Computers, Materials & Continua
DOI:10.32604/cmc.2021.013614
Article

Tech Science Press

High Security for De-Duplicated Big Data Using Optimal SIMON Cipher

A. Muthumari[1], J. Banumathi[2], S. Rajasekaran[3], P. Vijayakarthik[4], K. Shankar[5], Irina V. Pustokhina[6] and Denis A. Pustokhin[7,*]

[1]Department of Computer Science and Engineering, University College of Engineering, Ramanathapuram, 623513, India
[2]Department of Information Technology, University College of Engineering, Nagercoil, 629004, India
[3]Department of EEE, PSN College of Engineering (Anna University), Tirunelveli, 627152, India
[4]Department of Information Science and Engineering, Sir M. Visvesvaraya Institute of Technology, Bangalore, 56215, India
[5]Department of Computer Applications, Alagappa University, Karaikudi, 63003, India
[6]Department of Entrepreneurship and Logistics, Plekhanov Russian University of Economics, Moscow, 117997, Russia
[7]Department of Logistics, State University of Management, Moscow, 109542, Russia
[*]Corresponding Author: Denis A. Pustokhin. Email: dpustokhin@yandex.ru
Received: 30 August 2020; Accepted: 14 November 2020

Abstract: Cloud computing offers internet location-based affordable, scalable, and independent services. Cloud computing is a promising and a cost-effective approach that supports big data analytics and advanced applications in the event of forced business continuity events, for instance, pandemic situations. To handle massive information, clusters of servers are required to assist the equipment which enables streamlining the widespread quantity of data, with elevated velocity and modified configurations. Data deduplication model enables cloud users to efficiently manage their cloud storage space by getting rid of redundant data stored in the server. Data deduplication also saves network bandwidth. In this paper, a new cloud-based big data security technique utilizing dual encryption is proposed. The clustering model is utilized to analyze the Deduplication process hash function. Multi kernel Fuzzy C means (MKFCM) was used which helps cluster the data stored in cloud, on the basis of confidence data encryption procedure. The confidence finest data is implemented in homomorphic encryption data wherein the Optimal SIMON Cipher (OSC) technique is used. This security process involving dual encryption with the optimization model develops the productivity mechanism. In this paper, the excellence of the technique was confirmed by comparing the proposed technique with other encryption and clustering techniques. The results proved that the proposed technique achieved maximum accuracy and minimum encryption time.

Keywords: Cloud security; deduplication; clustering; optimization; big data; dual encryption

1 Introduction

Cloud computing has become a challenging platform nowadays which raises a considerable impact on the Information Technology (IT) industry and community activities [1]. Cloud computing applications provide services in addition to internet, and enable programming datacenters to render those services [2]. Moreover, cloud computing is a colossal system of computers combined together as a single computer. The size of cloud computing is developing and increasing in a consistent manner over the years. In any case, cloud computing possesses several advantages such as cost savings, economically efficient operations, improved collaboration, increase in scale and readiness by empowering worldwide computing representation and internet infrastructure [3]. There are different kinds of cloud storages based on accessibility, funds, scalability, and so on. Global nations have begun transitioning important information to cloud storage space owing to their improved security. With increasing patterns observed in data age rates, it is a monotonous errand that is intended for the storage of cloud suppliers to provide well-organized capacity [4]. Cloud computing is a promising and a cost-effective approach that supports big data analytics and advanced applications in the event of forced business continuity events, for instance, pandemic situations. To handle massive information, clusters of servers are required to assist the equipment which enables streamlining the widespread quantity of data, with elevated velocity and modified configurations. Security is the major challenge in the concept of cloud-based big data analytics [5].

Cloud data safety should be guaranteed and the readiness of servers to get upgraded for huge storage zones in an organization is important, especially using data encryption techniques. However, a major problem is that servers cannot store important information in reduplication technique when there is excess of encrypted data [6]. In this regard, the use of Deduplication process securely completes data encryption in cloud. This is a benchmark task to be accomplished in cloud. Data deduplication technique is modernized as a straightforward storage optimization technique in secondary servers [7]. Google Drive, Amazon S3 and Dropbox, are different storage suppliers who utilize data deduplication process. To accomplish upgradability, encryption is performed to save general information towards secure deduplication [8]. Advanced capacity and guaranteed security are the requirements of encryption and deduplication techniques [9,10].

In this paper, cloud big data security is proposed using dual encryption technique. The study uses clustering model to analyze the deduplication process hash function. Multi kernel Fuzzy C Means (MKFCM) helps cluster data, which is then stored in cloud on the basis of confidence data encryption procedure. Then, the confidence finest data is implemented in homomorphic encryption data by following SIMON Cipher (OSC) technique. This secure dual encryption technique combining the optimization model enhances the presentation. Finally, the findings of the work are compared with other encryption and clustering techniques.

2 Literature Review

Venila et al. [11] studied Map-Reduce structure for safeguarding big data security in cloud. Some methods utilize restricted footage anonymization to support privacy. Information is processed in alignment with examination, sharing, and mining. The review article focused on global footage anonymization to secure data privacy over Big Data using Map Reduce under cloud. This strategy replaced the individual detectors of composed data sensor, by confusing the values prior to saving it as a recognized storage.

Celcia et al. [12] analyzed multiple intermediate datasets which are utilized to capture the engendering of private and delicate data among other data sets. Privacy leakage of these

intermediate data sets was measured. By considering the importance of data holder's privacy, the upper bound constraint of the privacy leakage was determined. This research work handled big datasets. The authors encrypted the data as the study exposed the cost of safeguarding privacy in cloud, if privacy is severely breached.

Sookhak et al. [13] examined Remote Data Auditing (RDA) techniques. The local data storage capacity reduced and the data was stored in cloud servers. For cloud storage mechanism, an efficient RDA method is required as per algebraic properties with communication and low computational costs. Furthermore, Divide-and-Conquer Table (DCT) can hold dynamic data processes including deletion, alterations, embedding and addition of the data. The scale data storage expense connects the proposed data structure with low computational cost.

