iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 7 **Issue:** VI **Month of publication:** June 2019

DOI: http://doi.org/10.22214/ijraset.2019.6069

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.177
Volume 7 Issue VI, June 2019- Available at www.ijraset.com

Secure Duplication Detection in Cloud using Chunk Based Technique

Pranil Bari[1], Pratik Ghare[2], Anurag Gajare[3], Dipak Bhageshwar[4]
[1, 2, 3, 4]*B.E. (Final Year), Dept. of Computer Engineering, Modern Education Society's College of Engineering, Pune, Maharashtra, India.*

Abstract: ***Cloud computing is an internet technology that operates both internet and central remote servers to control the applications and data. Cloud computing refers to the delivery of storage capacity and computing as a service to a heterogeneous community of end-receivers. Data De-Duplication is an effective technique to optimize the utilization of storage space backup by avoiding the redundancy. The main core of the Deduplication algorithms is chunking and hashing functions. It is also referred as Deduplication granularity. The analysis of these three methods show that the content approach for deduplication is bit slow but the accuracy is good as compared to file and block strategies.***

Keywords: Third Party Authenticator, AES Algorithm, RSA Algorithm, SHA 512 Algorithms, Deduplication.

I. INTRODUCTION

A. Overview of De-Duplication.

Data De-duplication identifies the duplicate data to remove the redundancies and reduces the overall capacity of data transferred and stored. [4,5] De-duplication often called as "intelligent compression" or "single-instance storage" which is the method of reducing storage needs by eliminating redundant data. Only one unique instance of the data is actually retained on storage media, such as disk or tape. Redundant data is replaced with a pointer to the unique data copy. For example, if an organization webmail system might contain 50 instances of the same one megabyte (MB) file attachment. [2] If the webmail platform is backed up or archived, all 50 instances are saved, requiring 50 MB storage space. With data de-duplication, only one instance of the attachment is actually stored. Each subsequent instance is just referenced back to the one saved copy. In this example, a 50 MB storage demand could be reduced to only one MB. Data deduplication offers three benefits. First, lower storage space requirements will save money on disk expenditures. Second, efficient use of disk space also allows for longer disk retention periods and reduces the need for tape backups [1, 2 3]. Third, it also reduces the data that must be sent across a WAN**.**

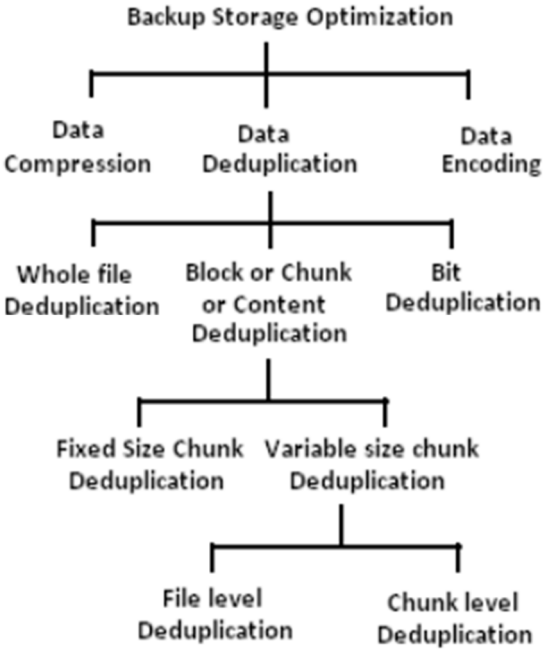


Figure: - De-duplication methods [2].

De-Duplication Techniques the optimization of backup storage technique is shown in The Data de-duplication can operate at the whole file, block (Chunk), and bit level. Whole file de-duplication or Single Instance Storage (SIS) finds the hash value for the entire file which is the file index. If the new incoming file matches with the file index, then it is regarded as duplicate and it is made pointer to existing file index. [9,11] Block De-duplication divides the files into fixed-size block or variable-size blocks. For Fixed-size chunking, a file is partitioned into fixed size chunks for example each block with 8KB or 16KB. The unique ID is then compared with a central index. If the ID exists, then that data block has been processed and stored before. Therefore, only a pointer to the previously stored data needs to be saved. If the ID is new, then the block is unique. The unique ID is added to the index and the unique chunk is stored. Block and Bit de-duplication looks within a file and saves unique iterations of each block or bit.

B. Chunk Level De-Duplication – DDDFS.

Detecting duplicates is Chunk level de-duplication. Data Domain De-duplication File System (DDDFS) is a file system which performs chunk level de-duplication. It supports multiple access protocols [14, 15, and 16]. Whenever a file to be stored, it is managed by the interfaces such as Network File System (NFS), Common Internet File System (CIFS) or Virtual Tape Library (VTL) to a generic file service layer. File service layer manages the file metadata using Namespace index and forwards the file to the content store. Content store divides the file into variable sized chunks. Secure Hash Algorithm SHA-1 finds the hash value for each variable size chunk, which is Chunk ID [11, 12, 16, and 17]. Chunk store maintains a chunk index for duplicate chunk detection. In this chunk level de-duplication, the efficiency of duplicate detection is high but the throughput of the de-duplication is low. So this method can be used for the applications with locality of reference between the data streams in the cloud storage.

