INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X
Impact factor: 4.295
(Volume3, Issue3)
Available online at www.ijariit.com

A Novel Technique to Remove Duplicacy of Files and Encrypt the Data Files Symmetrically In Cloud Environment

Sandeep Kaur
Yadavindra College of Engineering and Technology, Punjab
Sandeep.malke10@gmail.com

Abstract: With the help of cloud computing users are allowed to store, retrieve and share their data from anywhere. Cloud computing provides sharing of hardware, software and infrastructural storage to different users at a time. Encryption of cloud is a need of new technology for data security because clouds have data of different clients. Deduplication is one more parameter for cloud computation. In this paper, efficient method for removing duplicacy is proposed. Specifically, the proposed method uses hashing algorithms to remove duplicacy by calculating digest of files which takes less time than parsing method. Also, here encryption is done with the help of AES and Blowfish.

Keywords: Encryption, Security, AES, SHA1, MD5.

I. INTRODUCTION

Cloud Computing

Today we are moving very fast in a new digital era where we can store our data and perform our too costly computations remotely, on extremely powerful servers — the "cloud". Cloud computing, is recognized as on-demand computing, which is considerate of internet-based computing that administers shared processing of resources and required data to computers and other devices on their demand.Cloud computing can also be defined as storing data and accessing data and several programs over the internet instead of our computer's hard drive.

So we can say cloud computing becoming, a new technology that fulfills user's requirement for computing several resources for different networks, servers, storage and application without acquiring them physically.

One of the biggest advantages of the cloud environment is its workability to take and release resources according to the user's requirement for access of data and also provides brilliant infrastructure i.e. Transparency, Scalability, Monitoring and Security The widespread of cloud computing make it much cheaper and more powerful. Cloud computing increases the productiveness of the whole shared hardware and software resources, and services to various users and reallocates users' source demands dynamically [1].

Some serious challenges to privacy are also presented by cloud computing:

1. **Loss of user's Control:** Loss of control in cloud computing means the situation that cloud reduces the users' control on their data when they move or transfer the data from their own local servers to any remotely situated cloud servers.

2. **Issues regarding Virtualization:** Virtualization means to abstract computing resources from physical constraints logically. Virtual machine (VM) is the example of virtualization technology.

3. **Managerial Issues:** In cloud computing security biggest managerial challenges are the proper management all these technical solutions. The introduction of several vulnerabilities may occur if technical solution after being implemented is not managed properly.

4. **Multi-Tenancy regarding Issues:** "The practice of putting multiple tenants over the single physical hardware to reduce costs to the various user by leveraging economies of scale" is the meaning of multi-tenancy. It implies sharing of different computational resources, services, and applications with other tenants, hosted by the same physical or logical platform at the provider's premises.

5. **Reduced Transparency:** it ensures a reduced level of transparency because even if security facts about an organization are available, but they are not displayed in an organized and easily understandable manner for the user.

DATA DEDUPLICATION

This technique is used to increase the storage utilization and can also be used in network data transfers to decrease the number of bytes that must be sent over the network. In computing, data deduplication is a specialized method or data compression technique for removing duplicate copies of data which is repeated several numbers of times. We can also call this as intelligent compression of data and storage of single instance of data.
Removal of duplicate date can be at the file level and at block level

Removing duplicate at file level: File level Deduplication means taking the complete file which is compared with other identical files. If a proper match is found, the action performed to remove located duplicate copies.

Removing duplicate at Block level: It is performed over blocks. Files are firstly divided into a number of different blocks and store a single copy of each block. All these blocks are then divided into a variable or fixed size blocks.
Whole File Hashing, Sub File Hashing, and Delta Encoding are some data deduplication approaches.

WFH (Whole File Hashing)

Whole File Hashing function is applied to the complete file. Various cryptographic hash functions like MD5, SHA-1 or RC5 are used. These functions are used to find out whole replicate files. Less metadata overhead and fast with low computation are the Advantages of whole file hashing.

