iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 3 **Issue:** XII **Month of publication:** December 2015

DOI:

www.ijraset.com

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

International Journal for Research in Applied Science & Engineering Technology (IJRASET)

Secure Data Deduplication in Cloud

Ms. Reshma D. kapadi[#1], Mr. Pramod Patil[*2], Mr. Prashant v. Raut[#3]
[#1] *Department of Computer Engineering, NMIET, Pune, India.*
#2 Department Of Information Technology, NMIET, Pune, India.
#3 Department of Computer Science & Engineering, PCST, Bhopal, India.

Abstract— Secure deduplication is a system for discarding duplicate copies of limit data, and offers security to them. To lessening storage space and exchange information move limit in circulated stockpiling deduplication has been an unquestionably comprehended methodology. Thus simultaneous encryption has been generally get for secure deduplication, essential issue of making blended encryption sensible is to capably and constantly manage a giant number of joined keys. The key thought in this paper is that we can take out duplicate copies of limit data and most distant point the mischief of stolen data in case we decrease the estimation of that stolen information to the attacker. This paper makes the first attempt to formally address the issue of fulfilling beneficial and strong key organization in secure deduplication. We first present an example approach in which each customer holds a free expert key for scrambling the joined keys and outsourcing them. In any case, such a gage key organization arrangement creates countless with the growing number of customers and obliges customers to dedicatedly secure the master keys. To this end, we propose Dekey, User Behavior Profiling and Decoys advancement. Dekey new advancement in which customers don't need to manage any keys isolated however rather securely spread the blended key shares over distinctive servers for insider assailant. As a proof of thought, we execute Dekey using the Ramp puzzle sharing arrangement and demonstrate that Dekey secures confined overhead in sensible circumstances. Customer profiling and impersonations, then, fill two needs. Beginning one is tolerating whether data access is endorsed when peculiar information access is distinguished, and second one is that mixing up the assailant for fake information.

Keywords: Cloud, Security, Encryption.

I. INTRODUCTION

Distributed computing gives vast virtualized arrangement of activity to customer as organizations over the whole web while disguising the stage and executing unobtrusive components. Dispersed capacity organization is the organization of evergreen growing mass of data. To make data organization versatile in conveyed registering, deduplication has been a standard strategy. Data weight technique is used for getting rid of the duplicate copies of repeated data in appropriated stockpiling to diminish the data duplication. This system is used to improve stockpiling utilize besides be associated with system data trades to diminish the amount of bytes that must be sent. Keeping various data copies with the practically identical substance, deduplication wipes out overabundance data by keeping one and just physical copy and imply other tedious data to that copy. Data deduplication happens record level and moreover square level. The duplicate copies of indistinct archive discard by record level deduplication .For the square level duplication which wipes out duplicates bits of data that happen in non-undefined reports. Despite the way that data deduplication takes an impressive measure of preferences, security furthermore insurance concerns develop as customers' tricky data are talented to both insider and outsider attacks. In the traditional encryption giving data security, is clashing with data deduplication. Traditional encryption requires different customers to encode their data with own keys. Copies of indistinct archive discard by record level deduplication .For the square level duplication which wipes out duplicates bits of data that happen in non-undefined reports. Despite the way that data deduplication takes an impressive measure of preferences, security furthermore insurance concerns develop as customers' tricky data are talented to both insider and outsider attacks. In the traditional encryption giving data security, is clashing with data deduplication. Traditional encryption requires different customers to encode their data with own keys.