II. LITERATURE REVIEW

A. Proofs of Ownership" (PoW)

S. Halevi, et al proposed the deduplication systems as the notion of "proofs of ownership" (PoW) in which a client can prove to a server depending on Merkle trees and the error-control coding that it indeed has uploading without a copy of a file but their scheme may not assurance the freshness of the proof in every challenge. Additionally, this scheme has to make Merkle Tree on the encoded data and it has inherently inefficient and not consider about data privacy.

B. Third Party Auditor (TPA)

M. Bellare, et al. introduced, Confidentiality can be protected by converting predictable message into unpredictable form. The Server aided encryption for de-duplicated storage recommends different security mechanisms. One new concept established to generate the file tag for duplicate check and also for Key server (Third party auditor). Wee Keong Ng, et al. introduced a new notion that the author calls private data deduplication protocols are formalized in the context of two-party computations. The private data deduplication protocols has been analyzed and proposed as a feasible result. The simulation based framework is provably secure protected by the proposed private data deduplication protocol. The hash function is collision-resilient, and the discrete logarithm is hard and the erasure coding algorithms E can erasure up to α-fraction of the bits in the presence of malicious adversaries. S.Sadeghi, et al. suggested a novel encryption scheme that affords the vital security for both unpopular data and popular data. For unpopular data suggested another two-layered encryption scheme with stronger security while supporting deduplication process. The traditional conventional encryption is mainly performed for popular data that are not specifically sensitive. Thus, they achieved enhanced trade between the security and efficiency of the out-sourced data. D. Harnik, et al. presented a cloud storage services generally utilize deduplication, which eradicates redundant data by storing only a single copy of each block or file. Deduplication saves the bandwidth and space requirements of data storage services and it have most effective when applied across multiple users. Deduplication can be utilized as a covert channel by malicious software communicates with its control center and firewall settings at the attacked machine. Cloud storage providers are suspect to stop using this technology because of the high savings offered by cross-user deduplication so that they suggest easy mechanisms that enable cross-user deduplication while greatly saving the risk of data leakage.

C. Sparse Indexing

Camble, et al. proposed the "Sparse Indexing" deduplication system which uses a different approach to avoid the chunk lookup disk bottleneck. Sparse Indexing permits to save a chunk multiple times if the similarity based system is not able to detect the segments, which already have stored the chunk. At this point, the chunks are consecutively grouped into segments and that segments are utilized to search similar existing segments by using a RAM based index, whereas stores only a small fraction of the already stored chunks. Hence, Sparse Indexing is considered a member of the class of approximate data deduplication systems.

D. Object-Based Storage Devices (OSDs).

M. Armbrust, et al analyzed to point out a few security problems with convergent encryption, whereas, recommending two protocols and a security model for protecting data deduplication. Whereas the two models are like same and slightly vary in security properties. These two approaches protect via authenticated and anonymous to secure deduplication data and that can be applied to single server storage and distributed storage. In the earlier, single server storage, clients interact with a single file server that stores both metadata and data. Afterward, metadata is stored on an independent metadata server, and data is stored on a series of object-based storage devices (OSDs).