DE (Delta Encoding)

This approach is designed for and to use the mathematical term "delta". This stands for "change encountered" or "rate at which change is occurred" in an object. difference between the target and source object is shown by using delta encoding. It is used normally when SFH does not produce efficient results but there is a strong enough similarity between two items and chunks that storing the difference would take less space than storing the non-duplicate block.

SFH (Sub File Hashing)

This deduplication technique uses complete file which is divided into a number of smaller sections and after that data deduplication check is done on these sections. Fixed size chunking and variable size chunking are types of secure file hashing.

TF-IDF

During information retrieval, tf–idf, means term frequency–inverse document frequency, is a numerical statistic that defines the importance of the word in a document or in a collection or corpus. It is considered as a weighting factor for the retrieval of information and text mining. The number of times a word appears in the document tf-idf value increases proportionally but is offset by the frequency of the word in the corpus, which helps to adjust for the fact that some words appear more frequently in general.

Search engines often use variations of the tf–idf weighting scheme as a central tool in scoring and ranking of a document's relevance given a user query.

II. RELATED WORK

In 1996 Oded Goldreich and Rafail Ostrovsky had proposed the basic scenario of searchable encryption they hidden the access paradigm to intercept software pirates copy software without paying anything. Hence, various logarithmic intercommunication leads inadequate between users and servers in practical applications. In 2015 Yuhong Liu had found security and privacy limitation in cloud computing, and classified several solutions that are present, as compared to their advantages and disadvantages, and suggested future research directions. To make sure that cloud environment is protective and convincing it is compulsory to describe all the challenges of solutions that are currently present and also suggest some directions for future scopes
Also in 2004 Boneh and G. D. Crescenzo established the prime procedures of symmetric searchable encryption(AES) and construction of PEKS(public encryption with a keyword search) on the basis of bilinear maps for the email gateway to finding the sender whose emails are sent without knowing the email contents was explained. In AES, for the users and servers do not need the security channel the security of key and it is easier to govern the key, but computation load is increased to a greater extent is the biggest disadvantage and as a result, research proposed some alternately transformed method aimed to various application scenarios.
Cheng Guo in 2015 had organized a strategy which protects and recover the Top-N files by using keyword-based searching and asymmetric encryption was done over encrypted data. Hence, sensitive information is centralized into the server by outsourcing data to cloud servers, which is the great threat to sensitive information's privacy. So to ensure privacy, the user encrypts sensitive data before outsourcing it to the cloud server and ultimately retrieves the most consistent Top-N files between the whole data.
And in 2015 Rachana Chavda and Rajanikanth Aluvalu[9] reviewed various encryption-based access control model for enhancing cloud security along with their limitations. They will be concluding with a proposed access control model to enhance cloud security. They can increase security on access of the data in the cloud. Moreover in order to restrict the use of data from the third party they provide encryption on the data. Since this new computing technology requires the user to entrust their valuable data to cloud providers, so security and privacy concerns on outsourced data were increased.
Guojun Wang, Qin Liu and Jie Wu in 2010 proposed a strategy to help the organisation to competently share intimate data on cloud servers. This objective was attained by first adjoining the hierarchical identity-based encryption (HIBE) system and the ciphertext-policy attribute-based encryption (CP-ABE) system and then making a performance-expressivity tradeoff, ultimately applying proxy re-encryption and lazy re-encryption to their strategy.

N. R. Anitha Rani proposed a scheme in 2016 which the recognized data duplication is gained with reduced overhead when compared with traditional deduplication. The attention was on different deduplication methods with a protective cloud. To restrict the unrecognized user from accessing the data Symmetric key encryption is used. Data confidentiality is preserved using Convergent key encryption algorithm. Data deduplication is the eradication of repeated data within a current environment. For bandwidth optimization and to advance storage space in cloud storage environment data deduplication is used.

Organization.

The rest of this paper is organized as follows: Section III introduces steps of proposed scheme. In Section IV includes implementation information. The discussion, results and corresponding optimization of the proposed scheme are described in section V. Finally, there is a conclusion and future scope is given in Section VI.

