and Efficient User AuthenticationScheme Based on Password and Smart Card for Multiserver Environment. Securityand Communication Networks, 2018, 1–13. doi:10.1155/2018/9178941

[12]. Win, T. Y., Tianfield, H., &Mair, Q. (2018). Big Data Based SecurityAnalytics for Protecting Virtualized Infrastructures in Cloud Computing. IEEETransactions on Big Data, 4(1), 11–25. doi:10.1109/tbdata.2017.2715335

[13]. Jun Wu, Kaoru Ota, Mianxiong Dong, Jianhua Li, Hongkai Wang (2018),Big Data Analysis-based Security Situational Awareness for Smart Grid, IEEETransactions on Big Data, Vol. 4, Issue.3, PP. 408-417.

[14]. Lu, C. C., & Tseng, S. Y. (2002, July). Integrated design of AES (Advanced Encryption Standard) encrypter and decrypter. In Proceedings IEEE International Conference on Application-Specific Systems, Architectures, and Processors (pp. 277-285). IEEE.

[15]. Del Rosal, E., & Kumar, S. (2017). A fast FPGA implementation for triple DES encryption scheme. Circuits and Systems, 8(09), 237.

[16]. AbdElminaam, D. S., Abdual-Kader, H. M., &Hadhoud, M. M. (2010). Evaluating The Performance of Symmetric Encryption Algorithms. Int. J. Netw. Secur., 10(3), 216-222.
[17]. https://www.unb.ca/cic/datasets/nsl.html
[18]. Suthaharan, S. (2016). Support vector machine. In Machine learning models and algorithms for big data classification (pp. 207-235). Springer, Boston, MA.
[19]. Jiang, L., Zhang, L., Li, C., & Wu, J. (2018). A correlation-based feature weighting filter for Naive Bayes. IEEE transactions on knowledge and data engineering, 31(2), 201-213.
[20]. Mirjalili, S. (2019). Genetic algorithm. In Evolutionary algorithms and neural networks (pp. 43-55). Springer, Cham.
[21]. Walczak, S. (2018). Artificial neural networks. In Encyclopedia of Information Science and Technology, Fourth Edition (pp. 120-131). IGI Global.