International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 04 | Apr 2020 www.irjet.net p-ISSN: 2395-0072

Secure Encrypted Data Deduplication using Hashing Technique in Cloud

Kanika S. Gandhi[1], Devashree S. Patekar[2], Garima M. Virulkar[3], Kulshree S. Damle[4], DurveshSingh Thakur[5], Prof. Ms. Shwetambari G. Pundkar[6]

[1-5]U.G. Student, Department of Computer Science &Engineering, Prof. Ram Meghe Institute of Technology & Research, Amravati, India
[6]Associate Professor, Department of Computer Science &Engineering, Prof. Ram Meghe Institute of Technology & Research, Amravati, India

---

Abstract –

Cloud storage is a remote storage benefit, where consumers can upload and download their data anywhere and anytime. In data processing and data storing cloud Computing idea which performs an essential role in information technology. However, it raises effects regarding privacy and data secrecy because all the data are stored in the cloud storage and this is a subject of concern for users, and it affects their keenness to use cloud storage services. To prevent the willingness and illegal access from unauthorized users, the data stored in the cloud should be secured. There comes a data security concept called as Encryption. On the other hand, Privacy and the security of the data is stored in the cloud in the encrypted or cipher-text format. To maintain the users, a cloud storage server typically performs a specialized data compression technique (data deduplication) to eliminate duplicate data because of the storage space is not infinite. By this, only the encrypted data are going to be stored in the cloud which reduces the usage of the storage devices up to a great extent. We have a lot of deduplication plans which escape the duplicate data, but the main issue with those plans is poverty of security and poverty of tractability for the safe data access control. Due to these two issues, very few of them are taken into practice. In this, to deduplicated the encrypted data to allow secure data access control we used a scheme known as Attribute-based Encryption. Data deduplication technique provides the cloud users to control their cloud storage space virtually by avoiding storage of regular data's and save bandwidth. The data are finally stored in cloud server namely Cloud Me. To make certain data confidentiality the data are stored in an encrypted type using Advanced Encryption Standard (AES) algorithm. Data deduplication, which makes it possible for data owners to share a copy of the same data, can be performed to reduce the consumption of storage space. Due to the above issues, there is research on encrypted data deduplication. In this manuscript, we propose an encrypted data de-duplication mechanism which makes the cloud storage server be able to eliminate duplicate cipher texts and improves the privacy protection.

Key words – Cloud Computing, Encryption, Data storage, Decryption, Deduplication management, Cipher-text, Security, Perfect hashing, etc.

1. INTRODUCTION

There have been people eager to forecast the future of technology ever since the first technologies were flourished. Since the formation of computers, however, the outbreak in both the pace of alteration and the volume of prognostication on all sorts of technology trends is seen. The digital transformation is far more significant wherever the cloud computing is concerned. Cloud computing is the third wave of digital revolution. Cloud computing has completely changed the way of business-and their consumers-for storing and accessing the data. Cloud computing is empowering the companies to leverage the cloud to innovate cheaper and faster. There are 3 types of cloud-high, low & medium level. Based on this cloud computing can be broken up into three main services: Software as a service (SaaS), Infrastructure as a service (IaaS), and platform as a service (PaaS). These three services make up a rack space calls the cloud computing stack.

The expeditious development in big data and cloud computing has changed the user's terminology to the tackle the vast information. cloud computing immensely opens the doors for data providers who want to outsource their data to the cloud without revealing

their precise data to the foreign parties [10]. One threatening challenge of today's cloud storage services is the management of gross amount of data. The estimated amount of data generated in 2017 and 2018 is 30 zettabytes and 33 zettabytes respectively [17]. And there is a threat of expanding more amount of data in upcoming years. Till the end of 2020, it is expected that volume of the data will reach a peak of 40 trillion gigabytes, concerning report of IDC [6].

To resolve the above issue data deduplication is broadly practiced in cloud services providers, which eliminate multiple duplicate data to improve storage utilization. Paradisiac ally, the increasing value of data needs cost-efficient storage, to end this mutually the deduplication technique can be used. The method of Deduplication is holds unique information compression strategy to remove excess information. This decreases the rate of transmission and loading space in framework. This method of Deduplication finds out the replica of the data. It spares just one duplicate of the information and emphatically utilize consistent pointers for copied information. This whole technique actually removes the unnecessary copies of encrypted data in the cloud. For feasible cloud storage such as Dropbx, Mozy, Memopal and lessened maintenance cost this deduplication is the technique applied to user data [6].

Actually, there are two kinds of data duplication which are: file and block. From a meta context, block - level duplication is achieved at the aggregate level by duplicating the blocks of data that subsume the volume. While file-deduplication works, as its name implies, at the file level. If the duplicate files are in the duplication domain, they are single-instanced. The file duplication is generally considered as coarse level of duplication and block level, fine grained. As far as such type of block level deduplication is bothered, it often yields more considerable outcomes than file level dedupes. Block level block dedupe works even on just similar file but file dedupe works on whole identical file. Hence we have implemented block level deduplication in our paper. For example, in case where multiple edits of document are maintained, a file may have several copies each of with few words changed. File deduplication wouldn't work on these but block deduplication may be able to duplicate at block level.