Miao et al. [14] proposed the most proficient method to outline Deduplication mechanism that opposes savage power assault. The method minimized 'put stock' in supposition of Key Server (KS), a multiple-key server mechanism. This mechanism was introduced according to the basic principles of sightless signature threshold. The study conducted upon security demonstrated that the proposed scheme was secure with respect to future safety display as the scheme was able to oppose the beast force attack, regardless of the possibility that a set key server's quantity may get tainted. On the off chance, a secret input could be spilled upon stipulation which results in the corruption of all K-CSPs. Along these lines, the system is able to give an additional grounded security and it can efficiently oppose a disconnected animal force assault.

A message-locked encryption model, termed as block-level message-locked encryption, was proposed by few researchers. The excess of metadata storage space is block-labeled from encapsulated block keys. For block label comparison, a huge computation was performed on the overhead by BL-MLE [15]. The content-characterized chunking and bloom filter technologies incorporate a secure, comparability-based data deduplication method. This method was able to reduce computation overhead. In the presence of comparative documents, deduplication operations were performed that can minimize the computation overhead. While establishing comparative computation overhead, coveted security objectives were accomplished by the model with excellent performance assessments.

Cloud-of-Clouds is a deduplication-assisted essential storage model [16]. Deletion code and replication schemes were combined to form multiple cloud storage suppliers in DAC and the data blocks were stored. Both deletion and replication code mechanisms exhibited more benefits and reference characteristics with novelty. The data block was stored using eradication code mechanism with highly referenced data blocks. Performance enhanced via DAC whereas lightweight model execution of DAC was demonstrated and compared with current schemes to achieve low cost.

Pietro et al. [17] proposed a model to overcome data security issues and deduplication. The authors introduced a narrative Ownership (POW) system proof using all the highlights in cutting-edge arrangement with just a fraction of its overhead. The scheme demonstrated the proposed security method on the basis of hypothetical data, as opposed to suspicious computation. Further, feasible optimization methods were also suggested to enhance the performance. At last, the superiority of the proposed method was established using broad benchmarking.

3 Problem Identification

- Cloud Service Provider (CSP) gives service storage and it cannot be relied upon completely as the substance of accumulated information is of interest for many. At the same

time, it needs to perform genuinely on data storage in a sequential manner and increase business benefits.
- Incompetence and poor scalability result from conventional privacy storage with huge data sets.
- Conventional privacy protection schemes could not tackle everything rather than focus on a single issue only; especially when it comes to security and privacy aspects, the schemes were unrealistic. Custom-made privacy safeguarding schemes are cost-incurring and are difficult to actualize.
- Reducing the computational load in the set of directors and achieving frivolous authenticator cohort serves are major problems.
- The research works conducted earlier carried out M-clustering process manually which also had mismatch of information. Simultaneously, both privacy of information and security are compromised in conventional methods.

4 Proposed Work

The main aim of the proposed model is to store big datasets in cloud under high security. In this work, various approaches are utilized by producing Deduplication dataset with data encryption and decryption processes. In the beginning, the dataset is encrypted using a mapper to make a key. Then the mapper forms the dataset as groups, which feeds the data into duplication form according to matching score value. The process of deduplication restricts multiple entries or repetitive data. Deduplication stages are completed according to the amount of issue. Subsequently, these datasets are subjected to clustering. In this process, MFCM is utilized to cluster the data which allows a data to be robust via double or more groups. These clustered data are then encrypted through dual encryption, i.e., homomorphic and SIMON Cipher (OSC). For data decryption, the encrypted data is securely accessed since the approved information holders can get the data with the help of symmetric keys. To enhance the presentation of SIMON cipher, Modified Krill Herd Optimization (KHO) method is utilized. The Modified KHO method builds the encrypted key and provides the most significant privacy-preserved information. Again, when a value is equivalent to or more prominent by threshold value, the production is encouraged in the direction of minimizer. The information is then stored in cloud server using a reducer. The proposed technique is shown in Fig. 1. The proposed model provides high security and is well-organized. The security aspects of the model are appropriate for vast data Deduplication. The proposed big data security is described in-depth in the following sections.

4.1 Map-Reduce (MR) Framework

Map-Reduce system is utilized for big data analytics across different servers. Map-Reduce algorithm includes two significant responsibilities which are mentioned herewith.

Map phase: In this segment, soil data is obtained as input and divided into every map task and M Map task executes the comparable data.

Reduced phase: In this segment, soil data is clustered and separated from preceding segment using Standard Component Analysis (SCA). Secondly, the final aim obtains the output alike input map and unites those data tuples into light group of tuples. Fig. 2 shows the general model of map-reduce structure. Data reduction process consists of four elements, which include:

Input: Database involved

Map: Data sorting and filtering

Reduction: Similarity and missing values are redistributed to the mapped data.

Output: Information reduction

For most of the part, catalog is the outcome desired by Map-Reduce clients to acquire. Map and functions' reduction are specified by clients, based on particular applications.

