International Journal of Emerging Technology in Computer Science & Electronics (IJETCSE)
ISSN: 0976-1353 Volume 23 Issue 8 –NOVEMBER 2016.

A VITAL INVESTIGATION ON DEDUPLICATION IN CLOUD

K.B M Pavan Kumar[#1] and V.Chandra Sekhara Rao[*2]

[#] *Department of CSE, ADITYA College of Engineering, Surampalem, E.G.Dist., A.P, India*

[*] *Department of CSE, ADITYA College of Engineering, Surampalem, E.G.Dist., A.P, India*

***Abstract*— Deduplication is one of the important data compression techniques used to eliminate duplicate copies of the repetitive data, reducing the amount of broad storage space, in order to conserve bandwidth and it has been used in cloud storage. While supporting to Deduplication, in order to protect the confidentiality of sensitive data, it has been proposed to encrypt the data prior to outsource convergence encryption technology. It is trying to analyse with different traditional Deduplication systems. In this paper, a differential traditional approach has been implemented, further considered in the duplicate check in addition to the data itself as well as security analysis is also done on various traditional approaches. We display a few new deduplication developments supporting, to approve copy check in half-breed cloud engineering. Security investigation shows that our plan is secure as far as the definitions indicated in the proposed security scheme. As a proof of our view, we execute a model of our proposed copy check plan and conduct test bed verifications by our model.**

***Index Terms*—Cloud, data Deduplication, data security**

I. INTRODUCTION

While hiding the details of platform and implementation cloud computing provides "virtual" resources, seemingly infinite to the user as a service across the entire internet. Today's cloud service provider offers both high availability storage and large-scale parallel computing resources at relatively low cost. As cloud computing becomes widespread, an increase in the amount of data stored in the cloud and is shared by the user with the given authority to define access of the stored data. One of the important issues of the cloud storage service is the management of the volume to continue to increase the data. To scalable data management in the cloud computing, Deduplication [1], it has become a well-known technique, which recently has attracted more and more attention. Data de-duplication is a special data compression technology to eliminate repeated duplicate copies of the data in the storage.

Technology can be applied to transfer the data into network in order to reduce the number of bytes to be used and it must be sent in order to improve the storage utilization. Instead, to hold a plurality of data copy with the same contents, to save the physical copy of the only one, by referring to the other redundant data to the copy, to eliminate redundant data. Deduplication, can be done by any of the file-level or block level. In the case of file-level deduplication eliminates duplicate copies of the same file. For block level Deduplication, a block of data to be generated in the non-identical files.

Data deduplication, brings a lot of advantages, for the users of sensitive data, both insider and outsider of susceptible attack, which will cause security or privacy issues. Coming to Traditional encryption, while providing the confidentiality of the data, there is no de-duplication and data compatibility. Especially with the conventional encryption, to encrypt the data with your own key, you need a different user. In this way, different copies of the same data of the user, is created, it will lead to different cipher texts. With Convergence encryption [2], it is possible to get deduplication, and it has been proposed to enforce data confidentiality. It decrypts the copy of the convergence keys and data obtained by computing a cryptographic hash value of the content of the encryption / data copy. After the key generation and data encryption, user holds the key, and sends the encrypted text to the cloud. The basic Encryption operation is a deterministic, since it is derived from the data or content, the same data of the copy having is the same convergence key, and therefore to generate the same cipher text. In order to prevent unauthorized access, a secured proof of possession protocol [3] is required to prove that the user is actually overlapped with the same file that was detected. After the evidence, by using the same file, without any action from the subsequent user to upload the same file, it will provide a pointer from the server. The user by the owner of the data corresponding to their convergence key can be decoded and can be download both the pointer from the server, and also the encrypted file from the cloud.

II. DESCRIPTION ON LITERATURE REVIEW:

Bellare et al. [4] Dropbox, such as, of Mozy, and the only other such cloud storage service provider by storing a copy of each file that was uploaded, and then run the deduplication in order to save space. The client, conventionally, need to encrypt the file, however, the savings will be lost. Message lock encryption, to resolve this tension. However, they will essentially subject to brute-force attacks that can be used to recover files that fall into a known set. They proposed to provide architecture secure deduplication of storage to resistance to brute-force attack, you have to realize it in a system called DupLESS. In DupLESS, to encrypt the bottom

of the message-based key obtained from the key server through the PRF protocol that the client is not aware of it. By having this technique, it assures the strong confidentiality of the data or content. That is by converting the original message to cipher text, thus we can able to protect the confidentiality of the data. In order to solve the problem, they had to use a different third-party medium called the key server in order to generate a tag to check the duplication of the file.

Stanek et al. [5] Recent years have witnessed the trend of leveraging cloud-based services for large scale content storage, for both processing, and distribution. Security and privacy are among top concerns for the public cloud environments. Towards these security challenges, they proposed and implemented, one technique called OpenStack Swift, a new client-side deduplication scheme for securely storing and sharing outsourced data via the public cloud. The originality of that proposal is twofold. First, it ensures better confidentiality towards unauthorized users. That is, every client computes a peer data key to encrypt the data that he intends to store in the cloud. As such, the data access is managed by the data owner. Second, by integrating access rights in metadata file, an authorized user can de-cipher an encrypted file, only with his private key. For popular data that are not particularly sensitive, the traditional conventional encryption is performed. Another two layered encryption scheme with stronger security while supporting deduplication is proposed for unpopular data. In this way, they achieved better tradeoff between the efficiency and security of the outsourced data. Here they are failed to handle data maintenance as well as encryption key.

J. R. Douceur [2] Farsite distributed file system provides availability by replicating each file into a multiple number of desktop computers. Since this replication consumes significant storage space, it is important to reclaim the used space where it is possible. Measurement of over 500 desktop file systems shows that nearly half of all consumed space that is only occupied by duplicate files. They present a mechanism to reclaim space from this incidental duplication to make it available for controlling the file replication. That mechanism includes: convergent encryption, which enables duplicate files to be co-selected into the space of a single file, even if the files are encrypted with different user's keys; and SALAD, a Self-Arranging Lossy Associative Database for aggregating file content and location information in a decentralized, scalable and fault-tolerant manner. Large-scale simulation experiments show that the duplicate-file co-selecting system is scalable, highly effective, and fault-tolerant. Here they provided effective data privacy mechanisms using convergent encryption techniques.

Bellare et al. [7] Message Lock Encryption (MLE), formally a new encryption primitive to call a key as it is derived from the message under encryption and decryption are to be performed. MLE is, how to achieve security to the target message from the current number of cloud storage providers. This approach is, for privacy protection, we provided a definition of both forms. You can call this integrity as the tag integrity. Based on this infrastructure, this approach is to both practical and theoretical contribution. In practical terms, analysers of the natural family of this approach provides ROM security, MLE scheme includes the deployment scheme also. Challenges in theory, is acted as a solution of the standard model, this approach is deterministic encryption, and under different assumptions for secure hash function of correlation input. To implement this scheme for different classes of the message source, make the connection with the paradigm of the extract. Here they considered the application to get better security of outsourcing the storage to achieve space efficiency.

Xu et al. [8] proof of ownership schemes, one of the owner of the same file F is, in the set of bounded leakage, he / she can prove that the cloud storage that owns the file F in a robust and efficient manner location F on a certain amount of files efficiently extractable and information has been leaked. Following this work, we proposed and listed the benefits, such as, the deduplication method of security towards client-side: It has a scheme [3], for a while External enemies and curiosity of the users in both of the cloud storage server, should be a threat to protect the confidentiality of the data. This method has been safe with respect to any distribution with sufficient minimum entropy [3]. It is specific to the particular type of distribution of the input file. Here they are considering the problem of key management and block-level deduplication and address issues showed secure convergence encryption as efficient encryption.

Li et al. [6] deduplication of data is a technique for eliminating duplicate copies of the data, in order to reduce the broad storage space and uplift the bandwidth those are used in cloud storage. As it is a promising and resulting challenge to make the deduplication safe in the cloud storage. Although the convergence encryption has been widely adopted for the safe de-duplication. In practical, convergence encryption making the important issue that to manage a huge number of efficient and reliable convergence keys. That is, each user has to encrypt the convergence key, introduced from the baseline approach that holds the master key that is independent for outsourcing the data in the cloud. However, the key management system of such a baseline approach, is to generate multiple keys for huge number of users. The key, is required for the user to protect the master key that is only for the dedicated purpose. For this purpose, there is De-key (Decryption key), but for to use this there is a need to manage any key on their own, instead of using convergence key. Such a particular key is not distributed and not to be sharable on multiple servers. Security analysis indicates that it is safe in terms of the definitions specified in the security model De-key was proposed. As a proof of this concept, we will implement the De-key using the lamp secret dispersion method, De-key has shown that the limited overhead in a realistic environment. Here after they encrypt the file, by distributing these keys across multiple servers, it has taken up the issue of key management at the block-level deduplication.

III. SYSTEM ARCHITECTURE:

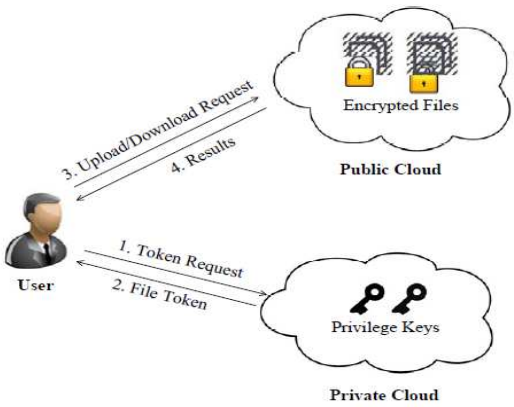

Fig 1: Basic implementation model

A. *Proposed Approach Details:*

In this model we are mainly implementing how the proposed system is doing our task in a sequential way. To achieve that we have to implement cloud service provider task, user task and private cloud task in a deterministic way. Each task is having their own significant functionality to achieve efficiency in a better way. Those are as follows.

B. *Cloud Service Provider (CSP)*

In this step, we develop a Cloud Service Provider module. This is an entity that provides a data storage service in public cloud. The S-CSP (Storage Cloud Service Provider) provides the data outsourcing service and stores data on behalf of the users. To reduce the storage cost, the S-CSP eliminates the storage of redundant data via deduplication and keeps only unique contents of the data. In this project, we assume that S-CSP is always online and has abundant storage capacity and computation power.

C. *Data Users Module:*

A user is an entity that wants to outsource data from the S-CSP and access the data later. In a storage system that supports deduplication, the user only uploads unique data but does not upload any duplicate data to save the bandwidth, even if he/she wants to upload the same content it does not allowed to upload the file of contents, which may be owned by the same user or different users. In the authorized deduplication system, each user is issued a set of privileges in the setup of the system. Each file is protected with the convergent encryption key and privilege keys to realize the authorized deduplication with differential privileges.

D. *Private Cloud Module*

Compared with the traditional deduplication techniques in the field of cloud computing, this is a new entity introduced for facilitating user's security by providing cloud service. Specifically, since the computing resources of data at user/owner side is restricted and the public cloud is not fully trusted in practice. Then the private cloud is able to provide data to user/owner with an execution environment and provides infrastructure to work as an interface between user and the public cloud. The private keys for the privileges are managed by the private cloud, who is responsible for the file token requests from the users. The interface offered by the private cloud allows user to submit files and queries to be securely stored and computed respectively. After the proper response of the private cloud only, the user can manipulate the data.

E. *Deduplication System Approach*

We consider several types of privacy checks we had the need to protect the data, that is,

i) Unforgeability of tag duplicate-check: There are two types of adversaries, that is, external adversary and internal adversary. The external adversary can be viewed as an internal adversary without any privilege. If a user has privilege p, it requires that the adversary can't forget and output a valid duplicate tag with any other privilege p' on any file F, where p does not match p'. Furthermore, it is also required that if the adversary does not make a request of tag with its own privilege from private cloud server, it cannot forget and output a valid duplicate token with p on any file F that has been queried for.

ii) Confidentiality of the file: It should be achieved by encrypting the data to be stored in the public cloud to access later.

IV. EVALUATION APPROACHES:

In this section we are going to discuss about various approaches how they achieve the goal of avoiding the duplication of data in the cloud. The evaluation of all these approaches is based on deduplication ratio. Our evaluation highlights file token generation, share the token against authentication of the user to access the services provided by the model.

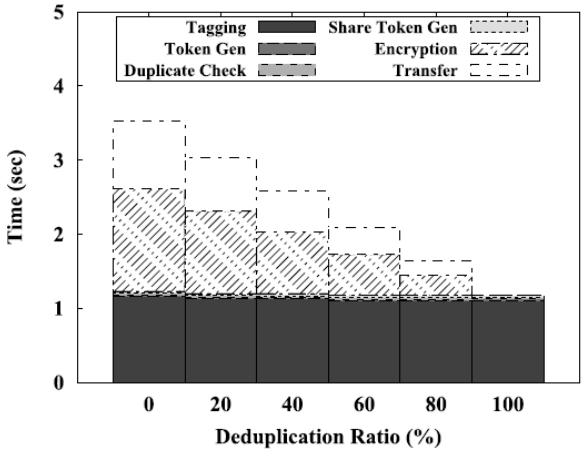

Figure 2: Time breakdown for different deduplication ratio.

To evaluate the effect of deduplication ratio, assume a cloud consists of 50 MB to 100 MB of files, each of which prepares two unique data sets. You must first upload the first set as the first upload. For the second upload, as our unique file, from the initial settings, such as duplicate files and the rest of the files from the second set, depending on the given

deduplication rate, part of the 50 files selected then, Upload and average time for encryption is to upload the second set to be skipped in the case of duplicate files and it shows the time required for both , which is reduced with the increase of the deduplication rates in Figure 2. If a duplicate is found, the search will decrease the time required for duplicate check to be completed. The amount of time spent to upload the deduplication rate of a file at 100 percent is only of 33.5 percent with a specific file in the list.

V. CONCLUSION:

In this paper, we aim to perform analyses on the problems of duplication with differential accesses in cloud computing. Unlike the existing data deduplication system, private cloud has been implemented as a proxy that allows the data of the owner /user performs a duplicate check with safe differential authority. Such architecture is attracting by getting practical attention from the researchers.

Data owner, while the operation of the data only has been managed by the private cloud, by taking advantage of the public cloud, we are outsourcing their data in the cloud storage. The new de-duplication system support differential overlap check in the S-CSP and has been proposed, under this hybrid cloud architecture that exists in the public cloud. Users are only allowed to perform the duplicate check of data on files and marked with the corresponding authority if it is a mark by having duplication, it omits the contents without encryption.

REFERENCES

[1] S. Quinlan and S. Dorward, “Venti: A new approach to archival storage,” in Proc. 1st USENIX Conf. File Storage Technol., Jan. 2002.
[2] 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. Int. Conf. Distrib. Comput. Syst.,2002, pp. 617–624.
[3] S. Halevi, D. Harnik, B. Pinkas, and A. Shulman-Peleg, “Proofs of ownership in remote storage systems,” in Proc. ACM Conf. Comput.Commun. Security, 2011, pp. 491–500.
[4] M. Bellare, S. Keelveedhi, and T. Ristenpart, “Dupless: Serveraided encryption for deduplicated storage,” in Proc. 22nd USENIX Conf. Sec. Symp., 2013, pp. 179–194.
[5] 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.
[6] 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., http://doi.ieeecomputersociety.org/10.1109/TPDS.2013.284, 2013.
[7] 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.
[8] 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.

Mr. K.BM Pavan Kumar is a student of ADITYA COLLEGE OF ENGINEERING surampalem.Presently he is pursuing his M.Tech in Comuter Science and Engineering from this college.He received his B.Tech from G.V.R&S COLLEGE OF ENGIENERING AND TECHNOLOGY, affiliated to Acharya Nagarjuna University, Guntur in the year 2012.His area of interest includes Computer Networks, Network Security and current trends and techniques in Computer Science.

Mr. V. Chandra Sekhara Rao is an excellent teacher, he is working as a Senior Assistant Professor in ADITYA COLLEGE OF ENGINEERING. He is pursuing his PhD in JNTU-KAKINADA. He received his Master of Technology degree from JNTU Kakinada. He is having 10 years of experience in teaching. To his credit, he had several publications and conferences. His area of interests are Data warehousing & Data Mining and recent trends in Computer Science.