IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 1, January 2016

Secured and Efficient Cloud Storage Data Deduplication System

Sumedha A Telkar (S. A. Maindakar) [1], Dr M Z Shaikh [2]

Lecturer, Computer Engineering, BVIT, Navi-Mumbai, India [1]

Principal, BVCOE, Navi-Mumbai, India [2]

Abstract: During the last decade, cloud computing technology becomes an attractive trend of leveraging cloud based services for large scale content storage, processing and distribution. Thus, data deduplication becomes more and more a necessity for cloud service provider. Also security and privacy are top concern for the public cloud. Aiming to address the above storage and security challenges, this paper makes the attempt to formalize the notion of secured and efficient cloud storage system. We developed a prototype consist of client side deduplication for the security, storing and sharing outsourced data using the hybrid cloud. As a proof of our framework, we implement prototype and demonstrate that the incurred overhead is very limited in realistic environment.

Keywords: Deduplication, convergent encryption, differential privileges, hybrid cloud.

I. INTRODUCTION

With the continuous and exponential increase of the number of users and the size of their data, data deduplication becomes more and more a necessity for cloud storage providers. By storing a unique copy of duplicate data, cloud providers greatly reduce their storage and data transfer costs. These huge volumes of data need some practical platforms for the storage, processing and availability and cloud technology offers all the potentials to fulfil these requirements. Data deduplication is referred to as a strategy offered to cloud storage providers (CSPs) to eliminate the duplicate data and keep only a single unique copy of it for storage space saving purpose [Fig.1].

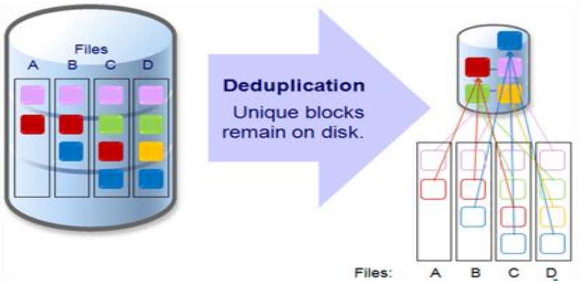

Fig.1. Data-Deduplication

By keeping a single copy of repeated data, data deduplication is considered as one of the most promising solutions to reduce the storage costs, and improve users experience by saving network bandwidth and reducing backup time. The advantages of deduplication unfortunately come with a high cost in terms of new security and privacy challenges. Deduplication [12, 14] can take place at either the file level or the block level. For file level deduplication, it eliminates duplicate copies of the same file. Deduplication can also take place at the block level, which eliminates duplicate blocks of data that occur in non-identical files.

Deduplication can be performed at different locations [1, 4, 6, 8, and 10] [Fig.2].

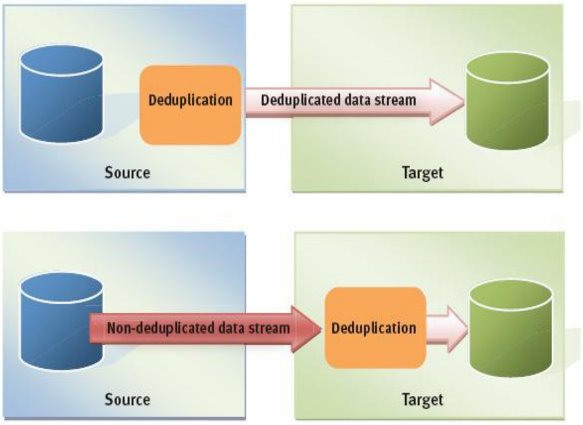

Fig.2. Deduplication at source and at target location

Depending on the participating machines and steps in the customized deduplication process, it is either performed on the client machine (*source-side*) or near the final Storage-server (*target-side*). In the former case, duplicates are removed before the data is transmitted to its storage. Since that conserves network bandwidth, this option is particularly useful for clients with limited upload bandwidth.

Convergent encryption [1, 7, and 11] has been proposed to enforce data confidentiality while making deduplication feasible. It encrypts/decrypts a data copy with a convergent key, which is obtained by computing the cryptographic hash value of the content of the data copy. After key generation and data encryption, users retain the keys and send the ciphertext to the cloud. Since the encryption operation is deterministic and is derived from the data content, identical data copies will generate same convergent key and hence the same ciphertext. To prevent unauthorized access, a secure proof of ownership protocol [3, 5] is also needed to provide the proof that the user indeed owns the same file when a duplicate is found. After the proof, subsequent users with the same file will be

DOI 10.17148/IJARCCE.2016.5175

provided a pointer from the server without needing to upload the same file. A user can download the encrypted file with the pointer from the server, which can only be decrypted by the corresponding data owners with their convergent keys. Thus, convergent encryption allows the cloud to perform deduplication on the ciphertexts and the proof of ownership prevents the unauthorized user to access the file.

An effective solution provided by Hybrid cloud [2, 9 and 13] is to split a task, keeping the computation on the private data within an organization's private cloud while moving the rest to the public commercial cloud [Fig.3]. Basically, the deployment of a cloud is managed in-house (Private Cloud) or over a third-party location (Public Cloud). While, for various reasons, it is deployed as an integrated private-public cloud (Hybrid Cloud).In private cloud configuration an organization may have control over its infrastructure or delegate that to a third party, being physically on-site or off-site. Securing the in-house cloud infrastructure is controllable and requires no need for extra trust mechanisms. Public cloud implementation is a model in which a service provider, third-party, offers public services on pay-per-use manner. Some of the benefits of this model are the economies of scale, ability to have short-term usage and greater resources utilization

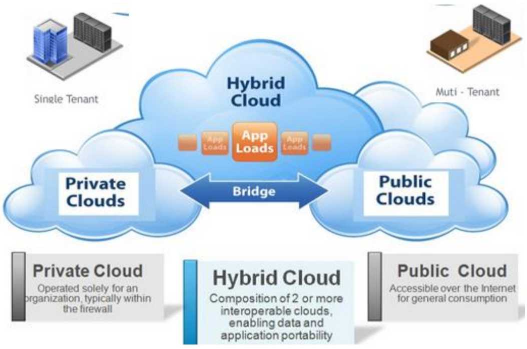
Single Tenant
Hybrid Cloud
Muti - Tenant
App Loads
App Loads
App Loads
Private Clouds
Bridge
Public Clouds
Private Cloud
Operated solely for an organization, typically within the firewall
Hybrid Cloud
Composition of 2 or more interoperable clouds, enabling data and application portability
Public Cloud
Accessible over the Internet for general consumption

Fig.3. Hybrid Cloud

II. RELATED WORK

Yuan et al. [8] proposed a deduplication system in the cloud storage to reduce the storage size of the tags for integrity check. Kaaniche, N [10] proposed client-side deduplication and implements Symmetric encryption for enciphering the data files and, Asymmetric encryption for metadata filesTo enhance the security of deduplication and protect the data confidentiality, Li et al. [4] addressed the key management issue in block-level deduplication by distributing these keys across multiple servers after encrypting the files. Bugiel et al. [2] provided an architecture consisting of twin clouds for secure outsourcing of data and arbitrary computations to an untrusted commodity cloud. Zhang et al. [9] also presented the hybrid cloud techniques to support privacy-aware data-intensive computing resulting in protection of sensitive data from public cloud. Bellare et al. [1] showed Data confidentiality by transforming the predictable message into unpredictable message. Introduced key server as third party to generate the file tag for duplicate check. Puzio et al. [11] implements an additional encryption operation and an access control mechanism as metadata manager to handle key management. Stanek et al. [6] presented a novel encryption scheme that provides differential security for popular data and unpopular data.

For popular data: the traditional conventional encryption is performed. For unpopular data: Another two-layered encryption scheme with stronger security while supporting deduplication is proposed. Xu et al. [7] also addressed the problem and showed a secure convergent encryption for efficient encryption, without considering issues of the key-management and block-level deduplication. Halevi et al. [3] proposed the notion of "proofs of ownership" (PoW) for deduplication systems, such that a client can efficiently prove to the cloud storage server that he/she owns a file without uploading the file itself. Ng et al. [5] extended PoW for encrypted files, but they do not address how to minimize the key management overhead.

III.EXISTING SYSTEM

Different from traditional deduplication systems, in the existing system [Fig.4] the differential privileges of users are considered in duplicate check besides the data itself [13].To support authorized deduplication, the tag of a file F will be determined by the file F and the privilege. To show the difference with traditional notation of tag, we call it file token instead. To support authorized access, a secret key kp will be bounded with a privilege p to generate a file token.

Let ϕ′(F,p) = TagGen(F, kp) denote the token of F that is only allowed to access by user with privilege p. In another word, the token ϕ′(F,p) could only be computed by the users with privilege p. As a result, if a file has been uploaded by a user with a duplicate token ϕ′(F,p) then a duplicate check sent from another user will be successful if and only if he also has the file F and privilege p. Such a token generation function could be easily implemented as H(F,kp), where H(_)denotes a cryptographic hash function.

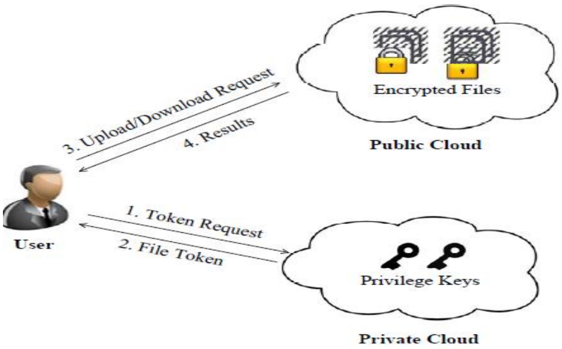

Fig.4. Existing Deduplication using Hybrid cloud

IV. PROPOSED SYSTEM

We proposed further extension of the existing system [Fig.4][13] by modifying the token generated at private cloud by including the Proof of Ownership(PoW) of files also along with the two factor authentication-one time password(2FA-OTP), for enhanced security. The steps performed in execution of our proposed system as shown below are:-

DOI 10.17148/IJARCCE.2016.5175

1. User profiling: Client registration and log-in
2. Session password: Token generation and verification
3. Client initiates file transfer (upload/download).
4. File-upload: check for duplicate
5. If duplicate at any of three level, create file pointer and store in CSP
6. If no duplicate found, store encrypted file in CSP.
7. File-download: PoW to decrypt and download file.
8. Server sync with client and completes file upload/download process.

V. IMPLEMENTATION

For implementation we preferred ASP.NET C# language, Visual studio framework and Windows O.S. Platform as it provides inbuilt server called IIS. ASP.NET provides inbuilt MSDN managed code to support cryptographic hashing algorithm needed to perform encryption and decryption. IIS (Internet information services) server of Windows allows creating and deploying ASP.NET application, which helps in to host our prototype web application on local network as well as on public network.

Our implementation of proposed deduplication system, in which we merge the twin cloud (private + public) for achieving security, provides straight away two modules:

1) for User/Client and

2) Admin/Server. User is not aware of hybrid cloud whereas Admin serves and monitors the cloud-user's activities.

We implemented Playfair cipher's modified version (mixing the concept of pair-based attribute scheme) to generate authentication token as 2FA-OTP(Two Factor Authentication-One time Password), communicated to user's registered e-mail id and mobile no. using SMTP Protocol and free SMS service for authorized logging purpose. Client module consists of:-

1. User registration,
2. Token generation and verification,
3. File-Upload: duplicate check for file /block /byte level and if file are not duplicate then encrypt and transfer it to pubic CSP
4. File-download: request module to download files uploaded by other users.

Admin module consists of:

1. Login operation
2. Monitoring all registered user's logging activities to CSP

VI. EVALUATION

We evaluated our prototype by conducting experiments in a LAN, where each machine equipped with Intel Core-2 Duo 2.93 GHz CPU, 4GB RAM, Windows 7 Professional 32-bit Operating System. The LAN machines are connected with 100Mbps Ethernet network. As our proposed system involves byte-level duplication check along with file and variable-size block-level too, we evaluated the deduplication system on basis of different factors:

1) File-type

2) Unique-file uploads time [Fig. 7]

3) Deduplication ratio, by breaking down our process into minimal steps as:

a) Duplicate check at 3 levels(file/block/byte) [Fig.5 and Fig.6]

b) Creating file pointer if duplicate found

c) Encryption of file, if not duplicate and

d) Finally transfer or upload to CSP.

With increase in file size, the time spent on duplicate check, encryption and transfer increases. We evaluated the effect of time to upload unique files by uploading 150 1MB unique files of different types. For every type, time remains constant for file encryption and upload. To evaluate the deduplication ratio, we uploaded same set of 150 1MB files again. Here, in case of duplicate files, encryption and uploading time would be skipped, thus achieving deduplication ratio 100%. The time required to assured duplication in either of the level is mentioned.

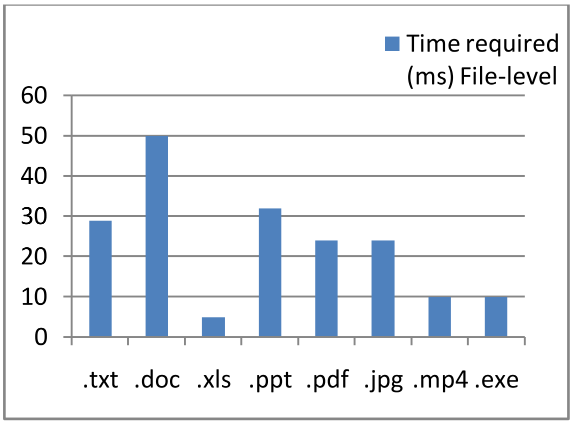

Fig.5. Deduplication time factor at File levels

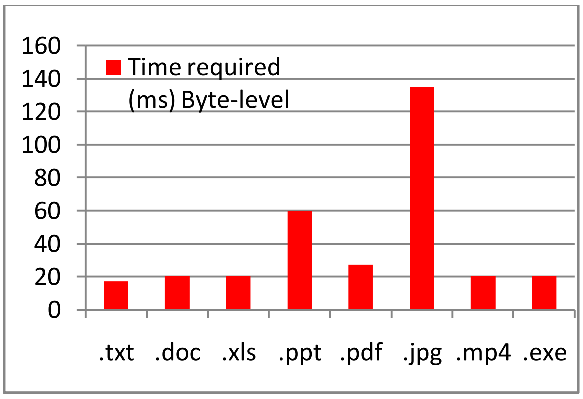

Fig.6. Deduplication time factor at Byte levels

DOI 10.17148/IJARCCE.2016.5175

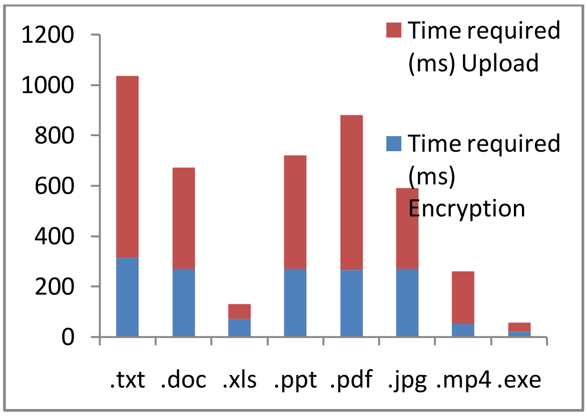

Fig.7. Time required uploading unique file of different type

VII. SUMMARY

Thus, we concluded that our proposed prototype is secured in terms of upload and download data operation on hybrid cloud, and also achieving 100% deduplication ratio for cloud storage.

VIII. CONCLUSION

The notion of authorized data de-duplication technique is specialized data compression technique which eliminates redundant data as well as improves storage and bandwidth utilization. Convergent encryption technique is proposed to enforce confidentiality during de-duplication, which encrypt data before outsourcing.

Security analysis demonstrates that the schemes are secure in terms of insider and outsider attacks. To better protect data security, we present Two Factor Authentication scheme (2FA) of user along with PoW of files, to address problem of authorized data de-duplication, in which the duplicate-check tokens of files are generated by the private cloud server with private keys.

REFERENCES

[1] M. Bellare, S. Keelveedhi, and T. Ristenpart. “Dupless: Serveraided encryption for deduplicated storage”. In USENIX Security Symposium, 2013.[2] Attention, shoppers : Store is tracking your cell, New York Times. References

[2] 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.

[3] 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

[4] J. Li, X. Chen, M. Li, J. Li, P. Lee, andW. Lou. “Secure deduplication with efficient and reliable convergent key management”. In IEEE Transactions on Parallel and Distributed Systems, 2013.

[5] 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.

[6] J. Stanek, A. Sorniotti, E. Androulaki, and L. Kencl. “A secure data deduplication scheme for cloud storage”. In Technical Report, 2013.

[7] 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.

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

[9] K Zhang, X Zhou, Y Chen and X Wang, “Sedic Privacy-Aware Data Intensive Computing”

[10] Kaaniche, N. ; Inst. Mines-Telecom, Telecom Sud Paris, Evry, France; Laurent,M.A “Secure Client Side Deduplication Scheme in Cloud Storage Environments”

[11] Puzio, P. ; SecludIT, Sophia-Antipolis, France ; Molva, R.; Onen,M.; Loureiro,S. “ClouDedup Secure Deduplication with Encrypted Data for Cloud Storage”

[12] Iuon-Chang and Po-Ching Chien, “Data Dedupliction Scheme for Cloud storage”. IJ3C, Vol. 1, No. 2 (2012)

[13] Jin li,yan kit li,xiaofeng chen,patrick p.c.lee,wenjing lou. “A hybrid cloud approach for secure authorized deduplication”. IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEM VOL: PP NO: 99 YEAR 2014.

[14] M. W. Storer, K. Greenan, D. D. E. Long, and E. L. Miller. “Secure data deduplication”. In Proc. of StorageSS, 2008.

[15]http://cyberintelligentsecurity.net/arkeia/products/data_deduplication.php

[16] http://www.uk.insight.com/en-gb/shop/acronis/deduplication

[17] ASP.NET C# tutorials by Mr. Kudvenkat at https://www.youtube.com/user/kudvenkat

[18] SQL Server tutorial by Mr. Kudvenkat at https://www.youtube.com/user/kudvenkat

[19] SQL Server tutorial at http://www.quackit.com/sql_server/tutorial/

[20] http://blog.p3infotech.in/2013/hybrid-cloud-model/

[21] https://www.linkedin.com/pulse/20140701143732-288460101-gartner-hybrid-cloud-cost-resiliency-benefits

[22] http://www.cloudinfinit.com/products/IaaS/hybrid

DOI 10.17148/IJARCCE.2016.5175