DOI: https://doi.org/10.46632/eae/1/1/5

Electrical and Automation Engineering
Vol: 1(1), 2022
REST Publisher; ISBN: 978-81-956353-5-1
Website: http://restpublisher.com/book-series/electrical-and-automation-engineering/

An Efficient Secure Data Deduplication and Portability in Distributed Cloud Server Using Whirlpool-Hct and Lf-Wdo

***A R Athira, P. Sasikala**
Vinayaka Mission’s KirupanandaVariyar Engineering College, Salem, Tamil Nadu, India.
*Corresponding author Email: parvathyretnakaran@gmail.com

Abstract. Distributed Cloud Computing Storage has come up as a service that can expedite data owners (DO) to store their data remotely according to their application or data or file environment. However, insecure data storage, high uploading bandwidth, integration issues of DCS has breached the trustworthiness of the user to store data. In order to conquer the challenge, the work has developed a data Deduplication and portability-based secure data storage in DCS. The work aids to remove unwanted data and selects the most relevant features to avoid data loss by using GK-QDA Feature Reduction Method and HFG feature selection method. The selected cloud server for the respective data or application is analyzed for redundant data by data duplication using a whirlpool hashing algorithm followed by a hash chaining algorithm. Finally, to minimize the integration issues while moving the encrypted data between the DCS, the work has developed an LF-WDO technique. An experimental analysis has showed an enormous result by achieving a computation time of 2987 ms as compared to the existing methods.

Keywords: Gaussian Kernel –Quadratic Discriminate Analysis (GK-QDA), hybrid forest genetic algorithm (HFG), Levy Flight – Wind Driven Optimization Algorithm (LF-WDO)

1. Introduction

As one of the momentous technology, the distributed storage used in cloud computing has enabled aggregate remote data storage. Hulk and ascendable cloud-based storage is provided for the users from the cloud vendors. [1, 2, 3]. However, the security affairs are still an obstacle for enterprises that caused by the operations on cloud side [4, 5]. Recently many works have been done related to distributed cloud storage (DCS) such as Mass Distributed Storage (MDS) [6, 7] using a Fully Homomorphic Encryption (FHE) and ABE as a security policy, Security-Aware Efficient Distributed Storage (SA-EDS) etc, [8, 9]. Even though many techniques have been developed, but still there remains to be a challenge of storing the data securely. The distributed cloud storage peculiarities result in more liability during data transmissions by malignant interventions or abuse activities [10]. After all the risks deriving from different network layers are somewhat fully addressed, therefore, it is a confront obstacle to efficiently secure distributed data in cloud systems. This work has proposed a data Deduplication and portability based protected data storage in distributed cloud computing (DCC) to provide a secure data storage in distributed cloud computing [11]. The carry-over of the paper is trace as follows: Section 2 reviews and discusses the related works based on secure data storage in DCC, Section 3 describes the proposed methodology, the speculative analysis of the proposed methodology is performed in section 4, and finally, Section 5 concludes the proposed method with future scope.

2. Literature Survey

Yibin Li et al. [12] developed a Security-Aware Efficient Distributed Storage (SA-EDS) model, which was mainly supported by the developed algorithms that included Alternative Data Distribution (AD2) Algorithm, Secure Efficient Data Distributions (SED2) Algorithm and Efficient Data Conflation (EDCon) Algorithm. Even though the approach provided with secure data storage but when the selection of cloud server was inaccurate. Esther Daniel et al. [13] presented an integrity verification and Deduplication of outsourced data with a lightweight auditing method. The developed scheme combined hashing and symmetric encryption with a renovated distributed hash table data structure, which enabled dynamic operations of the data efficient, also shortened the communication and computation struggle for integrity verification. The encryption scheme was not that secure to overcome external malicious attacks. Nabeil Eltayieba et al. [14] developed to provide secure data sharing accompanying the concept of blockchain with attribute-based signcryption in the cloud environment. The strategy satisfied the security obligations such as confidentiality and enforceability, of cloud computing. Further, by its nature wrong results returned as in the traditional cloud server this smart indenture solved the problem of cloud storage. But the scheme was highly complex and was inaccurate. Gagangeet Singh Aujla et al. [15] advanced secure storage, verification, and auditing (SecSVA) of big data in a cloud environment based on Kerberos-based identity verification and authentication, and Merkle hash-tree-based trusted third-party auditing on the cloud. Even though the approach enhanced with data Deduplication but portability issue leaded to slow down of the approach.