For better security and efficient handling of data, two models are used: client-side and server-side. Both can be applied to single server storage and distributed storage [8]. Both models use uniform scenario which depends on number of basic security techniques. For this we use block encryption to enable encryption which acquiesce deduplication of common chunks. A hash function is used by convergent encryption for getting the key value by converting the plain text into block of chunks. Any client encrypting the data will use the same key to decrypt the document.

When the users utilize cloud services, they hand over the control of their confidential data to the cloud service providers which can cause the risk of privacy leakage. Encrypted data is more secure to transmit over insecure network. To read the encrypted data he/she needs to have a secret key to decrypt message. Encrypting data has additional advantages assuring that messages should not be revoked during transmit of data and verifying the identity of the sender other than providing the confidentiality and privacy of a message. For the encryption and decryption, we use RC6 Algorithm.

RC6 is a symmetric key block cipher. The improvements of RC6 over RC5 include using four w-bit word registers, and introducing a quadratic equation into the transformation, integer multiplication as an additional primitive operation.

The evolution has provided a simple cipher yielding numerous evaluations and security in a small package. RC6 uses 128 bit block size and supports key sizes of 128, 192 and 256 bits. Also it uses fewer rounds and offers a higher throughput, data-dependent rotations, modular addition, and XOR operations.

Hashing is the practice of using an algorithm to map data of any size to a fixed length. This is called a hash value. For every block, hash value is unique. Now, whereas encryption is meant to protect data in transit, hashing is meant to verify that a file or piece of data has not been altered that it is authentic. Hash functions are useful and such functions are important cryptographic primitives used for things such as digital signatures and password protection and appear in almost all information security applications.

SHA stands for "Secure Hash Algorithm" it is a fingerprint that specifics the data. SHA-2 is a family of hashes and is available in a variety of lengths, the most popular being 256-bit. The hash function compares the computed "hash" to a known and expected hash value through which a person can also determine the data's integrity.

As we are using SHA-256 that means that the algorithm is going to output a hash value that is 256 bits, usually represented by a 64 character hexadecimal string. Whereas encryption is a two-way function hashing is a one-way function. A one-way hash the data cannot be generated from the hash, but can be generated from any piece of data.

SHA-256 consists of bitwise operations, modular additions, and compression functions. This is helpful in case an attacker hacks the database. Additionally, SHAs exhibit the avalanche effect where, when an input is changed slightly the output changes significantly.

To develop of the web application, we used ADO.NET this is a module used to establish connection between application and data sources. .NET is a framework to develop software applications. Moreover, it provides a broad range of functionalities and support.

ADO.NET Data sources can be such as SQL Server and XML. To connect, retrieve, insert and delete data ADO.NET consists of classes these classes. The ADO.NET classes are integrated with XML classes located into System.Xml.dll which are located into System.Data.dll and the components that are used for accessing and manipulating data are the .NET Framework data provider and the Data Set.

As we have developed a web application, the web pages are created using ASP.NET. It actually executes code on the server, code that can use databases and then produce html to the browser. ASP.NET with C# is used to develop the re-engineered supermarket management system, where ASP.NET is a reworking of the original Active Server Pages technology.

Along with ADO.NET and ASP.NET, we have used stored procedures in SQL. Stored procedures in SQL Server, stores the procedures program statements to perform operations in the database and return a status value to a calling procedure or batch. Structured Query Language (SQL) statements are set of stored procedure with an assigned name and are stored in a relational database management system as a group.

2. LITERATURE REVIEW

In this project, the notion of authorized data deduplication was proposed by Li et al. [10] to protect the data security by including differential advantages of users in the duplicate check. In this project we perform several new deduplication constructions supporting authorized duplicate check in hybrid cloud architecture where the duplicate-check tokens of files are generated by the private cloud server with private keys.

In order to tackle the problem that the unauthorized users can access the user information only by supplying the hash value. Halevi et al. [12] proposed the proof of ownership (PoW), which is an interaction protocol between client side and server side to verify the ownership of that client. In, the client and server create a Merkle Hash Tree (MHT) based on the source _le, and use a challenge-response model to verify the correct of MHT path provided by the client. Blasco et al. [5] proposed a system which is a bf-PoW scheme based on the bloom filter, to achieve the proof of ownership systematically which has the requirements of the certain tokens from the substantiated client. Through the security analysis, a wide range of benchmark tests and comparison of the already existing schemes, the proposed scheme greatly decreases the amount of both the client and the cloud server.

D. Harnik et al. [3] proposed a procedure that provides higher secure guarantees while slightly decreasing bandwidth savings, since deduplication offers considerable savings in both disk capacity and network bandwidth. Xia et al. [18] review the differences between data deduplication and the background of data deduplication and traditional data compression. Ng et al. [6] proposed a private data deduplication in data storage, where a client held a private data proves to a server stored a summary string of the data that he/she is the owner of that data without revealing further information to the server.

In this paper, Xu et al. [7] offers a cryptographic antediluvian to enlarge the security of client- side deduplication in the bounded percolate setting where the certain amount of efficiently-extractable information about any file is oozed. Encrypted deduplication has been deployed in commercial cloud environments and extensively analyzed in the literature to simultaneously achieve both data security and storage efficiency. Li et al. [2] proposed how the deterministic nature of encrypted deduplication makes it vulnerable to information leakage caused by analysis of the frequency.