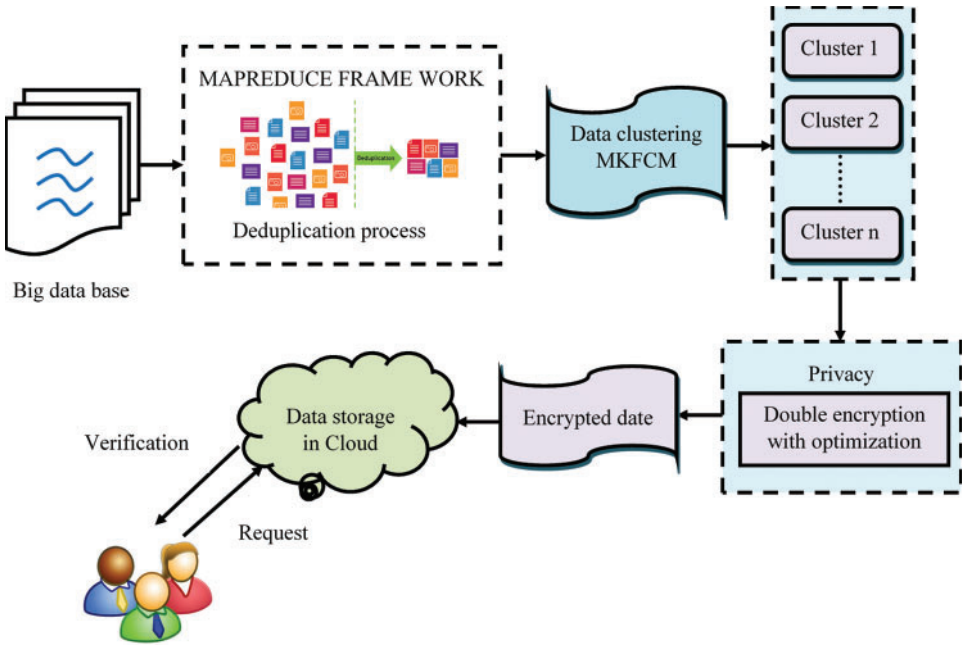

Figure 1: The block diagram for proposed high-security data in cloud

4.2 Data Deduplication Model

Data de-duplication is an optimization process which is used to eliminate specific excess data. In this process, every data proprietor is chosen and similar data is transferred. Then, a particular copy of duplicate data is shared, whereas the duplicate copies in the storage are expelled. Data deduplication basically refers to the storage of present data on the disk by a marvel that supersedes indistinct data in a record or identical locales of the document (comparable data). This cloud data deduplication process is performed using a hash function. In this process, the initial document is used first after which the pre-processed data is utilized to create the confirmed data. Data duplication occurs when the data carrier saves similar data which is already saved on CSP as shown in Fig. 3.

Hash-based data de-duplication strategies utilize a hashing algorithm to distinguish 'large piece' of information. Deduplication process evacuates not-so-special blocks. The actual process of data deduplication can be executed in a number of distinctive ways. Duplicate data is basically disposed by comparing two documents and finalizing on one established dataset, so that the other dataset, which is of no use further, is erased. Either variable or fixed length hash-based de-duplication breaks' data gives the 'chunks [18]' in which the hashing algorithm with 'chunk' procedure creates the hash. On the off chance, in new hash, the data is considered duplicate and does not get saved. The hash never lives at that point, whereas the stored data and hash list are refreshed by means of a novel hash.

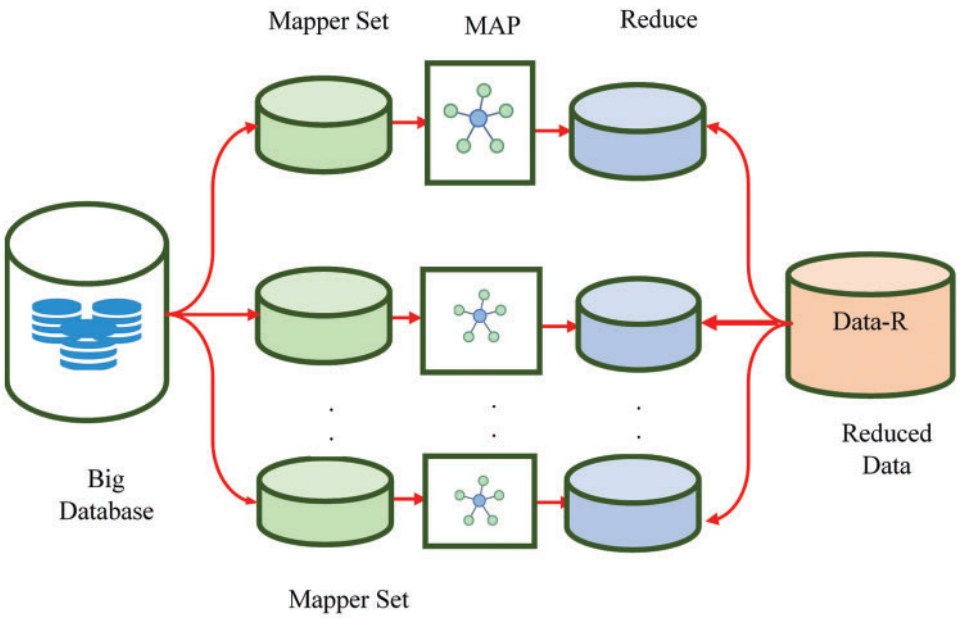

Figure 2: The flow diagram for map reduce framework

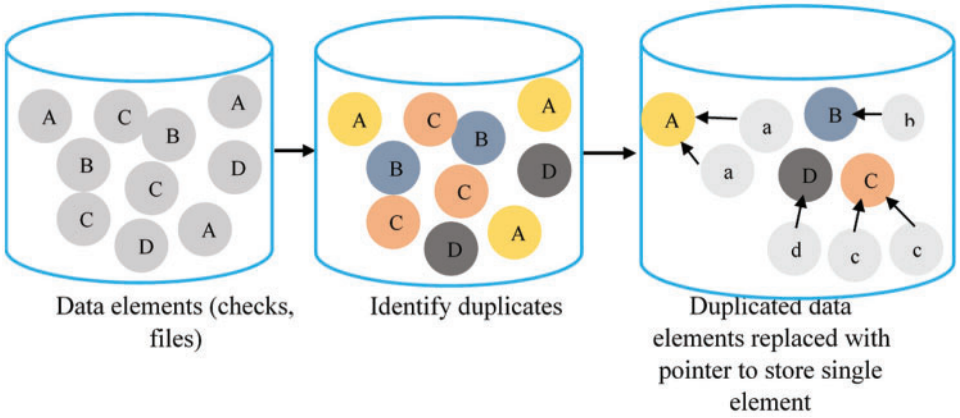

Figure 3: Data deduplication model

4.3 Multi Kernel Clustering Model for De-Duplicated Data

Clustering systems are one of the element invalid systems which are processed in forward direction. Clustering systems are main element in invalid systems that can be operated to position the feature in resemblance beam, among the data substance personalities. The presentation in verification procedure is developed and this clustering procedure is based on the aspect of stoppage.

Standard FCM Model

Hard C-means algorithm provides object classification and image area clustering by fuzzy C-means clustering algorithm.The minimization of criterion function is the fundamental element

in fuzzy C-means clustering alike hard K-means algorithm. The following equation explains the transformation of goal task.

$$O_l = \sum_{i=1}^{n} \sum_{j=1}^{c} u_{ij}^{l} \|x_i - c_j\|^2 \tag{1}$$

where, the real number m is higher than one. At cluster x_i, the degree of membership is u_{ij}^{l}. The d-dimension-measured information is j.

Steps for FCM

1) Initialize the duplicated data cluster centers.
2) Determine D-distance among cluster centers and data using the equation given below:

$$D(x_i, c_i) = \|x_j - c_i\|^2 \tag{2}$$

3) Compute the function of fuzzy membership u_{ij} using the equation given below:

$$u_{ij} = \frac{D(x_j, C_i))^{-1/(l-1)}}{\sum_{j=1}^{n} D(x_j, C_i))^{-1/(l-1)}}; \quad i = 1, 2, \ldots, C \tag{3}$$

4) Based on membership function, calculate fuzzy centers or centroid as:

$$C_i = \frac{\sum_{j=1}^{n} u_{ij}^{l} \left(x_j, C_i\right) x_j}{\sum_{j=1}^{n} u_{ij}^{l} K\left(x_j, C_i\right)}; \quad i = 1, 2, \ldots, C \tag{4}$$

Using FCM algorithm, the number of completed iterations is determined to understand the degree of membership correctness.

4.3.1 Multi Kernel FCM (MKFCM) Model

In the proposed duplicated data clustering process, multiple kernels are considered, i.e., Gaussian model-based Kernels such as K1 and K2. The projected multiple kernel fuzzy c-means (MKFCM) algorithm simultaneously finds the best degree of participation and ideal kernel weights for no-negative combination of kernels' arrangement. The objective function of MKFCM is shown in the condition given herewith.

$$O_l = \sum_{i=1}^{n} \sum_{j=1}^{c} u_{ij}^{m} \left(1 - K_m\left(x_j - C_i\right)\right) \tag{5}$$

Kernel Functions $\Rightarrow$

$$K_m = K1 * K2 \tag{6}$$

$$\text{where } K1 = \exp(-\|x_j - C_i\|^2/\sigma_1^2 \quad \text{and} \quad K2 = \exp(-\|x_j - C_i\|^2/\sigma_2^2 \tag{7}$$

In the above equation, $x_i \rightarrow$ Mean value of Neighbor duplicated data and $\sigma \rightarrow$ Variance of total duplicated data and finally the centroid equation as follows:

$$C_i = \frac{\sum_{j=1}^{n} u_{ij}^{m} K_m\left(x_j, C_i\right) x_j}{\sum_{j=1}^{n} u_{ij}^{m} K_m\left(x_j, C_i\right)}; \quad i = 1, 2, \ldots, C \tag{8}$$

By replacing the Euclidean distance, diverse kernels can be chosen in support of various situations. For clustering, appropriate Gaussian kernel is essential whereas the cloud stores the clustered data towards privacy or security procedures.

4.4 Dual Encryption Representation for Big Data Privacy

For process improvements, dual encryption is considered in the procedure involving big data security. Private information is decided in the direction of second-hand certification procedure in the result to encrypt the information. At present, homomorphic encryption strategy is initialized; however the same strategy is misused after second stage procedure. However, the encrypted information is present in excess effort to OSC process.

4.4.1 Homomorphic Encryption (HE)

Without knowing the private key, the encrypted data operation is executed by utilizing Homomorphic Encryption framework. Here, the secret key holder is a client [19]. The calculation for encrypted information is performed with Homomorphic Encryption model, due to data shortage. When a client decrypts the process given below, it explains the layout of encryption.

Steps for HE

Four functions are present in Homomorphic Encryption procedures.

H = {Decryption, Encryption, Evaluation, Key Generation}

Step 1: Key Generation: $Key\,Gen\,(i, j)$

Step 2: Select double huge prime numbers i and j randomly and compute $k = ij$.

Step 3: Encryption: $Enc\,(u, P_k)$

Step 4: Compute ciphertext $c = i^u \cdot r^k \bmod k^2 \quad \therefore c \in Z_{k^2}$.

Step 5: Decryption: $Enc\,(c, S_k)$

Step 6: Assume c is a cipher text towards decrypt where $c \in Z^*_{k^2}$.

Step 7: Compute plain text $p = \dfrac{L\left(c^{\alpha} \bmod k^2\right)}{L\left(i^{\alpha} \bmod k^2\right)} \bmod k$.

This technique yields similar outcomes after calculation, which it would have acquired, if the technique worked directly on crude information. These encryptions enable complex processes to execute on encrypted information, without compromising the encryption.

4.4.2 SIMON Ciphers

When the hardware is connected, lightweight block cipher is executed according to effective hash work. Varying structures of cipher family consist of low functions with various key sizes and block sizes. According to image pixels, every block and its key differ with the estimation of 16 block events that vary in the range of 32 to 128 bits. The cipher content blocks are brought to execute an event on plain text with fixed block size.

Characteristics of SIMON Cipher

- Because of security examinations, SIMON cipher comprises of nonlinear attributes which get directed in block data and size. One can think about a tree in which the differences can be utilized in fixed input. Few conceivable output differences are produced at each round with distinction.
- A solid structure of round capacity is exploited by one, thereby considering the basic attributes that are prolonged with more rounds on the rotten possibility.

- The key scheduled properties are killed with SIMON key calendars by employing round consistency and pixel quantity with regards to images.
- Based on lightweight block cipher, single key differential and singular key differential trademarks were found to be 15-round SIMON48.

The block cipher qualities are intended to be excellent. A minimum number of active S-boxes are ensured with quality, since the number finalizes the optimizer which in turn yields the optimum solution.

4.5 Designing Model

In wireless networks, cipher model is executed as DI security. There are some conditions associated with the model, such as decryption, encryption models, round and bit. The size {16 to 64} is represented using 2n-bit blocks with SIMON cipher. The following equation explains the same in detail.

$$Enc\ DI = Cipher^{i}_{q_l}, \ldots, Cipher^{n}_{q_1}, \quad i > 1 \tag{9}$$

'Round functions' are the functions of ciphers $C^{i}_{q_l}, \leq i \leq q$ and round keys are respective keys. The cipher is selected as the modified iterated block cipher to check, whether the functions are identical.

Round configuration: The round function utilizes 128-bit of plaintext as inputs in SIMON block cipher. In 68 rounds, the 128-bit cipher messages are produced with 128-bit key. The following steps explain the SIMON encryption operations.

- Two arbitrary bits of n-bit words perform the Bitwise AND activity.
- The bitwise AND task performs the activity of Bitwise XOR. One arbitrary is XOR-ed with the final value.
- Where, rotation count is y by means of $S^{y}(x)$ in bitwise revolution ROL.

The following equation explains SIMON round capacity for encryption.

$$RF(w_l, w_r, k_{round}) = \left(\left(S^{1}(w_l) \& S^{8}(w_l)\right) \oplus S^{2}(w_l) \oplus w_r \oplus k_{round}, w_l\right) \tag{10}$$

Eq. (11) describes the image information with the help of inverse function.

$$RF^{-1}(w_l, w_r, k) = \left(w_r, \left(S^{1}(w_l) \& S^{8}(w_l)\right) \oplus S^{2}(w_r) \oplus w_l \oplus k_{round}\right) \tag{11}$$

From Eq. (11), the left-most word is w_l, i.e., given block. The correct round key is k_{round} and the right-most word is w_r.

4.5.1 Key Generation

Each round key produces a key expansion of SIMON cipher from the master. From initial 128-bit master key, a total of 44 32-bit sized round keys is generated by the selected SIMON64/128 configuration. The previously saved n round of keys consolidate the given round, i. The activities are accompanied via key expansion.

To signify $a \oplus b$ as bitwise XOR.
where c is the rotation count which denotes $s^{-c}(a)$, as of right bitwise rotation ROR.

Eq. (12) explains the function of key expansion.

$$Key_i\left(k, c, z_j\right) = F\left(k_{i+3}, k_{i+1}\right) \oplus S^{-1}\left(F\left(k_{i+3}, k_{i+1}\right)\right) \oplus k_i \oplus c \oplus \left(z_j\right)_i \tag{12}$$

4.5.2 Key Optimization Model

The key optimization process in SIMON utilizes Krill Herd Optimization (KHO) method [20]. For example, two principle objectives are present in KHO process such as reaching sustenance and increasing krill thickness. The identification of nourishment and thickness expansion are conducted via crowding. The other mutation process, crossover, random diffusion, foraging activity and krill individual altogether induce the movement as the elements of KHO.

i. Objective function for OSC

During decryption process with key K, minimum number data is retrieved as the fitness function. Multiple key sets in TDES process generate the initial solution and the condition is satisfied using an optimal key.

$$F_i = MIN\left(\%\ of\ \ data\ \ Reterieved\right) \tag{13}$$

ii. New keys updating process

The discretionary dimensionality enables search by realizing an optimization algorithm. The n-dimensional decision space is used to sum-up the Lagrangian method.

$$\frac{dK_i}{dt} = M_i + A_i + P_i \tag{14}$$

Hence, the ith krill individuals are with physical diffusion, P_i. Krill individuals induce the motion, M_i and foraging motion, A_i.

i. Movement induced by other krill individuals

The local impact or area swarm thickness settles the course of movement of a krill individual in the advancement. The unpleasant swarm thickness and the objective swarm thickness are explained below:

$$M_i^{new} = M^{\max}\gamma_i + \omega_n M_i^{old} \tag{15}$$

The representation of $I_{\max}$ clarifies the documentations in above conditions as the huge rate. The movement, instigated by idleness weight ω_n, tends to be [0, 1].

ii. Foraging motion

Two effective parameters are used to figure out a similar scavenging development. The past experience is secondary one, while the initial sentence is the third. At *i*th krill individual, this development is communicated.

$$A_i = F_m\delta_i + \omega_m A_i^{old} \tag{16}$$

where, the scavenging velocity is denoted by F_m and the searching movement of dormancy weight is ω_m which tends to be [0, 1]. The best fitness of krill impact is $\delta_i{}^{best}$ and also the nourishment appealing is $\delta_i{}^{food}$. The searching rate calculation is deliberated to 0.02 (ms-1).

iii. Physical diffusion

Physical spread, with krill persons, is believed to be a sporadic procedure. This establishment expresses the degree so as to be a majority disgraceful scattering speed along with an uneven directional vector.

$$P_i = P^{\max}\lambda \tag{17}$$

Here, the maximum diffusion speed is $D^{\max}$. D is the random vector to random values [−1, 1].

iv. Crossover and mutation

For overall enhancement, the part of GA utilizes the crossover administrator as a suitable procedure. The mutation likelihood (Mr) manages the mutation.

$$cr = 0.2F_i \quad and \quad Mr = 0.5/F_{i\,best} \tag{18}$$

For global best, the mutation probability which utilizes new mutation probability is equivalent to zero. When the fitness value decreases, the global best increases.

Optimal keys are known to be 64 bits long, which are recognized in support of their compatibility whereas the adaptability is transferred in support of SIMON cipher. The anticipated technique encrypts the information to store in the cloud.

4.6 Cloud Storage Process

In view of the above dual encryption, the input duplicated information is encrypted. After encryption, the document is stored in the cloud which gains a structure with genuine client. Due to the secure double encryption mechanism, the confidentiality of information cannot be inferred directly. The legitimate message authentication or substantial signature is never produced with the advantage of proposed strategy.

