International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2017 IJSRCSEIT | Volume 2 | Issue 6 | ISSN : 2456-3307

A Theoretical Approach for Secure Authorized De-Duplication

Padma Punna, Racha Suresh
United Health Group Information Services Pvt. Ltd, India

ABSTRACT

Cloud computing is one of the technology which can be implemented for the users who are unable to have their complete infrastructure, unable to manage the security features and storage systems. For all of those people cloud computing is one of the best alternatives so that they will pay for what they use. But cloud computing is having its own difficulties. Cloud service providers provide service to multiple clients simultaneously and all these clients will use the cloud infrastructure for doing their operations. As this is used by multiple users and huge amount of data will be placed on the cloud. Data which is storing in the cloud may gets duplicated because of storing the data by multiple users. To eliminate redundant storage of data necessary steps should be taken by the cloud service provider. To eliminate redundant data a new approach is used which known as de-duplication. De-duplication is the process of eliminating the duplicated data from the cloud storage and brings the cloud storage structure to the consistent storage stucture. In this paper I am concentrating on the de-duplication process. But before this, to gain the knowledge on the cloud systems, I had some sort of survey on cloud computing and its services. While doing the survey I gets motivated with the de-duplication concept and decided to work on that. De-duplication is mainly done with the compression techniques like encryption and decryption methods. The main problem is that after encryption we use to have cipher text and we can't recognize what the data will be there in that and we are not even sure that data is duplicated or not. The process of de-duplication is applied after compressing the file and then it is encrypted and shared between the client and the service provider to even have the security for the data. In this paper I am proposig a new method for eliminating the duplicated data by providing the security to the data in the cloud storage.

Keywords : Hybrid, De-duplication, authorization, encryption, decryption, file, block, chunk.

I. INTRODUCTION

De-duplication is a data compression technique for eliminating redundant data. De-duplication can be organized into two different levels such as file level and block level, where file level de-duplication process into the entire file, where small change in a file makes different from current version of the file. Where as in block level data blocks are considered for de-duplication. In future stage the implementation of de-duplication could be based on location such as client side de-duplication or source side de-duplication. So, that ensuring client side de-duplication needs resource saving, in such cases only file hash values are sent to cloud server, if in case of duplicate existence.

However De-duplication is extensively used in various data management applications such as metadata management, backup level and storage optimization

Convergent Encryption

The convergent encryption generates encryption is applied on the file with any of the public key or private key. This encryption will convert the plain text to the cipher text and assigns encrypted key to the cipher data for avoiding future modification on existence file or data.

Proof of Ownership

De-duplication is a process where cryptographic hash function is applied to determine the similarity of data, once a duplicate copy found, then the new data is not considered but pointer to file ownership updated, proofing of ownership saves storage amount and resource utilization. When it comes the source side or client side process, data hash vales are computed at client side and send for duplicate check process. Whereas unauthorized users gain control to access hash values of a data for taking control on the. To prevent such kind of unauthorized users, a proof of ownership

considers user authentication to validate prover and verifier of ownership of file. Verifier calculates a short value of data D whereas, a prover has to compute short value of D and send it to verifier for asserting ownership of D

Till now so much of research work is going on in the field of de-duplication. After going through the literature I understand the importance of it and started working on the de-duplication process.

II. RELATED WORK

Wee Keong et.al [5] proposed first private data de-duplication protocols for private cloud for private data, this de-duplication protocol allows owners to store a data in a private cloud to formalize security technique on private data. This private de-duplication protocol is based on cryptography methods. Per them this de-duplication process is formalized the data on secured two party computational model. Based on their approach this protocol is more secured for private cloud data. They organized de-duplication process achieves these functions data correctness and soundness. This process needs more computational time to design secured framework using discrete algorithm and erasure coding algorithm

Another important approach designed by Pasquale Puzio et.al [6] is clouded which is based on additional encryption operation and access control mechanism for ensuring secured and efficient storage system. In addition, the clouded adopts key management system for data block level to organize de-duplication process. The major advantage this approach needs less computational cost and have less storage overhead. This approach mainly discussed with respective of data confidentially but it has limitation to support large amount of data storage process.

According to [7], the traditional encryption techniques are not feasible to organize de-duplication process in hybrid clouds, the main idea of the research is to achieve data confidentiality against outside adversaries by proposing cross user client side de-duplication on over client process. This scheme protects sensitive data against different data attacks. This scheme mainly designed for sensitive files or sensitive data objects. However, this schemes is good enough for sensitive files but it has limitations against data dictionary attacks.

In [8], the de-duplication in cloud storage organized using side channels in cloud services. According to Harnik et al.[8] most of the existing de-duplication schemes doesn't considered security against data leakage attacks. According to them the traditional security schemes combines the traditional key management techniques to invoke user level authentication to represent secured storage in clouds. The basic idea of this approach is to proposition of cross user de-duplication to obtain contents of files of other uses with the use of side channel. This scheme greatly reduces the data leakage by adopting cross user de-duplication. This scheme provides great advantage to mitigate data leakage process but it has limitation to ensure user data privacy

To improve data privacy in side channel process the JanStanek et al.[9] proposed an encryption scheme which ensures semantic security designed for unpopular data and presents weaker security for popular data with efficient storage and bandwidth. In this way, the de-duplication process organized for security for popular and unpopular data. To implement encryption scheme, they adopted tradition Diffie-Hellman. The basic drawback of this scheme to classify a data in terms of popular or unpopular is a tedious process and it produce more complex factor for large scale data process.

III. III. PROPOSED WORK:

This process will be implemented by the storage gateways. Whatever the data which is to be sent onto the cloud by the client must be checked for the duplicated entries by the storage gateways. This process has been done by the following steps.

1. Start process
2. Client wants to store the data onto the cloud and starts the process initiation to send it onto the cloud.
3. Upon receiving the data by the storage gateways, de-duplication process will be applied on the client data.

i. In the de-duplication process, which is applied by the storage gateways should be file level de-duplication process.

ii. By using the hash function, hash value will be calculated and the same will be used for

comparison for the existing hash value in the user cloud storage.

iii. If the hash value matches with the existing, then it is the duplicated value

Otherwise

i. It will be allowed to store on the cloud.

4. In this step Cloud Service Provider will implement the de-duplication process; in that de-duplication process we will be implementing the encryption and decryption process to provide the security to the hash values.
5. This de-duplication process should be block level and it should be of variable sized. With this variable size blocks will be achieving highest de-duplication ration when compared to the block sized de-duplication. But one complexity arises over here is calculation overhead. This can be overcome with the latest technologies i.e., with the hardware and software to increase the performance of the system.

i. Here the total cloud will be considered as the single unit and this will be divided into different blocks based on the user data.

ii. Dividing this into blocks will be based on the variable sized partitions

a. After applying the de-duplication process eliminate the duplicated data from the cloud storage.

6. To implement the security in the cloud structure, attribute based encryption and decryption is implemented. Here the hash value which is calculated while applying the de-duplication process will be encrypted and stored in the cloud server.
   a. With this encryption and decryption there will be storage overhead on the cloud server. To eliminate the storage overhead one process will be outsourced which is known as decryption.
   b. Decryption will be done at the client side by sharing the secret key with the client. This may reduce some processing overhead.
7. End process.

IV. AUTHENTICATION TECHNIQUE

The authentication process for each user begins with the generation of threshold signature. Let assume that there are n-number of users, who are willing to share their data into cloud, before accepting a data from the users, the cloud server verifies the user's signature to authenticate. To authenticate user, the cloud server computes authenticate based on user's registration information. Let assume a set of Member users X = {N1……N_{S2}) where N_{S2} represents identity of the ith (1<=I<=S1) member; c) a set of signer Y ={K1,……K_{S2}}where K is a subset of N and K_{S2} represents the identity of the jth (1<=I<=S2) member; d) a verifier v. The authentication technique is described as follows

- To generate a threshold signature for user data N, the following procedure describes threshold signature to authenticate a user to access data.
- **Step 1**: In the first step, user sends a request for threshold signature. This is started by one of the signers by sending a threshold generation request to cloud server along with user credential parameters such as (TK1… TK_{S2}).
- **Step 2**: In the next step the cloud server authenticator sends file token. For this the cloud server selects a random token $T_R \in Z *_q$ where 1<=R<=S_2 and forward to the corresponding signers by performing encryption operation
- **Step 3**: After that each user creates a signature: sig_{PKR}= $H_0(N).K_{PKR}$and calculates with a corresponding token: $T.sig_{PKR}$=$TA.sig_{PKR}$
- **Step 4**: After that start sending signature along with pseudonym public key. Here, each user sends the data to cloud server for future verification.

(N, Q_{PKR}, Q_{PK1}, T*sig_{PKR}, $HMAC_{PKR}$ (M,Q_{PKV},$T.sig_{PKA}$))

V. CONCLUSION AND FUTURE SCOPE

After careful analysis of the literature survey what I had on the de-duplication process, I understood the implementation of de-duplication process on the cloud. Theory which is surveyed till now does not have the concept of hybrid approach which is proposed for the implementation of de-duplication on the hybrid clouds. Another area which is concentrated in this work is security which is proposed by the use of ABE technique. Even in the implementation also the computation overhead is reduced in outsourcing the decryption process to the client. It make the proposed approach a fair and clean implementation which will be achieving the higher performance when compared to the existing techniques. In this report, whatever presented is written after the clear analysis of the

existing literature and the implementation of proposed hybrid approach for the de-duplication process is not done. So, soon it will be implemented and the results should be compared with the existing techniques such that the performance of this approach should be better than the existing approaches, if not based on the simulation results, some change can be implemented in the proposed technique to achieve the highest accuracy ratio.

VI. REFERENCES

[1]. AlexaHuth and James Cebula, 2011, Carnegie Mellon University. Produced or US CERT, a government organization. The Basics of Cloud Computing. Retrieved from https://www.us-cert.gov/sites/default/files/publications/CloudComputingHuthCebula.pdf

[2]. RajkumarBuyya, CheeShinYeo, SrikumarVenugopal, James Broberg, IvonaBrandic, 2008, Future Generation Computer Systems. Retrieved from http://www.buyya.com/papers/Cloud-FGCS2009.pdf

[3]. Saurabh Kumar Garg and RajkumarBuyya, Green Cloud computing and Environmental Sustainability. Retrieved from http://www.cloudbus.org/papers/Cloud-EnvSustainability2011.pdf

[4]. Naylor G. Bachiega, Henrique P. Martins, Roberta Spolon, Marcos A. Cavenaghi, Renata S. Lobato, AleardoManacero. Open Source Cloud Computing: Characteristics and an Overview. Retrieved from http://worldcomp-proceedings.com/proc/p2013/PDP3537.pdf

[5]. Wee Keong Ng, Yonggang Wen and Huafei Zhu. (2012). Private Data De-duplication Protocols in Cloud Storage. ACM, 9781450308571/ 12/03

[6]. Pasquale Puzio, RefikMolva, Melekonen, and Sergio Loureiro. (2014). Block-level De-duplication with Encrypted Data. Open Journal of Cloud Computing (OJCC) Volume 1, Issue 1, 10-18, ISSN 2199-1987

[7]. JiaXu, Ee-Chien Chang, and Jianying Zhou. (2013). Weak Leakage-Resilient Client-side De-duplication of Encrypted Data in Cloud Storage ∗. ACM 978-1-4503-1767-2/13/05. Retrieved from https://www.comp.nus.edu.sg/~changec/publications/2013_asiaccs.pdf

[8]. Danny Harnik, Alexandra Shulman-Peleg, and Benny Pinkas. (2013), Retrieved from http://www.pinkas.net/PAPERS/hps.pdf

[9]. Jan Stanek, Alessandro Sorniottiy, Elli Androulakiy, and Lukas Kencl. (2014). A Secure Data De-duplication Scheme for Cloud Storage. IBM Research, Retrieved from fc14.ifca.ai/papers/fc14_submission_5.pdf

[10]. Sun, Z., Shen, J., & Yong, J. (2011). DeDu: Building a de-duplication storage system over cloud computing. Computer Supported Cooperative Work in Design (CSCWD), 2011 15th International Conference on, 348-355.