III. PROPOSED SYSTEM

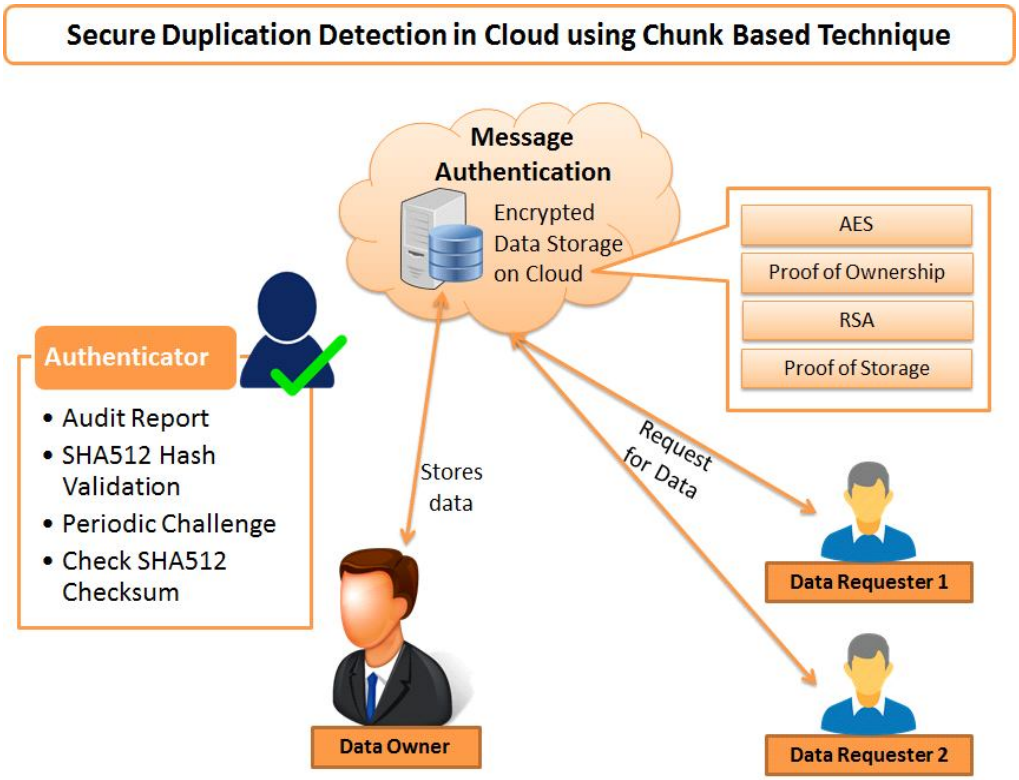


Figure:-System Architecture.

A. Proof of Ownership

Data Owner uploads document, metadata, checksum on cloud after encryption using keys from Data Owner and Cloud Service Provider. Also, a copy of metadata and checksum is sent to Auditor.

B. Data Access Via Permission Model

Registered users send access request and receive encrypted file if authorized. User calculates checksum to compare with original and reports to Data Owner if checksum mismatch occurs.

C. Prevention De-duplication

Avoid De-duplication

1) File Level

2) Block Level

Maintains the checksum of file data and block of file data and compare at the time of file upload to avoid De-duplication.

D. Proof of Storage by Third Party Authenticator (TPA)

Auditor Receives metadata after upload. Performs periodic or on-Demand integrity checks by sending challenges to Cloud Service Provider. On response from Cloud Service Provider, Auditor confirms response and reports status to Data Owner.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.177
Volume 7 Issue VI, June 2019- Available at www.ijraset.com

IV. ALGORITHMS USED

A. RSA Algorithm

The RSA includes both the public and the private keys and it can encrypt a message. The contrast keys are used for encryption and decryption process. This major reason for most widely uses RSA asymmetric.

1) Algorithm: It also assures some attributes that mainly includes; confidentiality, integrity, authenticity, and non-repudiation of the data storage.

TABLE I. RSA ALGORITHM [1]

Step 1	• Two prime numbers are selected as p *and q*. • For security purposes, integer's *p* and *q* should be chosen at random bases and should be similar in magnitude but it should be 'differ in length by a few digits to make factoring harder.
Step 2	• n = pq, which is the modulus of both the keys. • *n* is used as the modulus for both the public and private keys. Its length, usually expressed in bits, is the key length.
Step 3	• Calculate totient, totient = (p-1)(q-1)
Step 4	• Choose e such that e > 1 and coprime to totient which means gcd (e, totient) must be equal to 1, e is the public key.
Step 5	• Choose d such that it satisfies the equation; de = 1 + k (totient), d is the private key not known to everyone.
Step 6	• The ciphertext is calculated using the equation; c = m^e mod n Where m is the message.
Step 7	• With the help of c and d we decrypt the message using equation ; m = c^d mod n. Where d is the private key.

B. SHA Algorithm

The family of cryptographic functions includes SHA (Secure Hash Algorithms) plan for secure data storage.

Table II. SHA 512 ALGORITHMS [1]

Step 1	a. Length Value as well as Include Padding Bits: This step makes the input message an exact multiple of 1024 bits
Step 2	b. Initialize Hash Buffer with Initialization Vector: Firstly need to initialize the hash buffer with IV as the Initialization Vector, and then we can continue the process of first message block.
Step 3	c. Process Every 1024-bit (128 words) and Message Block Mi: Each and every message block is considered after 80 rounds of its processing.
Step 4	d. Final Step: Initially all the N blocks are processed and then afterwards hash buffer contents are message and understood.

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.177
Volume 7 Issue VI, June 2019- Available at www.ijraset.com

C. AES Algorithm

One of the most popular block cipher encryption algorithm is AES (Advanced Encryption Standard) Algorithm.

