International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 04 Issue: 09 | Sep -2017 www.irjet.net p-ISSN: 2395-0072

IMPLEMENTATION OF DE-DUPLICATION ALGORITHM

Nirmala Bhadrappa[1], Dr. G S Mamatha[2]

[1]*Department of ISE, R.V College of Engineering, Bengaluru, India*
[2]*Associate Professor, Dept. of ISE, R.V College of Engineering, Bengaluru, India*

ABSTRACT: *Data which are duplicated can be avoided using data de-duplication technique, and these techniques are used to reduce storage space. This also helps in reducing bandwidth and can be stored on cloud storage. These de-duplications are used to secure the data and have been a challenge for keeping the data securely. To avoid miss handling of data cloud convergent encryption technique is used. The duplication of data can be treated in two different methods. Firstly we will have to address the problem proficiently and should handle a huge count of convergent keys. Secondly data resource raises i.e. security and privacy. A third-party cloud service is proposed for confidentiality of data; reliability checking by access control mechanisms can be done both internal and external. As the duplication technique improves storage space, bandwidth and efficiency will be a conflict with the convergent encryption technique. So the convergent encryption technique requires key for their respective data to encrypt. The copies of same data and will be checked for data feasibility. Convergent encryption helps in encrypting and decrypting the data using a key guaranteeing same data to be duplicated on to itself. The key generation and data encryption technique helps to hold the key and send the cipher text to cloud service provider. Thus encrypting technique is used to determine identical copies and also to create similar key and the identical cipher text hence data is stored secured and only authorized user can access the information from the cloud service provider.*

KEYWORDS: data de-duplication, convergent encryption, de-duplication efficiency, SHA algorithm

1. INTRODUCTION

As we move forward with the development of enterprise data accelerates, the task of protecting and de-duplication becomes more challenging. The individualized computing frameworks like desktops, portable PCs, tablets, advanced mobile phones have turned out to be significant stages for various users, increasing the significance of data on these gadgets. We may lose data because of system failure or at times the data might be erased consequently or we may lose the data by losing the gadget or be lost by looting of gadget, yet people have enhanced the utilization of data protection and recovery tool in their individualized computing gadgets. Storage resources like Amazon S3 and Google storage take economic advantages to store the data on the cloud storage for users. Figure1. Replicates the data reinforcement for individual stockpiling, these have been outsourced so clients can oversee information much effectively without bothering about keeping up about the reinforcement. The clouds are always centralized so that they can be easily managed efficiently and disaster free. The clouds offer offsite storage for data backup. The data reinforcement for individual stockpiling in the cloud demonstrates a geographic division between customer and the service provider. In de-duplication the data redundancy is removed by generating hash key to the respective file and later these file is divided into smaller parts based on the number of lines in the file. This smaller parts can also be called has blocks. Hash key is also created to these blocks for de-duplication check.

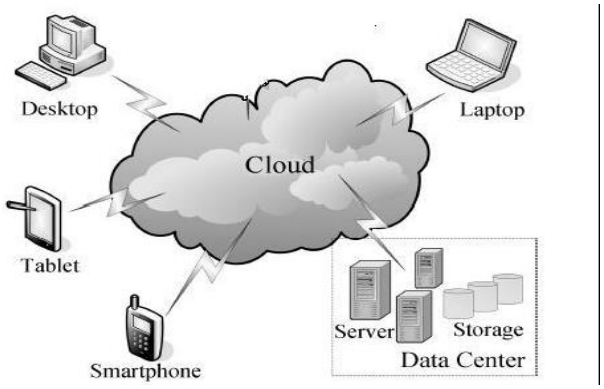


Fig-1: Shows cloud backup platform

The cloud concepts can be understood in more detail from the below:

1.1 DE-DUPLICATION

Data de-duplication is a technique of finding duplication of data in storage space. This technique used to improve bandwidth and utilization of storage that can also be used for data transfers over network and to decrease the number of bytes and file size. Data de-duplication technique identifies and removes the data which are not unique. Whenever any similar data match occurs, they are copied with a small reference. Based on file content or file name data de-duplication is performed.

Data de-duplication consists of following steps:

Step 1: Divide the input file into blocks based on number of lines in the file.
Step 2: Hash key value is generated for each block of file.

Step 3: The generated hash value is matched with stored hash value if it is matched then duplication removed.
Step 4: Duplicate data is replaced with reference object which is stored in database earlier

1.2 CONVERGENT ENCRYPTION

It is also known as content hashing. It is a cryptographic algorithm which generates alike cipher text from the plain text. Cloud computing has unique applications to remove duplication of files from storage, where the user is unaware or having rights to use to the keys. Convergent encryption is provided with authorization of a file in which an attacker can confirm whether the target possesses a specific file. The attack possesses a problem for users to store information which is publicly available/already held by the attacker. For an example consider the books which are banned or the files that cause copyright violation is a best sample for the same. If an argument could be made that a validation of a file attack is easily rendered unsuccessful then by simply adding a unique portion of data or a few arbitrary characters to the plain text before encryption; would cause the uploaded file to be unique and therefore results in a unique encrypted file. There are several samples to show convergent encryption scheme in a plain-text, which are broken down into small blocks, based on content of the text in the files, after that each block performs convergent encryption which may by mistake overcome any attempts at making the file unique by adding bytes at any place.

1.3 KEY MANAGEMENT

Key management is method for cryptographic keys in a cryptosystem to manage the files. The cryptosystem uses different types of key, so as to differentiate and compare. These keys may be asymmetric or symmetric. For both encryption and decryption of message alike keys are used in symmetric key algorithm. The selected keys must be distributed and securely stored for finding out duplication. In asymmetric key two different keys are used for encryption and decryption algorithm and they communicated with files once keys is found, it would be easy to manage them. The key management involves exchange of data, storage of data and use of them with required key.

2. FEATURES OF CRYPTOGRAPHIC

- **Confidentiality**
  Only authorized party should be able to access the information which is transmitted over network and not by the third party.
- **Authentication**
  The receiver should check the identity of the sender before accessing the information whether the information is sent from the authorized person or by an attacker
- **Integrity**
  The permission should not be given to everyone for modifying the transmitted data only an authorized party should be allowed to do so. Third party is not allowed to modify the data.
- **Non Repudiation**
  It ensures that neither the sender, nor the receiver of the information should be able to deny the transmission.
- **Access Control**
  Only approved persons will be able to access the data.

2.1 CRYPTOGRAPHIC ALGORITHM

Following are well known cryptographic algorithm

- **DES**: It stands for "Data Encryption Standard". It operates on 64 bit block of data using 56-bit key .It is symmetric key block cipher. In DES the algorithm and cryptographic key is applied simultaneously on a block of data rather than one bit at a time to encrypt a plaintext.
- **RSA**: RSA is designed by Rivest, Shamir and Adleman. It is a public-key system and it is an asymmetric cryptographic algorithm. It is used for encryption and decryption of message.
- **HASH**: A "hash algorithm" is also called as "message digest", or "fingerprint" and it is used for mapping the data size. Hash function returns the hash value. Hash value is stored in hash table
- **AES:** NIST as approved Advanced Encryption Standard and uses Rijndael block cipher.
- **SHA-1:** SHA-1 is a hashing algorithm, produces a digest of 160 bits (20 bytes) Same SHA-1 message digest is given for two different messages hence SHA-1 is recommended than MD5.
- **HMAC**: HMAC uses a key with an algorithm, algorithms such as MD5 or SHA-1 and it is one of the hashing algorithms hence in can also be referred as HMAC-MD5 and HMAC-SHA1.

3. LITERATURE SURVEY

Data de-duplication can be divided into two parts: de-duplication on unencrypted data and de-duplication on encrypted data. In the previous way, performs proof of ownership procedure in an effective and vigorous way. Though, in the last way, privacy of data is the essential security prerequisite make secure against third party as well as inside the cloud server. In this manner, the majority of the plans have been proposed to give information encryption, while as yet profiting by a de-duplication system, by empowering data proprietors to share the keys within the sight.

Data de-duplication over unencrypted data

Harnick et al. [1] exhibited how information de-duplication method can be utilized as a side channel that uncovers data to pernicious clients about the contents of documents of different clients. Halevi et al. [2] likewise presented a comparative assault situation on distributed storage that utilizations de-duplication over various clients. This is on account of just a little snippet of data about the information, in particular, its hash esteem, fills in as not just a record of the information to find data of the information among countless, yet in addition a proof that any individual whose the hash key is known esteem claims the comparing information. In this manner, any clients who can get the short hash an incentive for particular information can get to every one of the information put away in the distributed storage. Harnik et al. [3] projected randomized limits to maintain a strategic distance from an assault on distributed storage benefits that utilization server-side information de-duplication by ceasing information de-duplication. Be that as it may, their technique did not utilize customer side information ownership verifications to counteract hash control assaults. To conquer these assaults, Halevi et al. [4] presented and define the idea of proof of ownership; here the client demonstrates to a server that the document utilizing At that point, a test reaction convention between the server and the customer checks the possession. PoW is firmly identified with verification of retrievability [5] and evidence of information ownership [6]. In any case, confirmation of retrievability and information ownership regularly utilizes a pre-handling step that can't be utilized as a part of the information de-duplication strategy.

DATA DE-DUPLICATION OVER ENCRYPTED DATA

Keeping in mind the end goal to protect information security against inside cloud server and outside enemies, clients may need their information encoded. In any case, conventional encryption under various clients' keys makes cross client de-duplication inconceivable, since the cloud server would dependably observe distinctive figure writings, regardless of the possibility that the information are the same. United encryption, presented by Douceur et al. [7], is a hopeful answer for issue. Focalized encryption, an information proprietor infers an encryption key R <-F (N), where N is information or a document to be scrambled and F is a cryptographic hash function. Then, he figures the cipher text T <-P (R, N) through a piece figure P, erases N, and keeps just R in the wake of transferring C to the cloud storage. In the event that another client encodes a similar text, the alike cipher text T is created since it is deterministic. Therefore, on release of T from different clients later the underlying transfer, the server does not holds the record but rather refreshes data of data to demonstrate it has an extra proprietor. In the event that any genuine owner ask for and download T afterwards, they can encode with R. Be that as it may, united encryption experiences the following security flaws.

Xu et al. [8] likewise projected a spillage strong de-duplication plan to determine the information uprightness issue. This plan likewise empowers the information proprietor to scramble information with an arbitrarily chose key. At that point, the information encoded key is scrambled under a KEK derived from the information and conveyed to the next information proprietors after the proof of ownership procedure. On the off chance that a honest to goodness proprietor gets a cipher text, the honesty of the information can be reviewed by unscrambling the information encryption key with the same KEK.

4. WORKING

The Original Data block is chosen to out sourced into the CSP (cloud service provider). The file or block of file can be already present in cloud storage. The file that has to be uploaded to cloud service provider first checks whether the file or the block of file exists in the CSP

HASH KEY GENERATION

Based on file content hash key is generated. The tag, hash key generated is unique for each file. Key generation algorithm maps a data duplicate N to a convergent key R and it is based on the security parameter. The purpose of generating hash key is to encrypt the block of data with unique hash key.

ENCRYPTION OF FILE

Convergent encryption gives data confidentiality in de-duplication. The encrypt data, uses their own encryption keys for accessing i.e. The keys are derived from itself and hence, produce identical cipher text from identical files. Convergent encryption allows each person cloud storage in large amounts for a very low cost, It also offers privacy at its core. This privacy has concern with cloud storage services where de-duplicating data is done via convergent encryption, as de-duplication can be used to find users who are storing a file. Another reason is that it also finds if any other attacker also has a copy of the same file. For example, an oppressive government could find out users who are storing copies of banned books. The same is used to discover users who are storing copyrighted material, thus assuming direct access to the servers which is provided to the outside party. Private Key encryption is used to bypass de-duplication and hence forcing the cloud storage service to store a unique copy of the files. Here no “password reset” option is provided with convergent encryption, so if one forgets it the data will be lost for ever. The client decides a tag for the copy of data which can be utilized to recognize duplicates. The tags are same if two data copies are identical. To check whether there is any duplication, the client initially

sends the tag to server to check whether the copy exists or not. Encryption is done using advanced encryption standard algorithm. The encrypted file is stored in cloud service provider with hash key and tag.

PROCESS OF FILE UPLOADING

When user wants to upload a file "A" file level De duplication is performed first. The user computes file tag on the input file "A" Upon receiving, the auditor checks if there exists the same file with same tag on the cloud service provider. If auditor replies that there is file duplication, or there is no file duplication, if the user gets the reply that there is no file duplication, then duplication check is done with block- level. If there is file duplication then user checks for proof of ownership whether the same file "A" that is stored in the cloud service provider.

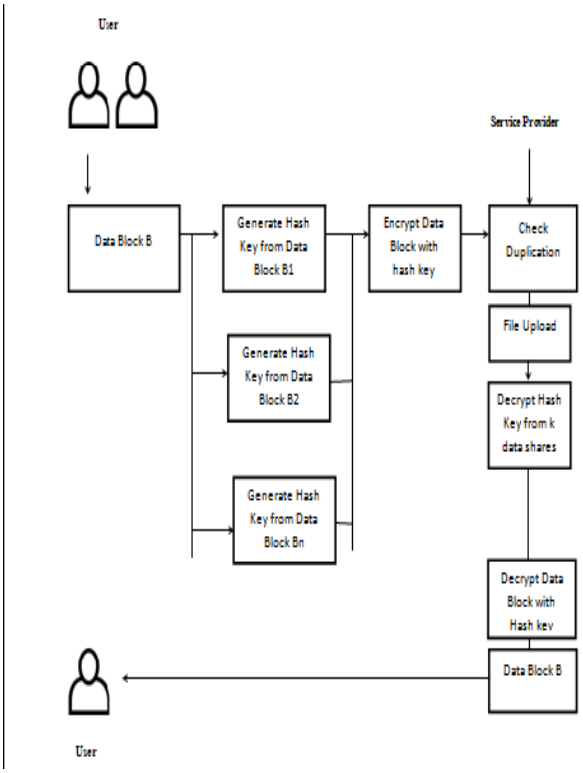


Fig 2: Architecture Diagram of de-duplication

PROCESS OF DE-DUPLICATION

Once the file is uploaded, upon receiving the auditor checks whether there exist the similar tag and the hash key for the corresponding file on the cloud service provider. If so, auditor replies to the user whether there is file duplication or there no file duplication. If the user receives the response "no file duplication" then it moves to the next level that is block level file duplication. If the auditor gets a reply as "file duplication" then the user runs proof of ownership i.e. it checks for the actual owner of the file on cloud service provider. If proof of ownership file is passed, the cloud service provider simply precedes a file pointer points to the file to the user, and no more information will be uploaded. If proof of ownership fails the upload operation on cloud service provider get terminated.

PROCESS OF FILE DOWNLOADING

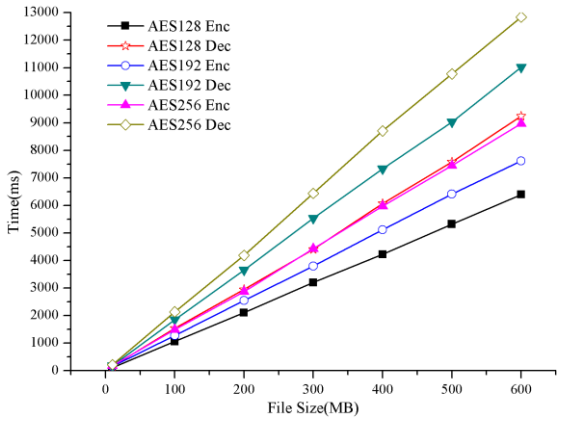
If in case user needs to download a file first along with a filename request is sent to a cloud service provider. Upon receiving, the request cloud service provider checks for the authentication of the user using secret key. If user is authenticated then the file to be downloaded will be in encrypted form using convergent key and corresponding tag and hash key of respective file can be decrypted.

DECRYPTION

The user downloads the file from the cloud service provider using hash key and tag , the user decrypt the file using decryption encryption standard.

EXPERIMENTAL ANALYSIS

When a user uploads the data with encryption algorithm, it is compared with two encryption algorithms such as DES and AES. All this depends on the basis of block size. DES has 64 bits block size and AES has 128 bits block size. So the number of blocks required to send over the network in DES is greater than that of AES. The AES analysis is more efficient than that of DES.



CONCLUSION

Maintaining encoded data with secure de-duplication is an important and significant aspect in practice on CSP. The proposed work maintains the encoded data in cloud with data de-duplication with on proof of ownership. Our proposed work can support updating of data flexibly and sharing of data with data de-duplication even though users are not online. Only authorized user can access the encoded data and get symmetric keys that is used for decryption hence it is much secured. The analysis of the performance and test conducted reviled that technique is secure, efficient,

suitable for data de-duplication. The results of our computer simulations showed the practicability of our proposed work.

REFERENCES

[1] M. Bellare, S. Keelveedhi, and T. Ristenpart, "DupLESS: Server Aided Encryption for Deduplicated Storage," Proceedings of the 22nd USENIX Conference on Security, 2013, pp. 179-194.

[2] Dropbox, "A File-Storage and Sharing Service," http://www.dropbox.com/.

[3] Google Drive, http://drive.google.com.

[4] Mozy, "Mozy: A File-storage and Sharing Service," http://mozy.com/.

[5] J.R. Douceur, A. Adya, W.J. Bolosky, D. Simon, and M.Theimer, "Reclaiming Space from Duplicate Files in a Serverless Distributed File System," Proceedings of IEEE International Conference on Distributed Computing Systems, 2002, pp. 617 -624

[6] G. Wallace, F. Douglis, H. Qian, P. Shilane , S. Smaldone, M. Chamness, and W. Hsu, "Characteristics of Backup Workloads in Production Systems," Proceedings of USENIX Conference on File and Storage Technologies, 2012, pp. 1-16.

[7]Z.O. Wilcox,"Convergent Encryption Reconsidered,"2011,http://www.mailarchive.com/cryptography@metzdowd.com/msg08949.html.

[8] G. Ateniese, K. Fu, M. Green, and S. Hohenberger, "Improved Proxy Re-Encryption Schemes with Applications to Secure Distributed Storage," ACM Transactions on Information and System Security, 9(1), 2006, pp. 1-30.

[9] Open de dup. http://opendedup.org/.

[10] D.T. Meyer and W.J Bolosky., "A Study of Practical De-duplication,"ACM Transactions on Storage, 2012, pp.1-20.