International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169
Volume: 2 Issue: 11 3574 – 3577

Survey on Secure Authorized De-duplication in Hybrid Cloud

Ms. Akanksha V. Patil
Student in Department of Computer Engg.
P.V.P.I.T., Bavdhan
Pune, Maharashtra, India
akanksha.v.patil@gmail.com

Mr. Navnath. D. Kale
Asst. Prof in Department of Computer Engg.
P.V.P.I.T., Bavdhan
Pune, Maharashtra, India
navnath1577@yahoo.co.in

***Abstract*—** Nowadays, cloud computing provides high amount of storage space and massive parallel computing at effective cost. As cloud computing becomes prevalent, excessive amount of data being stored in the cloud. However, exponential growth of ever-increasing volume of data has raised many new challenges. 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. To better protect data security, we present different privileges of user to address problem of authorized data de-duplication. We also present several new de-duplication constructions supporting authorized duplicate check in hybrid cloud architecture, which incurs minimal overhead compared to normal operation.

Keywords- *cloud computing, de-duplication, convergent encryption technique, authorized duplcate check*

*****

I. INTRODUCTION

Cloud computing is computing in which large groups of remote servers are networked to allow centralized data storage and online access to computer services or resources [9]. With cloud computing, large pools of resources can be connected through private or public network. In public cloud, services (i.e. applications and storage) are available for general use over the internet. A private cloud is a virtualized data center that operates within a firewall. In this research introduce mix of public and private cloud, hybrid cloud.

Cloud computing provides computation and storage resources on the Internet. Increasing amount of data is being stored in the cloud and it is shared by users with specified privileges, which defines special rights to access stored data. Managing the exponential growth of ever-increasing volume of data has become a critical challenge. According to IDC cloud report 2014, companies in India are making a gradual move from on-premise legacy to different forms of cloud. While the process is gradual, it has started by migrating certain application workloads to cloud [1]. To make scalable management of stored data in cloud computing, de-duplication [2] has been well known technique which becomes more popular recently. De-duplication is a specialized data compression technique, which reduce storage space and upload bandwidth in cloud storage. In de-duplication, only one unique instance of the data is actually on the server and redundant data is replaced with a pointer to the unique data copy.de-duplication can take place either at file level or block level.

From the user perspective, security and privacy concerns are arise as data are susceptible to both insider and outsider attack. We must properly enforce confidentiality, integrity checking, and access control mechanisms both attacks. De-duplication does not work with traditional encryption. User encrypts their files with their individual encryption key, different cipher text would emerge even for identical files. Thus, traditional encryption is incompatible with data de-duplication.

Convergent encryption [3] is a widely used technique to combine the storage saving of de-duplication to enforce confidentiality. In convergent encryption, the data copy is encrypted under a key derived by hashing the data itself. This convergent key is used for encrypt and decrypt a data copy. After key generation and data encryption, users retain the keys and send the cipher text to the cloud. Since encryption is deterministic, identical data copies will generate the same convergent key and the same cipher text. This allows the cloud to perform de-duplication on the cipher texts. The cipher texts can only be decrypted by the corresponding data owners with their convergent keys.

Differential authorization duplicate check is an authorized de-duplication technique where each user is issued a set of privileges during system initialization. This set of privileges specifies that which kind of users is allowed to perform duplicate check and access the files.

II. LITERATURE REVIEW

M. Bellare [5] design a system, DupLESS that combines a CE-type scheme with the ability to obtain message-derived keys with the help of a key server (KS) shared amongst a group of clients. The clients interact with the KS by a protocol for oblivious PRFs, ensuring that the KS can cryptographically mix in secret material to the per-message keys while learning nothing about files stored by clients. These mechanisms ensure that DupLESS provides strong security against external attacks and that the security of DupLESS gracefully degrades in the face of comprised systems. Should a client be compromised, learning the plaintext underlying

another client's cipher text requires mounting an online brute force attacks.

Aim of M. Bellare [6] is to formalize a new cryptographic primitive, Message-Locked Encryption (MLE), where the key under which encryption and decryption are performed is itself derived from the message. MLE provides a way to achieve secure de-duplication, a goal currently targeted by numerous cloud-storage providers. They provide definitions both for privacy and for a form of integrity that they call tag consistency. They provide ROM security analyses of a natural family of MLE schemes that includes deployed schemes. They make connections with deterministic encryption, hash functions secure on correlated inputs.

G. Neven [7] provides either security proofs or attacks for a large number of identity-based identification and signature schemes defined either explicitly or implicitly in existing literature. Underlying these is a framework that on the one hand helps explain how these schemes are derived and on the other hand enables modular security analyses, thereby helping to understand, simplify, and unify previous work. They also analyze a generic folklore construction that in particular yields identity-based identification and signature schemes without random oracles.

S. Bugiel [9] had propose an architecture and protocols that accumulate slow secure computations over time and provide the possibility to query them in parallel on demand by leveraging the benefits of cloud computing. The user communicates with a resource-constrained Trusted Cloud (either a private cloud or built from multiple secure hardware modules) which encrypts algorithms and data to be stored and later on queried in the powerful but untrusted Commodity Cloud.

j. li [11] had proposed Dekey, an efficient and reliable convergent key management scheme for secure de-duplication. Dekey applies de-duplication among convergent keys and distributes convergent key shares across multiple key servers, while preserving semantic security of convergent keys and confidentiality of outsourced data. They implement Dekey using the Ramp secret sharing scheme and demonstrate that it incurs small encoding/decoding overhead compared to the network transmission overhead in the regular upload/download operations

C. Ng [13] had present RevDedup, a de-duplication system designed for VM disk image backup in virtualization environments. RevDedup has several design goals: high storage efficiency, low memory usage, high backup performance, and high restore performance for latest backups. The core design component of RevDedup is reverse de-duplication, which removes duplicates of old backups and mitigates fragmentation of latest backups. They extensively evaluate our RevDedup prototype using different workloads and validate our design goals.

J. Stanek[14] had proposed a novel encryption scheme that guarantees semantic security for unpopular data and provides weaker security and better storage and bandwidth benefits for popular data. This way, data se-duplication can be effective for popular data, whilst semantically secure encryption protects unpopular content, preventing its de-duplication. Transitions from one mode to other take place seamlessly at the storage server side if a file becomes popular.

K. Zhang[22] had explained Commercial cloud services, such as the Amazon EC2, which enable their customers to process a large amount of data at a low cost. This benefit, however, comes with privacy risks: the computing tasks of organizations often involves sensitive data and therefore cannot be directly delegated to the public cloud without proper protection. Such protection cannot be expected from traditional secure outsourcing techniques, which often cannot handle the large amount of data such computation involves. A more practical solution is to split the computation so as to move the workload unrelated to sensitive data to the commercial cloud, while keeping the rest within an organization's private cloud. This hybrid computing paradigm needs to be supported by a new privacy-aware computation framework. To this end, they present Sedic, the first secure data-intensive computing system, in that paper. Their approach leverages the special features of MapReduce to schedule individual map tasks over a carefully planned data placement, in a way that the tasks within the private cloud only work on sensitive data and those on the public cloud only processes public data. As a result, all the workload that does not involve private information can be offloaded to the low-cost commercial cloud. To avoid an intensive data exchange between clouds, Sedic also automatically analyzes the reducer of a legacy MapReduce job to extract a combiner for aggregating the map outcomes on the public cloud. They implemented our techniques on Hadoop and evaluated our prototype on FutureGrid, a large-scale cloud test-bed. Their study shows that without jeopardizing user privacy, Sedic effectively outsourced a large amount of computing workload to the public cloud, fully preserved the scalability of MapReduce and also conveniently accommodated legacy computing jobs.

J. Xu [23] proposed growing need for secure cloud storage services and the attractive properties of the convergent cryptography lead us to combine them, thus, defining an innovative solution to the data outsourcing security and efficiency issues. Our solution is based on a cryptographic usage of symmetric encryption used for enciphering the data file and asymmetric encryption for meta data files, due to the highest sensibility of these information towards several intrusions. In addition, thanks to the Merkle tree properties, this proposal is shown to support data de duplication, as it employs

an pre-verification of data existence, in cloud servers, which is useful for saving bandwidth. Besides, our solution is also shown to be resistant to unauthorized access to data and to any data disclosure during sharing process, providing two levels of access control verification. Finally, we believe that cloud data storage security is still full of challenges and of paramount importance, and many research problems remain to be identified.

M. W. Storer[24] developed two models for secure de-duplicated storage: authenticated and anonymous. These two designs demonstrate that security can be combined with de-duplication in a way that provides a diverse range of security characteristics. In the models they present, security is provided through the use of convergent encryption.

This technique, first introduced in the context of the Farsite system, provides a deterministic way of generating an encryption key, such that two different users can encrypt data to the same cipher text. In both the authenticated and anonymous models, a map is created for each file that describes how to reconstruct a file from chunks. This file is itself encrypted using a unique key.

In the authenticated model, sharing of this key is managed through the use of asymmetric key pairs. In the anonymous model, storage is immutable, and file sharing is conducted by sharing the map key offline and creating a map reference for each authorized user.

In evaluation, they had analyzed the security of each model with regard to a number of security compromises. They found that the system is mostly secure against external attackers. Further, the security threats that our models do not explicitly guard against can be addressed through the addition of standard secure communications techniques such as transport later security. Security compromises by a malicious insider are largely mitigated from the design's avoidance of server side encryption. Since insiders are never exposed to plain-text or encrypted keys, their ability to change metadata values in an undetectable way is greatly diminished. Security is even more apparent in the chunk store where the content addressed nature of secure chunks intrinsically makes the detection of malicious changes quite noticeable. Finally, they examined the information leaks resulting from key compromises and found that the most severe security breaches result from the loss of the client's key. The damage in the event of such a key loss is confined, however, to the user's files.

III. PROPOSED SYSTEM

In traditional encryption, user encrypts their files with their individual encryption key, different cipher text would emerge even for identical files. Thus, traditional encryption is incompatible with data de-duplication.

In Proposed methodology, Convergent encryption has been used to enforce data confidentiality. Data copy is encrypted under a key derived by hashing the data itself. This convergent key is used for encrypt and decrypt a data copy. After key generation and data encryption, users retain the keys and send the cipher text to the cloud. Since encryption is deterministic, identical data copies will generate the same convergent key and the same cipher text. This allows the cloud to perform de-duplication on the cipher texts. The cipher texts can only be decrypted by the corresponding data owners with their convergent keys.

To better protect data security, paper makes the first attempt to formally address the problem of authorized data de-duplication. Furthermore, such unauthorized users cannot decrypt the cipher text even collude with the S-CSP. Security analysis demonstrates that that system is secure in terms of the definitions specified in the proposed security model.

In such an authorized de-duplication system, each user is issued a set of privileges during system initialization each file uploaded to cloud is also bounded by a set of privileges to specify which kind of users is allowed to perform the duplicate check and access the files.

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 different privileges of user 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] Komal Puri, "IDC Cloud Report 2014: Cross-Vertical Demand Side Perspective", Nov 19, 2014, [Online] Available

[2] S. Quinlan and S. Dorward. "Venti: a new approach to archival storage". In Proc. USENIX FAST, Jan 2002.

[3] 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 Proc. ICDCS, 2002, pp. 617-624.

[4] S. Halevi, D. Harnik, B. Pinkas, and A. Shulman-Peleg, ''Proofs of Ownership in Remote Storage Systems,'' in Proc. ACM Conf. Comput. Commun. Security, Y. Chen, G. Danezis, and V. Shmatikov, Eds., 2011, pp. 491-500..

[5] M. Bellare, S. Keelveedhi, and T. Ristenpart. Dupless: Serveraided encryption for deduplicated storage. In USENIX Security Symposium, 2013

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

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

[8] 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.

[9] 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.

[10] 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, pages 617–624, 2002

[11] 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..

[12] libcurl. http://curl.haxx.se/libcurl/.

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

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

[15] R. S. Sandhu, E. J. Coyne, H. L. Feinstein, and C. E. Youman. Role-based access control models. IEEE Computer, 29:38–47, Feb 1996.

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

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

[18] Z. Wilcox-O'Hearn and B. Warner. Tahoe: the least-authority filesystem. In Proc. of ACM StorageSS, 2008.

[19] 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.

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

[21] K. Zhang, X. Zhou, Y. Chen, X.Wang, and Y. Ruan. Sedic: privacyaware 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.

[22] K. Zhang, X. Zhou, Y. Chen, X.Wang, and Y. Ruan. Sedic: privacyaware 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.

[23] 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.

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