III THE PROPOSED METHODOLOGY

Proposed work uses hashing algorithms that are a secure hash algorithm and MD5 approach for removing duplicity rather than conventional parsing method which uses tf-idf for ranking of files and consumes more time. The step of the proposed work is explained as below:

1) The initial step is to input all the text files on client side which may consist duplicate data
2) In next step comparison between the proposed approaches, SHA1, MD5 and existing parsing method (base method) for removing duplicity is done.
3) SHA1, MD5 calculates the digest of all the text files. The files having same message digest are duplicate files which are ranked below in the list of all files.
4) In parsing method duplicate files are removed according to the score calculated using TF-IDF.
5) In next step time used to remove duplicity by three methods is calculated and compared.
6) Proposed method takes less time than the existing parsing method.
7) After that, all the files are ranked automatically. Deduplicated files are on the top of the list.
8) In the next step, text files are encrypted on the client side using blowfish and AES. Proposed method takes time and storage parameters into account and compares both the algorithms
9) Encryption time is calculated and then text files are uploaded to the cloud. Storage analysis is also done to keep check the memory required to store these files.
10) These encrypted files are downloaded by the client in order to decrypt these files and decryption time is calculated using both the algorithms.

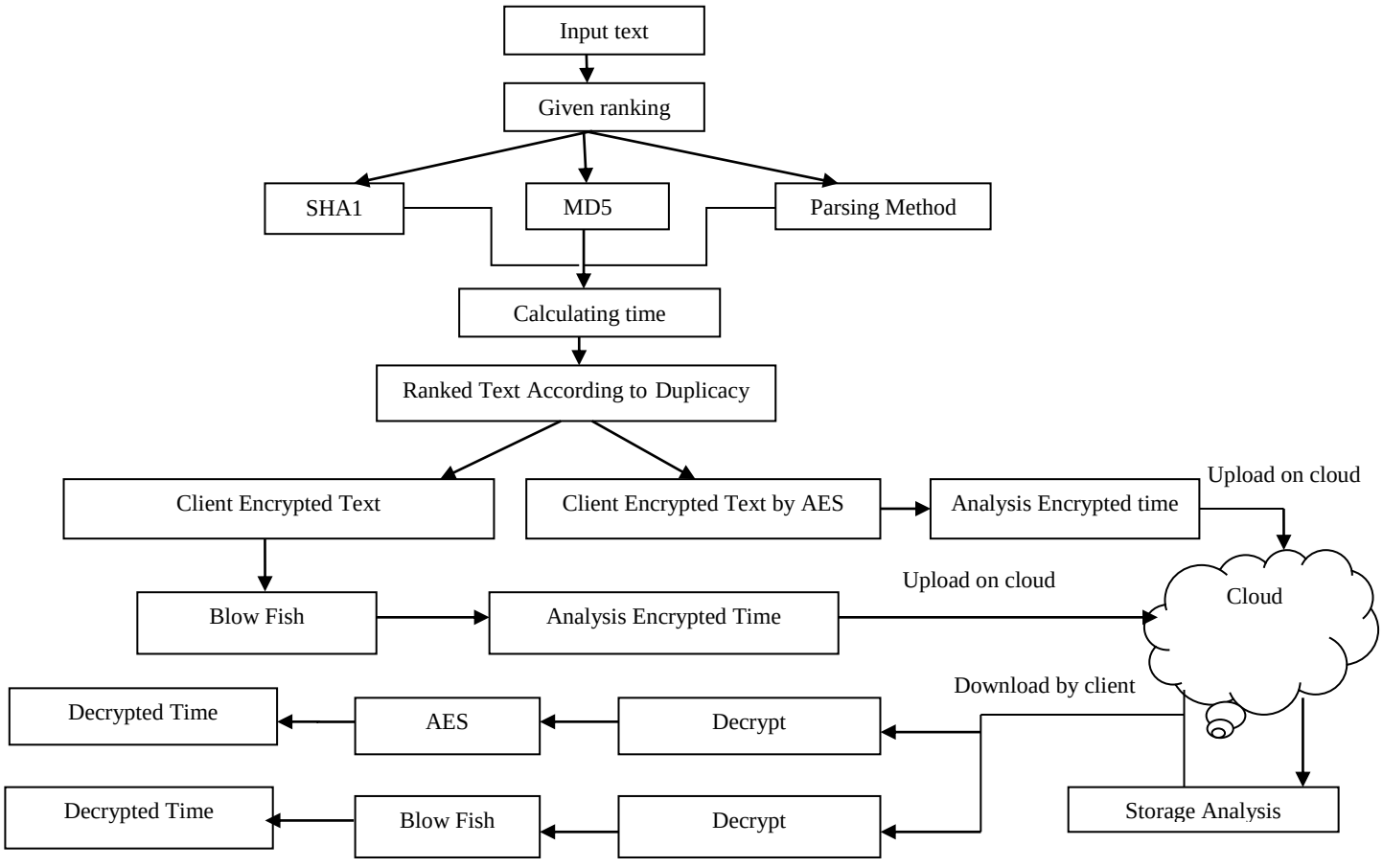


