Deduplication in Hybrid Cloud with Secure Data

Ashwini Shete
M B E Society's
College of Engineering,
Ambajogai Maharastra, India

B. M. Patil
M B E Society's
College of Engineering,
Ambajogai Maharastra, India

ABSTRACT
Deduplication is also called single instance technique, deduplication remove redundant data and stores original copy of data so it will saves the storage space to protect sensitive data. The data security and access to particular data is very much important in current days hence the features in deduplication have been widely used in cloud storage system. There was drawback in previous work where differential privileges of users are set permanently so any client can access any files which are not authorized to specific user. To overcome the drawback of the previous work, we are concentrated on client privileges which will grant or revoke the file access permissions to specific client. With deduplication technique, the client data is secured with advanced encryption algorithm rather than conventional encryption algorithm in hybrid cloud. In previous work convergent encryption technique was used with deduplication which does not provide the security to data. To better enhancement of the security of data, client has given specific permission to access the data, which is more enhanced security system for data security .Our simulation results shows that increase in security of data.

Keywords
Deduplication, Privileges Set, Hybrid Cloud, Conventional Encryption.

1. INTRODUCTION
Cloud computing is network of remote servers, which is hosted on the Internet to store, manage, and process the data instead of local server or a personal computers. Cloud computing is on-demand, pay-as-you-go access model for multiple customers. Cloud computing acting as a shared pool of computing resources e.g. servers, networks, storage, applications, and services. It can be continuously and easily accessed over the Internet. [1]Cloud computing is widespread, an large amount of data is being stored in the cloud and accessed by users with specified privileges. The management of the ever-increasing volume of data is critical challenge of cloud storage system, the data management makes possible through deduplication concept attracts more recently. In most of the organizations, the storage systems contain duplicate copies of data. The same file may be saved at different places by different users, which means multiple copies are getting created which has same data. Deduplication eliminates these extra copies which get created by saving only one copy of the data and replacing the other copies with pointers that lead back to the original copy. Deduplication technique is used in companies for backup and disaster recovery applications, this technique also used to make free space in the storage systems. Deduplication technique is having two levels, first is file level and second is block level. [2][3] The file level deduplication technique also called as Single Instance Storage (SIS) which will remove the duplicate copy of same file. Block level deduplication technique will reduces duplicated blocks of data that occurs in two different files. The Block-level deduplication makes more free space than SIS. J. R. Douceur et al. defines convergent encryption technique in which user encrypts or decrypts data with convergent key which is generated through cryptographic hash value from the content of data. For identical data copies same ciphertext will be generated hence convergent encryption is possible with deduplication technique. S. Halevi et al. defines proof of ownership protocol(POW) [11] which will secure the data from unauthorized access of data. The data in proof of ownership protocol, proofs that the user belongs to his own file and using own file when duplicate file is found. Once the proof get confirmed server will provide pointer to the same file to the user without need of upload to the same file.[4]

However, the old deduplication system does not support the differential authorization duplicate check, which is significant in most of the applications. In unauthorized deduplication systems, during system initialization set of privileges will be issues to each users. The privileges are defined (or bound) to each file at the time of uploading to the cloud to specify the type of user allowed to perform duplicate check and access of the files. User need to give his own file and own privileges as input before submitting the request for the file. [6]The copy of users own file and matched privileges in the cloud are the key inputs to find the duplicate file for the user. For example, In the company , access control is import factor and which needs to controlled by giving correct privileges to correct user. To realize the access control, company may provide the different privileges to one (or many) employee. For Effective management the data will be moved to the public cloud on storage server provider (S-CSP) with specific privileges and to store only one copy of the same file, the deduplication technique will be used. The privacy needs to be consider while giving access.[7] To realize the access, the control file needs to be encrypted and allowed duplication check with definite privileges. If we want to understand both differential authorization and deduplication check at the same time which may contradict each other's. Over the convergent encryption technique we cannot consider the differential privileges.[8]

2. LITERATURE SURVEY
Now a days lots of people store different types of data in large quantity such as personal data and corporate data on laptops, home computers, personal computers, tablets, smartphones. The data may be vulnerable to insider or outsider attackers also data may get lost due to hardware failure. Traditional backup is not the solution for this environment, this problem can be solved through one algorithm which is described by Paul Anderson et al namely secure encrypted deduplication algorithm. which will increase the speed of backups and eliminates storage requirement . This algorithm supports client -end use per user –encryption which is important for personal and confidential data. [1] Dropbox, Moze are the cloud storage service providers. Performs deduplication stores only one copy of each file protect sensitive data from

different attacks. Sriram Keelveedhi et al proposed a new architecture called Duplicateless Encryption for Simple Storage(Dupless) provides more security for deduplicated storage. Dupless is a group of clients connected to organization(company employee) encrypt their data with the help of key server. That is separate from storage server, clients are authenticate with the key server so they do not leak any information about their data. Key sever is inaccessible to attackers which provides security.[2]

To achieve secure deduplication Mihir Bellare et al defines new cryptographic Concept. Message locked Encryption (MLE) in which key is itself derived from message to perform encryption and decryption, also provides space-efficient secure outsourced storage maintains privacy and integrity of data. MLE scheme broadly divided in to two parts, practical analysis and theoretical analysis. In first scheme, analyse existing system and new variants, justifying others with proofs in the random-oracle-model (ROM).In second scheme, make connection with deterministic encryption. [3] Mihir Bellare et al.explored Identity based Identification (IBS) scheme having master public key and master Secret key. User can provide this authority with secret key based on his identity. user play the role of prover to prove his identity to verifier through the secret key. verifier knowing only the claimed identity of the prover and the master key of the authority. Identity based signature is different scheme based on the signature, user identity is prove through signature with his master public key. Now a days cloud computing is more cost effective for outsourcing storage and perform computations. For secure outsourcing data and computations based on tamper hardware and homomorphic encryption. But tamper hardware is not scalable solution and homomorphic encryption is inefficient. So due to these problems Stefan Nurnberges et al proposed twin cloud architecture. In this architecture, user communicates with the trusted cloud which is private cloud or built from multiple hardware modules which encrypts and verifies the data. The trusted cloud is used for security-critical operations in less time critical phase, where outsourced data is processed in parallel by commodity cloud.[4][5]

Ravi sandhu et al.defines Role based access control is a multi-user and multi -application system. RBAC is concept in which permissions are related with roles, and users, the users are assigned to appropriate role. so due that permissions are simply and greatly managed. In an organization various roles are created to perform job functions and roles are assigned to the users based on their qualification and responsibilities .users can easily switched from one role to another role as per the organization requirement. Roles can be granted new permissions as per the new application, permissions can be removed from the role as needed.[6] Now a days cloud storage systems are more popular, deduplication saves the cost by storing only single copy of repeated files. The client side deduplication perform deduplication and save bandwidth of uploading copies of the existing files to the server. In this work, Shai Halevi et al identify that client side deduplication allowing attacker to gain access file of other users, based on hash signature of the file. Attacker who knows hash signature of file easily prove to server that he (or she) is the owner of the file and server will allow he (or she) to download the entire file. To overcome such attacks, defines the concept of proof of ownership in which client efficiently prove to server that client holds the file.[7]

Deduplication eliminates duplicates and it introduces fragmentation which effects on the read performance. Chun-Ho Ng et al propose new concept reverse deduplication(revdup) which will reads the latest backups of virtual machines. The conventional deduplication removes duplicates from latest data whereas revdedup removes duplicates from old data. revdeduP achieves high efficiency, high backup throughput, and high read throughput.[8] Private data deduplication protocol is used for private data storage .Without see-through other information to the server, the deduplication protocol allows to keep private data to server with private data deduplication protocol technique. With the help of simulation-based framework, the security of private data protocol gets created. It is based on the standard cryptography assumption.[9]The hash function is the collision-resilient, the discrete algorithm is very hard and alpha-fraction bits present spiteful opponents in the present adversaries. Nevertheless multiple clients request to store the file on storage server, the data deduplication techniques store only single copy of the same file. In other words, if client send request to server for storage. The summary of string of file, Merkle-tree hash value. The server will check whether summary of string received and stored in database; if summary of string is not stored in the database then server will ask client to upload entire file to make the client as owner of that file. Otherwise server will inform client that no need to upload and assign particular client as owner of the file.

The cloud storage needs data integrity and storage efficiency which is important factor for any cloud system. There are many techniques available in the market some of them are efficient but having the certain limitation with it. To overcome these limitations we have to study and apply new technique which fulfils data integrity and storage efficiency features of cloud system appropriately. Storage efficiency will increase by removing unreasonable duplicate data from storage server, to achieve this proof of ownership technique will be used. Proof of Data Possession and Proof of Retrieve ability technique used for data integrity on cloud storage. Polynomial base authentication and homomorphic linear authentication are two novel scheme based technique. These technique are very much cost effective technique through which we can increase the productivity as well. As mentioned the cost effective techniques and increase in productivity is motive of all cloud services, One more technique used namely hybrid cloud computing technique, in this technique companies puts private data within companies cloud system and non-sensitive data will move to the public cloud system. By using this hybrid technique we are segregating the task to make it cost effective.

3. OVERVIEW OF EXISTING SYSTEM

A cross breed cloud is a distributed computing environment in which an association gives and deals with a few assets in-house and has others given remotely. For instance, an association may utilize an open cloud administration, for example, Amazon Simple Storage Service(Amazon S3) for chronicled information however keep on maintaining in house stockpiling for operational client information.

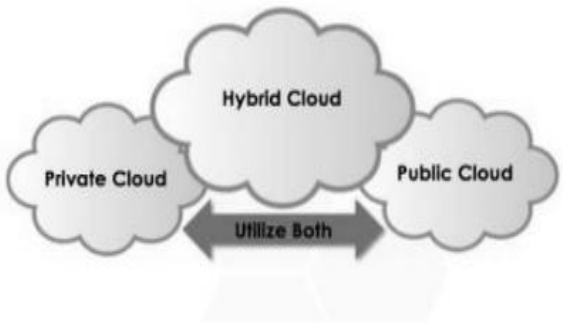

Fig 1: Hybrid Cloud Architecture

The idea of a half and half cloud is intended to conquer any difficulty between high control and high cost private cloud. The private cloud exceedingly callable, adaptable, minimal effort on Open cloud. Private Cloud is typically used to represent a Virtual machine group in which the equipment and programming is utilized and oversaw by a private substance. The idea of an open cloud, includes some type of membership based on asset pools in a facilitating supplier datacenter that uses multi-tenure. The term open cloud doesn't mean less security, yet rather it indicates to multi-tenure. The idea rotates intensely around network and information handiness. There are different use cases: asset burst-capacity for occasional interest, advancement and testing on a uniform stage without expending nearby assets, fiasco restoration, and obviously abundance ability to improve utilization of or free up neighborhood utilization. Virtual Machine has a key device for half and half cloud, it is called vCloud connector. It is a free module that permits the administration of open and private inside the vSphere customer. The device offers clients the capacity to deal with the console view, power status, and workloads tab offers the capacity to duplicate virtual machine formats to and from a remote open cloud advertising. Half and half cloud forsecure de-duplication at an abnormal state, our set of interest is an endeavor system, containing of a gathering of partnered customers (for instance, representatives of an organization) who will utilize the S-CSP and store information with de-duplication strategy. In this setting, de-duplication can be routinely utilized as a part of these settings for information support and fiasco restoration applications while incredibly shrinking storage room. Such frameworks are across the board and are frequently more appropriate to client record strengthening and synchronization applications than wealthier stockpiling reflections. There are three substances characterized in our framework, that is, clients, private cloud and S-CSP in broad daylight cloud. The S-CSP performs de-duplication by checking if the substances of two documents are the same and stores stand out of them. The entrance right to a document is characterized in view of an arrangement of benefits. The careful meaning of a benefit changes crosswise over applications. For instance, we may characterize a role based benefit as indicated by employment positions (e.g., Director, Project Lead, and Engineer), or we may characterize a time based benefit that determines a substantial time period (e.g., 2014-01-01 to 2014-01-31) inside which a document can be developed to A client, say Alice, might be doled out two benefits "Executive" and "get to right legitimate on 2014-01-01", so she can get to any document whose entrance part is "Chief" and available time period covers 2014-01-01. Every benefit is articulated to as a short message called token. Every document is connected with some record tokens, which indicate the tag with determined A client processes and sends copy check tokens to the general population cloud for approved copy check. Clients have entry to the private cloud server, a semitrusted outsider which will help in performing de- duplicable encryption by creating document tokens by asking for clients. We will clarify further the part of the private cloud server base. Clients are additionally provisioned with per-client encryption keys and approvals A. Engineering For Authorized De-duplication

3.1 Proposed System Architecture

In this paper, we will just consider the document level de-duplication for simplicity. In another word, we avoid an information duplicate to be an entire record and document level de-duplication which kills the capacity of any repetitive documents. Really, block level de-duplication can be effectively found from record level de-duplication, specifically, to transfer a document, a client first performs the record level copy check. On the off chance that the document is a copy, then all its squares must be copies also; something else, the client further performs the piece level copy check and distinguishes the remarkable squares to be transferred. Every information duplicate (i.e., a document or a piece) is connected with a token for the copy check. In S-CSP, This is a substance that gives an information stockpiling administration out in the open cloud. The S-CSP gives the information outsourcing administration and stores information in the interest of the clients. To reduce the capacity cost, the S-CSP takes out the capacity of excess information by use of de-duplication and keeps required information. In this paper, we accept that S-CSP is constantly online and has copious capacity limit and calculation power. In Data Users, A client is an element that needs to outsource information stockpiling

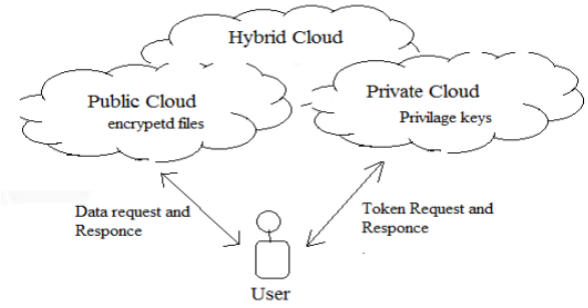

Fig 2: Proposed System Architecture

To the S-CSP and access the information later. In an ability to framework supporting de-duplication activity, the client extraordinary information, does not transfer any copy information to spare the transfer data transmission, which might be possessed by the same client or distinctive clients. In the approved de-duplication framework, every client is issued an arrangement of benefits in the setup of the framework. Every record is secured with the joined encryption key and benefit keys to understand the approved de-duplication with differential benefits. In Private Cloud. Contrasted and the customary de-duplication design in distributed computing, this is another element presented for encouraging client's protected utilization of cloud administration. In particular, since the registering assets at information client/proprietor side are limited and the general population cloud is not completely confided by and by, private cloud can give information client/proprietor with an execution situation and foundation functioning as an interface amongst client and people in general cloud. The private keys for the benefits are overseen by the private cloud, who answers the document token solicitations from the clients. The interface offered by the private cloud permits client to submit records and questions to be safely put away and figured separately. This is

a fresh plan for information de-duplication in distributed computing, which comprises of a twin mists (i.e., people in general cloud and the private cloud). Really, this cross breed cloud setting has pulled in more consideration as of late. For instance, an undertaking may utilize an open cloud administration, for example, Amazon S3, for filed information, however keep on maintaining in-house stockpiling for operational client information. Then again, the trusted private cloud could be a group of virtualized cryptographic co-processors, which are offered as an administration by an outsider and give the fundamental equipment based security elements to actualize a remote execution environment trusted by the clients. node movement. It also configures the location service with the termination timer for the cached entries.

4. RESULTS

The final results of the designed system are given below. From those results we get the detailed information to Check de-duplication and upload the files, Fetching the Signs using Hashing Algorithm, Checking for Duplication, file uploading, file downloading and attacker trying to attack(block) the cloud. Detailed procedure of the proposed system is given. Based on this we confirm that securely authorized de-duplication is successfully achieved with hybrid cloud approach. The output images as well as performance graphs given as below,

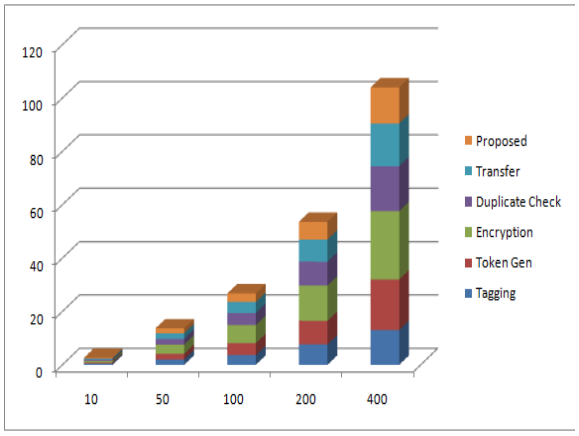

Fig 3: Time Breakdown for different fie size

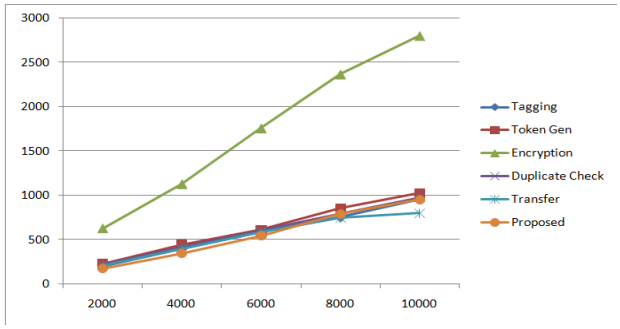

Fig 4: Time Breakdown for different no of files stored

5. CONCLUSION AND FUTURE WORK

The idea of approved information de-duplication was proposed to ensure the information security by including differential benefits of clients in the copy check. We additionally introduced a few new de-duplication developments supporting approved copy check in crossover cloud design, in which the copy check tokens of documents are produced by the private cloud present with private keys. Security examination shows that our plans are secure as far as insider and untouchable assaults indicated in the proposed security model. As a proof of idea, we executed a model of our proposed approved copy check plan and direct proving ground investigates our model. We demonstrated that our approved copy check plan brings about negligible overhead contrasted with joined encryption and system exchange. For future enhancement, It bars the security issues that may emerge in the down to earth sending of the present model. Additionally, it builds the national security. It spares the memory by de-copying the information and in this manner give us adequate memory. It gives approval to the private firms and secure the classification of the essential information

6. REFERENCES

[1] P. Anderson and L. Zhang, "Fast and secure laptop backups with encrypted de-duplication," in Proc. 24th Int. Conf. Large Installation.

[2] M. Bellare, S. Keelveedhi, and T. Ristenpart, "Dupless: Serveraided encryption for deduplicated storage," in Proc. 22nd USENIX Conf. Sec. Symp., 2013, pp. 179–194.

[3] M. Bellare, S. Keelveedhi, and T. Ristenpart, "Message-locked encryption and secure deduplication," in Proc. 32nd Annu. Int. Conf. Theory Appl. Cryptographic Techn., 2013, pp. 296–312.

[4] M. Bellare, C. Namprempre, and G. Neven, "Security proofs for identity-based identification and signature schemes," J. Cryptol. vol. 22, no. 1, pp. 1–61, 2009..

[5] Bugiel,SNurnberger,ASadeghian,T.Schneider,"Twinclouds: Anarchitectureforsecurecloudcomputing,"inProc.Worksh op Cryptography Security Clouds, 2011

[6] A D. Ferraiolo and R. Kuhn, "Role-based access controls, " in Proc. 15th NIST-NCSC Nat. Comput. Security Conf., 1992, pp. 554–563.

[7] S. Halevi, D. Harnik, B. Pinkas, and A. Shulman-Peleg, "Proofs of ownership in remote storage systems," in Proc. ACM Conf. Comput. Commun. Securi

[8] C. Ng and P. Lee, "Revdedup: A reverse deduplication storage system optimized for reads to latest backups," in Proc. 4t Asia-Pacific Workshop Syst., http://doi.acm.org/10.1145/2500727.2500731, Apr. 2013.

[9] W. K. Ng, Y. Wen, and H. Zhu, "Private data deduplication protocols in cloud storage," in Proc. 27th Annu. ACM Symp. Appl. Comput., 2012, pp. 441–446.

[10] J. Xu, E.-C. Chang, and J. Zhou, "Weak leakage-resilient clientside deduplication of encrypted data in cloud storage," in Proc. 8th ACM SIGSAC Symp. Inform., Comput. Commun. Security, 2013,pp. 195–206.

[11] J. Yuan and S. Yu, "Secure and constant cost public cloud storage auditing with deduplication," IACR Cryptology ePrint Archive, 2013:149, 2013.

[12] K. Zhang, X. Zhou, Y. Chen, X. Wang, and Y. Ruan, "Sedic: Privacy-aware data intensive computing on hybrid clouds," in Proc.18th ACM Conf. Comput. Commun. Security, 2011, pp. 515–526.

[13] Z. Wilcox-O'Hearn and B. Warner, "Tahoe: The least-authority filesystem," in Proc. ACM 4th ACM Int.

Workshop Storage Security Survivability, 2008, pp. 21–26.

[14] S. Quinlan and S. Dorward, "Venti: A new approach to archival storage," in Proc. 1st USENIX Conf. File Storage Technol., Jan. 2002, p.7.

[15] R. D. Pietro and A. Sorniotti, "Boosting efficiency and security in proof of ownership for deduplication," in Proc. ACM Symp. Inf.,Comput. Commun. Security, 2012, pp. 81–82.

[16] J. Li, X. Chen, M. Li, J. Li, P. Lee, and W. Lou, "Secure deduplication with efficient and reliable convergent key management," in Proc. IEEE Trans. Parallel Distrib. Syst.,

[17] Libcurl, (1997). [Online]. Available: http://curl.haxx.se/libcurl

[18] 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 Proc. 3rd Int. Workshop Secutiry Cloud Comput., 2011, pp. 160–167.

[19] J. Stanek, A. Sorniotti, E. Androulaki, and L. Kencl, "A secure data deduplication scheme for cloud storage," Tech. Rep. IBM Research, Zurich, ZUR 1308-022, 2013.

[20] M. W. Storer, K. Greenan, D. D. E. Long, and E. L. Miller, "Secure data deduplication," in Proc. 4th ACM Int. Workshop Storage Security Survivability, 2008, pp. 1–10.

[21] Z. Wilcox-O'Hearn and B. Warner, "Tahoe: The least-authority filesystem," in Proc. ACM 4th ACM Int. Workshop Storage Security Survivability, 2008, pp. 21–26.