For making the achievable deduplication and keep up the data mystery used united encryption framework. It encodes deciphers a data copy with a blended key, the data's substance copy got by enlisting the cryptographic hash estimation of. After the data encryption and key time process customers hold the keys and send the cipher text to the cloud. Since the encryption operation is determinative and is gotten from the data content, relative data copies will deliver the same centred key and consequently the same cipher text. A sheltered affirmation of ownership tradition is used to keep the unapproved get to besides give the confirmation to customer regarding the duplicate is found of the same record. With the explosive growth of digital data, deduplication techniques

are widely employed to backup data and minimize network and storage overhead by detecting and eliminating redundancy among data. Instead of keeping multiple data copies with the same content, deduplication eliminates redundant data by keeping only one physical copy and referring other redundant data to that copy. Deduplication has received much attention from both academia and industry because it can greatly improves storage utilization and save storage space, especially for the applications with high deduplication ratio such as archival storage systems. A number of deduplication systems have been proposed based on various deduplication strategies such as client-side or server-side deduplications, file-level or block-level deduplications. A brief review is given in Section 6. Especially, with the advent of cloud storage, data deduplication techniques become more attractive and critical for the management of ever-increasing volumes of data in cloud storage services which motivates enterprises and organizations to outsource data storage to third-party cloud providers, as evidenced by many real-life case studies [1]. According to the analysis report of IDC, the volume of data in the world is expected to reach 40 trillion gigabytes in 2020 [2]. Today's commercial cloud storage services, such as Dropbox, Google Drive and Mozy, have been applying deduplication to save the network bandwidth and the storage cost with client-side deduplication. There are two types of deduplication in terms of the size:

A. File-level deduplication File-level deduplication which discovers redundancies between different files and removes these redundancies to reduce capacity demands.

B. Block level deduplication Block level deduplication which discovers and removes redundancies between data blocks. The file can be divided into smaller fixed-size or variable-size blocks. Using fixed-size blocks simplifies the computations of block boundaries, while using variable-size blocks provides better deduplication efficiency.

II. RELATED WORK

M. Bellare, S. Keelveedhi, and T. Ristenpart, "Dupless: Server- aided encryption for deduplicated storage," in USENIX Security Symposium, 2013. Distributed storage administration suppliers, for example, Dropbox, Mozy, and others perform deduplication to spare space by just putting away one duplicate of each file transferred. Should customers expectedly scramble their files, on the other hand, investment funds are lost. Message-bolted encryption (the most noticeable indication of which is focalized encryption) re-understands this strain. On the other hand it is characteristically subject to animal power assaults that can recuperate files falling into a known set.

They proposed a construction modeling that ace vides secure deduplicated stockpiling opposing savage power assaults, and acknowledge it in a framework called DupLESS. In DupLESS, customers encode under message-based keys obtained from a key-server by means of a careless PRF convention. It empowers customers to store encoded information with an existing administration, have the administration perform deduplication for their sake, but accomplishes solid confidentiality guarantees. They have demonstrated that encryption for deduplicated stockpiling can accomplish execution and space reserve funds near that of utilizing the stockpiling administration with plaintext information. Utilizing the stockpiling administration with plaintext information.

J. R. Douceur, A. Adya, W. J. Bolosky, D. Simon, and M. Theimer, "Reclaiming space from duplicate files in a serverless distributed file system." in ICDCS, 2002, pp. 617-624. This framework gives accessibility by recreating every record onto numerous desktop PCs. Since this replication expends critical storage room, it is essential to recover utilized space where conceivable. Estimation of more than 500 desktop document frameworks demonstrates that about portion of all devoured space is possessed by copy records. They introduce an instrument to recover space from this accidental duplication to make it accessible for controlled document replication.

Their system incorporates (1) focalized encryption, which empowers copy records to combine into the space of a solitary document, regardless of the fact that the documents are scrambled with diverse clients' keys, and (2) SALAD, a Self-Arranging, Lossy, Associative Database for conglomerating document substance and area data in a decentralized, adaptable, issue tolerant way. Substantial scale recreation analyses demonstrate that the copy document mixing framework is versatile, exceedingly successful, and deficiency tolerant.

