Securing Cloud from Tampering and Duplication

Mr. Pratik Sawarkar
Computer Science & Engineering
NIT College of Engineering
Nagpur,India
pratik.sawarkar158@gmail.com

Ms. Sheetal Singh
Computer Science & Engineering
NIT College of Engineering
Nagpur,India
shits.singh.27@gmail.com

Ms. Priti Nitnaware
Computer Science & Engineering
NIT College of Engineering
Nagpur,India
priti.nitnaware@gmail.com

Ms. Rasika Tiwari
Computer Science & Engineering
NIT College of Engineering
Nagpur,India
rtiwari1795@gmail.com

Ms. Akriti Shrivastava
Computer Science & Engineering
NIT College of Engineering
Nagpur,India
akritishrivastava08@gmail.com

Ms. Pooja Dubey
Computer Science & Engineering
NIT College of Engineering
Nagpur,India
poojapddubey@gmail.com

***Abstract*—**Cloud computing is the most emerging technology today which is used by most of the social media sites to store the data. The data stored on the cloud is private data of the user so it must not be tampered by other entities. The previous system has worked on reducing the storage space by copying and archiving data but on the cost of reduced performance rate. We propose a system to enhance the storage space by performing deduplication on data and shuffling the data,between the number of directories within cloud after particular interval of time to avoid the tracking of data to enhance the security. The backup of the data will be taken timely into the back up directory. The proposed system will provide ease to use the cloud.

Keywords*- Tamperin, Deduplication, Shuffling, Security, Encryption.*

*****

I. INTRODUCTION

Nowadays with increase in population of people using the technology the use of storage space required is also increased. Many data providing sources are available like smartphones, tablets, pcs, and etc. Many of us now use more than one device to store data which we wish to retrieve at anytime from anywhere. Previously the technology that was used were connecting the sender and receiver through the physical means and hence the need of new technology became essential and resulted into the development of cloud. Nowadays with increase in population of people using the technology the use of storage space required is also increased. Many data providing sources are available like smartphones, tablets, pcs, and etc. Many of us now use more than one device to store data which we wish to retrieve at anytime from anywhere. Previously the technology that was used were connecting the sender and receiver through the physical means and hence the need of new technology became essential to a shared pool of configurable computing possessions such as servers, storage and applications.

II. PROBLEMS IN EXISTING SYSTEMS

File Confidentiality: During maintaining the file confidentiality, the data stored must not be allowed to be altered even by the administrator. The data thus generated must be protected from dictionary attacks design goal of file confidentiality requires preventing the cloud servers from accessing the content of files. The file should be stored in such a way that the data could not be tracked which is stored on cloud.

Secure Deduplication: Deduplication is a technique which prevents storage of any data which is already available instead of storing the redundant copies. However, the insiders are capable of leaking channel information. For example, a server passing the information to the client that the data which he wants to store is already available on the server and the data can either be a confidential one.

Encryption & Decryption: Encryption and decryption provides additional security to the data which needs to be on the stored data. There are two features of stored. A key is obtained from the data content and the encryption is performed on that data,in combination with the key. A label is attached to identify it as a duplicate copy.

Integrity Auditing: The main reason of implementing the integrity auditing is to provide the capability of verifying correctness whether the remotely stored copies are same, then their tags are same or not. Formally, a convergent encryption scheme is applied verifying the integrity 1. Public verification-in these verification is performed by anyone, not just the clients; 2.Stateless verification-it doesn’t require state information maintenance at the verifier side between the actions of auditing and data storage.

III. TAMPERING AND DUPLICATION

A) Tampering

The web application attacks that involved soliciting a website with manually entered data to generate an unexpected context. The protocol used for the communication on the web, HTTP protocol enables to convey parameters in the form of requests; it is done in the following several ways:

- Cookies;
- Form fields;
- URLs;
- HTTP headers.

It is very much important to understand that all these data transmission methods stated above can easily be manipulated .This manipulation is done by user and therefore, user data should not be considered reliable. Hence, security cannot be based on client verifications.

The data stored on the cloud can be tampered easily which results into the integrity loss. Because of these tampering, the location of the data where it has been stored can easily be tracked. If the location of the data on the cloud is known to any of the intruder then there may be possibility of loss of integrity of the data.

B) Deduplication

The data deduplication technique involves tracking of each data file and eliminate the file if more than one copy of it is found in the storage . It is the most adopted technique in minimizing the storage space utilization. The deduplication of data is very important for the shared storage. Deduplication can also be termed as a data reducing technique. Unlike the compression ,in which the data is compressed and all the data is kept in the compressed form .There are several ways to deduplicate the data.

IV. DATA DEDUPLICATION

There are number of ways to detect the duplicate data and remove it. All this leads to the point to reduce the size in order to save the storage. FIGURE I Shows the strategies that are used for data deduplication.

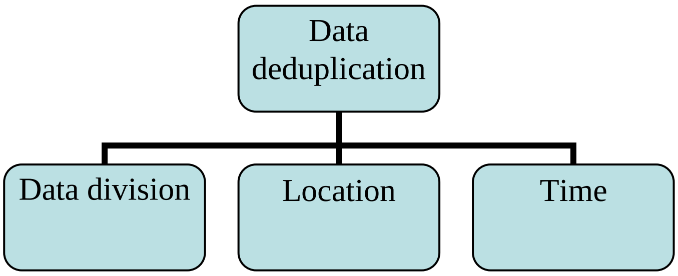

FIGURE I: Deduplication strategies

A. DATA DIVISION

In this method , data is divided into the sequence of bytes, then the blocks that are obtained used to test the redundancy .In deduplication only the unique block is stored There are three types of data division technique-

1. Whole file pass: In this procedure, the whole data is passed without dividing it into number of blocks. The compression is done with the hash indexed in the file if it matches, it counts it as a duplication.

2- Fixed size dividing: In this procedure, the data is divided into equal block sizes, which results into the fixed size block For example -2Kb, 4Kb, etc. To check whether there is any duplication or not, checksum is used in which only the unique checksum is stored in the storage. The drawback of this method is if the size of the data is large then it will divide it into a number of segments which results into the possibility of numerous errors.

3-Variable size dividing: In this method, the boundaries are decided according to the data size. The difference between this method and the fixed size method is that in this method the boundaries are not fixed. This method is efficient than the two previously given methods. This algorithm is the best for backup

B. LOCATION

In cloud, data can either be stored at the client side or at the server side .These are the two locations where the data can be stored and the deduplication is performed based on the locations at which the data has been stored

1- Client side: It is also called as the source. A special program is applied to detect the duplication on the database of the client to carry out the deduplication process. The main advantage of performing the deduplication at the client side is the less usage of bandwidth. Bandwidth is saved as only the unique data will be stored in the cloud.

2- Server side: In this, deduplication process takes place on the cloud servers .Once the data has been stored, then the server will handle and sort the data. The server then find the duplications and eliminate them. The main advantage of this procedure is reducing the number of overheads from clients.

C. TIME

Time plays the vital role in the field of processing and computing. If the duplicate files are eliminated, speed is made higher, then the processing time will be less. There are two types of deduplication methods that depends on the time. The first one is before storing the data to storage and the second one is after storing the data.

Before storing data and after storing data are also known as inline process and post process repeatatively.

V. PROPOSED METHODOLOGY

The proposed system works for reducing the need of storage and remote accessibility to the data. This requires a great amount of attention towards the security of the data. As cloud is now extensively used, the security of cloud becomes a key point of which care should be taken. Businesses that experience a data breach must inform their consumers if the hacker had the access to their personal information Operating system and application files are stored on a common physical infrastructure in a virtualized cloud environment and require system, file, and activity monitoring to offer the assurance and auditable evidence to enterprise customers that their resources have not been compromised or tampered with. Enterprises are often required to prove that their security compliance is in accord with policy , principles, and auditing practices, in spite of of the location of the systems at which the data resides. In cloud computing ,data is fluid and may exist in on-premises physical servers, on-premises virtual machines, or off-premises virtual machines that runs on cloud computing resources, and this will require some rethinking on the part of auditors and practitioners alike. Localized virtual machines and physical servers use the same operating systems , enterprise and web applications available in a cloud server environment, increasing the threat of an attacker or malware exploiting weakness in these systems and applications distantly. Virtual machines are vulnerable as they move between the private and the public cloud. A fully shared or partially shared cloud environment is expected to have a greater attack plane and as a result can be considered to be at greater risk than a dedicated resources environment. Many cloud servers hide the intruder's attempt to alter or to leak the confidential data to maintain their market value among the clients.

The increasing usage of cloud has increased the use of space needed. Among many files which are stored on cloud most of them are redundant files. This fact raises a technology namely de duplication.

Data de duplication is a specific technique for data compression, eliminating duplicate copies of already available data in cloud computing. This specific technique is used to help the efficient usage of storage available and can also be applied to network data transfers to decrease the amount of bytes that must be sent. In this deduplication process, unique chunks of data, or byte patterns, are recognized and stored during the analysis process. As the analysis continues, other chunks are compared to the already available copy and each time a match occurs, the redundant chunk is replaced with a small reference that points to the chunk that has been stored.It is known that the similar byte pattern might arise dozens, hundreds, or even thousands of times (the match frequency is rely on the size of the chunk), the amount of data that must be stored or transferred is large.

A. DEDUPLICATION STRATEGIES

Data de duplication technology is the technique to identify the data which seems to be duplicate and removal of the same which will reduce the memory utilization .Detection of the duplicate data is performed by checking the bits, file or the block. Data de-duplication technology uses mathematical logic to identify the duplicate data. One of these logics is to use "Hashing Algorithms". A hash index is a list in which every number is compiled. At present mainly the file level, block-level and byte-level deletion approach can be used for optimizing the storage capacity. [5]

i. File-level data de duplication strategy

File-level de duplication approach is also called as Single Instance Storage (SIS), checking the index back or archiving the file needs a list of comparisons between the available data .Updates on index should be performed regularly After small period of time. In this, index works as a remnant thorough which the data is accessed by assigning the pointers.

ii. Block-level data de duplication technology

In the block-level data de duplication technology, the data stream is divided into multiple blocks[0][1]. In this checking of the data block is performed , and determines if it has already met the same data before. If the block obtained is a new and was written to disk, then its id also needs to be stored in the index. This method pointer with a small-capacity alternative to the duplication of data blocks, rather than storing the repeated data blocks all over again and hence saving the available disk storage space. The hash value generated from each of the data can lead to conflict when hashing algorithm are applied(1).In hashing algorithms, the data blocks are checked to form the unique code. There are conflicts between the hash value generated but are of very less importance.

B. STUDY OF HASH ALGORITHMS

1. MD-2

MD2 generates a message digest of 128 bits.It is a cryptographic hash algorithm. It was published in August ,1989. It requires 18 rounds of its compression function to generate a 128 bit digest.The author of MD2 concludes, "MD2 cannot be considered as a secure one-way hash function anymore". In 2008, MD2 has further improvements on a pre-image attack having the time complexity of 273 compression function evaluations. In 2009, MD2 was shown to be vulnerable to a collision attack having the time complexity of 263.3 compression function evaluations.

2. MD-4

MD4 cryptographic hash algorithm generates fixed 128 bit message digest.It takes 48 rounds of its compression function. It was published in 1990.A collision attack published in 2007 can find collisions for full MD4 in less than 2 hash evaluation functions.

3. MD-5

MD-5 produces a message digest of fixed 128 bits. It performs 64 rounds. It was published in 1992. A 2013 attack by Xie Tao, Fanbao Liu, and Dengguo Feng breaks MD5 collision resistance in 218 times.

4. SHA-0

SHA-0 belongs from SHA family; it is another cryptographic hash algorithm generates a message digest of fixed 160 bits. It takes 80 rounds. It was published in 1993. A 2004 attack by Bihamet.Al breaks SHA-0 collision resistance at 241.

5. SHA-1

In 1995 SHA-1 was published.SHA-1 generates a message digest of 160 bits. It takes 80 rounds. For integrity purpose this algorithm is used frequently.Due to its time efficiency and robustness,it is it is most popular amog various hash algorithm.[2] Later on, a 2011 attack by Marc Stevens can create hash collisions with a complexity of 261 operations.

6. SHA-2

SHA-2 is a collection of different hash functions i.e. SHA-224, SHA-256, SHA-384 and SHA-512. None of them have proven completely breakable but still these algorithms are not preferred to ensure the integrity because they are not time efficient as SHA-1. It is found that, none of the hash algorithm is secure to ensure the integrity except SHA-2 but it is found that it is not time efficient. Many researchers have found these problems and proposed their own algorithms as a solution.

7. SHA-192

In 2009 SHA-192 was proposed. The authors of SHA-192 have proposed its own compression function which is similar to SHA-1, the only difference in SHA-1 and SHA-192 is that SHA-192 uses 6 chaining variable of 32 bits in its compression function which generates 192 bits output. To produce greater bit difference, advantagious properties of MD-5 and SHA-1 are combined. So the new algorithm SHA192 will be no longer susceptible to the collision. Here, A, B, C, D, E, F is the chaining variable. Each chaining variable holds 32 bits information. Initially all the chaining variable initialized with some value and during processing it changes its value and hold processing results and at last generates a result of 192 bits message digest.

8. SHA-192[1]

SHA 192[1] is another hash algorithm proposed in 2013. In this authors have proposed a new compression function to generate a message digest of 192 bits. Authors have combined the compression function of MD-5 and SHA-192 and take 64 rounds of compression function for each 512 bits message block

Algorithm name	Size of output	Rounds
MD-5	128	64
SHA0	160	80
SHA1	160	80

TABLE I: Comparison between different hash algorithm

C. FLOW CHART OF DEDUPLICATION

The flowchart mentioned below explains us the following operations performed. Here any file which needs to be uploaded on cloud first goes through the hashing value calculation procedure. For every hash value generated a hash index table is created from which the hash values are matched. Every data generates a unique hash code by matching this hash code value with the other data already available on cloud from the index table. It can be considered as a duplicate file if the code matches. Assigning pointer can be done to that data which is already available .Pointer helps in directly accessing the data. If the hash code value does not matches with any of the previously generated hash code value then that file only will be stored on cloud .This will reduce the storage space requirement for storing any file on cloud by removing the redundant data. This de duplication leads to fast processing of data during retrieval ,fetching and storage of data .The time required to store the data which is already available is also reduced by using pointers rather than storing the data physically.

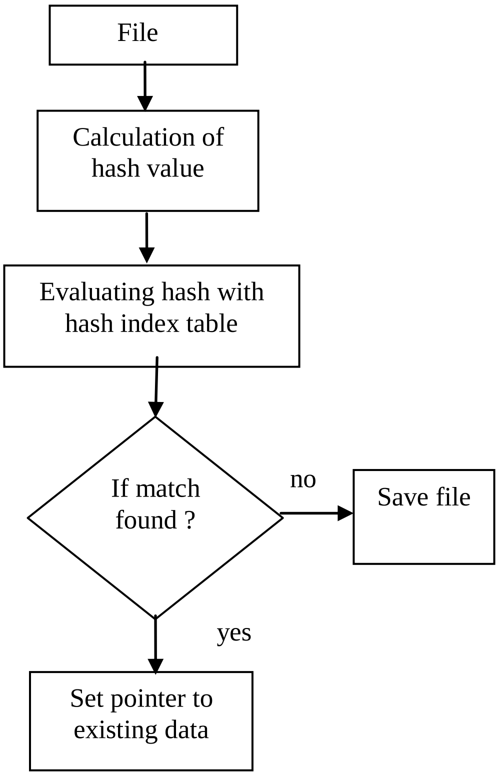
File
Calculation of hash value
Evaluating hash with hash index table
If match found ?
no
Save file
yes
Set pointer to existing data

FIGURE II: Flowchart of deduplication

VI. AES ENCRYPTION

Encryption technique is used to provide security to the data stored on cloud. In the below mentioned figure, the various approaches of hashing are represented and their usage percentage are described. There are various encryption techniques that are currently being used as follows:

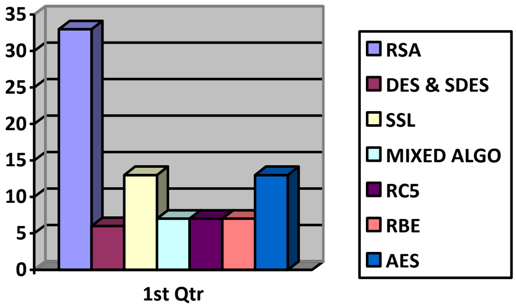

FIGURE III: Usage of various encryption techniques

RSA,DES,SDES,SSL 128 bit encryption, Mixed encryption algorithm, RC 5,RBE,AES.

It was originally called ‘Rijndael Cipher’ after the names of the developers .It was an entrant in a competition held by NIST in 1996 to find the new secured encryption method.

AES encryption is a method for scrambling data . A key is used to mix up data such that it can be securely stored or transferred over a network and only persons with the key can unscramble the data.

It is a symmetric key encryption algorithm. This means that the same key is used to scramble and unscramble the data.

AES algorithm is created by Vincent Richmen, John Daemen in 2001.Its key length is 128,129 and 256 bits and its block size is 128 bits.It is faster than any other algorithm.It provides excellent security

AES is a block cipher which encrypts 128 bits of data at a time. It treats the 16 bytes as a grid of 4*4.Messages which are longer than 128 bits are broken into blocks of 128 bits. Each block is encrypted separately using exactly the same steps.

If the message is not divisible by the block length, then padding is appended. To encrypt the message, we supply the message with a key. The AES encryption algorithm outputs unrecognizable data. To decrypt the message ,we supply the scramble the data and the sane key as before.

A. AES ANALYSIS

In present day cryptography, AES is widely adopted and supported in both hardware and software .Till date, no practical cryptanalytic attacks against AES has been discovered. AES has built in flexibility of key length which allows a degree of future proofing against progress in the ability to perform existive key searches. AES security is assured only if it is correctly implemented and good key management is employed.[1]

VII. SHUFFLING

Shuffling is a very common form of data obfuscation .It is a new technique that is introduced to avoid unauthorized access, tracking, tampering of data. There are various modes of creating temporary data and changing its location from on premises databases to the cloud.

Shuffling of data can be implemented by providing timestamps as soon as the data enters into the cloud. For that particular timestamp only, the data stays in a particular server.As soon as the time elapses,the data is moved onto some other servers.By implementing this methodology, we can avoid the tampering of data on the cloud.

VIII. CONCLUSION

In this study, the way of reducing the cloud storage are discussed. Deduplication is one of the various techniques used for optimization. Encryption methods are used to enhance the overall security of the data. The integrity of data is better maintained using encryption. The concept of using keys makes the data difficult to be extracted. In this, a new method is proposed depending on the time of data arrival to the cloud and shuffling the data according to timestamp allotted.Hnce this technique improves the storage capacity and performance.

REFERENCES

[1] Gurpreet Singh, Supriya, “A Study of Encryption Algorithms(RSA, DES, 3DES, AES) for Information security”, IJCA volume 67-No. 19, April 2013.

[2] Zuhair S. Al-Sagar, Mohd. S. Saleh, Aws Zuhair Sameen, “ Optimizing the Cloud Storage by Dta Deduplication:A Study”, IRJET. Vol 02 , December, 2015

[3] Akhila K , Amal Ganesh, Sunitha C , “A Study On Deduplication Techniques over Encrypted Data, ”4th International Conference on recent trends in Computer Science and Engineering.

[4] Uttam Thakore, “Survey of Security Issues in Cloud Computing, ”College of Engineering, University of Flourida.

[5] Qinlu He, Zhanhuai Li, Xiao Zhang, “Data Deduplication Techniques” , ICFITME 2010.