American International Journal of Research in Science, Technology, Engineering & Mathematics

Available online at http://www.iasir.net

ISSN (Print): 2328-3491, ISSN (Online): 2328-3580, ISSN (CD-ROM): 2328-3629

AIJRSTEM is a refereed, indexed, peer-reviewed, multidisciplinary and open access journal published by International Association of Scientific Innovation and Research (IASIR), USA

(An Association Unifying the Sciences, Engineering, and Applied Research)

Detection of De-Duplication Using SHA-512 and AES-256 in Cloud Storage

Tannu[1], Karambir[2]

[1,2]Department of Computer Engineering,

U.I.E.T, Kurukshetra University, Kurukshetra, Haryana, INDIA.

Abstract*: Security is the crucial challenge for the cloud and store secure data is a major concern nowadays so for security purpose the proposed scheme use cryptographic techniques for secure storage. A new challenge for cloud is overloading due to data deduplication. This paper includes a scheme for data de-duplication using SHA and AES.SHA512 is more secure than other hashing algorithms so the objective of using the SHA512 scheme for detect duplicate data and after removal of duplicate copies data send to the server. The purpose of using AES for security purpose and to encrypt data, store secure data in server using AES and update hash for the new uploaded file. The proposed scheme experimental results show the duplicate files, storage size, duplicate size and de-duplicate ratio and the proposed scheme approach for duplication detection also show reduction in storage size after duplicate checking and time also reduces due to storage size is equivalent to time in the proposed algorithm. The result also show the comparison of various hashing algorithms and show SHA512 is more secure and take less time than other hashing algorithms. The base scheme without deduplication checking scheme take more storage space and time also increases. The main purpose is to store efficient data on the cloud.*

Keywords*: Security, Cryptography, Hash, De-duplication, AES, SHA.*

I. Introduction

Cloud Storage is a place where data can be stored securely and user can also access services and applications in cloud, cloud computing and its storage solutions provide enterprises to store data in data centers [1].

A. Security concerns in Cloud

The major issue in adopting cloud is the security. The data stored in the cloud get increased every day and hence for the security purpose some mechanisms be used to ensure that the data is stored in secured manner without any unauthorized access[2]. Security for the data stored in the cloud environment is a wanted one. The major fear in this computing environment is the safety. Data de-duplication is also a major concern and to secure data by detecting duplicate data over cloud is a challenging one and solution is to use the cryptographic methods.

B. Cryptography techniques for secure storage

It is essential that a special care must be taken to protect the sensitive data when unauthorized access the data. A secure storage to must be achieved in cloud computing so use the cryptographic techniques for secure storage. The data is encrypted by the data owner before the data is uploaded to the cloud. Various encryption techniques have been proposed to secure the data even if the authorization fails and the attacker get hold of the data. Encryption is best way to make data storage on cloud secured, mainly encryption is of two types that is Symmetric - (AES) advanced encryption standard and different length of AES are 128, 192 and 256 bits.DES encrypts the data in a block of 64 bits and key length is 56 bits. Asymmetric encryption algorithms are RSA, Dieffie Hellman etc. [3]. Secure Hashing Algorithm (SHA) is also a cryptographic technique. The various SHA algorithms are SHA -1, SHA- 256 and SHA-512. The general idea of using the cryptography techniques SHA and AES technique is presented in the following subsections:

B1. Secure Hash Function Algorithm (SHA)

Secure Hash Algorithm is basically based on the concept of hash function and Hash function is a cryptography function that produces the hash value. Hashing is also used for check integrity and de-duplication and the proposed scheme consider the de-duplication detection by using secure hashing technique [11] .

(i) Message Digest Algorithm (MD5)

Message digest function is a cryptographic function that accepts a message of any length as input and returns as output a fixed-length digest.

(ii) Secure Hash Algorithm (SHA 512)

The SHA-512 produces message digest three times better than SHA-1.In the proposed scheme the SHA512 is used due to its features. SHA-512 faster and more secure than SHA-256 and the block size of SHA512 is 1024 bits [11].

(iii) Tiger160 Hashing Algorithm

Tiger is a cryptographic hash function. It is designed to be secure and run on 64-bit processors and it replaces MD4, MD5, SHA and SHA-1 in other applications. The versions are Tiger/128, Tiger/192and Tiger/160 [15].