Figure 1: Proposed System

IV IMPLEMENTATION

The proposed method uses secure hash algorithm and MD5 to locate duplicate files which reduce the time requirements to remove duplicate files as compared to parsing method which calculates score of each file using tf-idf.

V. EXPERIMENTAL RESULTS

In our proposed work our aim is to increase the utilization of cloud environment by removing the duplicate files. .In cloud computing security and deduplication of data is most important factor for data.

Performance parameters

Performance parameters are used for both deduplication and encryption process which describes the performance of proposed method.

Performance parameter for deduplication

Time: Time is the performance parameter used for the deduplication method. Proposed method takes less time to remove duplicity than existing parsing method.

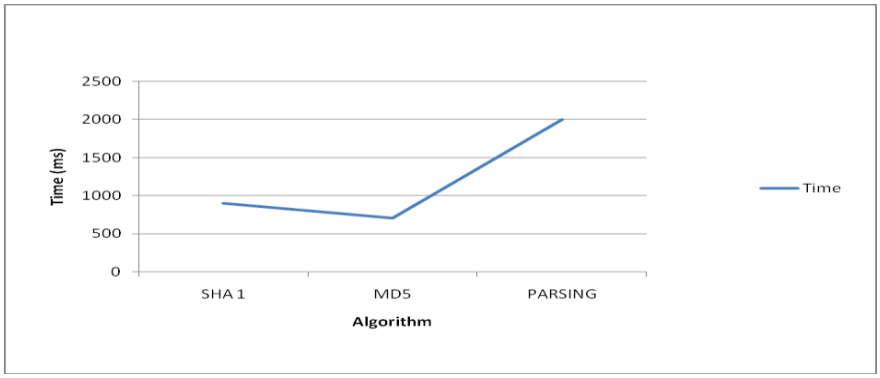
Performance parameters for encryption and decryption

Time: Time comparison of both the algorithm used for encryption.
Storage: Storage analysis is done after the files are downloaded from the cloud that is memory required to store files.

Table 1: Time comparison between proposed and traditional method for removing duplicate

Algorithm	Time (ms)
SHA 1	896
MD5	703
PARSING(base method)	2000

Comparison of the base method and the proposed method for removing duplicate is shown in table 1 with respect to time parameter used for the research. As it is clearly seen from the above table that removing duplicate using hashing algorithms takes less time as compared to conventional methods.



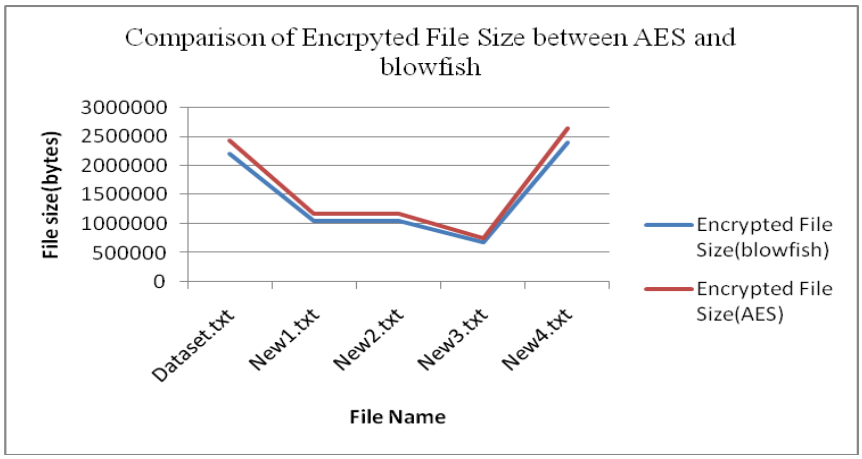
Graph 1: Comparison result graph

Graph 1 shows the time taken to remove duplicate files by various methods

Table 2: Comparison of blowfish and AES

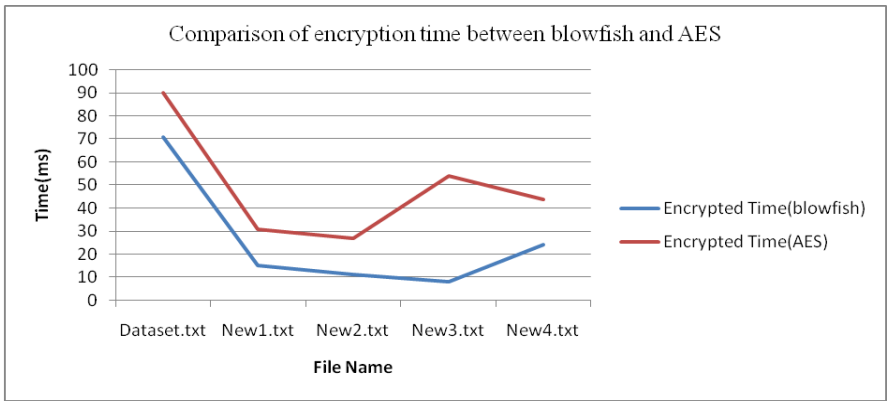
File Name	Input file size (bytes)	Encrypted file size in bytes (blowfish)	Encrypted file size in bytes(AES)	Encryption time in milliseconds (blowfish)	Encryption time in milliseconds (AES)	Decryption time in milliseconds (blowfish)	Decryption time in milliseconds (AES)
Dataset.txt	1216841	2204478	2433696	71	90	57	60
New1.txt	581469	1061085	1168962	15	31	13	18
New2.txt	581632	1062894	1169892	11	27	13	34
New3.txt	378754	687489	758754	8	54	12	16
New4.txt	1315331	2401145	2643300	24	44	35	65

Given table 2 show the comparison of blowfish and AES encrypted file size, encryption time and decryption time. As it is clearly seen that as compared to AES encrypted file size, encryption time and decryption time for various text files is less in blowfish algorithm.



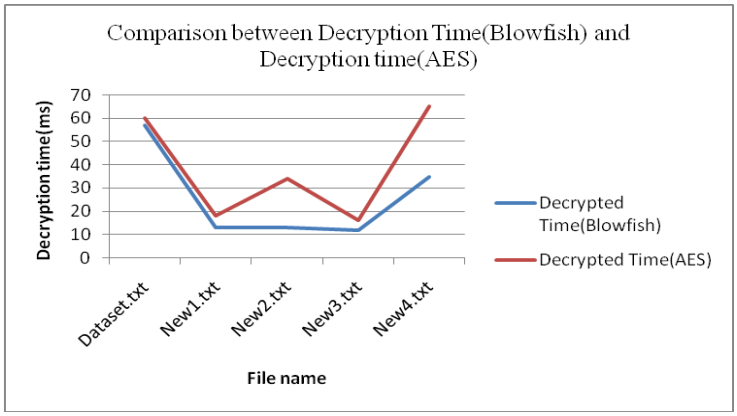
Graph 2: Comparison of Encrypted File Size of blowfish and AES algorithm

Given graph shows the comparison of encrypted text file size using blowfish and AES algorithm for different text files.



Graph 3: Comparison of Encryption time of blowfish and AES algorithm

Given graph shows the comparison of encryption time using blowfish and AES algorithm for different text files.



Graph 4: Comparison of decryption time of blowfish and AES algorithm

Given graph shows the comparison of decryption time using blowfish and AES algorithm for different text files.

CONCLUSION AND FUTURE SCOPE

In this work a method is implemented for removing the duplicate files using the SHA 1 and MD5 algorithm, ensuring the reduced time to deduplicate files being uploaded by the clients while using the cloud. The MD5 algorithm takes less time for removing duplicates of files and ranks the files according to their digest count. The Same digest means duplicate file and lowers the ranking. Encryption is done using diffie Hellman hybrid with AES and blowfish. Blowfish is more efficient than AES as it takes less encryption and decryption time and space for encrypted data also proposed technique takes less time for removing duplicate files than the parsing keyword based searching method. As the results show that the proposed approach is efficient.

In future further enhancements in the work on the more number of files deduplicated by using hadoop and parallel processing. Along with this various other encryption techniques with different block sizes can be combined to obtain more efficient results.

REFERENCES

[1] Anita N, Kumar, S. R., & Kumar, P. P. (2016). A survey on data redundancy check in a hybrid cloud by using convergent encryption. Indian Journal of Science and Technology, 9(4).

[2] Boneh, D., Di Crescenzo, G., Ostrovsky, R., & Persiano, G. (2004, May). Public key encryption with keyword search. In International Conference on the Theory and Applications of Cryptographic Techniques (pp. 506-522).Springer Berlin Heidelberg.

[3] Behl, A., & Behl, K. (2012, October). An analysis of cloud computing security issues. In Information and Communication Technologies (WICT), 2012 World Congress on (pp. 109-114). IEEE.

[4] Chen, M. Y., Liu, C. W., & Hwang, M. S. (2013, August). SecureDropbox: a file encryption system suitable for cloud storage services. In Proceedings of the 2013 ACM Cloud and Autonomic Computing Conference (p. 21). ACM

[5] Di Pietro, R., & Sorniotti, A. (2012, May). Boosting efficiency and security in proof of ownership for deduplication. In Proceedings of the 7th ACM Symposium on Information, Computer and Communications Security (pp. 81-82). ACM.

[6] Goldreich, O., & Ostrovsky, R. (1996). Software protection and simulation on oblivious RAMs. Journal of the ACM (JACM), 43(3), 431-473.

[7] Guo, C., Song, Q., Zhuang, R., & Feng, B. (2015, August). RSAE: Ranked keyword search over asymmetric encrypted cloud data. In Big Data and Cloud Computing (BDCloud), 2015 IEEE Fifth International Conference on (pp. 82-86). IEEE.

[8] Liu, Y., Sun, Y., Ryoo, J., Rizvi, S., & Vasilakos, A. V. (2015). A survey of security and privacy challenges in cloud computing: solutions and future directions. Journal of Computing Science and Engineering, 9(3), 119-133.

[9] Li, M., Yu, S., Zheng, Y., Ren, K., & Lou, W. (2013). Scalable and secure sharing of personal health records in cloud computing using attribute-based encryption. IEEE transactions on parallel and distributed systems, 24(1), 131-143.

[10] Wang, G., Liu, Q., & Wu, J. (2010, October). Hierarchical attribute-based encryption for fine-grained access control in cloud storage services. In Proceedings of the 17th ACM conference on Computer and communications security (pp. 735-737). ACM.

[11] Zha, Z. J., Yu, J., Tang, J., Wang, M., & Chua, T. S. (2014). Product aspect ranking and its applications. IEEE Transactions on Knowledge and Data Engineering, 26(5), 1211-1224.