5 Result

The proposed work with clustering model and double encryption was implemented using Java with JDK 1.7.0. The operation framework stage consisted of 1.6 GHz, 4 GB RAM with Intel (R) Core i5 processor configuration in Windows 10 operating system. The datasets, with different medical information, were used via Map-Reduce structure in cloud condition. The following section explains the database and comparative investigation.

5.1 Database Description

From UCI machine learning repository, the big data security procedure was validated by checking the medical database. In total, the maximum size of databases was 1,000,000 including breast cancer and Switzerland databases. Tab. 1 provides the details of the dataset employed.

Switzerland database: There were 76 traits present in the database and 14 subsets were utilized by means of distributed assessment. Particularly, ML researchers utilized a special case of database. The presence of coronary heart disease is referred to 'objective' field alludes.

Breast cancer database: Dr. Wolberg reported that the sampled clinical cases turned up periodically. The chronological information was gathered according to the database.

Table 1: Combined database

Size of the file (MB)	Execution time (ms)	Encryption time (ms)	Decryption time (ms)	Memory (bits)
1	37489	7865	5868	1225887
2	46215	13252	8564	1325588
3	50235	13658	11254	1579554
4	56689	16524	12332	1544712
5	621028	18256	16235	1768525

5.2 Proposed Technique (MKFCM-OSC) for Time And Memory Analysis

Tabs. 2 and 3 show time and memory analyses for the applied Switzerland and breast cancer databases and a combination of both, using the proposed method. The time and memory analysis, displayed in Tab. 1, used a combination of databases to support dissimilar file size in terms of MegaBytes (MB). The table clearly shows the time taken for data execution and the time taken to decrypt and encrypt the known information in a secure manner. Moreover, the decrypted data got stored in the cloud through memory allocation, with regards to file size. When a file was considered weighing 5 MB, the execution, encryption and decryption times were 621028, 18256 and 16235 ms respectively. Further, the memory allocation for 5 MB file size was 1768525 bits. Similarly, time and memory analyses of Switzerland database and breast cancer database are demonstrated in Tabs. 2 and 3.

Table 2: Switzerland database

File size (MB)	Execution time (ms)	Encryption time (ms)	Decryption time (ms)	Memory (bits)
1	36488	7651	5469	1215784
2	45215	11245	8654	1325586
3	48962	13658	10268	1478548
4	55698	15478	12447	1544754
5	61025	18646	14639	1655882

Table 3: Breast cancer database.

File size (MB)	Execution time (ms)	Encryption time (ms)	Decryption time (ms)	Memory (bits)
1	32092	7513	3912	1214276
2	42313	11103	8218	1320926
3	47951	13639	7596	1475240
4	54759	15293	7635	1540292
5	56969	18234	10216	1651941

Fig. 4 explains the proposed and existing techniques with respect to encryption time analysis. Encryption time analysis, with number of mappers, segregates the big data. The proposed

MKFCM approaches yielded better results than previous Fuzzy C means (FCM) techniques. The encryption time was analyzed corresponding to various amount of mappers such as 5, 10, 15, 20 and 25. For the mapper 15, the encryption time for FCM was 12000 ms and for MKFCM, it was 12500 ms. Compared to existing techniques, the time used for data encryption was high in MKFCM technique.

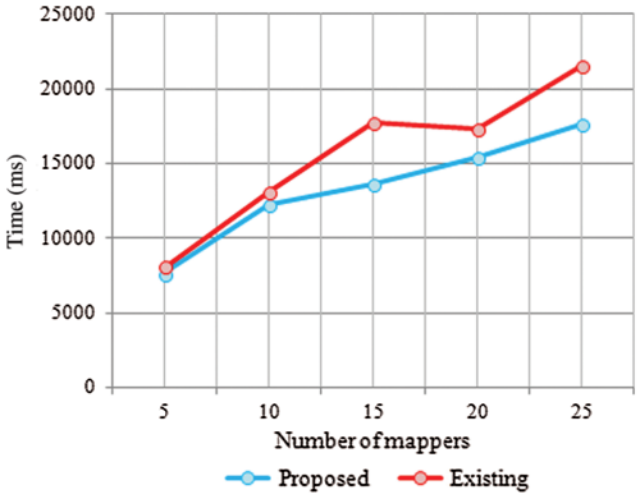

Figure 4: Encryption time analysis

5.3 Duplicated Data Analysis

Tab. 4 shows the duplicated data for Switzerland data and Breast cancer data. The proposed Hash function-based framework accepted 'contribution' as a document and transferred the data in an encrypted shape. After this, the proposed framework checked the data for duplication in whole big data. The table demonstrates the percentage investigation of duplicated data for both Hash function and Authorized Party-based approach. When Switzerland data was utilized, the proposed Hash function achieved 44.65% duplicated data, a value low than the current approved collection strategy. In breast cancer database, the proposed technique accomplished 44.74% duplicated data from 1MB document measure. Further, the percentage of duplicated data was investigated for 2, 3, 4 and 5 MB data as well.

Table 4: Percentage of duplicated data for the proposed technique

File size (MB)	Switzerland data (%)		Breast cancer data (%)	
	Hash function based	Authorized Party based	Hash function based	Authorized party based
1	44.65	49.85	44.74	42.77
2	46.28	51.82	47.11	45.80
3	48.69	47.80	49.66	48.64
4	46.39	49.50	46.82	45.29
5	51.15	55.15	52.05	50.51

5.4 Analysis of Clustering Techniques

The accuracy analysis of clustering techniques such as FCM and MKFCM for Switzerland database is shown in Fig. 5. The graph clearly depicts the accuracy of clustering approaches for different file sizes such as 1, 2, 3, 4 and 5 MB. For 1 MB file size, the proposed MKFCM achieved 84.76% accuracy, while the existing FCM accomplished 82.72% accuracy. For 2 MB file size, the proposed MKFCM achieved 85.56% accuracy, while FCM accomplished 83.62% accuracy. Similarly, the accuracy was analyzed for the file sizes such as 3, 4, and 5 MB. The graphical representation reveals that the proposed technique achieved maximum accuracy compared to existing clustering techniques.

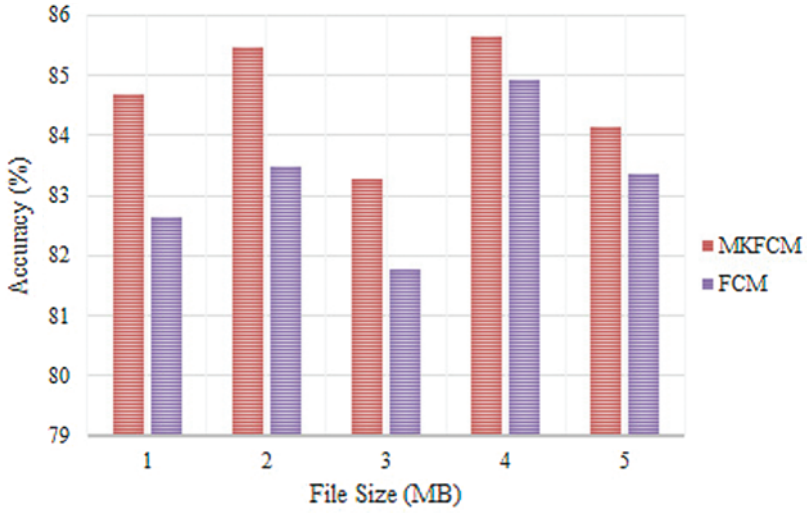

Figure 5: Comparative analysis of Switzerland database

Fig. 6 compares the analysis of clustering techniques such as FCM and MKFCM in which the breast cancer database was used. The graph clearly depicts the accuracy of clustering approaches for different file sizes such as 1, 2, 3, 4 and 5 MB. For 1 MB file size, the proposed MKFCM achieved 84.06% accuracy, while the existing FCM accomplished 82.98% accuracy. For 2MB file size, the proposed MKFCM achieved 85.36% accuracy and the FCM accomplished 84.22% accuracy. Similarly, the accuracy was analyzed for file sizes such as 3, 4, and 5 MB. From the graphical representation, it can be concluded that the proposed technique accomplished maximum accuracy in breast cancer database compared to existing clustering techniques.

5.5 Comparison of Encryption Techniques

Fig. 7 demonstrates the comparative examination of various encryption techniques such as Triple Data Encryption Standard (TDES), Advanced Encryption Standard (AES), Data Encryption Standard (DES), Homomorphic Encryption (HE) and OSC. For security reasons, double encrypted big data was considered in association with the optimization model. Further, the self-assurance optimized data executed the Homomorphic Encryption technique to encrypt data by OSC and stored the encrypted data in cloud. A graph was plotted based on time taken for encrypting the data vs. diverse record measures. For 1 MB record measure, AES took 8000 ms for encrypting the data. Similarly, DES took 7500 ms, TDES took 7600 ms, HE consumed 8200 ms

and the proposed OTDES took 7200 ms to encrypt the data. To conclude, the proposed OSC accomplished the least time for encrypting big data compared to other encryption processes.

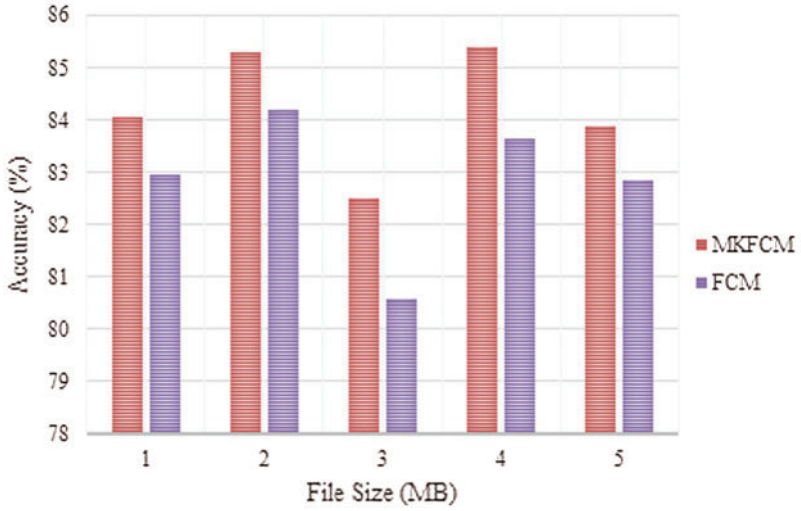

Figure 6: Comparative analysis of breast cancer database

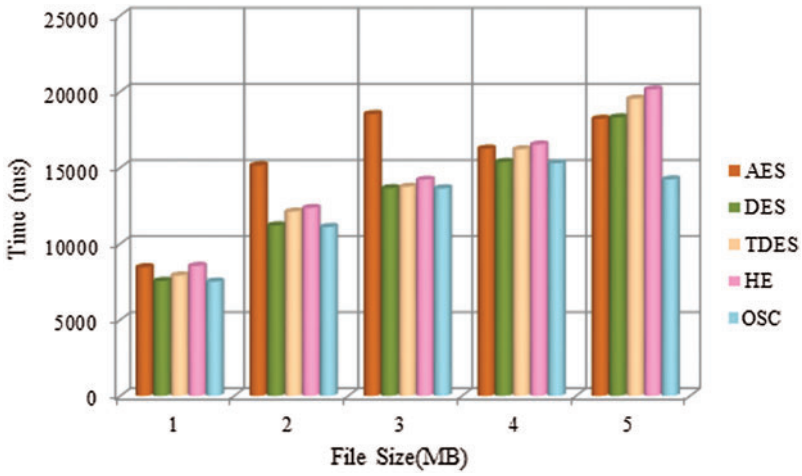

Figure 7: Comparative analysis of different encryption techniques

6 Conclusion

The proposed de-duplication method on encrypted big data in cloud computing was analyzed by optimal double encryption approach. The framework reduced the measure of capacity required by the cloud service providers. Notwithstanding the double encryption technique, de-duplication process was proposed in which the hash function was used to transfer the data in an encrypted format. After this, data duplication was checked and other functions such as modification, deletion, and de-duplication of the data were performed. De-duplication process expelled the repetitive blocks. Here, MKFCM clustering model was used to analyze the de-duplication process. MKFCM

algorithm identified the best degree of participation and ideal kernel weights for non-negative combination of arrangement of bits. The key optimization process was accomplished by OSC during when the KHO technique was used. In view of the above processes, the proposed technique encrypted the data and stored the data on cloud. The findings of the study infer that the proposed double encryption scheme ensured enhanced the authentication accuracy and security compared to other techniques. However, the authors recommend future researchers to improve the performance of the proposed model using lightweight cryptographic techniques.

Funding Statement: The author(s) received no specific funding for this study.

Conflicts of Interest: The authors declare that they have no conflict of interest to report regarding the present study.

References

[1] S. N. Mohanty, K. C. Ramya, S. S. Rani, D. Gupta, K. Shankar *et al.,* "An efficient lightweight integrated block chain (ELIB) model for IoT security and privacy," *Future Generation Computer Systems*, vol. 102, pp. 1027–1037, 2020.

[2] K. Shankar, M. Elhoseny, E. D. Chelvi, S. K. Lakshmanaprabu and W. Wu, "An efficient optimal key based chaos function for medical image security," *IEEE Access*, vol. 6, pp. 77145–77154, 2018.

[3] B. Rashidi, "High-throughput and lightweight hardware structures of HIGHT and PRESENT block ciphers," *Microelectronics Journal*, vol. 90, pp. 232–252, 2019.

[4] F. Khelifi, "On the security of a stream cipher in reversible data hiding schemes operating in the encrypted domain," *Signal Processing*, vol. 143, pp. 336–345, 2018.

[5] K. Shankar and M. Elhoseny, "An optimal lightweight rectangle block cipher for secure image transmission in wireless sensor networks," in *Secure Image Transmission in Wireless Sensor Network (WSN) Applications. Lecture Notes in Electrical Engineering*, vol. 564, pp. 33–47, Cham: Springer, 2019.

[6] S. Thakur, A. K. Singh, S. P. Ghrera and M. Elhoseny, "Multi-layer security of medical data through watermarking and chaotic encryption for tele-health applications," *Multimedia Tools and Applications*, vol. 78, no. 3, pp. 3457–3470, 2019.

[7] K. Shankar and M. Elhoseny, "An optimal light weight cryptography—Simon block cipher for secure image transmission in wireless sensor networks," in *Secure Image Transmission in Wireless Sensor Network (WSN) Applications. Lecture Notes in Electrical Engineering*, vol. 564, pp. 17–32, Cham: Springer, 2019.

[8] H. Tao, M. Z. A. Bhuiyan, A. N. Abdalla, M. M. Hassan, J. M. Zain *et al.,* "Secured data collection with hardware-based ciphers for IoT-based healthcare," *IEEE Internet of Things Journal*, vol. 6, no. 1, pp. 410–420, 2019.

[9] B. K. Das and R. Garg, "Security of cloud storage based on extended hill cipher and homomorphic encryption," in *2019 Int. Conf. on Communication and Electronics Systems*, Coimbatore, India, pp. 515–520, 2019.

[10] K. Shankar and M. Ilayaraja, "Secure optimal k-nn on encrypted cloud data using homomorphic encryption with query users," in *2018 Int. Conf. on Computer Communication and Informatics*, Coimbatore, India, pp. 1–7, 2018.

[11] S. Venila and J. Priyadarshini, "Scalable privacy preservation in big data a survey," *Proceedings of Procedia Computer Science*, vol. 50, pp. 369–373, 2015.

[12] Celcia and Kavitha, "Privacy preserving heuristic approach for intermediate data sets in cloud," *International Journal of Engineering Trends and Technology*, vol. 9, no. 5, pp. 235–241, 2014.

[13] M. Sookhaka, A. Gani, M. K. Khan and R. Buyya, "Dynamic remote data auditing for securing big data storage in cloud computing," *Journal of Information Science*, vol. 380, pp. 101–116, 2017.

[14] M. Miao, J. Wang, H. Li and X. Chen, "Secure multi-server-aided data deduplication in cloud computing," *Journal of Pervasive and Mobile Computing*, vol. 24, pp. 129–137, 2015.

[15] J. F. Liu, J. Wang, X. Taoc and J. Shen, "Secure similarity-based cloud data deduplication in the ubiquitous city," *Journal of Pervasive and Mobile Computing*, vol. 41, pp. 1–12, 2016.

[16] S. Wua, K. C. Li, B. Mao and M. Liao, "DAC: Improving storage availability with deduplication-assisted cloud-of-clouds," *Future Generation Computer Systems*, vol. 74, pp. 190–198, 2017.

[17] R. D. Pietro and A. Sorniotti, "Proof of ownership for deduplication systems: A secure, scalable, and efficient solution," *Journal of Computer Communications*, vol. 82, pp. 71–82, 2016.

[18] M. U. Tahir, M. R. Naqvi, S. K. Shahzad and M. W. Iqbal, "Resolving data de-duplication issues on cloud," in *2020 IEEE Int. Conf. on Engineering and Emerging Technologies*, Lahore, Pakistan, pp. 1–5, 2020.

[19] K. Shankar and S. K. Lakshmanaprabu, "Optimal key based homomorphic encryption for color image security aid of ant lion optimization algorithm," *International Journal of Engineering & Technology*, vol. 7, no. 9, pp. 22–27, 2018.

[20] W. Chen, Z. Shao, K. Wakil, N. Aljojo, S. Samad *et al.,* "An efficient day-ahead cost-based generation scheduling of a multi-supply microgrid using a modified krill herd algorithm," *Journal of Cleaner Production*, vol. 272, 122364, 2020.