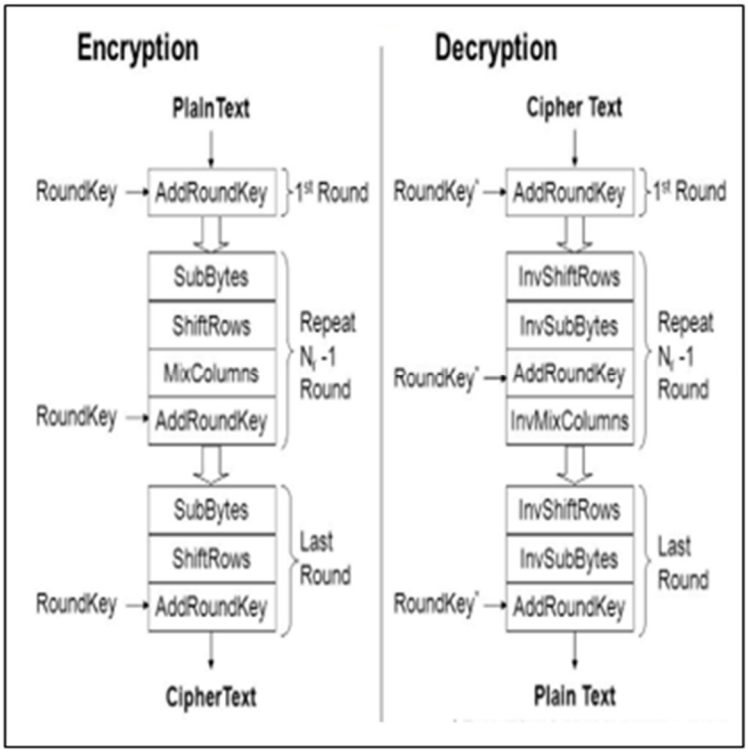
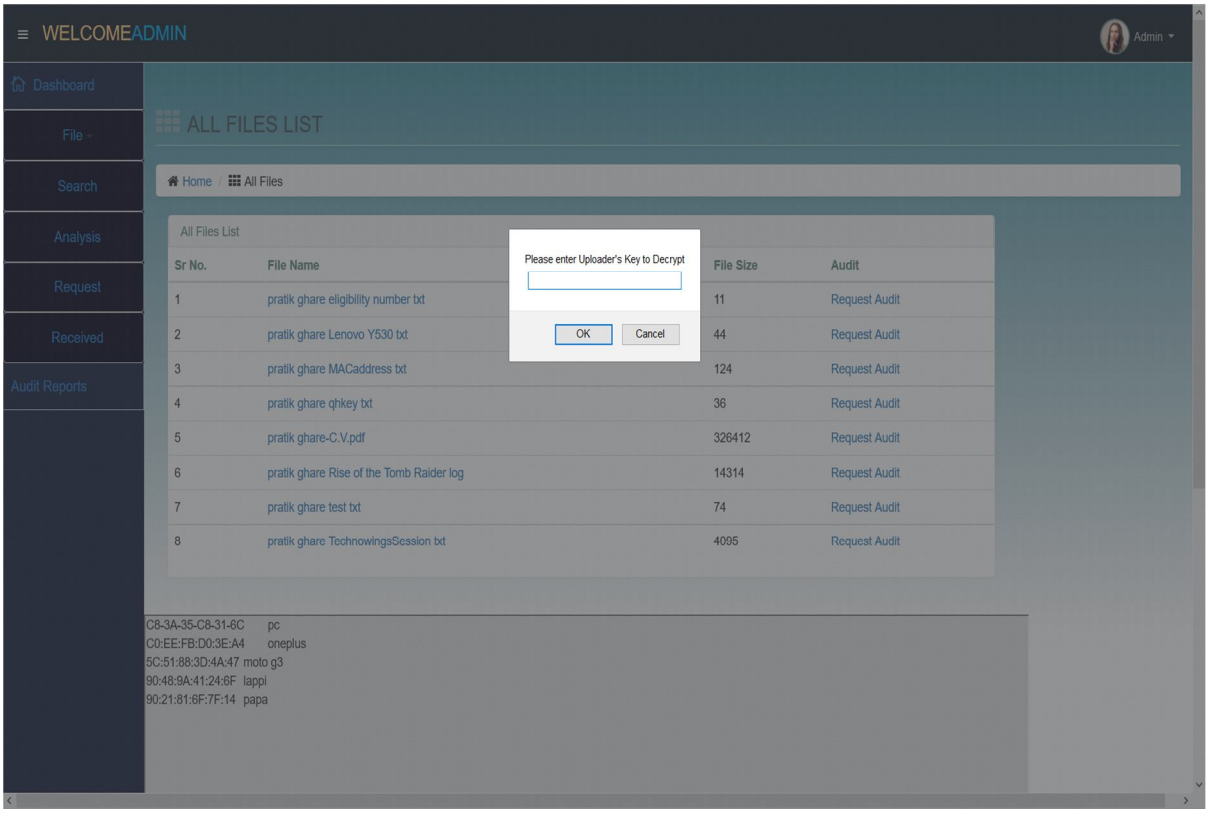