A. D. Santis and B. Masucci, "Multiple ramp schemes," IEEE Transactions on Information Theory, vol. 45, no. 5, pp. 1720-1728, Jul. 1999. An incline plan is a convention to disseminate a mystery s picked in S among a set P of n members in a manner that: (1) sets of members of cardinality more noteworthy than or equivalent to k can reproduce the mystery s; (2) sets of members of cardinality not as much as or equivalent to t have no data on s, though (3) sets of members of cardinality more noteworthy than t and not as much as k may have "some" data on s. In this correspondence they examine various incline plans, which are conventions to share numerous privileged insights among a set P of members, utilizing distinctive slope plans. Specifically, they demonstrate a tight lower bound on the offers extent held by every member and on the merchant's arbitrariness in numerous slope plans.

Message-locked encryption and secure deduplication, in EUROCRYPT, 2013, pp. 296-312. It's another cryptographic primitive, Message-Locked Encryption (MLE), where the key under which encryption and decoding are performed is itself gotten from the message. MLE gives an approach to accomplish secure deduplication (space-efficient secure outsourced stockpiling), an objective right now focused by various distributed storage suppliers. They give definitions both to protection and for a type of respectability that they call label consistency. In light of this establishment, they make both common sense and hypothetical commitments. On the pragmatic side, they give ROM security investigations of a characteristic group of MLE plans that incorporates conveyed plans. On the hypothetical side the test is standard model arrangements, and they make associations with deterministic encryption, hash capacities secure on connected inputs and the example then-extricate worldview to convey plans under diverse presumptions and for distinctive classes of message sources. Their work demonstrates that MLE is a primitive of both down to earth and hypothetical hobby.

A. Shamir, How to share a secret, Commun. ACM, vol. 22, no. 11, pp. 612-613, 1979. In this, its demonstrated to that industry standards to gap information D into n pieces in a manner that D is effortlessly reconstructable from any k pieces, however even finish learning of k - 1 pieces uncovers truly no data about D. This method empowers the development of hearty key administration plans for cryptographic frameworks that can work safely and dependably notwithstanding when adversities wreck a large portion of the pieces and security ruptures uncover everything except one of the remaining pieces.

S. Halevi, D. Harnik, B. Pinkas, and A. Shulman-Peleg, "Proofs of ownership in remote storage systems." in ACM Conference on Computer and Communications Security, Y. Chen, G. Danezis, and V. Shmatikov, Eds. ACM, 2011, pp. 491-500. Distributed storage frameworks are turning out to be progressively mainstream. A promising innovation that holds their expense down is deduplication, which stores just a solitary duplicate of rehashing information. Customer side deduplication endeavors to recognize deduplication opportunities as of now at the customer and recovery the data transfer capacity of transferring duplicates of existing _les to the server. In this work they recognize assaults that endeavor customer side deduplication, permitting an aggressor to get entrance to subjective size files of different clients in light of a little hash marks of these files. All the more specifically, an assailant who knows the hash mark of a file can persuade the capacity benefit that it possesses that file, consequently the server lets the aggressor download the whole file. (In parallel to their work, a subset of these assaults were as of late presented in the wild concerning the Dropbox _le synchronization administration.) To overcome such assaults,

They present the thought of confirmations of-possession (PoWs), which lets a customer efficiently demonstrate to a server that that the customer holds a file, as opposed to simply some short information about it. They formalize the idea of confirmation of proprietorship, under thorough security definitions, and thorough efficiency requirements of Petabyte scale stockpiling frameworks. They then present arrangements taking into account Merkle trees and specific encodings, and break down their security. They actualized one variation of the plan.

III.METHODS USED IN SECURE DEDUPLICATION

The primitives mentioned below have been used for secure deduplication of data in the cloud:

A. Symmetric Encryption

Symmetric encryption uses common secret key k to encrypt and decrypt information. The symmetric encryption scheme made up of three primary functions. These functions are mentioned below:

KeyGen SE (1x)^ k is the key generation algorithm that generates k using security parameter 1x;

Enc SE (k, M)^ C is the symmetric encryption algorithm that takes secret k, and message M and then outputs cipher text C, and

Dec SE (k, C) ^ M is a symmetric decryption algorithm that takes the secret k and cipher text C and then outputs original message M.