3. Proposed Secured Data Storage in Distributed Cloud Computing

In cloud storage, the data Deduplication method is used to reduce the upload bandwidth and storage zone by removing the data clones from the cloud service provider (CSP) but data Deduplication is a challenge in DCS. To overcome such challenges, we proposed a secure data Deduplication system and portability with distributed cloud server as shown in figure 1.

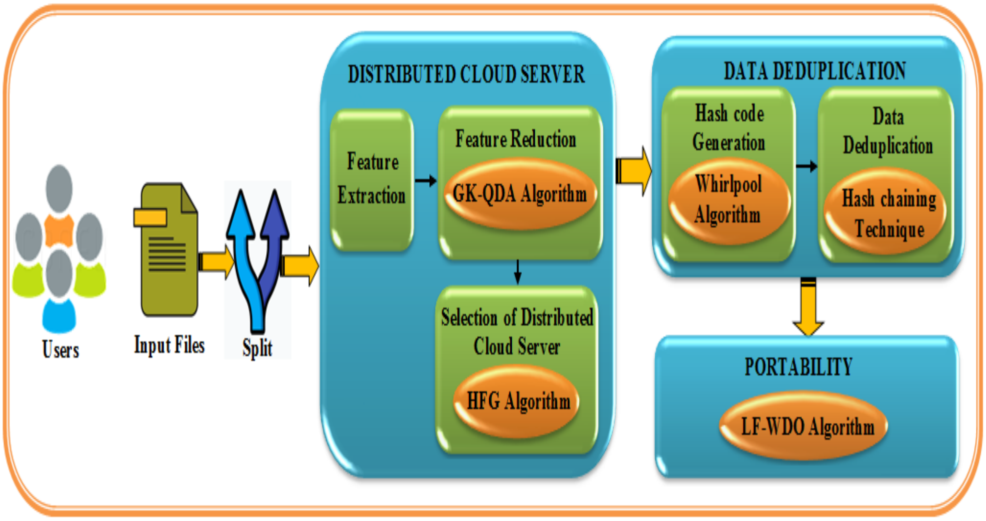

FIGURE 1. Proposed secure data storage in DCS

4. Distributed Cloud Server

The distributed cloud server provides storing the data and retrieving it at the time of use. But storing and retrieving is a difficult task in distributed cloud computing due to mass data, irrelevant data, complex data, etc. The work has developed a Gaussian Kernel -QDA and Forest Genetic Algorithm for reduction of irrelevant features from extracted features and selection of distributed cloud servers. Feature extraction The initial phase of the proposed work that extricate the various features from file and distributed cloud server. The feature like Task Cost, Speed, Weight, etc. parameters is extracted from the splitted file. In the same way, the features or accessories information like CPU resources, memory resources, and storage resource Processing speed and cycle are extracted from the distributed cloud server. Feature Reduction Feature Reduction contributes towards reducing the unwanted features from the extracted features of both file as well as distributed cloud server. As the existing feature reduction technique led to some amount of data loss; to conquer this problem, the work has developed a Gaussian Kernel -QDA Algorithm.

$$\Phi_{\kappa}(\Re) = \log \pi_{\kappa} - \frac{1}{2}\varphi_{\kappa}^{t} \sum_{\kappa} \varphi_{\kappa} + \Re^{T} \sum_{\kappa} \varphi_{\kappa} - \frac{1}{2}\Re_{\kappa}^{t} \sum_{\kappa} \Re - \frac{1}{2}\log\left|\sum_{k}\right| \tag{1}$$

$$\forall(\Re) = \sqrt{\frac{a}{\pi}.e^{-\omega.\Re^{2}}} \tag{2}$$

Where φ_{κ} and $\sum_{k}$ denotes the mean and covariance matrix, $\Phi_{\kappa}(\Re)$ is the given input features, $\forall(\Re)$ is the Gaussian kernel function that converts the input features into correlation matrix. The techniques obtain the reduce feature:

$$\Im_{r}^{*} = [\Re_{1}, \Re_{2}, \Re_{3}, \Re_{4}, ...\Re_{n}] \tag{3}$$

5. Selection of Distributed Cloud Server

The proposed work is proffered by the feature selection process to tap the cloud server that is by selecting most of the paramount features from the condensed features to reduce the model accuracy, eliminate the insignificant features that may also lead to more computational time. To ameliorate the extracted feature, the work has developed a Hybrid Forest Genetic Algorithm (HFG) stated below:

Step1: Initialize the forest or file features ($\Re_{t} \in Rand$) with the random trees or cloud server ($\Re_{T}$) which consists of $(D+1)$ dimensions vector of feature $\Im$. Initially, the age of the trees is set to zero ($\Re_{T}^{age} = 0$), thereafter, the local solution $\Re_{T} = LSC(\Re_{t})$ is evaluated within the range of $\Re_{t} \in \Gamma\{0,1\}$

Step2: For evaluating the local solution crossover $\Re_t \to \lambda_c[\Re_1, \Re_2, \Re_3, \Re_4, ... \Re_n]$and mutation is performed on the trees and the mutated value is selected $\lambda_m(r_t) \to [r_1, r_2, r_3, r_4, ... r_n]$. Thereafter increment the Tree age by 1.

Step3: Now global seeding is done for the selected tree and random selection of tree is done using GSC parameters $(\Re_T = GSC(r_t))$. Then, the value of each selected variable will be negated (changing from 0 to 1 or vice versa). Based on the trees global search is done over the global space and best tree is updated. Thus, End of the iteration after obtaining a best tree (cloud server) as a substrate of the best selected features (file features) $\Im_s = [r_1^*, r_2^*, r_3^*, r_4^*, r_5^*, ...r_n^*]$.

6. Data Deduplication

Data Deduplication in Cloud server is generated by the hash code for the appropriate split file using Whirlpool Hashing Algorithm. Next, the cloud server checks the hash value availability using the Hash Chaining Technique in the distributed cloud server. If it is available, then the cloud server refers to the stored file location path. If it is unavailable, then the cloud server performs the compression and encryption to store the data. Initially, the selected file features are converted into the hash function using whirlpool given by the file $[r_1^*, r_2^*, r_3^*, r_4^*, r_5^*, ...r_n^*]$, the whirlpool hash function is given by:

$$\begin{aligned} &\lambda_m(r_t) \to [r_1, r_2, r_3, r_4, ... r_n] \\ &\mathrm{H}_0' = [r_1^*, r_2^*, r_3^*, r_4^*, r_5^*, ...r_n^*] \end{aligned} \tag{4}$$

$$\mathrm{H}_i' = \Psi(\mathrm{H}_{i-1}', r_1^*) \oplus \mathrm{H}_{i-1}' \oplus r_1^* = \mathrm{int}\ ermediatevalue \tag{5}$$

$$\mathrm{H}_s' = f(\mathrm{H}_{i-1}', r_{i-1}), \quad 1 \le i \le L \tag{6}$$

Where, H_0' is the function that is going o generate hash value H_i' is the hash value generating for the given features, Ψ is the constant function, H_{i-1}', r_1^* denotes the previous hash value chaining with the blocks r_1^* and at last $\mathrm{H}_i' = \mathrm{H}_s'$ is checked whether the same hash value is obtained or not. portability Cloud application portability provides an ability to move encrypted data between cloud servers with a minimum level of integration issues with a particular time interval to improve the security of the encrypted data. This phase was done by using Levy Flight – Wind Driven Optimization Algorithm. To replace the selection of random position and velocity, the levy flight technique will be replaced in the WDO algorithm.

Step 1: Initially the feature follows Newton's second law of motion based on temperature and air pressure. Now, based on the moving air force constant, the equation is formulated, such as pressure gradient force $\Omega_{pg} = -\nabla \mathrm{P} \delta v$, gravitational force $\Omega_g = r\delta v\bar{a}$, Coriolis force $\Omega_c = -2\aleph \times \bar{u}$, friction force $\Omega_f = -r\alpha\bar{u}$.