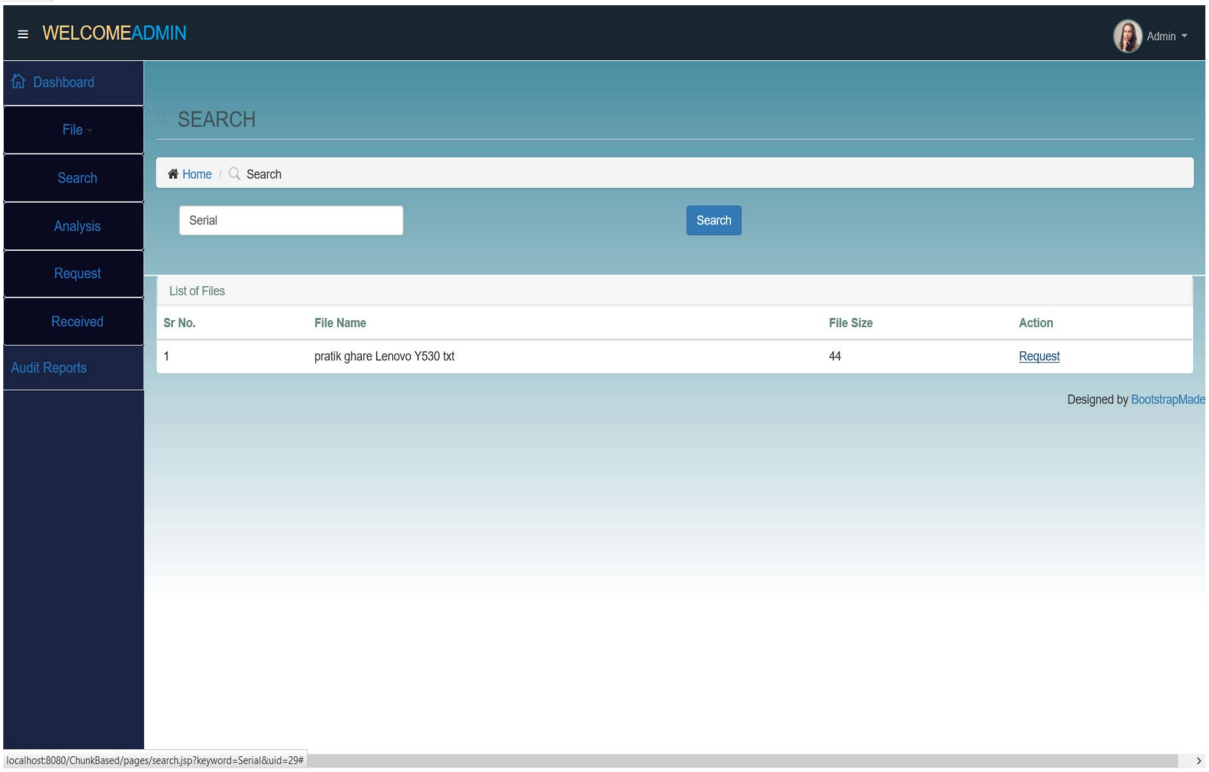
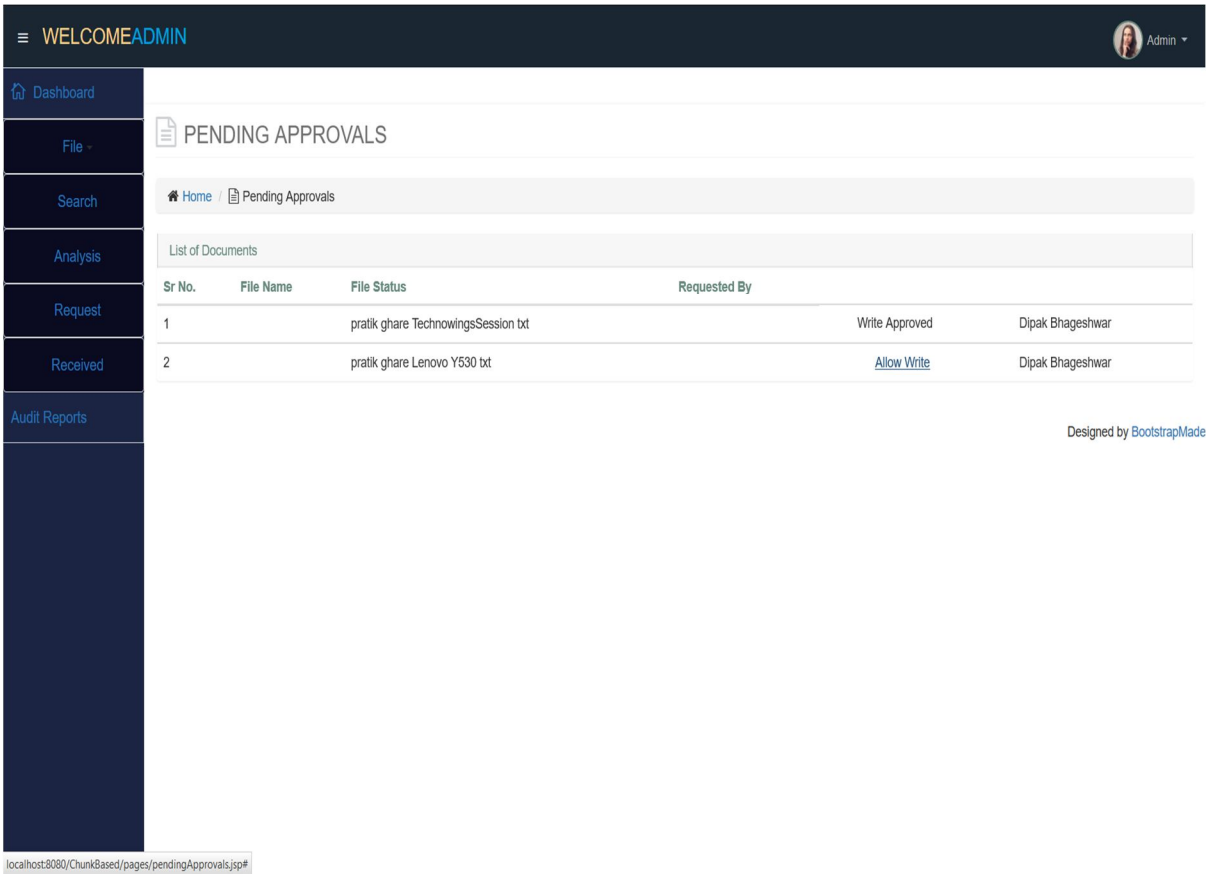