In the Storer's et al. [8] paper that have developed two models for secure deduplicated storage which are anonymous and substantiated. These two of the models demonstrate that the security can be amalgamated with deduplication in a way that provides a multiple security characteristics range. The security is provided through the use of convergent encryption in the models that they have presented. A map is created for each file that narrates how to rebuild a file from blocks in both the anonymous and authenticated models. To prevent information leakage, several solutions have been proposed. However, these solutions are based on a strong assumption that all individual files are independent of each other. Shin et al. [9] proposed a storage GW-based secure client-side deduplication protocol. A storage GW is a network appliance that provides access to the remote cloud server and simplifies interactions with cloud storage services, and is used in various cloud service delivery models such as public, private, and hybrid cloud computing. The proposed solution, by utilizing the storage GW as an important component in the system design, achieves greater network efficiency and architectural flexibility while also reducing the risk of information leakage.

SHOBANA et al. [4] offers system that compress the data by removing the duplicate that is same data, to assure the privacy of the confidential data during deduplication. To encrypt the data before deploying encryption method is used.

MihirBellare et al. [21] proposed a paper on Duples: Server helped encryption for de-copied capacity. This paper for the most part highlights on Data trustworthiness and capacity which are two principles essential requirement for distributed storage. Evidence of Retrievability (POR) and Proof of Data Possession (PDP) systems assures information uprightness for distributed storage. The individual examining of these developing undertakings can be blundering. This plan of open key in view of homomorphism direct authenticator, which permit TPA to play out the evaluating without requesting the nearby duplicate of information and along these lines drastically decreases the correspondence and calculation expenses when differed with the clear information examining perspective. Jewie Yuan et al. [22] proposed a paper on Secure and Constant Cost Public Cloud Storage Auditing with De-duplication. This paper conveys the arrangement that bolster skilled and secure information trustworthiness evaluating with capacity deduplication for distributed storage. This sorts out issue with novel plan in light of strategies including polynomial-based confirmation labels and homomorphic straight authenticators. Evidence of Retrievability (POR) and Proof of Data Possession (PDP) are the course of actions for information honesty and capacity proficiency for distributed storage. Verification of Ownership (POW) is the method that exile redundantly copied information from the server intensifying capacity proficiency in a secured way. This security of the proposed plot in view of the computational Diffie-Hellman issues, the Static Diffie-Hellman issue and t-solid Diffie-Hellman issue.

Yang et al. [20] another secret method for more reliable possession. Provable duty for in de-duplication in distributed storage by utilizing remote information checking method. The far site circulated document framework gives approachability by repeating each record onto different desktop PCs. Since this replication demolish huge storage room, it is essential to recover utilized space where feasible.

In the following project Zhang et al. [11] offered two anonymous CP-ABE schemes in which a similar part is added before the decryption part with fast decryption by introducing a new technique called match-then-decrypt into the decryption. For the purpose of the fast decryption, to allow aggregation of pairings during decryption, special attribute secret key components are generated, and hence the unidentified decryption only involves small and constant number of pairings

Gigi's et al. [23] this paper widens data flow testing techniques to Web applications, and presents a proposed perspective to data flow testing of ASP.NET Web applications. It considers the data flow analysis of ASP.NET Web applications, which have different structure than traditional programs. Yingyang et al. [24] design has been fruitfully applied in system design. It brings a great ease to secondary development staff and escalate the system design process, improve the system design performance. Stored procedure router has a certain reuse. When invoking stored procedure coded by PL/SQL, it is convenient to use stored procedure router. In this article, Zhang et al. [25] mainly discusses the application of distributed database technology in research-oriented platform building. Based on the .NET Remoting technology in C# solves the remote communication problems. Then solve the system's data consistency problem combing Based on the above research, the developed comprehensive research-oriented platform can effectively manage various branch course data and improve users' using efficiency and accuracy.

Bhaskar et al. [26] this paper comprises of a straightforward outline to accomplish a secured de-duplication system in half and half cloud. The outline comprises of couple of modules and two stages. The main stage is encryption and second stage is deduplication.

However, the above data deduplication schemes do not take into account the key updating and user revocation. Kwon et al. [13] proposed a new deduplication scheme with multimedia data, which is based on randomized convergent encryption and privilege-based encryption to achieve authorized reduplication and user revocation. Hur et al. [14] proposed a novel data reduplication scheme for the server-side, which uses the randomized convergent encryption algorithm and ownership group key distribution technique to achieve the authorized access and support security reduplication with ownership changes dynamically. Ding et al. [15] proposed a secure encrypted data reduplication scheme, which exploits the homomorphic encryption algorithm to achieve security data reduplication, and supports ownership check and user revocation. However, the above schemes exploit the homomorphic encryption and proxy re-encryption with high computation cost.

3. PROPOSED SYSTEM

3.1 Objectives of Proposed System

Specifically, the contributions of this paper can be summarized as below:

- We motivate to save cloud storage and preserve the privacy of data holders by proposing scheme to manage encrypted data storage with reduplication.
- Our scheme can flexibly support data sharing with reduplication even when the data holder is offline, and it does not intrude the privacy of data holders.
- We aim an effective address to verify data ownership and verify duplicate storage with secure challenge and huge data support.
- We combine cloud data reduplication with data access control in a simple way, thus resign data reduplication and encryption.
- We verify the security and assess the execution of the proposed scheme through analysis and simulation. The results show its efficiency, effectiveness and applicability.

