IJCTA, 9(26) 2016, pp. 187-192
© International Science Press

A Survey on Analysis of Efficient Deduplication in Cloud Computing Environment

Francina Sophiya D.[1] and Swarnalatha P.[2]

ABSTRACT

Cloud computing has arrived as one of the fastest-growing segments of the Information technology industry which provides variant services such as software, platform and infrastructure for internet users. The ability to leverage economies of scale, open source software, geographic distribution and automated systems to drive down costs makes cloud computing an attractive option for businesses. With the potentially infinite storage space offered by cloud providers, users tend to use as much space as they can and vendors regularly look for techniques aimed to minimize redundant data, maximize space savings and to make data manageable. A technique which has been adopted to manage large redundant data is Deduplication which plays a key role in Cloud Computing services. In this paper, we address redundancy issues in Cloud Computing environments and three encryption methods for data Deduplication over cloud have been discussed. The paper studies existing challenges and the various Deduplication approaches.

Keywords: Redundancy, Deduplication, Cloud Computing, Cloud Providers, Encryption, File-level Deduplication, Block-level deduplication, Attribute Based Encryption, Convergent Encryption, Message-Locked Encryption.

1. INTRODUCTION

With the infinite storage space that is offered by cloud service providers, users tend to use as much space as they can and vendors constantly look for technique to minimize redundant data and maximize space savings. A technique which has been widely adopted is cross-user deduplication. The simple idea behind deduplication is to store duplicate data (either files/blocks) only one time. Therefore, if a user wants to upload a file (or block) which is already stored, the cloud provider will add the user to the owner list of that file (or block). Deduplication has proved to obtain high space and money saving. And many cloud storage providers are currently adopting it. Deduplication has proved to achieve high cost reduction, reducing up to 90-95 percent storage needs for backup applications and up to 68% in standard systems. Clearly, the savings, which can be passed back directly or indirectly to cloud users, are significant to the economics of cloud business. [8] Being data deduplication applied in cloud technology can reduce the data storage ,size and save network bandwidth, the dynamicity of data in cloud storage systems are different from backup and archive systems, which brings challenges for implementation of data deduplication in cloud storage systems.

2. BACKGROUND WORK

In this section, we present some important observations drawn from previous and our analysis of the vendor problem in cloud storage, showing how to make an efficient deduplication that allows very appealing reductions in the usage of storage resources.

Pasquale *et al.* [11] present the inherent security exposures of convergent encryption and propose ClouDedup, which preserves the combined advantages of deduplication and convergent encryption**.**

[1] Research Scholar, SCOPE, VIT University, Vellore, TN, India, *E-mail: sophiyadominic08@gmail.com*

[2] Associate Professor,SCOPE, VIT University, Vellore, TN, India, *E-mail: pswarnalatha@vit.ac.in*

Zheng et al. [1] propose a scheme based on attribute-based encryption (ABE) to deduplication encrypted data stored in the cloud and support secure, effective and efficient data access control.

Fatema et al. [10] extended there work for enterprise data deduplication framework is composed of three steps: In the first step, the data and its metadata is indexed in such a way as to ensure complete data privacy against a semi-honest cloud service provider. The second step consists in performing the multi-user private keyword searchable encryption on the encrypted data within a particular enterprise, keeping the searches and the resulting files secret from the cloud service provider. Step 3 makes use of a strategy to support data sharing between users, by utilizing the existing metadata, the indexing structures, and the searchable encryption scheme.

Xiaolong et al. tried a new approach combining client-side deduplication and target-side deduplication for cloud storage system. The contribution of this paper is twofold. Firstly, file-level and chunk-level duplication are detected and eliminated locally by Client and globally by Metadata Server (MS) to improve the deduplication ratio. Secondly, we put forward the Delay Dedupe strategy, a delayed target-deduplication scheme based the chunk level deduplication and the access frequency of chunks in the Snodes.

N. Lakshmi et al. [4] proposed system uses ALG data dedupulication technique, which avoids data redundancy saving space ultimately increasing the efficiency of the data storage. This system also uses the standard AES algorithm for data encryption thus adding data security. Finally this also uses the RSS key method, which is generated dynamically in a unique way, which is far secured than the privilege keys.

N. Jayapandian et al. [5] takes a solution called Convergent Encryption (CE). CE is a deterministic symmetric encryption scheme in which the key K is derived from the message M itself by computing K = H(M) and then encrypting the message as C = E(K;M) = E(H(M);M) where H is a cryptographic hash function and E is a file cipher.Using CE, any user holding the same message will produce the same key and cipher text, enabling deduplication. And suggested MLE for small files.

Jan et al.[9] found a secure deduplication scheme for encrypted data that has dynamic ownership management capability. There proposed scheme is constructed based partially on a randomized convergent encryption scheme in order to randomize the encrypted data, which renders the proposed scheme secure against the chosen-plaintext attack while still allowing deduplication over the data. There proposed scheme is further integrated into the re-encryption protocol for owner revocation. The owner revocation is executed by re-encrypting the outsourced cipher text and selectively distributing the re-encryption key to valid owners by the cloud server.

3. DATADEPUBLICATION CHALLENGES

Data deduplication is a technique used for reducing the amount of storage space that an organization requires to save. In many organizations, the storage systems contain duplicate copies of many data. Consider the example where the same file may be saved in several different places by different users, or two or more files that aren’t identical may still include much of the same data. Deduplication eliminates the additional extra copies by saving just one copy of the data and replacing the other copies with pointers that lead back to the original copy. Organization frequently use deduplication in backup and disaster recovery applications, but it can be used to free up space in primary storage that leads efficient storage management.

In simple, deduplication takes place on the file level; that is, it eliminates duplicate copies of the same file. This kind of deduplication is known as file-level deduplication or single instance storage (SIS). Deduplication can also take place on different levels the block level, eliminating duplicated blocks of data that occur in non-identical files. Block-level deduplication frees up more space than SIS, and a particular type known as variable block or variable length deduplication has become most popular. Often the phrase “data deduplication” is used as block-level / variable length deduplication.

The data deduplication can reduce the amount of disk or tape that the organization is in need, and in turn reduces costs. NetApp reports says that there is different cases in which deduplication can reduce storage requirements up to 95 percent, and the type of data that you're trying to deduplicate and the amount of file sharing your organization does will influence your own deduplication ratio. While deduplication can be applied to data stored on disk, the relatively much costs of disk storage make deduplication a very popular option for disk-based systems. Eliminating extra copies of data saves cost not only on direct disk hardware money, but also on related costs, like electricity cost, cooling expense, maintenance charge, floor area, etc.

Deduplication can also reduce the amount of network bandwidth required for backup processes, and in many cases, it can speed up the backup and recovery process.

4. EVALUATION

Our scheme works in a system containing three types of entities [1]:

1. A CSP that offers a storage service. A cloud service provider is a company that offers service component of cloud computing – typically Infrastructure as a Service (IaaS), Software as aService (SaaS) or Platform as a Service (PaaS) – to other business. Cloud service providers are sometimes referred to as cloud providers or CSPs.
2. A data owner that stores its data at the CSP (assume as only one data owner for one data M); and
3. Data holders (ui, i = 1 . . . n) that are eligible data users and could save the same data as the data owner at the CSP.[1]

Our proposed system design data deduplication framework is composed of three steps: In the first step, the data and its metadata is indexed in such a way as to ensure complete data privacy against a semi-honest cloud service provider. The second steps consists in performing the multi-user private keyword searchable encryption on the encrypted data in cloud, keeping the searches and the resulting files secret from the cloud service provider. Step 3 makes use of a strategy to support data sharing between users, by utilizing the existing metadata, the indexing structures, and the searchable encryption scheme.[10]This ensures a data confidentiality in the duplication.

(A) Primitive Function System

KeyGen(F) :The key generation algorithm takes a file content F as input and outputs the convergent key ckFof F;

Encrypt (ckF;F) : The encryption algorithm takes the convergent key ckFand file content F as input and outputs the ciphertextctF;

Decrypt (ckF; ctF): The decryption algorithm takes the convergent key ckFand ciphertextctFas input and outputs the plain file F;

TagGen(F) :The tag generation algorithm takes a file content F as input and outputs the tag tagFof F.

(B) File Uploading

Phase 1 (cloud client '! cloud server): client performs the redundancy check with the cloud server to confirm if such a file is stored in cloud storage or not before uploading a file. If there is a duplicate, another protocol called Proof of Ownership will be run between the clients and the cloud storage server. Otherwise, the following protocols (including phase 2 and phase 3) are run between these two entities.

Phase 2 (cloud client ’! auditor): client uploads files to the auditor, and receives a receipt from auditor.

Phase 3 (auditor ’! cloud server): auditor helps generate a set of tags for the uploading file, and send them along with this file to cloud server.

(C) Encryption method for deduplication

Firstly, this ensures data confidentiality through Attribute based encryption, symmetric key encryption, and PKC. The original user data is encrypted using symmetric encryption with DEK, which is then encrypted using the Encrypt Key algorithm under access policy AP. Assuming that the symmetric key algorithm is secure (for example, using a standard algorithm such as AES), the scheme’s data confidentiality merely relies on the security of the Encrypt Key algorithm.[1]

Secondly, in Convergent encryption users check the Convergent keys from each data set or original data and encrypt the data copy with the generated convergent key. Users also add the tag for the data so that the tag will helps to detect the duplicate data. By using converged key generation algorithm to encrypt the user data. This will ensure the security, ownership and authority of the data. [7]. Convergent encryption provides a viable option to enforce data confidentiality while realizing deduplication. It encrypts/decrypts a data copy with a convergent key, which is derived by computing the cryptographic hash value of the content of the data copy itself [8]. After key generation and data encryption, users retain the keys and send the cipher data to the cloud. Since encryption is deterministic, identical data copies will generate the same convergent key and the same cipher data. This makes the cloud to perform deduplication on the cipher texts. The cipher texts can only be decrypted by the corresponding data owners using their convergent keys.

1. Thirdly, Message-Locked Encryption, A standard message-locked encryption scheme consists of five algorithms, Setup, KeyGen, Enc, Dec, and TagGen.
2. Setup: takes 1λ, returns a public parameter P;
3. KeyGen: takes P and a message M, returns a messagederivedkey K;
4. Enc: takes P, message-derived key K and message M,returns a ciphertext C;
5. Dec: takes P, message-derived key K and ciphertextC,returns a message M;
6. TagGen: takes P and ciphertext C, returns a tag T.

5. PERFORMANCE ANALYSIS

Convergent encryption suffers from some weaknesses which have been widely discussed in the literature [11], As the encryption key depends on the value of the plaintext, an attacker who has gained access to the storage can perpetrate the so called “dictionary attacks” by comparing the ciphertexts resulting from the encryption of well-known plaintext values from a dictionary with the stored cipher texts.[1] Indeed, even if encryption keys are encrypted with users’ private keys and stored somewhere else, the potentially malicious cloud provider, who has no access to the encryption key but has access to the encrypted chunks (blocks), can easily perform offline dictionary attacks and discover predictable files. This issue arises in [8] where chunks are stored at the storage provider after being encrypted with convergent encryption.[11][12].Since this has direct access to server data are less secure.

In attribute based encryption the complexity of implementation is high and it is not scalable and flexible for large volume of data but this supports large number of duplicate copies so it is much suitable for small data that has much copies. Since it is implemented in symmetric key policy the key distribution and computation is the bigger problem and could damage when compressed.

So, we suggest A new cryptographic primitive called Message-Locked Encryption (MLE)which subsumes Convergent Encryption.[5] Although MLE schemes can perform secure deduplication of encrypted data, they were proposed originally for file-level and target-based deduplication. We could extend an MLE scheme for secure DLSB-deduplication of large files by performing MLE on each data block (i.e., treating a data block as a file) and employing an existing Proof of Ownership scheme (e.g., the PoW scheme in).[13][1]MLE is faster scheme and safe from attacks.

There are too many researchers have been done to secure duplication check of data on cloud. In the cloud storage, data Deduplication has two methods present in existing system. First method of the data Deduplication is perform as post processing method [7] In this which data is first store on the storage device and then duplication check is applied on the data. The use of this method is there is no need to wait for calculating the hash function and the speed of storage not get downgrade. The main drawback with this system is that if storage capacity of the device is low then the file storage. problem of this the post processing method is not useful at all because it checks the file after storing it on the cloud server. Second method of the duplication check is the inline duplication check. It is check when new entries are to be added to the database the duplication of the file. It will checks for the block level duplication of the file [7][11] before adding the new entry or new data to the database. This method has some drawback such as each time need to calculate the hash function which may lead to slower throughput of the storage device. Another method of duplication check is source data Deduplication in which data duplication is done at the side of the source. The file duplication is check before it get uploaded on the cloud.[8].

The data that needs to be stored is first preprocessed if necessary. In certain cases the preprocessing takes up a lot of time based on the type of processing that is implemented. Data cleaning, Normalization, Data hiding, Structuring of data, etc are some of the preprocessing steps available. The next step is the chunking. The process of splitting the given data into any blocks or chunks of data is called as chunking. This is the crucial step in the deduplication process. This is because, based on the size of each chunk the number of duplicate data changes. Based on the initial data, the chunk size should be fixed in such a way that the obtained chunks have large number of duplicates and thus the storage size will be reduced as much as possible.

Cryptographically efficient scheme provable ownership of the file (POF) can be used, for a client to prove to the server that it indeed has the file. We achieve the efficient goal by relying on dynamic spot checking, in which the client only needs to access small but dynamic portions of the original file to generate the proof of possession of the original file makeable reduce the burden of computation on the client and minimizing the I/O between the client and the server. At the same time, by utilizing dynamic coefficients and randomly chosen indices of the original files, our scheme mixes the randomly sampled portions of the original file with the dynamic coefficients to generate the unique proof in every challenge. This technique guarantees the targeted security.

6. CONCLUSIONS

Managing data with deduplication is an important practice for achieving a successful cloud service, especially for large data storage. In this paper, a practical scheme that manages the encrypted large volume data in cloud with deduplication based on ownership challenge has been discussed. This survey will provide researcher and developer, an idea on privacy in deduplication, hype and challenges which intern facilitates them to evaluate and improve the existing and new deduplication techniques. The real time scenarios with their algorithm implementation have been dealt as a future work of the paper.

REFERENCES

[1] Zheng Yan, Mingjun Wang, and Yuxiang Li, Athanasios ,V. Vasilakos and Lulea., "Encrypted Data Management with Deduplication in Cloud Computing" *IEEE Cloud Computing IEEE computer society*, **3(2)**, 28-35, 2012.

[2] Kirubakaran R, Mano Prathibhan C, Karthika C., “Cloud Based Model for Deduplication of Large Data” *IEEE International Conference on Engineering and Technology (ICETECH)*, 1-4, 2015.

[3] Xiaolong Xu, Qun Tu., “Data deduplication mechanism for cloud storage systems” *International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery*, 286-294, 2015.

[4] N. Lakshmi Pritha, N.Velmurugan, Dr. S.Godfrey Winster, A.Vijayaraj., “Deduplication Based Storage and Retrieval of Data from Cloud Environment” *International Conference on Innovation Information in Computing Technologies(ICIICT)*, 1-6, 2015.

[5] N.Jayapandian, Dr.A.M.J.Md.Zubair Rahman, I.Nandhini., “A Novel Approach for Handling Sensitive Data with Deduplication Method in Hybrid Cloud” *IEEE International Conference on Control System, Computing and Engineering*, 1-6, 2015.

[6] Junbeom Hur, Dongyoung Koo, Youngjoo Shin, Kyungtae Kang., “Secure Data Deduplication with Dynamic Ownership Management in Cloud Storage” *IEEE Transactions on Knowledge and Data Engineering*, **99**, 1-6, 2016.

[7] Mane VIdya Maruti and Prof. Mininath K.Nighot., “Authorized Data Deduplication Using Hybrid Cloud Technique” *International Conference on Energy Systems and Applications (ICES), Pune, India*, **26(5)**, 1206-1216, 2015.

[8] Zheng Yan, Senior Member, IEEE, Wenxiu Ding, Xixun Yu, Haiqi Zhu, and Robert H. Deng., “Deduplication on Encrypted Big Data in Cloud” *IEEE Transactions On Big Data*, **2(2)**, 138-150, 2016.

[9] Jan Stanek, Alessandro Sorniotti, Elli Androulaki, and Lukas Kencl., “A Secure Data Deduplication Scheme for Cloud Storage” *Lecture Notes in Computer Science*, **8437**, 99-118, 2014.

[10] Fatema Rashid, Ali Miri, Isaac Woungang., “Secure Enterprise Data Deduplication in the Cloud” *IEEE Sixth International Conference on Cloud Computing*, 367-374, 2013.

[11] Pasquale Puzio, Refik Molva, Melek O nen, Sergio Loureiro., “ClouDedup: Secure Deduplication with Encrypted Data for Cloud Storage” *IEEE CloudCom*, 2013.

[12] Mihir Bellare Sriram Keelveedhi, Thomas Ristenpart., “Message-Locked Encryption and Secure Deduplication” *A preliminary version of this paper appears in the proceedings of Eurocrypt*, 1-29, 2013.

[13] R.Manjusha, R.Ramachandran., “Comparative Study of Attribute Based Encryption Techniques in Cloud Computing” *International Conference on Embedded Systems (ICES)*, 116-120, 2014.