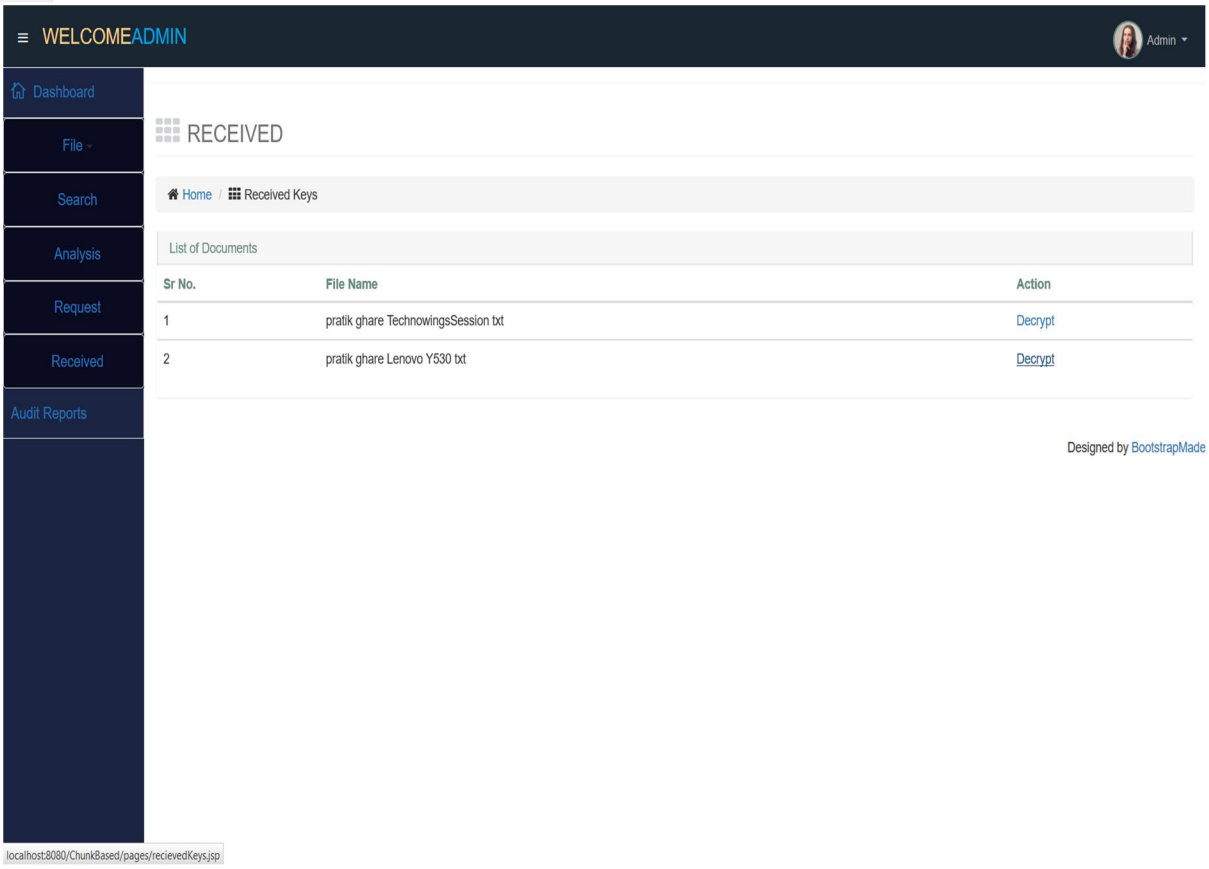
Figure: - AES Algorithm Stepwise.