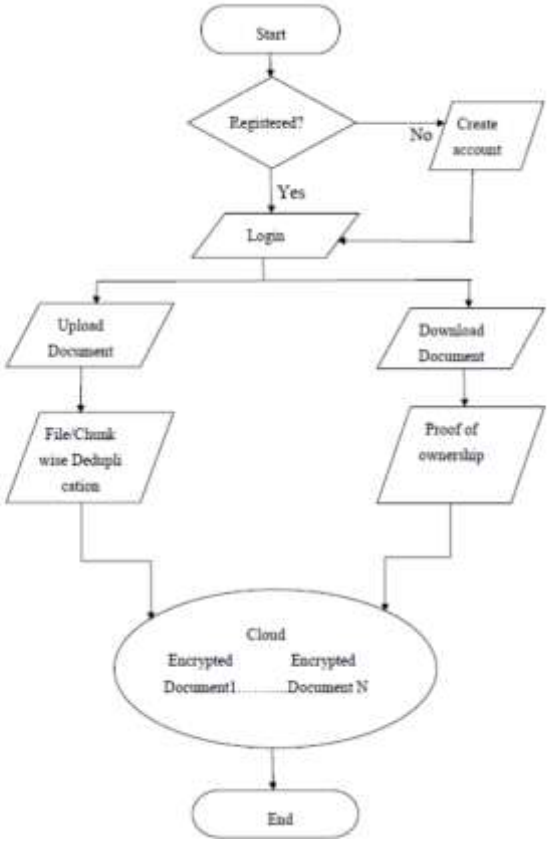


Fig-1: Flowchart of System

The proposed system detects duplicate files or data stored on the cloud. Security and privacy of the file is maintained by storing the files in encrypted format. This system will result in storing only those files which have unique content. For developing this system, we have used some hashing algorithm as well as encryption algorithm. The idea behind this project is to save the space on cloud and also the time of the user. The overview of the proposed reduplication system consists of the following three phases: authorized reduplication (Phase 1), proof of ownership (Phase 2), and encryption (Phase 3).

System construction shown in the flowchart describes the working of the system. The working of the proposed reduplication system is mainly divided into three phases as authorized reduplication, proof of ownership, and encryption.

In Phase 1 that is in authorized reduplication, first it will check if the user is authorized or not. The authorized user can only upload or download the documents/files from the cloud. While uploading the file, the system will check whether the content of the file is similar to those files which are already present on the cloud. This is performed using Deduplication technique. Data reduplication or Single Instancing essentially refers to the elimination of redundant data. Data reduplication is one of the transpiring techniques that can be used to enhance the use of existing storage space to store a large amount of data. Basically, data reduplication is removal of redundant data. In the proposed system, reduplication is performed page wise. The system will check for the duplicate file.

Following are the steps to perform reduplication:-

1. Divide the input data into blocks or "chunks."
2. Calculate a hash value for each block of data.
3. Use these values to determine if another block of the same data has already been stored.
4. Replace the duplicate data with a reference to the object already in the database.

We have performed Data Deduplication using SHA-256 Algorithm. This algorithm consists of the following steps:

Step 1: Append padded bits

The message is filled so that its length is congruent to 448, modulo 512. This padding is single 1 bit added to the end of the message, followed by as many zeros are required so that the length of bits equals 448 modulo 512.

Step 2: Append length

A 64-bit representation of the message's length is appended to the result. This step is to make the message length an exact multiple of 512 bits in length.

Step 3: Parsing the message

The padded message is parsed into N 512-bit message blocks, $M^{(1)}$, $M^{(2)}$,..., $M^{(N)}$,by appending 64-bit block. Step 4: Initialize Hash

Value The initial hash value, $H^{(0)}$ is set, consist of eight 32-bit words, in a hexadecimal form.

Step 5: Prepare the message schedule

SHA256 uses a message schedule of sixty-four 32-bit words. The words of the message schedule are labeled W_0, W_1,..., W_{63} [1].

$$Wt = \begin{cases} M_t^{(t)} & 0 \le t \le 15 \\ \sigma_1^{(256)}(W_{i-2}) + W_{i-7} + \sigma_0^{(256)}(W_{i-15}) + W_{i-16} & 16 \le t \le 63 \end{cases}$$

Where:

$$\sigma_1^{(256)}(W_{i-2}) = ((W_{i-2})\ ROTR\ 17) \oplus ((W_{i-2})\ ROTR\ 19) \oplus ((W_{i-2})\ SHR\ 10)$$

$$\sigma_0^{(256)}(W_{i-15}) = ((W_{i-15})\ ROTR\ 7) \oplus ((W_{i-15})\ ROTR\ 18) \oplus ((W_{i-15})\ SHR\ 3)$$

Fig-2: Step 5 of SHA

Step 6: Initialize the eight working variables, a, b, c, d, e, f, g, and h, with the (i-1)st hash value for t=0 to 63:

$$
\begin{aligned}
T_1 &= h + \Sigma_1^{(256)}(e) + Ch(e,f,g) + K_1^{(256)} + W_t \\
T_2 &= \Sigma_0^{(256)}(a) + Maj(a,b,c) \\
H &= G \\
G &= F \\
F &= E \\
E &= d + T_1 \\
D &= C \\
C &= B \\
B &= A \\
A &= T_1 + T_2
\end{aligned}
$$

Where:

$$
\begin{aligned}
\Sigma_1^{(256)}(e) &= (e\ ROTR\ 6) \oplus (e\ ROTR\ 11) \oplus (e\ ROTR\ 25) \\
\Sigma_0^{(256)}(a) &= (e\ ROTR\ 2) \oplus (e\ ROTR\ 13) \oplus (e\ ROTR\ 22) \\
Ch(e,f,g) &= (e \wedge f) \oplus (\sim e \wedge g) \\
Maj(a,b,c) &= (a \wedge b) \oplus (a \wedge c) \oplus (b \wedge c)
\end{aligned}
$$

Fig-3: Step 6 of SHA

Step 7: Output

After repeating steps one through four a total of N times, the resulting hash function is:

$$H_0^{(N)} \,||\, H_1^{(N)} \,||\, H_2^{(N)} \,||\, H_3^{(N)} \,||\, H_4^{(N)} \,||H_5^{(N)} \,||H_6^{(N}$$

Fig-4: Step 7 of SHA

After all input blocks from W^i have been used and we $\omega(63)$ has been created, we can create the new hash H^i such that each input block of H^i is the sum of the corresponding input block of H^{i-1} plus the corresponding input block of $\omega^i(63)$.

$H^i(j) = H^{i-1}(j) + \omega^i(63)(j)$ *where + is the addition modulo* 2^n. If other message blocks M^i remain, repeat the process if W^i was the last message schedule, then $H^i = H$ is message M's final hash or digest.

While downloading the file the authorized user will get a verification code sent on email. The email contains the primary key which helps in decrypting the file. When the user wants to download a file the phase 2 that is proof of ownership is executed. The file will download only using the primary key; this ensures the security of the documents/files.

Following the phase 1, if the file is not identical then phase 3 that is encryption is performed to store the file on the cloud. Encryption is a method that encodes a message or file so that it can be only be read by certain people. The proposed system uses Rivest cipher-6 algorithm for encryption and decryption.

There are two kinds of cipher among which RC6 is (Rivest Cipher 6) is a symmetric block cipher. RC6 proper has a block size of 128 bits and also supports various key sizes of 128, 192, 256 bits respectively. It uses data-dependent rotations, modular addition, and, XOR operation. It also provides better security than RC5.The algorithm use 4 registers (each of 32 bits). Due to use of fewer rounds it offers higher throughput.

The algorithm includes three steps: pre-whitening, post- whitening and loop of similar rounds. Like RC5, RC6 consist of three components such as encryption algorithm, Decryption algorithm, and a Key expansion algorithm. The other following specifications are RC6: w/r/b where w for word size, r for non-negative no rounds and b for byte size of encryption key. The seven primitive operations used in algorithm are: addition, subtraction, and multiplication operations use two's complement representations etc. yet the parallel assignment operation is primitive and fundamental. The block encryption process is depicted in above figure.

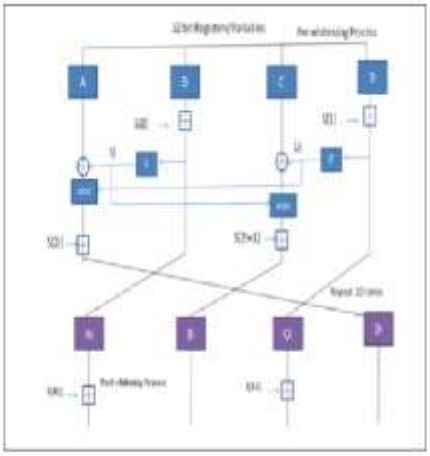


Fig-5: RC-6 Working

1)First the registers B & D go through pre-whitening loop, later the round r, which is designated by for loop in fig. Then registers B & D undergoes in quadratic equation and rotate ($\log_2$ w) bits towards left {t = (B x (2B + 1)) <<< $\log_2$ w, u = (D x (2D + 1)) <<< $\log_2$}.

2) The consequent value of B is XOR with A and D with C. This value t is added to round keys[2i] by rotating left u

bits, and occurring values of D& C are added to round key S[2i+1] by rotating left { A = ((A ⊕ t) <<< u) + S[2i] ,C = ((C ⊕ u) <<< t) + S[2i+ 1] (A,B,C,D) = (B,C,D,A)}. In the terminating stage of rounds, the values are muted by applying parallel assignment, to admit the AB calculation with CD, which increases the cryptanalytic intricacy. Lastly, registers A & C go through post-whitening loop. Yet encryption and decryption are same some differences occur.

3) Now the decryption process begins with pre- whitening loop for C& A. It runs in opposite direction for r rounds {(A, B, C, D) = (D, A, B, C) u = (D x (2D + 1)) <<< $\log_2 w$, t = (B x (2B + 1)) <<< $\log_2$ w}. Firstly, parallel assignment is applied and then quadratic equation is applied on D&B. The round key S[2i+1] is deducted from C value and final answer is rotated t bits. The round key S[2i] is removed from A which is rotated u bits. Thus registers C&A involve XOR operation with u and t respectively { C = ((C - S[2i + 1]) >>> t)⊕ u , A = ((A - S[2i]) >>> u)⊕ t}. After terminating the loop, D & B go through a post-whitening loop.