Step 2: Based on the equation all the forces are summed together and equated in equation 7 thereafter velocity and position of the air parcel is obtained using equation 8-10:

$$r\bar{u}\Delta t = (r\delta v\bar{a}) + (-\nabla \mathrm{P}\delta v) + (-r\alpha\bar{u}) + (-2\aleph \times \bar{u}) \tag{7}$$

$$\bar{u}_{new} = ((1-\alpha)\bar{u}_{old}) + a(-r_{old}) + \left[\left|\frac{\mathrm{P}_{\max}}{\mathrm{P}_{old}} - 1\right| \Omega \mathrm{T}(r_{\max} - r_{old})\right] + \left[\frac{-\ell u_{old}^{other\dim}}{\mathrm{P}_{old}}\right] \tag{8}$$

$$r_{new} = r_{old} + (\bar{u}_{new} \times \hbar_{lf}) \tag{9}$$

$$\hbar_{lf}[r_{new}(k)]r_{old} = \hbar_{lf}\left\lfloor \exp(-r_{new}|k|^{\beta}) \right\rfloor r_{old} \tag{10}$$

Where, $\bar{u}_{new}$ denotes the updated air velocity which depends upon the current air velocity $\bar{u}_{old}$, α r_{old} is the current search space features of the file with allocated cloud server, $\mathrm{P}_{\max}$ and P_{old} states the maximum pressure and pressure at the current location , Ω T and ℓ is the constants, $r_{\max}$ states the cloud server facing any problem for storing the file , $\hbar_{lf}$ represents the levy flight distribution for updating the time step , $[r_{new}(k)]$ is the probabilities of step addition of the random variables and $\beta \in 0,2$ .thus from the above technique the files are allocated to new cloud server(r_{new}) due to some issues in the existing servers(r_{old}).

7. Results and Discussion

The projected framework is validated based on various metrics along with various existing algorithms in order to observe the efficiency of the framework towards secure data storage in DCS. Performance analysis of the proposed FGA based on various metrics The analysis elaborates the evaluation based on waiting time, process time, response time, and turnaround time for the proposed FGA along with the existing horse optimization algorithm (HOA), Lion optimization algorithm (LOA), Genetic Algorithm (GA), Forest optimization algorithm (FOA). The evaluation is exhibited in table 1.

TABLE1 1. Evaluation of proposed FGA for different metrics

Metrics/techniques	Waiting Time	Process Time	Response Time	Turn Around Time
HOA	3452	3214	6754	6457
LOA	3124	3020	6186	6124
GA	2865	2651	5214	5647
FOA	2345	2124	4421	5214
Proposed FGA	2143	2005	4002	4876

Table 1 states that the scheduled FGA tends to achieve a waiting time of 2143 ms, the process time of 2005 ms, the response time of 4002 ms, and Turnaround time of 4876ms, which is faster and has reduced computational complexity as compared to the existing methods, which achieve a metrics value ranging between 2124 ms-6754ms.Performance analysis of the proposed whirlpool-HCT based on Hash code generation time The proposed whirlpool-HCT is analyzed based on the time taken for generating the hash code. The graphical analysis of the proposed whirlpool-HCT vs. Hash code generation time is exemplified in figure 2

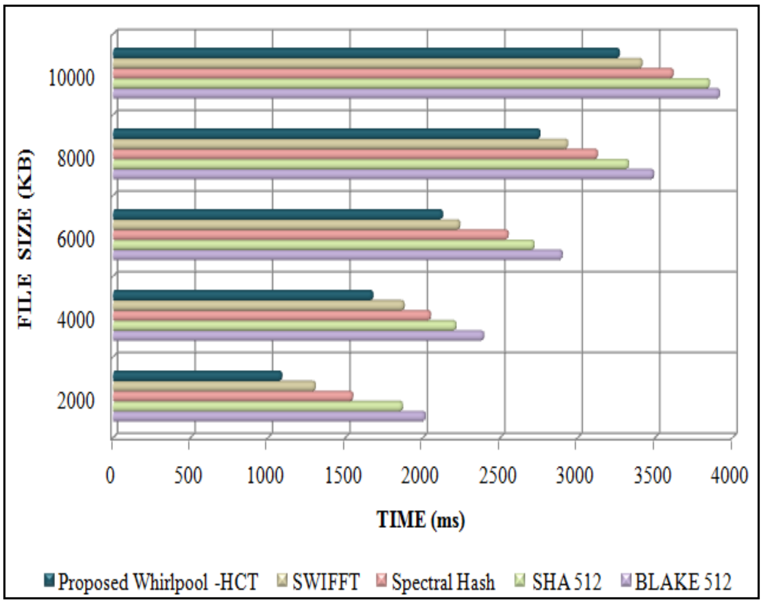

FIGURE 2. Graphical demonstration of proposed whirlpool-HCT based on hash code generation

From figure 2, it can be exemplified that the proposed Whirlpool-HCT algorithm takes an average time of 2180 ms for a file size of average 6000kb to generate a hash code, whereas the existing BLAKE512, SHA512, Spectral hash, and SWIFFT tends to achieve an average time of 2939ms, 6847ms, 2791ms, and 2351ms to generate a hash code for an average file size of 600kb. Thus, the proposed methods tend to consume less time as well as remain to be more secure to store data as compared to the existing methods. Performance analysis of the proposed LF-WDO based on Computation time Based on computation time, the proposed LF-WDO is validated along with the existing methods, such as Spider Monkey Optimization Algorithm (SMOA), Grey Wolf Optimization Algorithm (GWOA), Brownian Motion Bat Optimization Algorithm (BMBAT), and Wind Driven Optimization Algorithm (WDO). The graphical analysis of the proposed whirlpool-HCT vs. Hash code generation time is illustrated in figure 3.

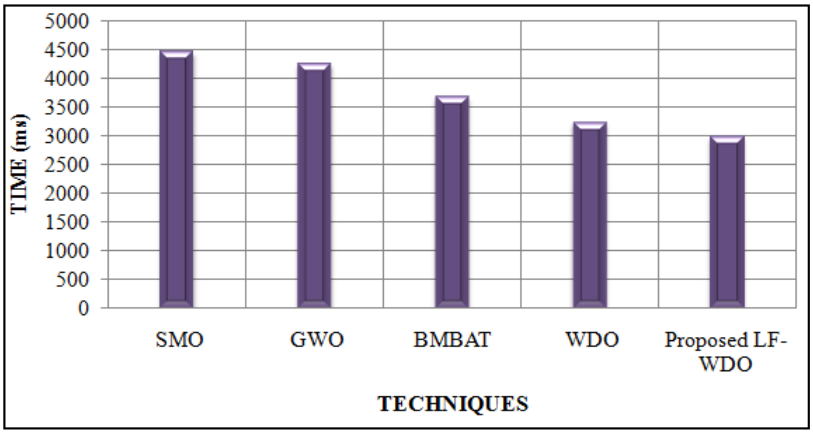

FIGURE 3. Graphical demonstration of proposed LF-WDO based on computation time

From figure 3, it can be illustrated that the proposed LF-WDO algorithm takes a Computation time of 2987ms to process the entire portability of CS, whereas the existing SMOA, GWOA, BMBAT, and WDO tends to achieve a computation time of

4468ms, 4257ms, 3687ms, and 3241ms, which comparatively consumes more time than the proposed algorithm. Thus, the proposed LF-WDO tends to be robust and secure to solve the integration issues faced between encrypted data and DC

8. Conclusion

The paper has developed a data Deduplication and portability-based secure data storage in distributed cloud computing. The effort has mainly concentrated on data Deduplication and portability of data in order to reduce the upload bandwidth, storage space, malicious attacks, integration issues, etc. The proposed methods reduce the irrelevant data of file or DCS and select the most appropriate features needed for allocating storage in DCS Using GK-QDA Algorithm and Forest Genetic Algorithm. Thereafter, copied data from CSP is removed using hash code generation and data duplication using the whirlpool algorithm following the hash chaining algorithm. Finally, for avoiding the portability issues, the work has developed an LF-WDO. Experimental analysis has achieved an better outcome while considering response time of 4002 ms and computation time of 2987 MS as correlated to existing expedient methods.

References

[1]. MahdiGhafoorian, DariushAbbasinezhad-Mood, and Hassan Shakeri, “A thorough trust and reputation based RBAC model for secure data storage in the cloud”, IEEE Transactions on Parallel and Distributed Systems, vol. 30, no. 4, pp. 778-788, 2018.

[2]. MuhammadUsman, Mian Ahmad Jan, and Xiangjian He, “Cryptography-based secure data storage and sharing using HEVC and public clouds”, Information Sciences, vol. 387, pp. 90-102,2017,10.1016/j.ins.2016.08.059.

[3]. Wei Liang, Yongkai Fan, Kuan-Ching Li, Dafang Zhang, and Jean-Luc Gaudiot, “Secure data storage and recovery in industrial blockchain network environments”, IEEE Transactions on Industrial Informatics, vol. 16, no. 10, pp. 6543-6552,2020.

[4]. Qinlong Huang, Yixian Yang, and MansuoShen, “Secure and efficient data collaboration with hierarchical attribute-based encryption in cloud computing”, Future Generation Computer Systems, vol. 72, pp. 239-249,2017,10.1016/j.future.2016.09.021.

[5]. Yongkai Fan, Xiaodong Lin, Gang Tan, Yuqing Zhang, Wei Dong, and Jing Lei, “One secure data integrity verification scheme for cloud storage”, Future Generation Computer Systems, vol. 96, pp. 376-385,2019,10.1016/j.future.2019.01.054.

[6]. Wei Zhang,Xiaohui Chen, Yueqi Liu, and Qian Xi, “A distributed storage and computation k-nearest neighbor algorithm based cloud-edge computing for cyber-physical-social systems”, IEEE Access, vol. 8, pp. 50118-50130,2020,10.1109/ACCESS.2020.2974764.

[7]. Mingzhe Wang, and Qiuliang Zhang, “Optimized data storage algorithm of IoT based on cloud computing in distributed system”, Computer Communications, vol. 157, pp. 124-131,2020,10.1016/j.comcom.2020.04.023.

[8]. RayapatiVenkataSudhakar and ChMalleswaraRaoT, “Security aware index based quasi–identifier approach for privacy preservation of data sets for cloud applications”, Cluster Computing, pp. 1-11,2020,10.1007/s10586-019-03028-7.

[9]. SattarFeizollahibaroughand MehrdadAshtiani, “A security-aware virtual machine placement in the cloud using hesitant fuzzy decision-making processes”, The Journal of Supercomputing, pp. 1-31,2020.

[10]. HamedTabrizchi, and MarjanKuchaki Rafsanjani, “A survey on security challenges in cloud computing: issues, threats, and solutions”, The journal of supercomputing, vol. 76, no. 12, pp. 9493-9532,2020.

[11]. Bijeta Seth, SurjeetDalal, VivekJaglan, Dac-Nhuong Le, Senthilkumar Mohan, and GautamSrivastava, “Integrating encryption techniques for secure data storage in the cloud”, Transactions on Emerging Telecommunications Technologies, pp. e4108,2020,10.1002/ett.4108.

[12]. Yibin Li, KekeGai, LongfeiQiu, MeikangQiu, and Hui Zhao, “Intelligent cryptography approach for secure distributed big data storage in cloud computing”, Information Sciences, vol. 387, pp. 103-115,2017.

[13]. Esther Daniel and VasanthiN. A, “LDAP: A lightweight deduplication and auditing protocol for secure data storage in cloud environment”, Cluster Computing, vol. 22, no. 1, pp. 1247-1258,2019.

[14]. NabeilEltayieb, RashadElhabob, Alzubair Hassan, and Fagen Li, “A blockchain-based attribute-based signcryption scheme to secure data sharing in the cloud”, Journal of Systems Architecture, vol. 102, pp. 101653,2020,10.1016/j.sysarc.2019.101653.

[15]. Gagangeet SinghAujla, RajatChaudhary, Neeraj Kumar, Ashok Kumar Das, and Joel JPC Rodrigues, “SecSVA: secure storage, verification, and auditing of big data in the cloud environment”, IEEE Communications Magazine, vol. 56, no. 1, pp. 78-85,2018.

[16]. Chinnasamy, Sathiyaraj, M. Ramachandran, Kurinjimalar Ramu, and P. Anusuya. "Study on Fuzzy ELECTRE Method with Various Methodologies." REST Journal on Emerging trends in Modelling and Manufacturing 7, no. 4 (2021).