(iv) Whirlpool Hashing Algorithm

Whirlpool is a cryptographic hash function and Whirlpool is a based on a modified Advanced Encryption Standard (AES) and Whirlpool returns a 512-bit message digest. Versions of Whirlpool hashing algorithm are 2000, 2003. The design of this hash function is very dissimilar than that of MD5 and SHA-1 [15].

B2. Advanced Encryption Standard algorithm (AES)

AES is a symmetric encryption algorithm. AES has plaintext size with 128bit which is generally its block size. AES key size is commonly having 3 types 128,192,256 [10]. In this paper AES is used for secure storage purpose and use key size for AES is 256 bit. [5].

II. Related Work

Zhifeng Xiao et al. [6] discussed the cloud computing environment and advanced research of cloud computing. It included some attributes of security in cloud environment such as integrity, confidentiality etc.

R K Karunavathi [7] discussed the cryptography technique advanced encryption algorithm and purpose was to secure data or to secure communication among data.

Zuhair S. Al-sagar et al. [8] described the data deduplication technique and removing the duplicate data .This paper discussed the deduplication concept used for optimization and using de-duplication technique it improved the storage capacity.

Renuka C. Deshpande and S. S. Ponde [9] discussed the de-duplication concept used in cloud and reduced the over space in network or detected the duplicate data and used the AES and SHA scheme for de-duplication concept or to gave access rights to server . The main purpose was to detect duplicate data and trim down the bandwidth.

Katakam Srinivasa Rao[10] described the data de-duplication technique by detecting duplicate data or detect same copies of data. This paper discussed various advantages of de-duplication concept it provide security or privacy concern when data were at risk.

Sangeeta Raheja et al. [11] discussed the various hashing techniques like SHA 1,SHA 256,SHA 512 and compared different hashing algorithms like SHA1,SHA256,SHA512.

Vishal R. Pancholi and Bhadresh P. Patel [12] discussed about the concept of security in cloud or to improve data in cloud server. This paper used AES algorithm that was to be based on permutation and substitution. AES uses 128,192 and 256 bit key that is highly secured.

Nivedita Shimbre and Priya Deshpande [13] discussed the SHA-1 techniques and hash coding techniques. This paper described the AES (Advanced Encryption Standard) algorithm to securely store the data on cloud and using cloud storage system it analyzed the data security problems.

III. Implementation Details

The data de-duplication concept has designed to detect the duplicate data in cloud. Data de-duplication technique has various advantages like provide security and privacy by detecting duplicate data when same file uploaded in the cloud. The cryptography techniques used for security purpose. In the proposed scheme to detect duplicate data by using cryptographic technique SHA (secure hash algorithm). The mainly purpose of using SHA algorithm for data de-duplication when user uploads a file if same content of file matches it detect the duplicate files and after removal of duplicate copies send data to the server and when content of file not match so it securely store data in server by using AES algorithm and purpose is to securely store data in server, generate secret key or to encrypt data and update hash of new uploaded file. Data De-duplication is important concept in cloud security and it detect duplicate files, duplicate size. The objective of this algorithm design is to detect duplicate data.

A. Data Deduplication

Data duplication occurs at the time when user tries to store the same data that has been stored already in cloud. This is checked by data owner through hash (token) comparison. If same file matches so it detects the duplicate file through hash comparison. A file is a data content file when examining the data of each file for de-duplication, and it is basically uses the hash value of the file as an identifier. If more than one files have the same hash value, they are supposed to have the same data content it means duplicate file occurs and when every new file uploaded by user so need to check the hash value first in order to make sure only unique data file is being stored [14].

B. Cryptography Techniques Used

Cryptography is a technique used for securing the information and also used for encryption and decryption. The purpose of using the SHA scheme in proposed work for de-duplication checking and the purpose of using AES to secure data and store encrypted data to the storage .This paper consider the technique SHA-512 and the key size 512 is more so it means no collision occur and if key length decreases so chances of replication occurs. SHA-512 hashing algorithm takes less time as compare to SHA 256 and SHA-1. In this algorithm design SHA-512 is used for detection of duplicate content, remove duplicate copies and after removal of deduplication send data to

the server. The MD5, Tiger160 and whirlpool2003 algorithms are used for the time comparison of various hashing algorithms with SHA 512, and the purpose of comparing the algorithms is to show the SHA512 is more secure and take less time for producing the hash value. The purpose of using AES in the proposed system is to generate secret key 256, or to securely store encrypted form of content in cloud.

C. Algorithm for Proposed Scheme

This algorithm has designed to check the duplicate content for each file. Suppose in the index table 'a' is name of the files and 'c' is duplicate files name when user uploaded a file the table entry is appended and dup is the duplicate files, dupc represents duplicate counter, ha represents hash of a , ca represents content of file a.

Proposed Algorithm (De-duplication_Store Enc_File Algorithm (hashmap ha, file a, file c))

1 *Start*
2 *Read File 'a' and store into ca.*
3 *Create hash of files content ca into ha.*
4 *Check files hash ha in hash-map for duplication.*
5 *If (ha exists)*
dupc++ // show no. of duplicates
add dup filename to dup_ table // show duplicate files
end
6 *Else increment in storage // total size is uploaded*
i. encrypt File AES
ii. Store encrypted file to storage // show server storage size
iii. Put Hash to file hash map // unique file is uploaded
end
7 *End*

This algorithm is proposed by the combination of SHA-512 and AES-256 for data de-duplication technique. The various steps of propose scheme involves in it.

1) File Reading- Firstly check the content in the form of bytes for one file at one time.
2) Hashing- In the second step check the each content (bytes) of file and compare the hash (token) of the content.
3) Duplicate checking- In the third step if in the directory the same name of file already exist so it means duplicity is there so increments in the duplicate counter and save the duplicate file and note the copy of duplicate file and store data to the server after deduplication. If in the directory same file not exist so it means the new file uploads so increments in storage it means the total size of file is uploaded in the server.
4) AES encryption- In this algorithm design when content of file not match so securely store the file content in encrypted form and to generate the secret key 256 and store encrypted file into the server.
5) Store - When unique file uploads in the cloud so it again creates the hash of file and store encrypted file to storage**.**
6) Update hash of new uploaded file **-** If the hash is not match for the new uploaded file so encrypts the content of file and store it and the result show the total storage size and if hash matches so again increments in duplicate counter and in the result it show the duplicate files and the duplicate size of file.

D. File operations

In the proposed algorithm design this paper includes the procedure for data duplication and when users uploading a file in cloud firstly check in which step the duplicate data occurs. Figure 1 show that in the client side first plain file is to be read byte by byte and role of SHA-512 is to create the hash of file and in duplication checking step is used to compare the hash values from one file to another. If file hash map matches it detects duplicate data and after removal of duplicate copies, sends data to the server. If file hash map not match so using AES-256 generate secret key, store encrypted data to the server.

A. Steps for uploading the file

1) User loads a file to server
2) Client Side Strategy –
i. Read plain files.
ii. Check De-duplication by hash match.
iii. Index of directory maintains hash for hashmap.
iv. In Figure1 Index 1contains the file name, file hash, secret key.
v. The directory contains test sets text files.

vi. If hashmap of files matches it detect duplication and after removal of duplicate copies store file to the server.

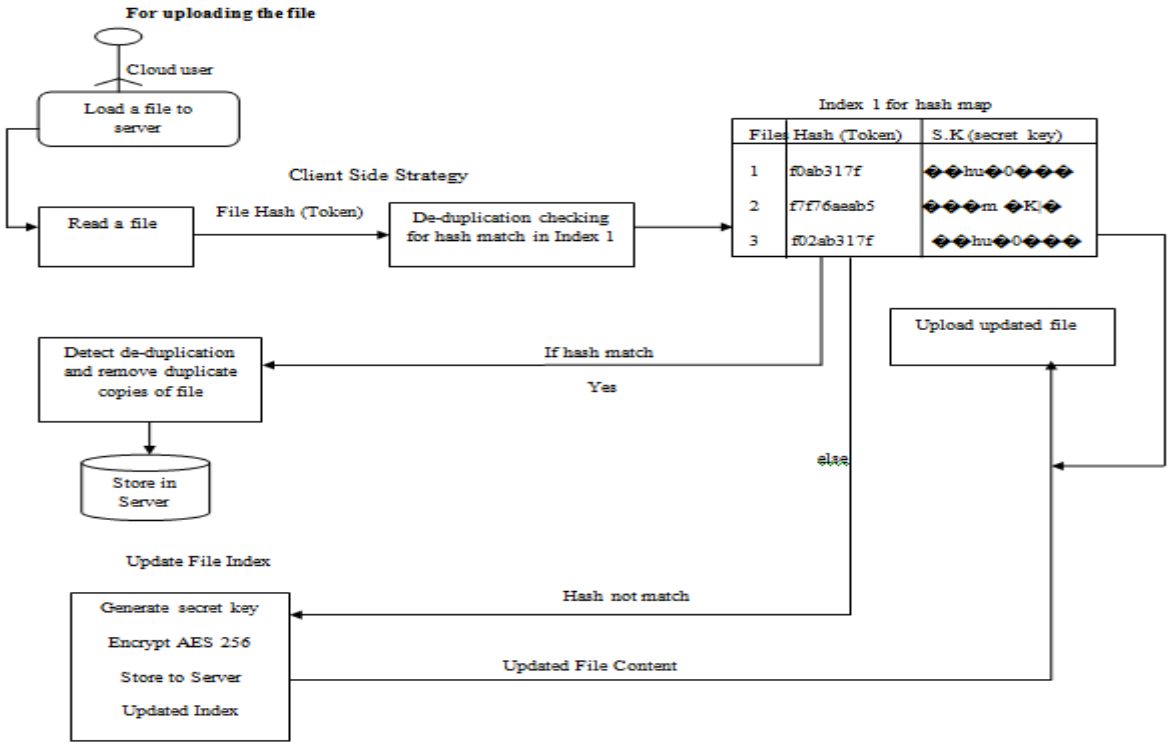
3) Server Side Strategy –

i. If hash map of file not match using AES 256 store encrypted file to server.

ii. The updated file index contains secret key and encrypt using AES 256 and store encrypted file securely to the server.

iii. Else update hash of new uploaded file.

Figure1 Proposed Framework



IV. Results and Analysis

This proposed scheme provides a secure approach for detect duplicate data in cloud and this proposed scheme use the SHA-512(secure hashing algorithm) and AES(Advanced encryption standard) for detecting duplicate data and get efficient results. SHA 512 is more secure and AES256 is the best encryption algorithm and both give the best results. The Net-beans IDE 8.1 Platform and Java language is used to accomplish the results .This proposed scheme gives the result of storage size, duplicate size and de-duplication ratio and reduces the storage size and time and for this proposed solution test the text (content) files for detecting the duplicate data and tested for 3 test sets in the base directory and under base directory client and server maintains the data. This proposed scheme also work for other files like for images, pdf, audio format files it detect duplication when user upload the same file which is already present on the cloud it show the duplicate results and experimental result also show the comparison of various secure hashing algorithms like message digest hashing MD5,SHA512 hashing ,tiger160 hashing and whirlpool2003 hashing. By comparing the time of these algorithms the result show SHA 512 takes less time and SHA 512 is more secure so in the proposed scheme SHA 512 is used because 512 key size is more so it means no collision occurs.

The base scheme (without deduplication checking) takes more time and storage size also increases and in this paper proposed scheme show for duplication checking and this approach take less time and storage size also reduces. This proposed scheme tested for 6 test sets size in KB's, MB's, GB and test sets give duplicate ratio 0.25, 0.35, 0.4, 0.5, 0.6, 0.70. The users check the duplication for any no. of files.

Notations:

De-Duplicate ratio: It show the percentage ratio of duplicate files and the formula used for proposed duplicate ratio.

Total Size- In client side directory the size of the test set before deduplication checking is the total size.

Storage Size- After removal of deduplication the size of the server side is the storage size.

Duplicate Size – The size of the duplicate files.

De-Duplicate Ratio computed by the use of formula :

De-Duplicate ratio= duplicate size/total size

Duplicate size computed by the use of formula:

Duplicate Size= Total Size – Storage Size

Table I. Result Table

Test Sets Size	Total Size (Bytes)	Storage Size (Bytes)	Duplicate Size (Bytes)	Proposed Deduplicate Ratio
995 KB	1019817	761794	258023	0.25
20 MB	20971520	13631488	7340032	0.35
100 MB	104857600	62914560	41943040	0.4
500 MB	524288000	262144000	262144000	0.5
700 MB	734003200	293601280	440401920	0.6
1 GB	1073741824	314572800	759169024	0.70

The outcome of proposed scheme is shown in Table I and the result shows proposed de-duplicate ratio for three test sets and show the duplicate files, total size, storage size and duplicate size.