The major agenda is to provide secure encrypted data with authorized deduplication in cloud. For this we have used RC6 algorithm. For secure data transposal cryptography is very important factor. The encrypted data is more protected to relay on insecure network and any illegal user won't be able to interpret the data if the user does not have a secret key to decrypt the message. For this plain text is converted into cipher text and after getting the key the message becomes readable for the user. For this RC6 algorithm is used.

There are several modules given to explain the system and the working procedure of the system how it will work.

1. Cloud Admin module

In this module the cloud admin can see the request of the users that want to use the service of the cloud. The cloud admin can accept/cancel the pending request, and create login credentials for the user. The cloud admin can also monitor the usage of the cloud by any user and also calculate the rent according to the usage. The cloud admin can track the payment done by the user in this module.

2. Company Admin module

In this module the company admin can request for membership to the cloud admin. Once the request is accepted, company admin can login from the given credentials.

The company admin can upload/download documents. Admin can specify the access permission attributes of the uploaded documents and can also change the access permission attributes. The company admin can track the usage and can make the payments according to the usage.

3. Employee module

In this module the employee can login/logout, upload/download document. The employee will get a secret key on email. By using it the document can be decrypted.

4. PERFORMANCE ANALYSIS

The Deduplication system is mainly developed to save space on the cloud. This system ensures that only single copy of a data gets stored on the cloud to save the space.

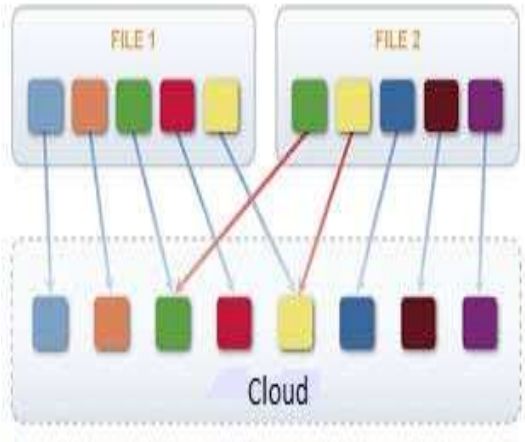


Fig-6: Overview of Deduplication

Above figure shows the working of Deduplication. Both the files contain some data. Consider that file 1 is already uploaded on the cloud and its entry in the database is also done. Now another user wants to upload file 2. This file contains some unique data and also some matching data that was already uploaded by the previous user. The system will now check for deduplication and will only upload unique data on the cloud. It will also create entry of unique data in the database. The duplicate data will not be uploaded on the cloud but in the database it will create a pointer to the data which is already uploaded.

reldocid	EncryptedDocSize	TotalDuplicationCount	savedSpace_in_KB
1010	214	6	1284
1017	100	1	100
1021	150	2	300
1024	170	1	170

Fig-7: Space Analysis

This is how Deduplication system will save space. This system will not only save space but it will also save time. The execution time of the system is comparatively less than regular time.

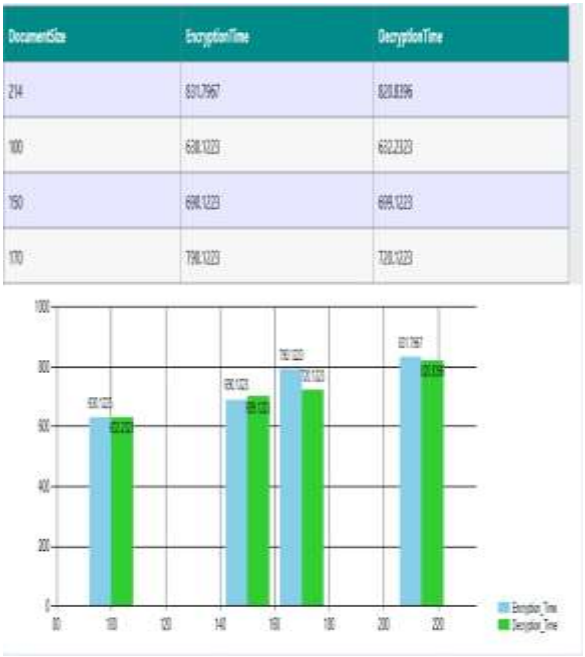


Fig-8: Time Analysis

5. ALGORITHM COMPARISON

Data Deduplication is a data reduction approach that not only reduces storage space, by reducing duplicate data but also minimizes the transmission of the redundant data which is divided into blocks, for each block hash value is generated. This is produced by SHA-256 Algorithm which is a member of SHA-2. It is a cryptographic hash function with digest length of 256 bits.

SHA-256 is essentially a 256-bit block cipher algorithm which encrypts the intermediate hash value using the message block as key. Hence there are two main components to describe they are the SHA-256 compression function, and the SHA-256 message schedule.

SHA-1 is prone to length extension attacks. In cryptography, SHA-1 is a cryptographic hash function which takes an input and produces a 160-bit (20-byte) hash value known as a message digest which is typically rendered as a hexadecimal number, 40 digits long.

TEST STRING	SHA - 1	SHA 256
hash value	d79c69966efe6297762 8f804bdaa8d0b823e09 e7	d13baa5b91ea95462b1d26 b3a3b1874b6be955af5a96 30d1d1d0ea9bb981bf0e
password	5baa61e4c9b93f3f0682 250b6cf8331b7ee68fd8	5e884898da2804715 1d0e56f8dc6292773603d0d6aab bdd62a11ef721d1542d8
cryptography	48c910b6614c4a0aa58 51aa78571dd1e3c3a66 ba	e06554818e902b4ba339f0 66967c0000da3fcda4fd7eb 4ef89c124fa78bda419

Fig-9: Comparison of example execution of SHA-1 and SHA-256.