B. Convergent Encryption

Convergent encryption provides the data confidentiality in the deduplication. A user derives a convergent key from the each original data copy and encrypts the data copy with convergent key. In addition, the user also derive tag for the data copy, such that to detect duplicates the tag will be used Here, we assume that the tag holds the property of correctness , i.e., if two data copies are the same, the tags of data also same. The user first sends the tag to the server side to check if identical copy has been already stored for detect

duplicates.

C. Proof of Ownership

The notion of proof of ownership (PoW) enables users to prove their ownership of data copies to the storage server. Specifically, The Proof of the ownership is implemented as a interactive algorithm run by user and storage server.

D. Identification Protocol

The identification of the protocol having two phases as follows:

1) Proof: The user can demonstrate his identity to a verifier by performing some identification proof which is related to his identity.

2) Verify: The verifier occurs verification with input of public information.

IV. PROBLEM FORMULATION AND SYSTEM IMPLEMENTATION

Various suggestions have been made to secure remote data in the Cloud using encryption and standard access controls. Most would concur the standard's greater part approaches have been shown to miss the mark from time to time for an arrangement of reasons, including insider ambushes, misconfigured organizations, broken utilization, surrey code, and the creative improvement of fruitful and present day attacks not envisioned by the implementers of security methodologies. Building a dependable circulated registering environment is inadequate, in light of the way that accidents continue happening, and when they do, and information gets lost, there is no genuine approach to get it back. Though deduplication technique can save the storage space for the cloud storage service providers, it reduces the reliability of the system. Data reliability is actually a very critical issue in a deduplication storage system because there is only one copy for each file stored in the server shared by all the owners. If such a shared file/chunk was lost, a disproportionately large amount of data becomes inaccessible because of the unavailability of all the files that share this file/chunk. If the value of a chunk were measured in terms of the amount of file data that would be lost in case of losing a single chunk, then the amount of user data lost when a chunk in the storage system is corrupted grows with the number of the commonality of the chunk.

In this work we intend to perform, we can take out duplicate copies of limit data and cutoff the mischief of stolen data if we reduce the estimation of that stolen information to the attacker. Affirming whether data access is endorsed when abnormal information access is recognized, and confusing the attacker with fake information. For the circumstance where the passage is successfully identified as an unapproved access, the Cloud security structure would pass on unbounded measures of sham information to the adversary, thusly securing the client's certified data. Here we propose the system that accomplishes the above mentioned goal. This scheme is a main tool in building Threshold Cryptosystems and other objects- secret sharing is a pretty big field, and we will spend several lectures describing it. The basic Secret Sharing Scheme is given by two algorithms: sharing (Share) and recovery (Rec). Rec is a deterministic algorithm which recreates the message from some or all of the shares. It operates in the way you would expect Share takes a message M and split it into pieces. Since M is secret, Share must introduce randomness (that is, Share is probabilistic). To indicate randomness we will use arrow ! :

There are two types of algorithm:

The Sharing Algorithm: Share (M)! (S1; S2;):

The Recovery Algorithm: Rec(S S1, S2,)=M'.

In sharing algorithm the secrets (s1, s2, sn) are distributed securely among servers 1 through n, and pub is a public share. (We include pub for the sake of generality but observe that it is often empty. If pub is present, we assume that it is authenticated, so no one can change it: it's just published in the sky.) In Recovery algorithm the correctness property of the algorithm says that for any message M, Rec (Share (M)) = M.

Two kinds entities will be involved in this deduplication system, including the user and the storage cloud service provider (S-CSP). Both client-side deduplication and server-side deduplication are supported in our system to save the bandwidth for data uploading and storage space for data storing.

User. The user is an entity that wants to outsource data storage to the S-CSP and access the data later. In a storage system supporting deduplication, the user only uploads unique data but does not upload any duplicate data to save the upload bandwidth. Furthermore, the fault tolerance is required by users in the system to provide higher reliability.

S-CSP. The S-CSP is an entity that provides the outsourcing data storage service for the users. In the deduplication system, when users own and store the same content, the S-CSP will only store a single copy of these ?les and retain only unique data. A deduplication technique, on the other hand, can reduce the storage cost at the server side and save the upload bandwidth at the user side. For fault tolerance and confidentiality of data storage, we consider a quorum of S-CSPs, each being an independent

Volume 3 Issue XII, December 2015
ISSN: 2321-9653

entity. The user data is distributed across multiple S-CSPs.

The applicants first generate the hash key for the data and uploads its data to the cloud. This data is encrypted with the help of hash value that has been already generated by the applicant. If the hash key is wrong then the data will not be encrypted and the process stops. The system will automatically prompts the error message to the applicant and applicant will automatically log out from the system.

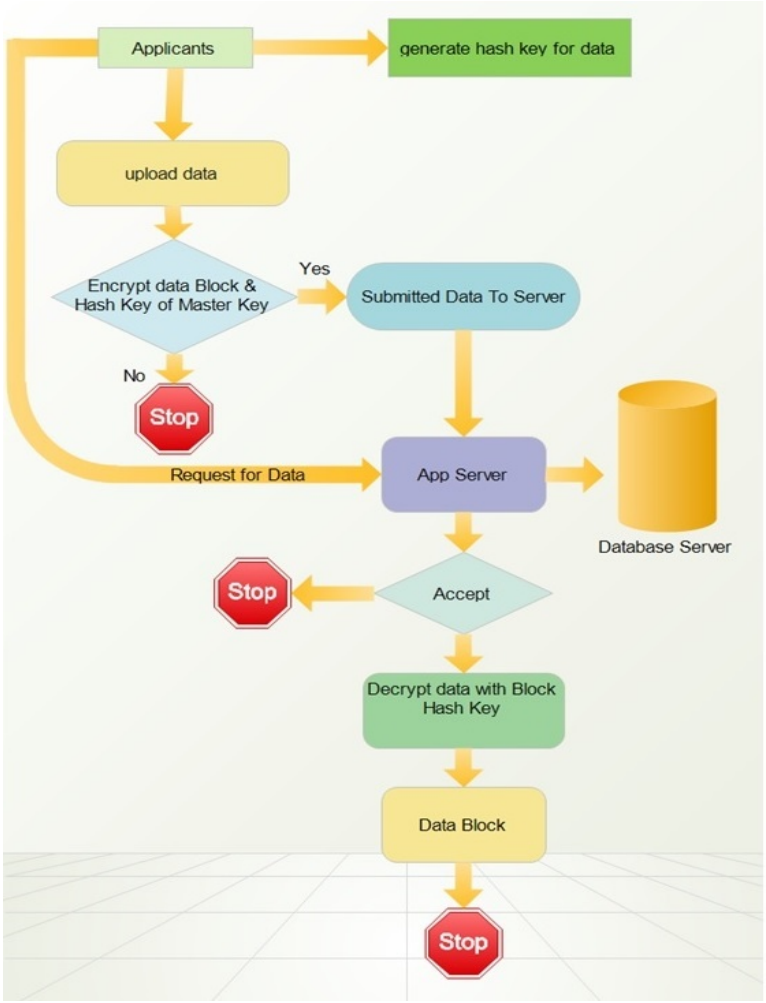


Figure 1. System Implementation

The encryption of data and further process is possible only if the generated and provided hash code matches. Then this encrypted data is submitted to the application server. The application server stores all the data to the database server. When the user requests to the data that has been stored on the database server in the encrypted form, the request is directly forwarded to the application server. Which in turn makes a query to the database server. The data provided to the applicant in the encrypted form.

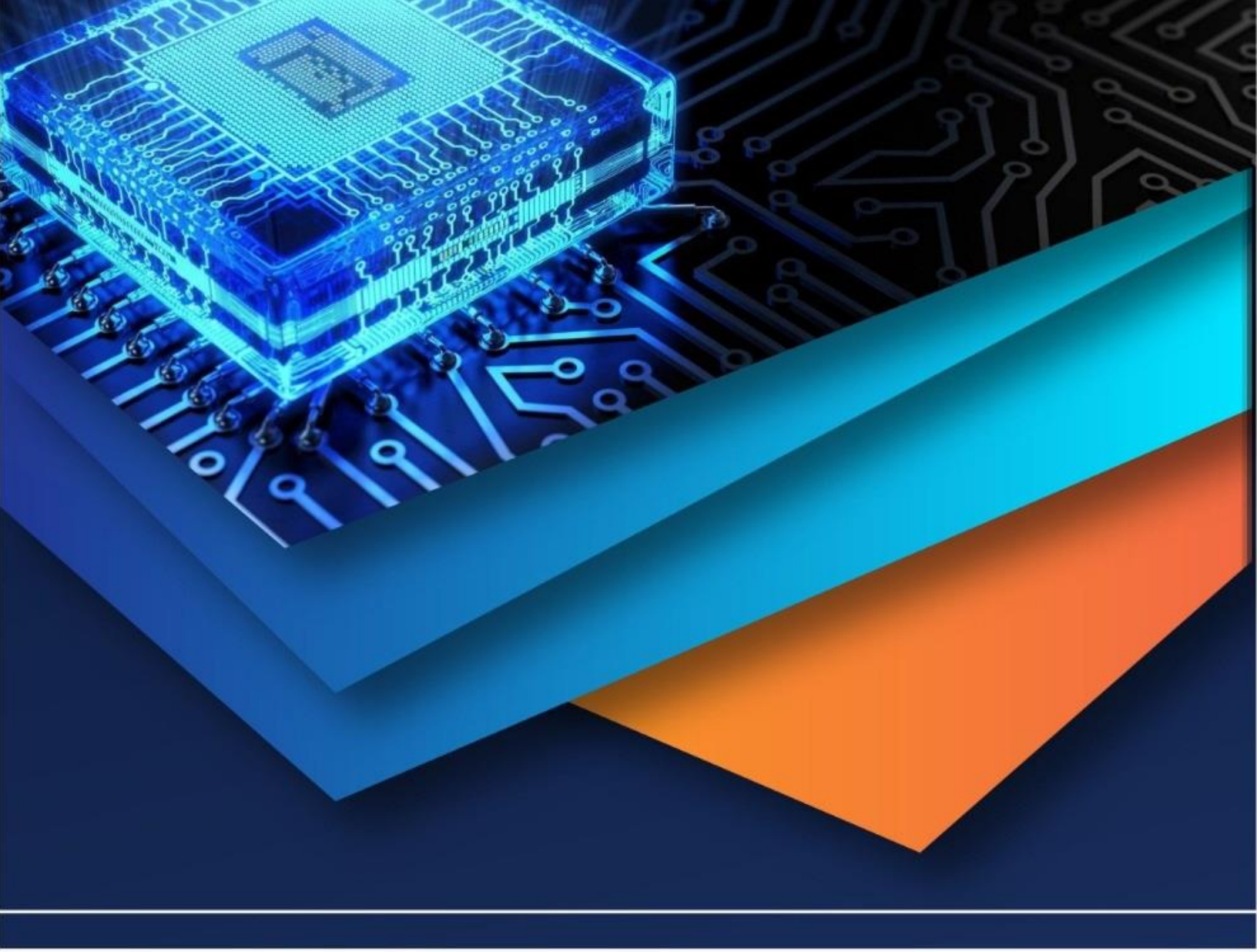
In order to decrypt the received data which is in encrypted form, the same hash value is used. The applicant has to provide the hash value then, the data is decrypted and provided to the applicant in the original form. If the hash value provided by the applicant at the time of decryption of data is wrong, then the applicant is unable decrypt the data.

V. CONCLUSION

The basic believed is that we set that ensured deduplication organizations can be executed given additional security highlights insider attacker on Deduplication and untouchable aggressor by using the disclosure of masquerade activity. The attacker's perplexity and the additional costs created to perceive real from false information, and the demoralization sway which, though hard to measure, expect a basic part in preventing masquerade development by threat unwilling aggressors. We set that the blend of these security components will give remarkable levels of security to the deduplication.

REFERENCES

[1] J. Gantz and D. Reinsel, The digital universe in 2020: Big data, bigger digital shadows, and biggest growth in the far east, http://www.emc.com/collateral/analyst-reports/idcthedigital-universe-in-2020.pdf, Dec 2012.
[2] M. O. Rabin, Fingerprinting by random polynomials, Center for Research in Computing Technology, Harvard University, Tech. Rep. Tech. Report TR-CSE-03-01, 1981.
[3] J. R. Douceur, A. Adya, W. J. Bolosky, D. Simon, and M. Theimer, Reclaiming space from duplicate ?les in a serverless distributed file system. in ICDCS, 2002, pp. 617624.
[4] M. Bellare, S. Keelveedhi, and T. Ristenpart, Dupless: Serveraided encryption for deduplicated storage, in USENIX Security Symposium, 2013.
[5] Message-locked encryption and secure deduplication, in EUROCRYPT, 2013, pp. 296312.
[6] G. R. Blakley and C. Meadows, Security of ramp schemes, in Advances in Cryptology: Proceedings of CRYPTO 84, ser. Lecture Notes in Computer Science, G. R. Blakley and D. Chaum, Eds. Springer-Verlag Berlin/Heidelberg, 1985, vol. 196, pp. 242268.
[7] A. D. Santis and B. Masucci, Multiple ramp schemes, IEEE Transactions on Information Theory, vol. 45, no. 5, pp. 17201728, Jul. 1999.
[8] OpenSSL Project. http://www.openssl.org/.
[9] M. Bellare, S. Keelveedhi, and T. Ristenpart. Dupless: Server aided encryption for deduplicated storage. In USENIX Security Symposium, 2013.
[10] J. Yuan and S. Yu. Secure and constant cost public cloud storage auditing with deduplication .IACR Cryptology ePrint Archive, 2013.
[11] J. Li, X. Chen, M. Li, J. Li, P. Lee, and W. Lou. Secure deduplication with efficient and reliable convergent key management. In IEEE Transactions on Parallel and Distributed Systems, 2013.
[12] M. Bellare, S. Keelveedhi, and T. Ristenpart. Message locked encryption and secure deduplication. In EUROCRYPT, pages 296– 312, 2013.
[13] J. Xu, E.-C. Chang and J. Zhou. Weak leakage-resilient client-side deduplication of encrypted data in cloud storage. In ASIACCS, pages 195–206, 2013.
[14] C. Ng and P. Lee. Revdedup: A reverse deduplication storage system optimized for reads to latest backups. In Proc. of APSYS, Apr 2013. W. K. Ng, Y. Wen, and H. Zhu. Private data deduplication protocols in cloud storage. In S Ossowski and P. 2012.
[15] R. D. Pietro and A. Sorniotti . Boosting efficiency and security in proof of ownership for deduplication. In H. Y. Youm and Y. Won, editors, ACM Symposium on Information, Computer and Communications Security2012.
[16] 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.
[17] 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.
[18] K. Zhang, X. Zhou, Y. Chen, X. Wang, and Y. Ruan. Sedic: privacy aware data intensive computing on hybrid clouds. In Proceedings of the 18th ACM conference on Computer and communications security, CCS'11, pages 515–526, New York, NY, USA, 2011. ACM.
[19] A. Rahumed , H. C. H. Chen, Y. Tang, P. P. C. Lee, and J. C. S. Lui. A secure cloud backup system with assured deletion and version control. In 3rd International Workshop on Security in Cloud Computing, 2011.
[20] M. Bellare, C. Namprempre , and G. Neven. Security proofs for identity-based identification and signature schemes. J. Cryptology, 2009.
[21] M. Bellare and A. Palacio. Gq and schnorr identification schemes: Proofs of security against impersonation under active and concurrent attacks. In CRYPTO, pages 162– 177,2002



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY