ISSN 23499842(Online), Volume 1, Special Issue 2(ICITET 15), March 2015
International Journal of Innovative Trends and Emerging Technologies

DUPLICATION AVOIDANCE MECHANISM FOR STORING DATA IN HYBRID CLOUD

CHANDRA KALA G [1], PRIYANKA P[2], SOWMYA C M[3], PAUL JASMINE RANI L[4],SARAVANAN A[5]
[1,2,3]UG [Scholar], [4] Assistant Professor, [5]Professor Department of Computer Science and Engineering,
Rajalakshmi Institute of TechnologyChennai, Tamil Nadu, India
chandusaru11@gmail.com prynka94@gmail.com sowmyamurugesan12@gmail.com
pauljasminrani.l@ritchennai.edu.in saravanan.a@ritchennai. edu.in

ABSTRACT-Data de-duplication has been widely used in the cloud storage to reduce the amount of storage space and save bandwidth size. It is one of the important technique for eliminating duplicate copy of repeating data. Duplication occurs when more than one user saves the same file or data in the same cloud server. While supporting de-duplication, the confidentiality of sensitive data should also be maintained. In our project, the efficiency is improved in de-duplication method to overcome the problem of authorized data de-duplication which supports hybrid cloud architecture. Instead of using the traditional de-duplication system, the privileges are considered in duplicate check before uploading the data in the cloud server. Here, size of the cloud server is reduced. The data security is provided by using encryption technique to encrypt the data before outsourcing, so that the data is consistent and reliable.

I. INTRODUCTION

Cloud computing provides seemingly unlimited

"virtualized" resources to users as services across the whole Internet, while hiding platform and implementation details. Today‟s cloud service providers offer both highly available storage and massively parallel computing resources at relatively low costs. As cloud computing becomes prevalent, an increasing amount of data is being stored in the cloud and shared by users with specified *privileges*, which define the access rights of the stored data. One critical challenge of cloud storage services is the management of the ever-increasing volume of data. To make data management scalable in cloud computing, deduplication has been a well-known technique and has attracted more and more attention recently. Data deduplication is a specialized data compression technique for eliminating duplicate copies of repeating data in storage. The technique is used to improve storage utilization and can also be applied to network data transfers to reduce the number of bytes that must be sent. 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. Existing data deduplication systems, the private cloud is occupied as a different to allow data owner/users to securely perform duplicate check with differential privileges. In architecture is practical and has involved much interest from researchers. The data owners only outsource their data storage by utilizing public cloud while the data operation is managed in private cloud.

II. DISADVANTAGES

Problems such a construction of authorized deduplication has several serious security problems, which are listed below. Each user will be issued private keys for their corresponding privileges. A restriction makes the authorized deduplication system unable to be widely used and limited. Second, the above deduplication system cannot prevent the privilege private key sharing among users. The users will be issued the same private key for the same privilege in the construction. So there is the chance the data repetition in cloud server.

III. PROPOSED SYSTEM

In using advanced deduplication system supporting authorized duplicate check. In this new deduplication system, a hybrid cloud architecture is introduced to solve the problem. To get a file token, the user needs to send a request to the private cloud server. The private cloud server will also check the user‟s identity before issuing the corresponding file token to the user. The authorized duplicate check for this file can be performed by the user by requesting the server before uploading this file. Based on the results of duplicate check done by server, the user either uploads this file or utilize the file that already exists in the public cloud.

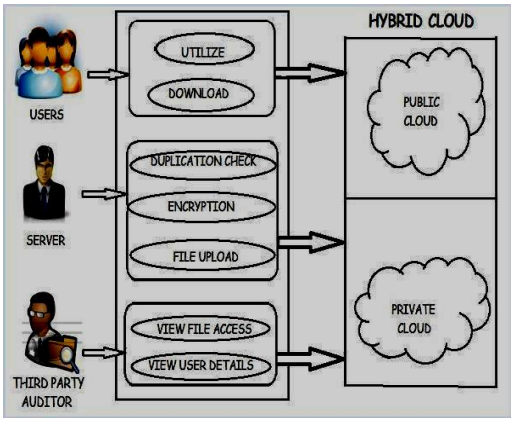
IV. ADVANTAGE

The storage space of the cloud server is reduced. Hence the data can be retrieved faster. This effectively increase network bandwidth. And duplication of the files can be avoided. And the confidentiality of the data is also maintained effectively.

V. SIGNIFICANCE

In this project, we aim to solve the problem of de-duplication with privileges in cloud computing. We consider a hybrid cloud architecture consisting of a public cloud and a private cloud.

VI. PROPOSED ARCHITECTURE



Why a Hybrid approach to cloud computing works best for now?

Companies have spent billions of dollars over the years building and maintaining backup disk arrays and secondary data centers to keep things running in the event that something goes wrong.
Cloud turns backup and recovery on its head, making it possible to provision back -up sites as needed, for pennies. Ironically, however, this goes against the gut instincts of many IT executives, who spend a lot of time worrying about data security and availability. Thats the view of David Nichols, principal and Americas CIO services leader for Ernst& Young IT Advisory Services, who has been working closely with companies across the globe to identify cloud opportunities. I recently had the opportunity to chat with Nichols to get his take on the pluses and minuses of cloud computing, and he observed the approach is still in its early stages. While "there is a lot of demand for cloud," he confirms, most companies do not yet have a formal strategy or end-goal. That‟s because most companies have not reached the point where at least 30% of their workloads are carried in the cloud, he says. Once an organization passes that 30% threshold, cloud starts to become a serious part of the business. And Nichols has heard a lot of arguments from both sides lately about the viability of cloud – the cloud not as secure as on-premises systems for data; the cloud is more secure than on-premises systems for data. On one side, he hears: ""„There‟s no way I can move mission-critical data off-premises. I can‟t allow stuff to not be within my four walls, how do we do disaster recovery, how do I make sure this stuff is restored?‟"

Still, he relates, other executives say they favor the cloud precisely because it is more secure. "Another

CIO told me that was exactly the reason he moved some of his stuff to the cloud. As he put it: „If all they do is data storage, they're going do it better than my people do it. That's all they do, every day. When your transmission breaks, are you going to take it to a generalist, or are you going to take it to a transmission specialist? Plus, cloud providers will have better procedures, more sophisticated and better data recovery procedures and more sophisticated firewalls.‟"" Perhaps the best approach is to have the best of both worlds. Nichols says a hybrid cloud strategy, which incorporates both off-site and off-site services, provides a "failsafe" model for enterprises. "Maybe there are some things you can do in the cloud model, and some things that you can‟t because maybe you don‟t feel as secure. But the cloud could be one more failsafe. Maybe you can get one or two or three more „9s‟ from a recovery perspective that you couldn‟t have gotten otherwise. Maybe this is a cheaper way to get there than it was to have to buy all the stuff and house yourself. And that‟s just on the data storage side." By relying on the cloud for backup, it "provides one more failsafe approach, instead of having to buy a sophisticated server and

RAID platform," Nichols points out. " Its a pretty cheap solution to back it up in the cloud someplace, and therefore know you „ll always have it in case something goes wrong." There are even rumblings from many IT executives that the cloud backup is quicker than the primary backup. There are cases in which cloud could function as the primary backup site, with on-premises backup as the failover environment. The "hybrid" approach Nichols alludes to may be more common within many enterprises than all-cloud environment, he predicts. For the most part, cloud is still an under-the-radar phenomenon, and companies have not developed formal cloud strategies. "They‟re not really sure what is the biggest impact," Nichols says. "They‟re not really sure how to go about it. Is this a cost-reduction exercise? Can we really use it to drive the business? Or is this only going to help us drive certain aspects of our corporate strategy, but not really moving the needle on our ongoing business? Until cloud workloads surpass that
30% mark in organizations, cloud will "remain hidden within operating models within IT organizations," he points out. In the long run, he predicts, cloud penetration within enterprises will reach 70%. As organizations move between the 30% and 70% points, expect to see widespread adoption of hybrid approaches to cloud computing – a blended strategy of using outside cloud services and internal private cloud. At the same time, he says, any and all new software development – both within enterprises and among vendors – is now taking place around a cloud model versus the traditional on-premises approach. "Right now, were at a pretty strong inflection point right now," he explains. "Very little, if anything, being built on traditional go-install-on-your-local-device model. Almost nothing going forward is going to be built within that traditional framework." In the next installment of my chat with

E&Y‟s David Nichols, we discuss how cloud is changing the roles and relationships of outsourcers, entrepreneurs, and IT professionals, with advice on

building a successful cloud relationship.

VII. CONVERGENT ENCRYPTION

Definition

Convergent encryption, also known as content hash keying, is a cryptosystem that produces identical cipher text from identical plaintext files. This has applications in cloud computing to remove duplicate files from storage without the provider having access to the encryption keys.

Illustration

The basic idea of convergent encryption (CE) is to derive the encryption key from the hash of the plaintext. The simplest implementation of convergent encryption can be defined as follows: Alice derives the encryption key from her **message** M such that K = H(M), where H is a **cryptographic hash function**; she can encrypt the message with this key, hence: C = E(K, M) = E(H(M), M), where E is a **block** applying this technique, two users with two identical plaintexts will obtain two identical cipher texts since the encryption key is the same; hence the cloud storage provider will be able to perform de-duplication on such cipher texts. Furthermore, encryption keys are generated, retained and protected by users. As the encryption key is deterministically generated from the plaintext, users do not have to interact with each other for establishing an agreement on the key to encrypt a given plaintext. Therefore, convergent encryption seems to be a good candidate for the adoption of encryption and de-duplication in the cloud storage domain.

VIII. CONCLUSION

The authorized data de-duplication is proposed to avoid the duplicate files in the cloud storage. Diversified privileges are given in order to protect the data privacy. A hybrid cloud architecture which is a combination of both public and private cloud is considered to solve the problem of duplication. Convergent encryption technique is used to avoid the duplicate files in the cloud storage. The file keys are generated and maintained by the server in the private cloud. Hence, the confidentiality of the data is maintained. In future, the files from the user system can also be requested and uploaded without affecting the confidentiality of the data.

X. REFERENCES

[1] OpenSSL Project. http://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. Keelveedhi, and Ristenpart. Dupless: Serveraided encryption for deduplicated storage. In *USENIX Security Symposium*, 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 for identity-based identification and signature schemes.
J.Cryptology,22(1–61, 2009. [6] 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.

[7] 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.

[8] 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.

[9] D. Ferraiolo and R. Kuhn. Role-ased access controls. In *15th NIST- NCSC National Computer Security Conf.*, 1992.

[10] GNU Libmicrohttpd. http://www.gnu.org/software/libmicrohttpd/.

[11] 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.

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

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

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

[15] 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.