SHA-1 and SHA-256 are of cryptographic functions designed to keep data secured and each of which was successively designed with increasingly stronger encryption in response to hacker attacks. SHA-1 has these two steps,

Step 1- Bits Padding:-

Add Padding to the end of the genuine message length is 64 bits and multiple of 512,

Step2- Appending length: -

In this step the excluding length is calculated. Whereas, SHA-256 consist of above two steps and additional more five steps they are

Step 3: Parsing the message,

Step 4: Initialize Hash,

Step 5: Prepare the message schedule,

Step 6: Initialize the eight working variables, a, b, c, d, e, f, g, and h, with the $(i-1)^{st}$ hash value for t=0 to 63,

Step 7: After repeating steps one through four a total of N times, the resulting hash function is generated with a formula.

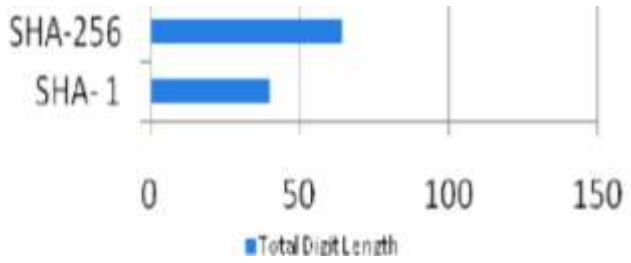
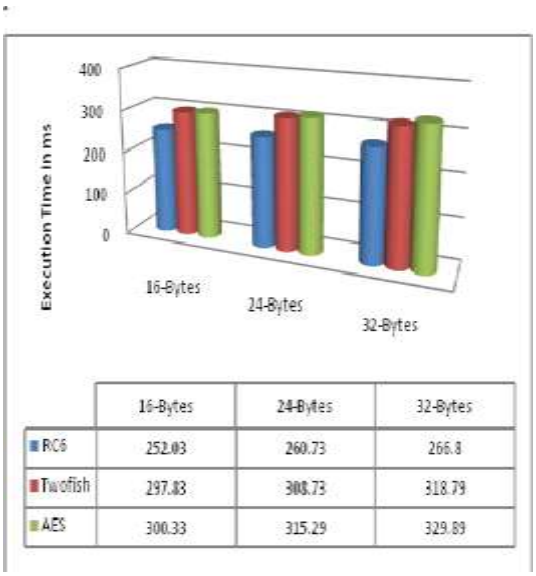


Fig-10: Comparison of length of the output digit between SHA-1 and SHA-256.

The three strings that are the hash value, password and cryptography are taken as sample string for generating massage digest for SHA 1, SHA-256. The above results show that SHA-256 is more secure than SHA-1.

Different algorithms which plunge in this category are Cipher Feedback mode (CFB), Counter mode (CTR), Galois Counter Mode (GCM) and algorithms which covered by this category of Symmetric key cryptography are: RC1, RC2, RC4, RC5, RC6, AES, DES, 3DES, CAST5,Twofish etc. This all algorithm is similar like RC6 but, the highest Speed/Area ratio can be achieved more in RC6.Hence this algorithm is far better and simplest to use and provides high performance and better security. The following fig below shows the comparison of three algorithms.



	16-Bytes	24-Bytes	32-Bytes
RC6	252.03	260.73	266.8
Twofish	297.83	308.73	318.79
AES	300.33	315.29	329.89

Fig-11: Comparison of RC6, Twofish, AES

The integer multiplication used in RC6 for 32 bits is more efficient.RC6 is 1.182 times quicker than Twofish algorithm and 1.91 times quicker than Rijindael algorithm for 16 bytes' key. The security provided by the RC6 algorithm is much secured.

6. CONCLUSIONS

Managing encrypted data with deduplication is essential and significant in practice for achieving a successful cloud storage service, primarily for data storage. In this paper, we scheduled a practical scheme to manage the encrypted data in cloud with deduplication based on ownership challenge. Our scheme can openly support data update and sharing with deduplication. Encrypted data can be securely approached because only certified data holders can obtain the symmetric keys used for data decryption. Thus this paper compresses the data by deleting the duplicate copies of equivalent data and it is widely used in cloud storage to save bandwidth and minimize the storage space. Deduplication eliminates duplicate data stored on Cloud. Thus by reducing the storage usage, cost will also be reduced automatically. To secure the privacy of sensitive data during deduplication, the encryption technique is used to encrypt the data before outsourcing.

REFERENCES

[1] S. G. Pundkar, G. R. Bamnote "Secure Sharing of Personal Records in Cloud using Encryption" Global Journal of Engineering Science and Researches May 2015.

[2] J. Li, C. Qin, P. P. C. Lee, and X. Zhang, "Information leakage in encrypted deduplication via frequency analysis,'' in Proc. 47th Annu. IEEE/IFIP Int conf. Dependable Syst. Netw., Jun. 2017, pp. 1_12.

[3] D. Harnik, B. Pinkas, and A. Shulman-Peleg, "Side channels in cloud services: Deduplication in cloud storage,''*IEEE Security Privacy*, vol. 8, no. 6, pp. 40_47, Nov./Dec. 2010.

[4] R. SHOBANA, K. SHANTHA SHALINI, S. LEELAVATHY and V. SRIDEVI "De-Duplication of Data in Cloud" Int. J. Chem. Sci.: 14(4), 2016

[5] J. Blasco, R. Di Pietro, A. Orfila, and A. Sorniotti, „„A tunable proof of ownership scheme for deduplication using bloom filters,"" in Proc. IEEE Conf. Commun. Netw. Secure. (CNS), Oct. 2014, pp. 481–489.

[6] W. K. Ng, Y. Wen, and H. Zhu, „„Private data deduplication protocols in cloud storage,"" in Proc. 27th Annu. ACM Symp. Appl. Comput., 2012, pp. 441–446.

[7] J. Xu, E.-C. Chang, and J. Zhou, „„Weak leakage-resilient client-side de-duplication of encrypted data in cloud storage,"" in Proc. 8th ACM SIGSAC Symp. Inf., Comput. Commun. Secure, 205, pp. 195–206.

[8] M. W. Storer, K. Greenan, D. D. E. Long, and E. L. Miller, "Secure data deduplication,'' in *Proc. 4th ACM Int Workshop Storage Secure. Survivability*, 2008, pp. 1_10.

[9] Y. Shin and K. Kim, "Differentially private client-side data deduplication protocol for cloud storage services,'' *Secure. Commun. Netw*, vol. 8, no. 12, pp. 2114_2123, 2015.

[10] J. Li, Y. K. Li, X. Chen, P. P. C. Lee, and W. Lou, "A hybrid cloud approach for secure authorized deduplication,''

[11] *IEEE Trans. Parallel Distrib. Syst.*, vol. 26, no. 5, pp. 1206_1216, May 2015.

[12] Y. Zhang, X. Chen, J. Li, D. S. Wong, H. Li, and I. You, "Ensuring attribute privacy protection and fast decryption for outsourced data security in mobile cloud computing,'' *Inf. Sci.*, vol. 379, pp. 42_61, Feb. 2017.

[13] S. Halevi, D. Harnik, B. Pinkas, and A. Shulman-Peleg, ``Proofs of ownership in remote storage systems,'' in *Proc. 18th ACM SIGSAC Conf. Comput. Commun. Secure*, 2011, pp. 491_500.

[14] H. Kwon, C. Hahn, D. Kim, and J. Hur, ``Secure deduplication for multimedia data with user revocation in cloud storage,'' *Multimedia Tools Appl.*, vol. 76, no. 4, pp. 5889_5903, 2017.

[15] J. Hur, D. Koo, Y. Shin, and K. Kang, ``Secure data deduplication with dynamic ownership management in cloud storage,'' *IEEE Trans. Knowl. Data Eng.*, vol. 28, no. 11, pp. 3113_3125, Nov. 2016.

[16] W. Ding, Z. Yan, and R. H. Deng, ``Secure encrypted data deduplication with ownership proof and user revocation,'' in *Proc. Int. Conf. Algorithms Archit. Parallel Process.* Cham, Switzerland: Springer, 2017, pp. 297_312.

[17] S. More, P. Shinde, S. Shaikh, V. Gunjal, S. Chavan, A. Kalia, V. Kolhe, "Secure Data Retrieval in Ad-Hoc Network using RC6 Algorithm" IJESC Vol. 6 issue no. 5, 2016.

[18] W. Xia, H. Jiang, D. Feng, F. Douglis, P. Shilane, Y. Hua, M. Fu, Y. Zhang, and Y. Zhou, ``A comprehensive study of the past, present, and future of data deduplication,'' *Proc. IEEE*, vol. 104, no. 9, pp. 1681_1710, Sept 2018.

[19] S. G. Pundkar, Dr. G. R. Bamnote "Access of Encrypted Personal Record in Cloud" International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169 Volume: 3 Issue: 1.

[20] C. Yang, J. Ren, and J. F. Ma. Provable ownership of file in deduplication cloud storage. Proc. IEEE Global Common. Conf., 2013.

[21] MihirBellare, SriramKeelveedhi, Thomas Ristenpart. Message-Locked Encryption and Secure Deduplication. A preliminary version of this paper appears in the proceedings of Euro crypt 2013. This is the full version. March 2013.

[22] Jiawei Yuan, Shucheng Yu, Secure and Constant Cost Public Cloud Storage Auditing with deduplication IEEE conf. Oct 2013

[23] Moheb R. Girgis, Alaa I. El-Nashar, Tarek A Abd El-Rahman, Marwa A. mohammed "An ASP.NET Web Applications Data Flow Testing Approach" International Journal of Computer Applications Volume 153 No 8 November 2016

[24] TengYingyang, Ye Zing gang, "Design and Implementation of Stored Procedure Router Based on Dynamic SQL" 2009

[25] *Chun-e Zhang, Hui-xian Tao, Li-juan Yang, "The Study of Research-Oriented Learning Platform Building Based on Distributed Database"*

[26] Kameswari Bhaskar, R. Sathiyavathi ,Jayashree R, L. Mary Gladence, V. Maria Anu, "A NOVEL APPROACH FOR SECURING DATA DE-DUPLICATION METHODOLOGY IN HYBRID CLOUD STORAGE"ICIIECS