V. RESULT AND DISCUSSION









VI. ADVANTAGES

A. Bit De-duplication done exact de-duplication and it is more efficient since it eliminates redundancy.

B. In variable size chunking, the impact on the systems performing the inspection and recovery time is less. The efficiency of identifying the duplicate is high.

C. Fixed-size chunking is conceptually simple and fast since it requires less processing power due to the smaller index and reduced number of comparisons.

VII. APPLICATIONS

Data security Application over cloud.

VIII. CONCLUSION

It is highly desirable to improve the private cloud backup storage efficiency by reducing the de-duplication time. De-duplication eliminates the redundant data by performing only the single copies of data in storage space and it is also essential for utilizing cellular networks, wired communication, backup services process, wireless communication, etc., to save the amount of data in storage and to rapid up the backup process. It is highly desirable to improve the private cloud backup storage efficiency by reducing the de-duplication time.

REFERENCES

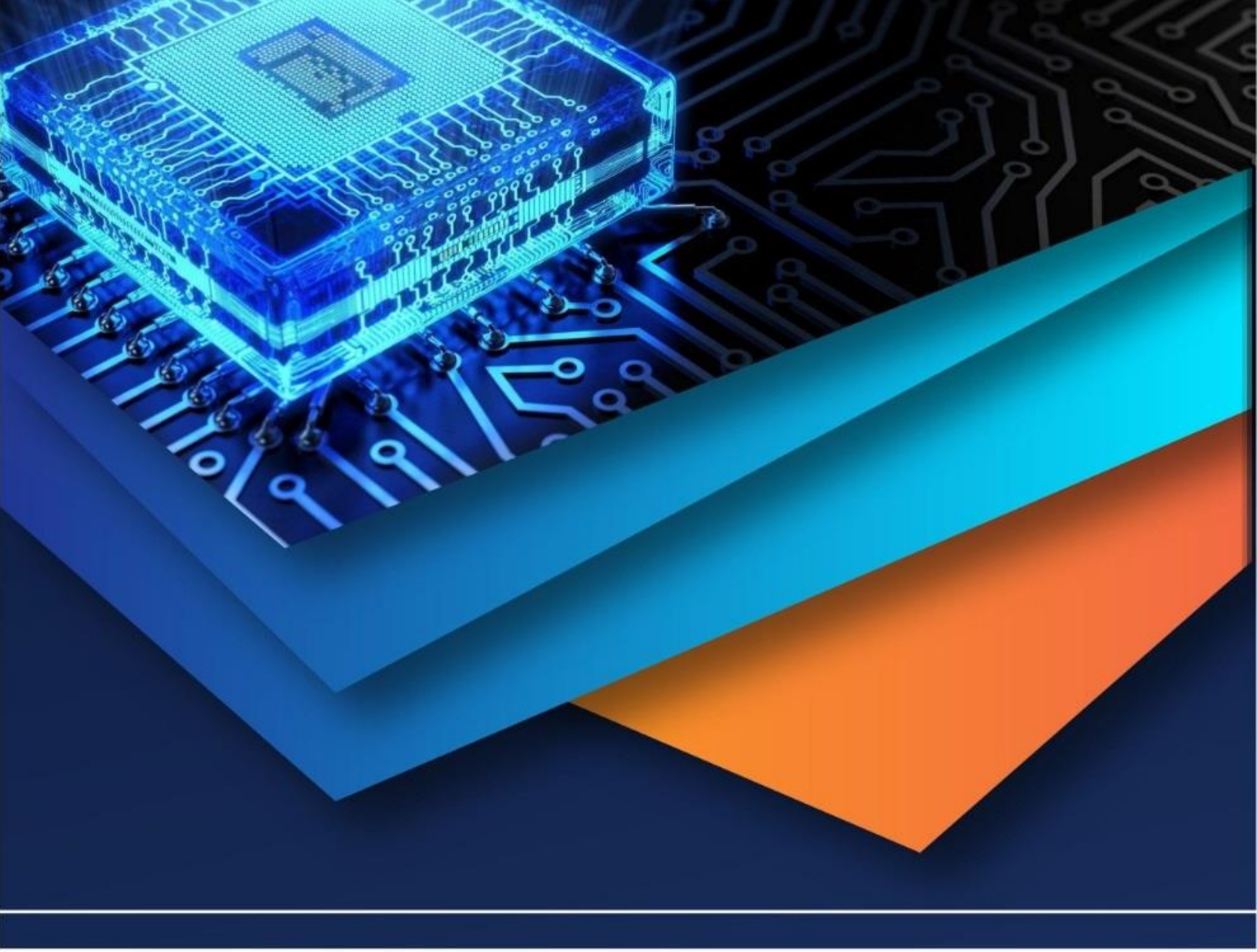
[1] W. K. Ng, Y. Wen, and H. Zhu. Private data deduplication protocols in cloud storage. In S. Ossowski and P. Lecca, editors, Proceedings of the 27th Annual ACM Symposium on Applied Computing, pages 441–446. ACM, 2012.

