International Journal of Research (IJR)
e-ISSN: 2348-6848, p- ISSN: 2348-795X Volume 2, Issue 09, September 2015
Available at http://internationaljournalofresearch.org

Stopping the Duplication of Data Storage in Public & Private Cloud Storage using S-CSP

Chittampally Ramya[1]& M.Mohanrao[2]
[1]M-Tech Dept. of CSE Megha Institute of Engineering & Technology for Women
[2]Assistant ProfessorDept. of CSE Megha Institute of Engineering & Technology for Women

Abstract –
In recent days, cloud computing became an emerging service model which provides highly available storage and massively parallel computing. Cloud storage enables users to out- source their data backup over remote cloud providers. Managing ever increasing growth of data, is became a great headache. To address this problem, data deduplication technique is introduced which eliminates redundant data copies by keeping a physical copy. For security concerns encryption becomes a necessary before updating data into the cloud. Since these are two challenges that we focused in this paper. For achieving deduplication along with data security, secure hashing algorithm is used.
Keywords:-Deduplication; Private Cloud; S-CSP; Data Users

1. INTRODUCTION

Cloud computing is the delivery of computing services over the Internet. Cloud services allow individuals and businesses to use software and hardware that are managed by third parties at remote locations. Data deduplication, an effective data compression approach that exploits data redundancy, partitions large data objects into smaller parts, called chunks, represents these chunks by their fingerprints, replaces the duplicate chunks with their fingerprints index lookup, and only transfers or stores the unique chunks for the purpose of communication or storage efficiency. Source Deduplication that eliminates redundant data at the client site is obviously preferred to target deduplication due to the former's ability to significantly reduce the amount of data transferred over wide area network with low communication bandwidth. A hybrid cloud is a combination of different methods of resource pooling (for example, combining public and community clouds).Cloud services is popular Cloud services are popular because they can reduce the cost and complexity of owning and operating computers and networks. Since cloud users do not have to invest in information technology infrastructure, purchase hardware, or buy software licenses, the benefits are low up-front costs, rapid return on investment, rapid deployment, customization, flexible use, and solutions that can make use of new innovations. Some other benefits to users include scalability, reliability, and efficiency. Scalability means that cloud computing offers unlimited processing and storage capacity. The cloud is reliable in that it enables access to applications and documents anywhere in the world via the Internet. Cloud computing is often considered efficient because it allows organizations to free up resources to focus on innovation and product development. Another potential benefit is that personal information may be better protected in the cloud. Specifically, cloud computing may improve efforts to build privacy protection into technology from the start and the use of better security mechanisms. ALG-Dedupe outperforms the existing state-of-the-art source deduplication schemes in terms of backup window, efficiency and cost saving for its high deduplication efficiency and low system overhead. Thus, the basic idea of ALG-Dedupe is to effectively exploit this application difference and awareness by treating different types of applications independently and adaptively during the local and global deduplication.

International Journal of Research (IJR)
e-ISSN: 2348-6848, p- ISSN: 2348-795X Volume 2, Issue 09, September 2015
Available at http://internationaljournalofresearch.org

2. RELATED WORK

First process is the input file selection. Select the input file. Then input file has been split into the data chunks. After the data chunks has been loaded into the database, based on the input file. Then calculate the backup window size. In ALG-Dedupe filters out these tiny files in the file size filter before the deduplication process, and groups data from many tiny files together into larger units of about 1 MB each in the segment store to increase the data transfer efficiency over WAN. Figure 1 system architecture An Application aware Local-Global source deduplication scheme that not only exploits application awareness, but also combines local and global duplication detection, to achieve high deduplication efficiency by reducing the deduplication latency to as low as the application-aware local deduplication while saving as much cloud storage cost as the application-aware global deduplication. Most of the files in the PC dataset are tiny files that less than 10 KB in file size, accounting for a negligibly small percentage of the storage capacity. The statistical evidences about 63 percent of all files are tiny files, accounting for only 1.9 percent of the total storage capacity of the dataset. To reduce the metadata overhead. The deduplication efficiency of data chunking scheme among different applications differs. Depending on whether the file type is compressed or whether SC can outperform CDC in deduplication efficiency, The file can be classified into three main categories: compressed, static compressed files, and dynamic uncompressed files. The dynamic files are always editable, while the static files are uneditable in common. To strike a better tradeoff between duplicate elimination ratio and deduplication, An deduplicate compressed files with WFC, separate static uncompressed files into fix-sized chunks by SC with ideal chunk size, and break dynamic uncompressed files into variable-sized chunks with optimal chunk size using CDC based on the Rabin finger printing to identify chunk boundaries. While there are benefits, there are privacy and security concerns too. Data is travelling over the Internet and is stored in remote locations. In addition, cloud providers often serve multiple customers simultaneously. All of this may raise the scale of exposure to possible breaches, both accidental and deliberate. Concerns have been raised by many that cloud computing may lead to "function creep" — uses of data by cloud providers that were not anticipated when the information was originally collected and for which consent has typically not been obtained. Given how inexpensive it is to keep data, there is little incentive to remove the information from the cloud. Some other benefits to users include scalability, reliability, and efficiency. Scalability means that cloud computing offers unlimited processing and storage capacity. The cloud is reliable in that it enables access to applications and documents anywhere in the world via the Internet. Cloud computing is often considered efficient because it allows organizations to free up resources to focus on innovation and product development.

A hybrid cloud is a combination of different methods of resource pooling (for example, combining public and community clouds). Since cloud users do not have to invest in information technology infrastructure, purchase hardware, or buy software licenses, the benefits are low up-front costs, rapid return on investment, rapid deployment, customization, flexible use, and solutions that can make use of new innovations. Depending on the location where redundant data is eliminated. The deduplication can be

International Journal of Research (IJR)
e-ISSN: 2348-6848, p- ISSN: 2348-795X Volume 2, Issue 09, September 2015
Available at http://internationaljournalofresearch.org

categorized into source deduplication that applies data deduplication at the client site and target deduplication that eliminates redundant data at the backup server site. Since data backup for personal computing in the cloud storage environment implies geographic separation between the client and the service provider.

3. Proposed System

There are three entities defined in our system, that is, users, private cloud and S-CSP in public cloud as shown in Fig. 1.Hybrid cloud. An environment comprised of two or more of the above specified cloud computing deployment models in a manner where they are bound together using technology that supports application, service or data portability, migration and interoperability.

• Public cloud. An environment provisioned for open use by the general public.

• Private cloud. An environment provisioned for exclusive use by a single organization comprising multiple (internal) consumers. Sometimes called an enterprise cloud, private clouds can be on premise or off-premise, managed internally or managed by a third-party provider. In this paper, we will only consider the file level deduplication for simplicity. In another word, we refer a data copy to be a whole file and file-level deduplication which eliminates the storage of any redundant files. Actually, block-level deduplication can be easily deduced from file-level deduplication, which is similar . Specifically, to upload a file, a user first performs the filelevel duplicate check. If the file is a duplicate, then all its blocks must be duplicates as well; otherwise, the user further performs the block-level duplicate check and identifies the unique blocks to be uploaded. Each data copy (i.e., a file or a block) is associated with a token for the duplicate check.

• **S-CSP**. This is an entity that provides a data storage service in public cloud. The S-CSP provides the data outsourcing service and stores data on behalf of the users. Fig 1- Block diagram of proposed system To reduce the storage cost, the S-CSP eliminates the storage of redundant data via deduplication and keeps only unique data. In this paper, we assume that S-CSP is always online and has abundant storage capacity and computation power.

• **Private Cloud**. Compared with the traditional deduplication architecture in cloud computing, this is a new entity introduced for facilitating user's secure usage of cloud service. Specifically, since the computing resources at data user/owner side are restricted and the public cloud is not fully trusted in practice, private cloud is able to provide data user/owner with an execution environment and infrastructure working as an interface between user and the public cloud. The private keys for the privileges are managed by the private cloud, who answers 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. This is a novel architecture for data deduplication in cloud computing, which consists of a twin clouds (i.e., the public cloud and the private cloud). Actually, this hybrid cloud setting has attracted more and more attention recently. Under the assumption, two kinds of adversaries are considered, that is, 1)external adversaries which aim to extract secret information as much as possible from both public cloud and private cloud; 2) internal adversaries who aim to obtain more information on the file from the public cloud and duplicate-check token information from the private cloud outside of their scopes. Such adversaries may include S-CSP, private cloud server and authorized users

• **Data Users:** A user is an entity that wants to outsource data storage to the S-CSP and access the data later. In a storage system supporting

International Journal of Research (IJR)
e-ISSN: 2348-6848, p- ISSN: 2348-795X Volume 2, Issue 09, September 2015
Available at http://internationaljournalofresearch.org

deduplication, the user only uploads unique data but does not upload any duplicate data to save the upload bandwidth, 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 theconvergent encryption key and privilege keys to realize the authorized deduplication with differential privileges.

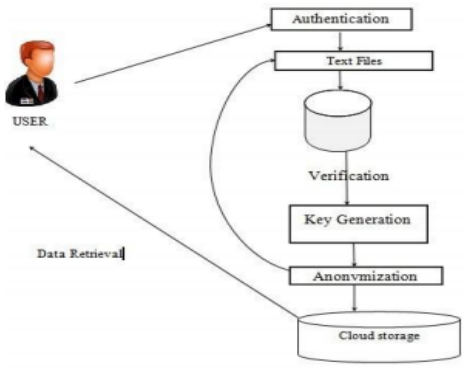

Figure 1: System Architecture

We consider cloud storage system as shown in Fig. 1, which involves data owners (user), the private cloud storage (database), and the public cloud storage (cloud storage). The user firstly get authenticate. Only authenticate user can enter in the system. Authenticate key is stored in the private cloud. File data is then transferred towards public cloud. For securely B. System Modules 1) User Authentication: Authentication is accepting proof of identity given by a credible person who has first-hand evidence that the identity is genuine. Attribute comparison might be vulnerable to forgery. In general, it relies on the facts that creating a forgery indistinguishable from a genuine artifact requires expert knowledge, that mistakes are easily made, and that the amount of effort required to forgery is considerably greater than the amount of profit that can be gained from it. Only the privileged user is allowed to store the content on the cloud and allowed to process the further procedures. 2) Key Generation: Key generation is the process of generating key for checking the file duplication which occur on the cloud normally by enabling SHA-1 Algorithm for the key generation, normally individual key will be generated for each file, even when the same file is altered the key for the same file be changed for each individual file and it also very secure than the other algorithm such as HMAC which is used in the proposed system since SHA-1 uses an iterative algorithm. It generates digests by first splitting content into blocks of 64 bytes and, one after the other, combining those blocks together to generate the 20 byte digest. 3) File transmission: A unique key will be generated for each individual file and each file will be checked with cloud whether it is previously exist or not in the cloud if it already present the current file will not be uploading an error signal will be a raised. This will reduce the amount of storage space which occurs on the cloud and also bandwidth will be reduced and also these terms to reduce the cost of cloud whenever user needs to store inside the cloud. 4) Data Anonymization: Data anonymization is normally done to add security inside the cloud whenever the file is not duplicated the file which we want to store inside the cloud will checked with generated key by using SHA-1 algorithm and the file will get encrypted and stored on the cloud. 5) Data De-Anonymization and Downloading: Whenever the user wants its encrypted cloud data to get viewed or downloaded, the user want to specify the key which the user enters during the authentication section once user enters the specified it will checked with the database and file content will be decrypted and it will be downloaded from the cloud.

International Journal of Research (IJR)
e-ISSN: 2348-6848, p- ISSN: 2348-795X Volume 2, Issue 09, September 2015
Available at http://internationaljournalofresearch.org

4. Experimental Results

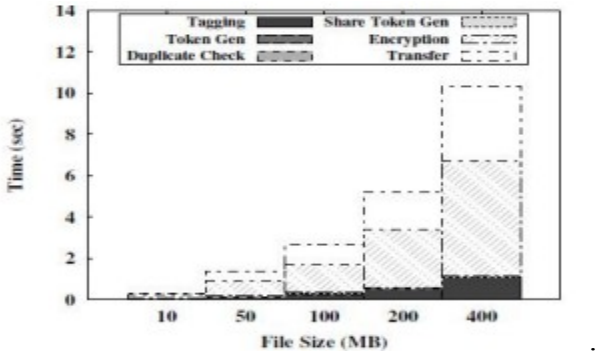

Fig-2: Time Breakdown for Different File Size

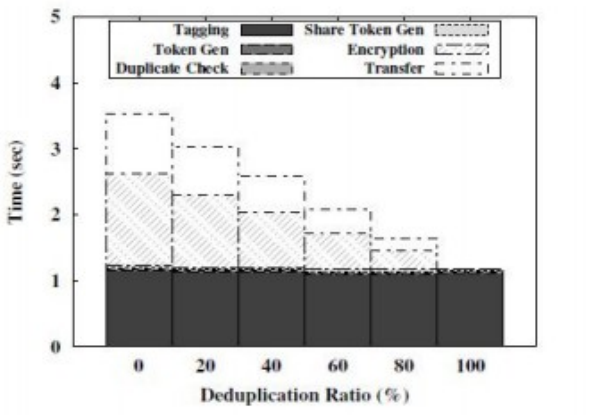

Fig-3: Time Breakdown for Different Duplication Ratio

5. CONCLUSION

In this paper, the notion of authorized data deduplication was proposed to protect the data security by including differential privileges of users in the duplicate check. We also presented several new deduplication constructions supporting authorized duplicate check in hybrid cloud architecture, in which the duplicate-check tokens of files are generated by the private cloud server with private keys. Security analysis demonstrates that our schemes are secure in terms of insider and outsider attacks specified in the proposed security model. As a proof of concept, we implemented a prototype of our proposed authorized duplicate check scheme and conduct testexperiments on our prototype. We showed that our authorized duplicate check scheme incurs minimal overhead compared to convergent encryption and network transfer.

6. REFERENCES

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

[2] 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.

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

[4] 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.

[5] 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

[6] 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 Parallel and Distributed Systems*, 2013.

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

[8] 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.

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

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

Author Profile

Dr.Shaik Abdul Muzeer

Professor & Principal

Megha institute of Engineering & Technology for Women

Dr.S.A.Muzeer, at present working as a principal of Megha institute of engineering & Technology has completed his PG and P.HD in Electronics & Communication Engineering and published around 25 Papers in National & International Journals. His area of research is Digital signal processing and Bio-medical engineering