De-Duplication of Data in Cloud Storage

Jessia.E.Joby
Student, Department Of Information Technology
Velammal Engineering College, Chennai.
R.Manju
Student, Department Of Information Technology
Velammal Engineering College, Chennai.
J.Sathya Priya
Assistant Professor, Department Of Information Technology
Velammal Engineering College, Chennai.

----------**ABSTRACT**----------

Data DE-Duplication is a technique forcompressing the redundant data which is widely used in cloud storage. Convergent encryption technique has been proposed to encrypt the data before outsourcing. To provide a better security these papers address the problem of authorized data de-duplication. Different from traditional DE-duplication systems, the privileges of users are further considered in duplicate check besides the data itself. We also present several new DE-duplication constructions supporting authorized duplicate check in public cloud architecture. Security analysis demonstrates that our schemes are secure in terms of the definitions specified in the proposed security model. We implement a prototype of our proposed authorized duplicate check scheme and conduct test bed experiments using our prototype. Our proposed work of authorized duplication checking enables to achieve less overhead.

Keywords- **Cloud Computing**

I. INTRODUCTION

Cloud Computing has attracted a lot of attentions in recent times. The media as well as analysts are generally very positive about the opportunities that Cloud Computing is offering. In May 2008, Merrill Lynch (2008) estimated the cost of Cloud Computing to be three to five times for business applications and more than five times for consumer applications. According to the Gartner press release from June 2008, Cloud Computing will be the “no less influential than e-business” (Gartner 2008a). The positive attitude towards the influence of Cloud Computing resulted in optimistic Cloud-related market forecasts. In October 2008, IDC (2008b) forecasted that almost three fold growth of spending on Cloud services until 2012, reaching $42 billion. Same analyst firm reported that the cost advantages associated with the Cloud model becomes even more attractive in the economic downturn (IDC 2008b). Positive market prospects are also driven by the expectation that Cloud Computing might become the fundamental approach towards the Green IT. Despite of broad coverage of Cloud Computing in commercial press, there is still no common agreement on what exactly Cloud Computing is and how it relates to Grid Computing. To gain an understanding of what Cloud Computing is, we first look at several existing definitions of the terms. Based on those definitions, we identify key characteristics of Cloud Computing. Then we describe the common architecture and components of Clouds in detail, discuss opportunities and challenges of Cloud Computing, and provide a classification of Clouds. Finally, we make a comparison between the Grid Computing and Cloud Computing.

II. EXISTING SYSTEM

Many data de-duplication mechanisms have been proposed for efficient data de-duplication in order to safe storage space.[11] Current issue for data de-duplication is to avoid full-chunk indexing to identify the incoming data which is new, which is time consuming process. [10] Some data chunks may be read frequently in period of time, but may not be used in another time period. Some datasets may be frequently accessed or updated by multiple users at the same time, while others may need the high level of redundancy for reliability requirement.

III. PROPOSED SYSTEM

The convergent encryption techniques have been proposed to encrypt the data before outsourcing. For better protect data security, this paper makes the first attempt to formally address the problem ofauthorized data de-duplication.

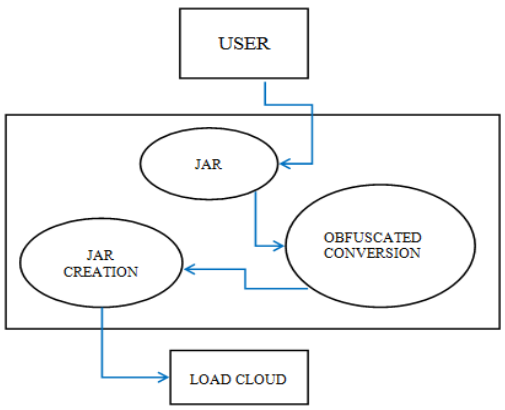

Fig.1- Workflow

Different from traditional de-duplication systems, the differential privileges of users are further considered in duplicate check besides the data itself. We also present several new de-duplication constructions supporting authorized duplicate check in hybrid cloud architecture. Security analysis demonstrates that our scheme is secure in terms of the definitions specified in the proposed model. As a proof of concept, we implemented the prototype of our proposed authorized duplicate check scheme and conducted test bed experiments using the prototype. We show that our proposed authorized duplicate check scheme shows minimal overhead compared to normal operation.

IV. ALGORITHMS

The algorithms used in our system are AES and MD5 which is used for encryption, decryption and to reduce the size of a data.

4.1. AES Algorithm

To protect the client's privacy, we apply the anonymous AES in branching programs. To reduce the decryption complexity due to the use of AES, we apply recently proposed decryption outsourcing with privacy protection to shift

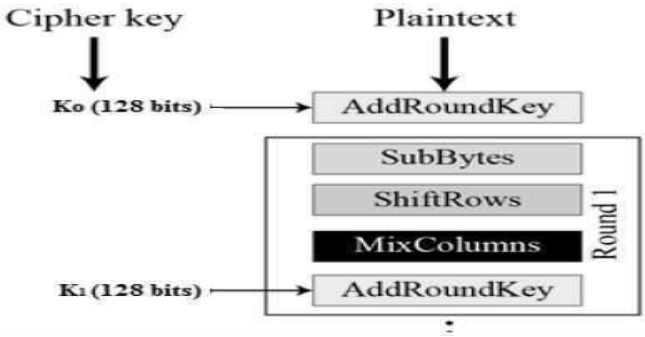

Fig.2- AES Operation.

client's pairing computation to the cloud server. The adversary launches Key Generate algorithms to query for as many private keys as he wants, which correspond to attribute setsA1, . . . ,Aq being disjoint in charged by all authorities {Ak }, but none of these keys satisfy.Besides, he also conducts arbitrarily many computations using the public and secret keys that he has (belonging to compromised authorities).

4.2. MD5 Algorithm

The MD5 message-digest algorithm uses the hashfunction for producing a 128-bit value. One basic requirement of any cryptographic hash function is that it should be computationally impossible to find two distinct messages which hash to the same value.

V. MODULES

The process of DE duplication will take place under the four modules.

- User Interface.
- de duplication in cloud.
- duplicate hash value check.
- to access files.

5.1. User Interface

In the User Module we can register and login the account in a secured way. Users have authentication and security to access the detail which is presented in the ontology system. Before accessing or searching the details the user should have the account in that otherwise they should register first.

5.2. De Duplication In Cloud

To support authorized de-duplication, the tag of a file F will be determined by the file F and the privilege. To show the difference with traditional notation of tag, instead we call it file token. To support authorized access, a secret key kp will bounded with a privilege p to generate a file token. Let $\phi' F;p$ = Tag Gen(F, kp) denote the token of F that is only allowed to access by user with privilege p. In another word, the token $\phi' F;p$ could only be computed by the user with privilege. As a result, if a file has been uploaded by an user with a duplicate token $\phi' F;p$, then a duplicate check sent from another user will be successful if only he also has the file F and privilege p. Such a token generation function could be easily implemented as H(F, kp), where H(_) denotes a cryptographic hash function.

5.3. Duplicate Hash Value Check

We consider several types of privacy that we need protect, they are i) enforceability of duplicate-check token: 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 privileges. If a user has privilege p, it requires that the adversary cannot forge and produce output as valid duplicate token with anyother privilege p' on any file F, where p does not match p'. Furthermore, it also requires that if theadversary does not make a request of token with its own privilege from private cloud server, it cannot forge and output as valid duplicate token with p on any F that can been queried.

5.4. To Access Files

Once the key request was received, the sender will send the key or he can ignore it. With the key and request id which was produced at the time of sending key request, the receiver can decrypt the message.

VI. ADVANTAGE

- The challenge of cloud storage is the management of huge volume of data.
- The user is only allowed to perform the duplicate check for files marked with the corresponding privileges.
- We present an advanced scheme to support stronger security by encrypting the file with differential privilege keys.
- It provides 50% faster backups.
- By using this technique 20-40% of impact on the server will get reduced.
- Impact on the network will also get reduced for about 80-90%.
- Reduce the storage size of the tags for integrity check. To enhance the security of de-duplication and to protect the data confidentiality.

VII. CONCLUSION AND FUTURE ENHANCEMENTS

The de-duplication on encrypted files along with preserving confidentiality and security is highly demanded for organizations and even individuals when storing the files under public third-party cloud storage providers. In this paper, we have integrated erasure correcting code technique with de-duplication system. By utilizing the verification of erasure coded data, our scheme achieves error correction and redundancy. Data corruption has been detected and it can be re-generated during the file retrieval stage even if any of the distributed servers has been attacked. The proposed methodology works better with any files like text, image or video, but it needs to test for huge file system. The system can improve much better if the reed Solomon code for encoding & decoding is extended for huge files of over 1000MB.

The Google might not release enough white documents for Google Sites during our development period. However, the proposed service can still efficiently backup specific web site that is constructed by Google Sites and let users save their cloud web pages in their local storage. The related and extensive services will be popular for users who care their own cloud data. Hence, it is necessary for cloud vendors to harmonize a series of commercial standards for further engineering utilization. Some functions are not implemented in the CBS v1.0. However, the extensive functions, such as uploading to generic web server, graphical user interface, and comparison of similarity, will be appended in the future version.

REFERENCES

[1] B. Zhu, K. Li, and R. H. Patterson, “Avoiding the disk bottleneck in the data domain deduplication file system.” In6thUSENIX Conference on File and Storage Technologies (FAST 08), 2008, pp. 1–14.

[2] K. Srinivasan, T. Bisson, G. R. Goodson, and K. Voruganti, “iD-edup: latency-aware, inline data deduplication for primary storage.” in 11th USENIX Conference on File and Storage Technologies(FAST’12), 2012, pp. 1–14.

[3] B. Mao, H. Jiang, S. Wu, andL.Tian , “Pod:Performance oriented i/o de-duplication for primary storage systems in the cloud,” in Parallel and DistributedProcessing Symposium, 2014 IEEE 28thInternational.IEEE, 2014, pp. 767–776.

[4] J.An and D. Shin, “Offline de-duplication aware block separation for solid state disk,” in 11th USENIX Conference on File and Storage Technologies (FASR’ 13),2013.

[5] A. Wildami E. L. Miller and O. Rodeh, “Hands: A heuristically arranged non-backup in-line deduplication system,” in Data Engi-neering (ICDE), 2013 IEEE 29th International Conference on IEEE,2013,pp. 446-457.

[6] C. Constantinescu, J. Glider, and D. Chambliss “Mixing de-duplication and compression on active data sets,” in Data Compression Conference (DCC), 2011. IEEE, 2011, pp. 393–402.

[7] El-Shimi, R. Kalach, A. Kumar, A. Ottean, J. Li, and S.Sengupta, “Primary data deduplicationlarge scale study and system design,” in 2012 USENIX Annual TechnicalConference (ATC’12), 2012, pp. 285–296.

[8] Tin TheinThwel, Ni LarThein, “An Efficient Indexing Mechanism for Data Deduplication ,“,2009 InternationalConference on the Current Trends in Information Technology(CTIT),pp.22-2.