[2] Iuon –Chang Lin, Po-ching Chien ,"Data Deduplication Scheme for Cloud Storage" International Journal of Computer and Control(IJ3C),Vol1,No.2(2012)

[3] Bugiel, S., N¨urnberger, S., Sadeghi, A.-R., Schneider,T.: Twin Clouds: An architecture for secure cloud computing (Extended Abstract). In: Workshop on Cryptography and Security in Clouds (WCSC 2011), March 15-16 (2011)

[4] S. Halevi, D. Harnik, B. Pinkas, and A. Shulman-Peleg. Proofs of ownership in remote storage systems. In Y. Chen, G. Danezis, and V. Shmatikov, editors, ACM Conference on Computer and Communications Security, pages 491–500. ACM, 2011.

[5] S. Bugiel, S. Nurnberger, A. Sadeghi, and T. Schneider. Twin clouds: An architecture for secure cloud computing. In Workshop on Cryptography and Security in Clouds (WCSC 2011), 2011.
[6] M. Armbrust, A. Fox, R. Griffith, A. D. Joseph, R. Katz, A. Konwinski,G. Lee, D. Patterson, A. Rabkin, I. Stoica, and M.Zaharia, "A view of cloud computing," Communication of the ACM, vol. 53, no. 4, pp.50–58, 2011.
[7] Wen Xia, Member,Hong Jiang "DARE: A Deduplication-Aware Resemblance Detection and Elimination Scheme for Data Reduction with Low Overheads", ,IEEE TRANSACTIONS ON COMPUTERS, VOL. 65, NO. 6, JUNE 2016.
[8] Zheng Yan, Wenxiu Ding,Xixun Yu,"Deduplication on Encrypted Big Data in Cloud",IEEE TRANSACTIONS ON BIG DATA, VOL. 2, NO. 2, APRIL-JUNE 2016.
[9] Rongmao Chen,Yi Mu,"BL-MLE: Block-Level Message-Locked Encryption for Secure Large File Deduplication",IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY.
[10] Xue Yang, Rongxing Lu, Kim Kwang Raymond Choo, Fan Yin, and Xiaohu Tang, "Achieving Efficient and Privacy-Preserving Cross-Domain Big Data Deduplication in Cloud" IEEE 2017.
[11] Chunlu Wang, Jun Ni, Tao Xu, Dapeng Ju "TH_Cloudkey: Fast, Secure and lowcost backup system for using public cloud storage" IEEE 2013.
[12] Jan Stanek, and Lukas Kencl," Enhanced Secure Thresholded Data Deduplication Scheme for Cloud Storage". IEEE 2016.
[13] Yuan Zhang, Chunxiang Xu, Hongwei Li, Kan Yang, Jianying Zhou, and Xiaodong Lin, "HealthDep: An Efficient and Secure Deduplication Scheme for Cloud-Assisted eHealth Systems" 2018 IEEE.
[14] Huijun Wu, Chen Wang, Yinjin Fu, Sherif Sakr, Kai Lu, Liming Zhu, "A Differentiated Caching Mechanism to Enable Primary Storage Deduplication in Clouds" 2017 IEEE.
[15] Youngjoo Shin, Dongyoung Koo, Joobeom Yun and Junbeom Hur, "Decentralized Server-aided Encryption for Secure Deduplication in Cloud Storage" IEEE 2017.
[16] Z. Li, X. Zhang, and Q. He, Analysis of the key technology on cloud storage, in International Conference on Future Information Technology and Management Engineering, 2010, pp. 427428.
[17] D. Harnik, B. Pinkas, and A. Shulman-Peleg. Side channels in cloud services: Deduplication in cloud storage. IEEE Security & Privacy, 8(6), 2010.
[18] Mohamed Adel Serhani, Chafik Bouhaddioui, Rachida Dssouli," Big Data Quality: A Quality Dimensions Evaluation",ResearchGate, DOI:10.1109/UIC-ATC-ScalCom-CBDCom-IoPSmartWorld.2016.0122
[19] P. Anderson and L. Zhang. "Fast and secure laptop backups with encrypted deduplication". In Proc. of USENIX LISA, 2010.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY