IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 18, Issue 1, Ver. III (Jan – Feb. 2016), PP 74-79
www.iosrjournals.org

Self Assured Deduplication System for Authorized Deduplication

Anudeep Para[1], Mahesh N[2]
[1]*Student of M.Tech (CSE)*
[2]*Asst. Prof, Department of Computer Science and Engineering, QIS Institute of technology, Ongole*

Abstract: *A hybrid cloud is a coalescence of public and private clouds bound together by either standardized or proprietary technology that alters information plus application movability. Proposed method aiming to expeditiously resolving ye quandary from deduplication on derivative favors in remote location computing. A hybrid remote location structure lying of a populace remote location plus a individual remote location plus ye information owners simply source their information storage by utilizing public cloud while the information operation is managed in private cloud. To build information management scalability in cloud computing, deduplication has been a very well-kenned technique recently is use. Deduplication reduces your bandwidth requisites, expedites the data transfers, and it keeps your cloud storage needs to a minimum. Proposed method demonstrate respective incipient deduplication expressions fortifying sanctioned duplicate assure inside hybrid remote location structure. To hold the secrecy of information ye convergent encoding proficiency holds made up used to encrypt ye information afore source. Sanctioned deduplication system support differential sanction duplicate check. As a proof of concept, a prototype is implemented in sanctioned duplicate check scheme and conduct test bed experiments utilizing prototype, sanctioned duplicate check scheme incurs minimal overhead compared to mundane operations.*

Keywords: *Deduplication, Proof of Ownership, Convergent Encryption, Key Management.*

I. Introduction

To make information management scalable in cloud computing, deduplication has been a well-kenned technique plus has magnetized more plus more care recently. Information deduplication is a specialized information compression technique for rejecting duplicate replicas of reiterating information in memory. The technique is used to ameliorate memory utilization plus can withal be used to network information transfers to reduce ye number of bytes that must be sent. In lieu of keeping multiple information copies with ye similar content, deduplication excretes redundant information by holding only 1 physical copy plus referring other redundant information to redundant imitate. [1] Deduplication can carry lay at ye data records level or ye chunk level. For data records level deduplication, infotech rejects repeat facsimiles from ye like data records. Deduplication can adscititiously choose home astatine ye chunk level, which excretes double chunks from information that occur in non-identical data records.

Albeit information deduplication brings an plethora of profits, protection plus secrecy pertains stand up while utilizer’s sensitive information are sensitive to some insider plus foreigner approaches .Traditional encoding, while supplying information confidentiality, is uncongenial with information deduplication. Concretely, natural encoding desires different utilizer’s to encipher their information with their possess keys. Thus, very information replicas of different utilizers will lead to different ciphertexts, building deduplication infeasible. Convergent encryption has been suggested to enforce information confidentiality while building deduplication feasible. Infotech cipher text/normal text a information copy with a confluent key [2] , which is incurred through calculating the cryptanalytic hash measure from ye message from ye information imitate. Afterward key propagation plus information encoding, utilizer’s hold ye key values plus send out ye ciphertext to ye remote location. Afterwards ye encryption procedure is deterministic plus is derived from the information content, identical l information copies will engender the same convergent key plus hence the same ciphertext. To avert wildcat access, a insure proof of ownership protocol is withal needed to supply the proof that the utilizer indeed owns ye Lapp data file whenever a double is detected. Afterward ye proofread, subsequent utilizer’s on ye Lapp data file volition be supplied an arrow of ye waiter less wanting to transfer ye like data file. A utilizer can download ye cipher text records with ye arrow of ye host, which can alone be decoded by ye representing information owners with their focused keys [3]. Hence, convergent encryption sanctions ye remote location to perform deduplication on ye ciphertexts plus ye proof of ownership obviates ye unauthorized utilizer to get at ye data files.

II. Related Work

Hybrid cloud can be built utilizing any technology it changes granting to unlike vendors. Key constituents In many of the situations, implementation of the hybrid cloud has a comptroller that will hold track of all placements of private and public clouds, IP address, servers and other resources that can run systems efficiently.

Data deduplication is 1 of consequential information compression techniques for rejecting duplicate replicas of reiterating information, and has been widely used in cloud memory to reduce the sum of memory space plus preserve bandwidth [4]. To forfend ye confidentiality of sensitive information while fortifying deduplication, Cloud computing provides ostensibly illimitable “virtualized” resources to users as accommodations across the whole Internet, while obnubilating platform and implementation details. Today’s cloud accommodation providers offer both highly useable storage plus massively parallel calculating resources at relatively low costs. As remote location computing turns prevailing, a incrementing number from information makes up restored in ye remote location and shared by utilizer’s with designated favors, which determine the approach corrects of ye memory information.

2.1 Proposed Method

Hybrid cloud can be built utilizing any technology it changes granting to unlike vendors. Key components In many of the situations, implementation of the hybrid cloud has a comptroller that will hold track of all positions of private and public clouds, IP address, servers plus early resources that can run systems efficiently.

Some of the key components include

- Orchestration manager plus cloud purveying for storage, populace cloud resources which includes virtual machines and networks, the private and public clouds, which are not compulsorily compatible or identical.
- Synchronization element and Data transfer expeditiously replace information among private plus public clouds.
- Changing configurations of storage, network and some early resources are constituting crossed by configuration monitor.[4]

In the Fig 1, the simplest view of hybrid cloud is allowed for, a single off-premises public cloud plus on-premises private cloud is within the Enterprise Datacenter is shown plus public cloud demonstrates the safe association to store information on to the cloud is denoted by the arrow:

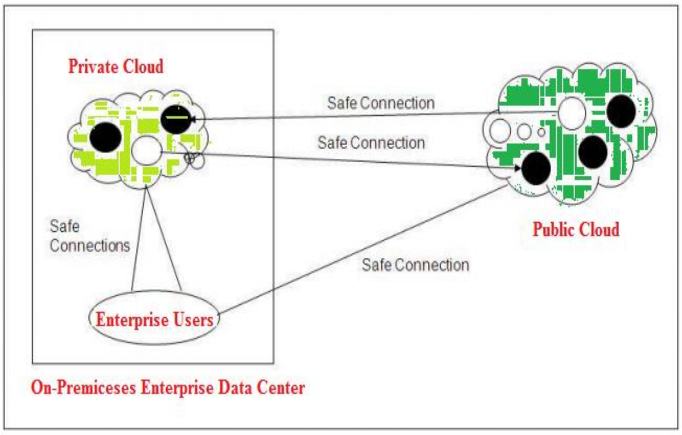

Fig 1: Hybrid Cloud Environment.

The ebony circles shows active virtual server images and white circles shows virtual server images which have been migrated by utilizing safe connections. The arrows designate that the direction of migration. Utilizing safe connections initiative utilizers are linked to ye clouds, which can be secure HTTP browsers or virtual private networks (VPNs) .A hybrid cloud could additionally can consist of multiple public or/and private clouds. [3]

Data de-duplication has many patterns. Generally, there is no one best way to enforce information de-duplication across an entire an organization. Instead, to maximize the gains, systems may spread more than one de-duplication strategy [15]. It is very essential to understand the backup and backup challenges, when culling de-duplication as a solution.

We have introduced hybrid cloud architecture in our aimed deduplication scheme [5]. The private keys for exclusive right will not be supplied to utilizer’s directly, which will be held plush plus led by ye private cloud server rather. In this manner, the utilizer’s cannot contribution these private keys of favors in this suggested structure, which be tokens that it can avoid ye privilege key distributing amongst utilizers in the over straight structure. To get a data file keys, ye utilizer inevitably to ship a call for to ye individual remote location waiter [6]. Ye suspicion from such building can be reported as follows. To perform the duplicate check for some data file, the utilizer wants to get ye data file keys on ye individual remote location waiter [8]. The Individual remote location waiter will additionally assure ye utilizer's individuality afore publishing ye representing data file keys to ye utilizer. The sanctioned double assure as such information data file bum be did through ye utilizer on ye populate remote location afore transferring this data records. Predicated on ye answers of double assure, ye utilizer either uploads this data file or runs PoW [10].

III. Implementation

Afore affording our construction of ye deduplication scheme, we determine an binary cognation R = f((p, p′)g because comes. Given 2 privileges p plus p′, we verbally show that p corresponds p′ if plus only if R(p, p′) = 1.

3.1 System Setup:

An identification protocol _ = (Proof, Verify) is additionally determined, where Proof plus swear constitute ye proof plus check algorithm severally. Moreover, for apiece one utilizer U exists surmised to have a mystery key skU to execute ye identification with waiters. Postulate that utilizer U features ye favor adjust PU. It additionally formats a PoW set of rules POW for ye data records ownership proof. The private cloud server will control a table which shops each utilizer's public information pku plus its representing privilege set PU.

3.2 File Uploading:

Suppose that a information proprietor requires to transfer plus apportion an data records F on user's whose privilege belongs to ye set PF = fpjg. The information owner demands act with ye secret remote location afore doing duplicate assure with ye S-CSP. Information owner does an recognition to try out infoteches individuality on secret tokens skU. If it is communicated, ye secrete remote location waiter testament get ye representing favors PU of ye utilizer of its memory table list. The utilizer calculates plus sends ye information data records tag ϕF = TagGen(F) to ye secrete remote location waiter, who will return fϕ′ F;p_ = Tag Gen(ϕF , kp_)g back to the utilizer for total p_ gratifying R(p, p_) = 1 plus p 2 PU. Then, the utilizer will act plus ship ye file token fϕ′ F;p_ g to y S-CSP.

- If an double data is detected by ye S-CSP, ye utilizer continues proof of ownership of this data file with ye S-CSP. If the cogent evidence is passed, the utilizer will be assigned a pointer, which approves him to access ye file.
- Otherwise, if no duplicate is found, the utilizer computes the encrypted file CF = EncCE(kF , F) with ye convergent key kF = KeyGenCE(F) plus uploads (CF , fϕ′ F;p g) to ye cloud server. The convergent key kF is stored by the utilizer locally.

3.3 File Retrieving:

Guess a utilizer requires to getting a data records F. It beginning sends out an call for plus ye data records name to the S-CSP. Upon getting the request plus data file designation, the S-CSP will assure whether ye utilizer is worthy to download F. If failed, the S-CSP sends back an terminate signal to the utilizer to denote ye data getting from network loser. Differently, ye S-CSP afford the representing ciphertextCF [13]. on experiencing ye ciphered information from ye S-CSP [11], the utilizer utilizes ye key kF memory topically to recuperate ye pristine €file F [14].

We implement a prototype of the proposed authorized deduplication system, in which we model three entities as separate C++ programs. A Client program is used to model the data users to carry out the file upload process. A Private Server program is used to model the private cloud which manages the private keys and handles the file token computation. A Storage Server program is used to model the S-CSP which stores and deduplicate files. Our implementation of the Client provides the following function calls to support token generation and deduplication along the file upload process.

• File Tag (File) - It computes SHA-1 hash of the File as File Tag;
• Token Req (Tag, Use rid) - It requests the Private Server for File Token generation with the File Tag and User ID;
• DupCheckReq(Token) - It requests the Storage Server for Duplicate Check of the File by sending the file token received from private server; Share TokenReq(Tag, {Priv.}) - It requests the Private Server to generate the Share File Token with the File Tag and Target Sharing Privilege Set;
• File Encrypt(File) - It encrypts the File with Convergent Encryption using 256-bit AES algorithm in cipher block chaining(CBC) mode, where the convergent key is from SHA-256 Hashing of the file;

• File Upload Freq (File ID, File, Token) – It uploads the File Data to the Storage Server if the file is Unique and updates the File Token stored. Our implementation of the Private Server includes corresponding request handlers for the token generation and maintains a key storage with Hash Map.

• Token Gen (Tag, User ID) - It loads the associated privilege keys of the user and generate the token with HMAC-SHA-1 Algorithm.

Methods Used In Secure Deduplication:

Following are the secure primitive used in the secure deduplication

(i) Symmetric Encryption

Symmetric encryption uses a common secret key k to encrypt and decrypt information. A symmetric encryption scheme made up of three primary functions.

• Key Gen SE (1λ) → k is the key generation algorithm that generates k using security parameter 1λ .

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

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

(ii) Convergent Encryption

Convergent encryption provides data confidentiality in deduplication. A user derives a convergent key from each original data copy and encrypts the data copy with the convergent key. In addition, the user also derive tag for the data copy, such that to detect duplicates 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 the data also same. The user first sends the tag to the server side to check if the identical copy has been already stored for detect duplicates[7].

(iii) Proof of Ownership

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

IV. Experiment Results:

Our evaluation focuses on comparing the overhead induced by authorization steps, including file token generation and shared token generation, against the convergent encryption and file upload steps. We evaluate the overhead by varying different factors, including 1) File Size 2) Number of Stored Files 3) Deduplication Ratio 4) Privilege Set Size. We break down the uploading process into 6 steps, 1) Tagging 2) Token Generation 3) Duplicate Check 4) Share Token Generation 5) Encryption 6) Transfer. For each step, we record the start and end time of it and therefore obtain the breakdown of the total time spent. We present the average time taken in each data set in the figures

File Size

To evaluate the effect of file size to the time spent on different steps, we upload 100 unique files (i.e., without any deduplication opportunity) of particular file size and record the time break down. Using the unique files enables us to evaluate the worst-case scenario where we have to upload all file data [12]. The average time of the steps from test sets of different file size are plotted in Figure 2. The time spent on tagging, encryption, upload increases linearly with the file size, since these operations involve the actual file data and incur file I/O with the whole file.

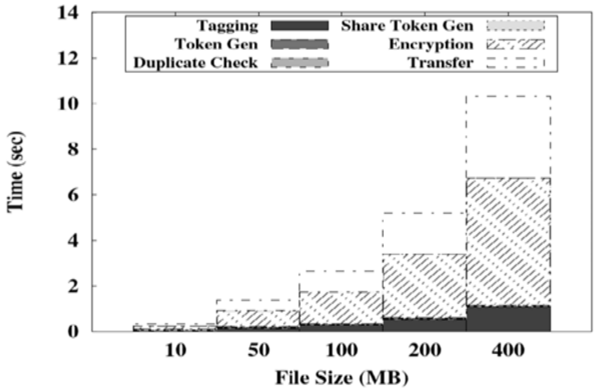

Fig 2: Time Breakdown for Different File Size

Number of Stored Files

To evaluate the effect of number of stored files in the system, we upload 10000 10MB unique files to the system and record the breakdown for every file upload. From Figure 3, every step remains constant along the time. Token checking is done with a hash table and a linear search would be carried out in case of collision.

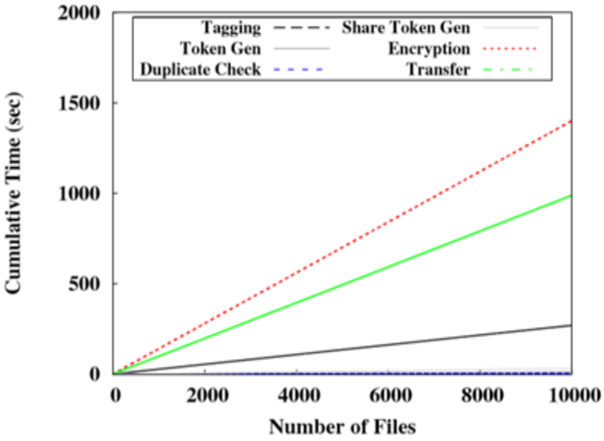

Fig 3: Time Break Down For Different Number of Files

Deduplication Ratio

To evaluate the effect of the deduplication ratio, we prepare two unique data sets, each of which consists of 50 100MB files. We first upload the first set as an initial upload. For the second upload, we pick a portion of 50 files, according to the given deduplication ratio, from the initial set as duplicate files and remaining files from the second set as unique files. The average time of uploading the second set is presented in Figure 4.

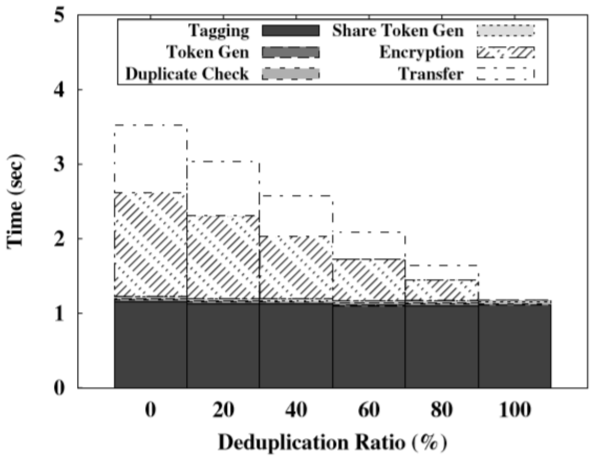

Fig 4: Time Break Down For Different Deduplication Ratio.

V. Conclusion

The cerebration of sanctioned information deduplication was suggested to ascertain the information security through counting differential gains of clients in ye double copy check. The presentation of a elite incipient deduplication growths fortifying sanctioned duplicate re-create in hybrid cloud architecture, in that ye duplicate assure tokens of documents are caused by ye private remote location waiter holding secrete keys. Security check presents that ye methods are assure regarding insider plus outsider assaults detailed in the suggested security model. As an issue verification of conception, ye formulated model of ye proposed punished duplicate imitate check method plus tested ye model. That showed the sanctioned duplicate copy check method experience minimum overhead comparing convergent encryption and data transfer.

References

[1] OpenSSLProject. ttp://www.openssl.org/.

[2] P.Anderson and L. Zhang. Fast and secure laptop backups with encrypted de-duplication.In Proc. of USENIX LISA, 2010.

[3] M.Bellare,S. Keelveedh , and T. Ristenpart . Dupless: Server aided encryption for deduplicated storage. In USENIX SecuritySymposium, 2013.

[4] M. Bellare, S. Keelveedhi, and T. Ristenpart. Message-locked encryption and secure deduplication. In EUROCRYPT, pages 296–312, 2013.

[5] M. Bellare, C. Namprempre, and G. Neven. Security proofs foridentity-based identification and signature schemes. J. Cryptology,22(1):1–61, 2009.

[6] M. Bellare and A. Palacio. Gq and schnorr identification schemes:Proofs of security against impersonation under active and concurrentattacks. In CRYPTO, pages 162–177, 2002.

[7] S. Bugiel, S. Nurnberger, A. Sadeghi, and T. Schneider. Twinclouds: An architecture for secure cloud computing. In Workshopon Cryptography and Security in Clouds (WCSC 2011), 2011.

[8] J. R. Douceur, A. Adya, W. J. Bolosky, D. Simon, and M. Theimer.Reclaiming space from duplicate files in a serverless distributedfile system. In ICDCS, pages 617–624, 2002.

[9] D. Ferraiolo and R. Kuhn. Role-based access controls. In 15thNIST-NCSC National Computer Security Conf., 1992.

[10] GNULibmicrohttpdhttp://www.gnu.org/software/libmicrohttpd/.

[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 arallel and Distributed Systems, 2013. libcurl. http://curl.haxx.se/libcurl/.

[13] C. Ng and P. Lee. Revdedup: A reverse deduplication storage system ptimized for reads to latest backups. In Proc. of APSYS,Apr 2013.

[14] 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.

[15] R. D. Pietro and A. Sorniotti. Boosting efficiency and securityin proof of ownership for deduplication. In H. Y. Youm and Y. Won, editors , ACM Symposium on Information,Computer and Communications Security, pages 81–82. ACM, 2012.

AUTHORS PROFILE

ANUDEEP PARA is Pursuing M.Tech (Computer Science and Engineering),QIS Institute of Technology Ongole, Prakasam Dist, Andhra Pradesh, India.

MAHESH N currently working as Asst. Professor in QIS Institute of technology, in the Department of Computer Science and Engineering, Ongole, PrakasamDist, Andhra Pradesh, India.