Figure 2 De-duplicate Ratio Graph

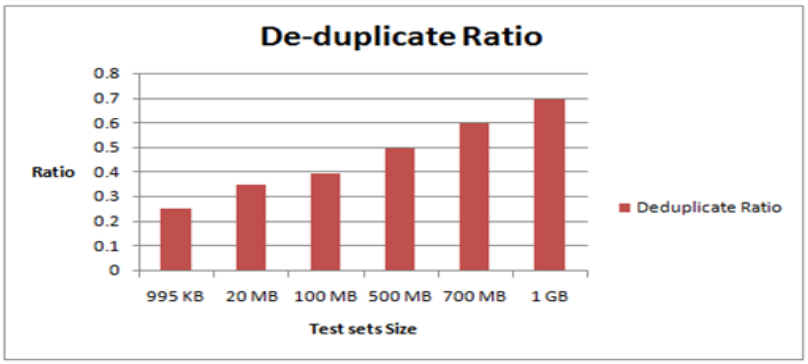


Figure 2 show the proposed de-duplicate ratio for test sets in the form of graph and the ratio increases according to the computed test sets size.

Figure 3 Total Size, Storage Size, Duplicate Size (bytes) Graph

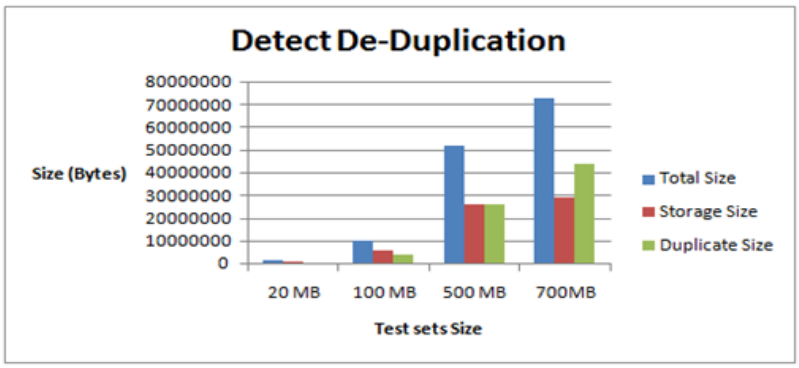


Figure 3 show the results of file size (total size, storage size and duplicate size) for three test sets. The total size (bytes) is more than storage size (bytes) and duplicate size.

The total size for test set (text files) is more and after duplication detection the no. of files reduces so storage size also reduces and using AES store encrypted data to the server. When user uploads the file if same file is already

in the cloud so it means duplicate file is there so it removes the duplicate files and store the one copy of same file into the server. The proposed scheme approach computes the number of duplicates, duplicate files and duplicates size for three test sets.

Figure 4 Time Comparison of Various Hashing Algorithms

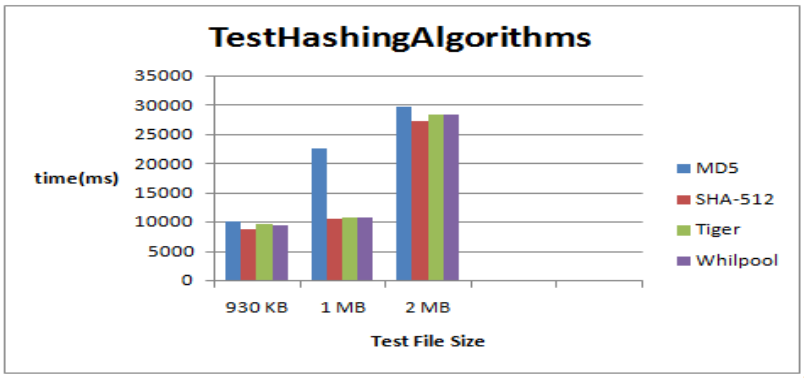


Figure 4 show the graph time comparison of hashing algorithms and show SHA512 take less time than other hashing algorithms.

Figure 5 Time Comparison of AES Algorithms

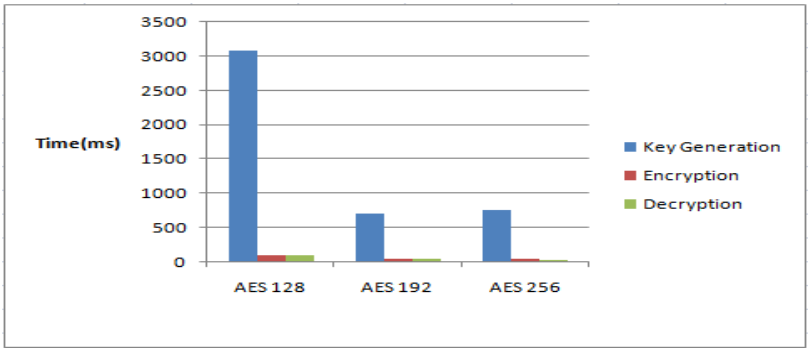


Figure 5 show the time comparison of AES 128, AES192, AES 256 for file size 1 MB and show AES 256 save more energy and time than others.

Figure 6 Storage Size Graph

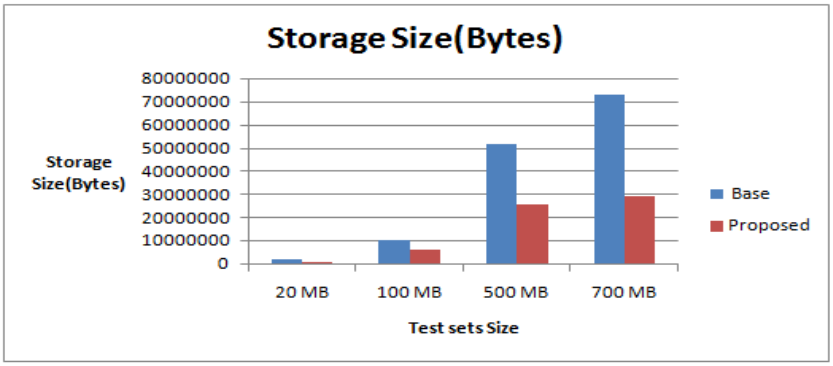
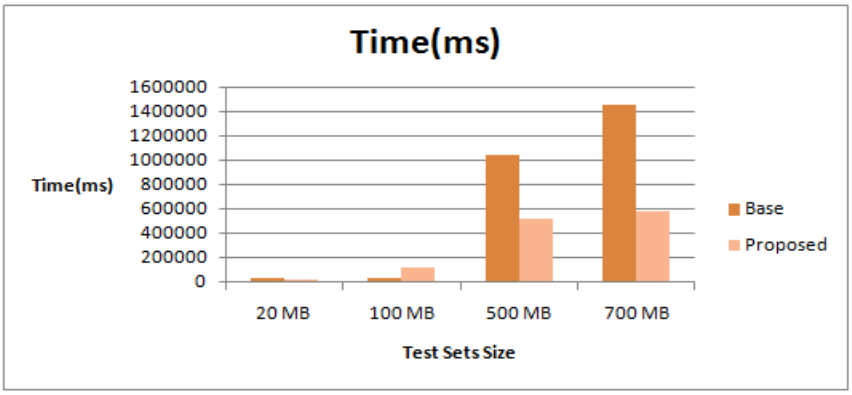


Figure 6 shows result of storage size and test for three test sets. The result show storage size increases for base(without de-duplication checking) scheme and the results show for Proposed scheme (duplication checking) after duplication detection it take less storage space than Base scheme.

Figure 7 outcome shows Storage Server Time (ms). Base scheme (without de-duplication checking) it take more storage space and take more time to store data due to storage size is equivalent to time in the proposed algorithm and after duplication detection storage size reduces due to no. of files reduces so take less time to store data for proposed scheme.

Figure 7 Storage Server Time (ms) Analysis



V. Conclusion

By applying proposed algorithm detection of duplicate data occurs and after removal of duplicity store data to server using SHA-512 scheme. To store secure encrypted data to storage server using AES-256 scheme.Results were occurred using test sets for proposed scheme. The proposed scheme results evaluate the total size, storage size, duplicate file size and de-duplicate ratio and test duplicity for six test sets and show the duplicate ratio increases by 0.25, 0.35, 0.4,0.5,0.6,0.70 and also compared the time of various hashing algorithms like MD5,SHA-512,Tiger,Whirlpool and time comparison of AES algorithms, the result illustrate that SHA512 take less time and more secure than other hashing algorithms and AES256 saves more time and provide high security than others. The proposed scheme results also illustrate the reduction in storage size and take less time to store efficient data and the base scheme without duplication checking takes more storage space and also takes more time to store data. In future work, the proposed scheme will be applied for block level de-duplication and in big data to solve the challenge of de-duplication and solve the problem of breakdown data by integrity check; will try to use different algorithms to solve the problem of de-duplication.

VI. References

[1] Ahmed Albugmi, Madini O. Alassafi and Robert Walters, Gary Wills, "Data Security in Cloud Computing", IEEE fifth International Conference on Future Generation Communication Technologies, Luton, UK, 20,pp. 55-59,2016.
[2] Suraj R. Pardeshi , Vikul J. Pawar and Kailash D. Kharat, "Enhancing Information Security in Cloud Computing Environment using cryptographic techniques", IEEE International Conference Communication and Electronic Systems, Coimbatore, India, pp.330-336,2016.
[3] Jayachander Surbiryala, Chunlei Li and Chunming Rong, "A Framework for Improving Security in Cloud Computing", 2nd IEEE International Conference on Cloud Computing and Big Data Analysis, Chengdu, China, pp. 260-264, 2017.
[4] Manreet Kaur and Jaspreet Singh, "Data De-duplication Approach based on Hashing Techniques for Reducing Time Consumption over a Cloud Network", International Journal of Computer Applications, Volume 142,Issue 5,pp.4-10,2016.
[5] Daniyal M. Alghazzawi , Syed Hamid Hasan and Mohamed Salim Trigui, "Advanced Encryption Standard-cryptanalysis", IEEE International Conference on Computing for Sustainable Global Development ,New Delhi, India, pp. 660-667,2014.
[6] Zhifeng Xiao and Yang Xiao, "Security and Privacy in Cloud Computing", IEEE Communications Surveys and Tutorials, Volume 15, Issue 2, pp. 843 – 859, 2013.
[7] Manjesh. K.N and R K Karunavathi,"Secured High throughput implementation of AES Algorithm", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 5,pp. 1193-1198,2013.
[8] Zuhair S. Al-sagar, Mohammad S. Saleh and Aws Zuhair Sameen," Optimizing the Cloud Storage by Data De-duplication", International Research Journal of Engineering and Technology , e-ISSN: 2395 -0056 ,Volume 02, Issue 09 ,pp. 2524-2527,2015.
[9] Renuka C. Deshpande and S. S. Ponde," De-duplication Using SHA-1 and IBE with Modified AES", International Journal of Science and Research, Volume 6,Issue 2,pp. 1886-1889,2016.
[10] Katakam Srinivasa Rao, "A Survey on Authorized De-duplication Techniques in Cloud Computing", International Journal of Engineering Science and Computing, March, Volume 6, Issue 3, pp. 2102-2105, 2016.
[11] Sangeeta Raheja, Shradha Verma and Nisha Raheja, " Review and Analysis of Hashing Techniques", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 4, Issue 5,pp. 296-298,2014.
[12] Vishal R. Pancholi and Bhadresh P.Patel, " Enhancement of cloud computing security with secure data storage using AES", International Journal for Innovative Research in Science & Technology, Volume 2 , Issue 09,pp. 18-21,2016.
[13] Nivedita Shimbre and Priya Deshpande, "Enhancing Distributed Data Storage Security for Cloud Computing Using TPA and AES algorithm", International Conference on Computing Communication Control and Automation,PuneIndia,Volume12,Issue6,pp. 35-39,2015.
[14] Fatema Rashid, Ali Miri and Isaac Woungang, "A Secure Data Deduplication Framework for Cloud Environments", IEEE tenth Annual International Conference on Privacy, Security and Trust, Paris, France,pp.81-87,2012.
[15] Saurabh Sindhu and Divya Sindhu, Cryptographic Algorithms: Applications in Network Security, International Journal of New Innovations in Engineering and Technology, Volume 7,Issue 1, pp.18-28,2017.