International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2013): 4.438

Security Analysis and Deduplication Using Convergent Algorithm

Rajkumar N Pagadi[1], Dr. Shubhangi D C[2]

[1]P.G.Student, Department of Computer Science and Engineering, VTU RO Post Graduate Centre, Kalaburagi, Karnataka (India)

[2]Professor & Course Co-ordinator., Department of Computer Science and Engineering, VTU RO Post Graduate Centre, Kalaburagi, Karnataka (India)

Abstract: ***Data compression techniques such as data deduplication is used for removing duplicate copies of repeating data, it has been extensively used in cloud storage so as to reduce storage space and save bandwidth. To preserve confidentiality of sensitive data while supporting deduplication. For that purpose the convergent encryption technique has been proposed so as to encrypt the data before outsourcing. To better protect data security, the aim is to formally address the problem of authorized data deduplication. We also consider the differential privileges of users are considered in duplicate check besides the data. To support authorized duplicate check in a hybrid cloud architecture we have several new deduplication constructions. In order to demonstrate that the scheme is secure in terms of the definitions specified in the proposed security model security analysis is performed. we implement a prototype of our proposed authorized duplicate check scheme and conduct testbed experiments using our prototype. The proposed authorized duplicate check scheme sustains minimal overhead when compared to normal operations.***

Keywords: Deduplication, Convergent encryption key management, Duplicate check, proof of ownership.

1. Introduction

Cloud Computing, the long-held dream of computing as a utility, has the potential to transform a large part of the IT industry, making software even more attractive as a service and shaping the way IT hardware is designed and purchased. Developers with innovative ideas for Internet services no longer require the large capital outlays in hardware to deploy their service or the human expense to operate it. Cloud computing consists of hardware and software resources made available on the Internet as managed third-party services. Data deduplication is a technique for removing duplicate copies of storage data. Identical expressing challenge of today's cloud storage services is the management of the increasing volume of data. Although Data Deduplication brings a lot of benefits, as the user's sensitive data are susceptible to both insider and outsider attacks this results in security and privacy concerns. To make data management scalable in cloud computing, deduplication has been a well-known technique and has gained more attention recently. Instead of keeping multiple data copies with the same content, deduplication keeps only one physical copy in order to eliminate redundant data and referring other redundant data to that copy. The proof of ownership protocol is needed to provide proof that the user actually has the same file when a duplicate is found and to prevent unauthorized access In order to save cost and efficiently management, the data will be moved to the storage –cloud service provider (SCSP) in the public cloud with specified privileges and the deduplication technique will be applied to store only one copy of the same file. For privacy consideration, some files will be encrypted and allowed the duplicate check by employees with specified privileges to realize the access control.

2. Literature Survey

Literature survey is the most important step in software development process. Before improving the tools it is compulsory to decide the economy strength, time factor. Once the programmer‘s create the structure tools as programmer require a lot of external support, this type of support can be done by senior programmers, from websites or from books.

P. Anderson and L. Zhang "Encrypted de-duplication with fast and secure laptop backups". Now people store huge amount of personal and corporate data on laptops and on home computers. These usually have poor or discontinuous connectivity, and are vulnerable to theft and hardware failure. since conventional backup solutions are not convenient to this context, and also the backup regimes are often restricted. This paper describes the algorithm which takes advantage of the data which is common between users in order to increase the speed of backups, and also reduce the storage requirements and for confidential personal data the algorithm supports client-end per-user encryption.

S. Keelveedhi, M. Bellare, and T. Ristenpart, proposed a method "Message-locked encryption and secure deduplication". A new cryptographic primitive, Message-Locked Encryption (MLE), where encryption and decryption are performed under the key is which is obtained from the message. MLE provides a way to attain secure deduplication (space-efficient secure outsourced storage). We provide definitions both for privacy and for a form of integrity that we call tag consistency. Based on this base, we make the practical and theoretical contributions. On the practical side, we provide the security analyses of ROM for a natural family of MLE schemes that contains deployed schemes. On the theoretical side the challenge are the standard model

solutions and we aim to deliver schemes under different assumptions and for classes of message sources.

M. Bellare, C. Namprempre, and G. Neven Security proofs for identity-based identification and signature schemes. This paper provides either security proofs or attacks for a large number of identity-based identification and signature schemes defined either explicitly or implicitly in existing literature. Here is a framework that helps to explain how the schemes are derived and also permit modular security analyses, which helps to understand, simplify, and unify previous work. We study the generic folklore construction that provides identity-based identification and signature schemes without random oracles.

C. Ng and P. Lee. Rev de dup "A reverse deduplication storage system optimized for reads to latest backups **".**Deduplication effectively eliminate duplicates, yet degrades read performance since it introduces fragmentation. We propose Rev De dup, a deduplication system that optimizes reads to the latest backups of virtual machine (VM) images using reverse deduplication. Rev De dup removes duplicates which are in the old data, thereby shifting fragmentation to old data while keeping the layout of new data as sequential as possible.

3. System Architecture

In this new deduplication system, hybrid cloud architecture is introduced to solve the problem. The private keys for privileges will not be issued to users directly, which will be kept and managed by the private cloud server instead. In this way, the users cannot share these private keys of privileges in this proposed construction,

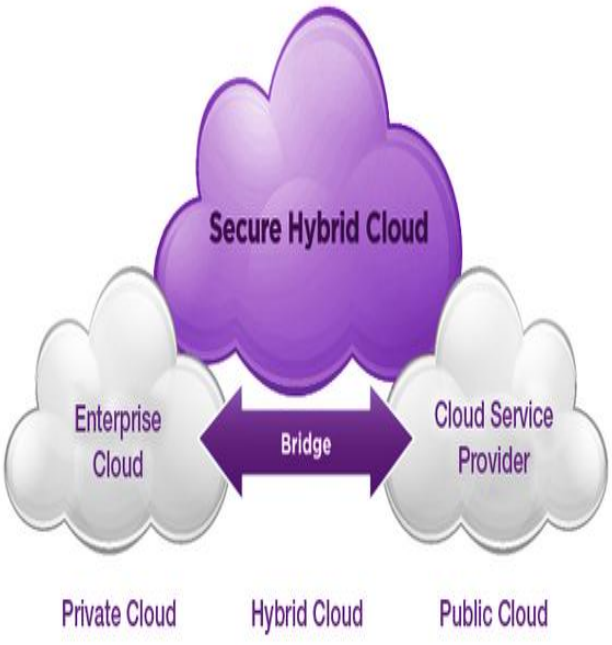


Figure1: Architecture for secure hybrid cloud

which means that it can server. The intuition of this construction can be described as follows. To perform the duplicate check for some file, Token to the user. The authorized duplicate check for this file can be performed by the user with the public cloud before uploading this file. Based on the results of duplicate check, the user either uploads this file or runs pow.

4. Methodology

In aiming at efficiently solving the problem of deduplication with differential privileges in cloud computing, we consider a hybrid cloud architecture consisting of a public cloud and a private cloud. Unlike existing data deduplication systems, the private cloud is involved as a proxy to allow data owner/users to securely perform duplicate check with differential privileges. Such architecture is practical and has attracted much attention from researchers.

The data owners only outsource their data storage by utilizing public cloud while the data operation is managed in private cloud. A new deduplication system supporting differential duplicate check is proposed under this hybrid cloud architecture where the SCSP resides in the public cloud.

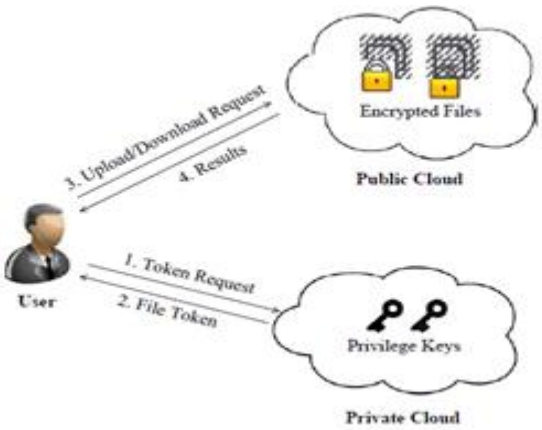


Figure 2: Architecture for secure data Deduplication

The user is only allowed to perform the duplicate check for files marked with the corresponding privileges. Furthermore, we enhance our system in security. Specifically, we present an advanced scheme to support stronger security by encrypting the file with differential privilege keys. In this way, the users without corresponding privileges cannot perform the duplicate check. Furthermore, such unauthorized users cannot decrypt the cipher text even collude with the S-CSP. Security analysis demonstrates that our system is secure in terms of the definitions specified in the proposed security model. Finally, we implement a prototype of the proposed authorized duplicate check and conduct testbed experiments to evaluate the overhead of the prototype. We show that the overhead is minimal compared to the normal convergent encryption and file upload operations.

Algorithm

Symmetric encryption: - Symmetric encryption uses common secret key κ to encrypt and decrypt information. A symmetric encryption scheme consists of three primitive functions.

Step 1: $KeyGen_{SE}$ $(1^{\lambda}) \rightarrow k$ is the key generation algorithm that generates k using security parameter 1^{λ}.

Step 2: Enc_{SE} (k, M) $\rightarrow$ C is the symmetric encryption algorithm that takes the secret k, and message M and then outputs the cipher text C.

Step 3: Dec_{SE} (k, C) $\rightarrow$ *M* is the symmetric decryption algorithm that takes the secret *k* and cipher text *C* and then outputs the original message *M*.

Convergent encryption: - Convergent encryption provides data confidentiality in deduplication. A user or data owner derives a convergent key from each original data copy and encrypts the data copy with the convergent key. In addition, the user also derives a *tag* for the data copy, such that the tag will be used to detect duplicates. Here, we assume that the tag correctness property holds, i.e., if two data copies are the same, then their tags are the same. To detect duplicates, the user first sends the tag to the server side to check if the identical copy has been already stored. Note that both the convergent key and the tag are independently derived, and the tag cannot be used to deduce the convergent key and compromise data confidentiality. Both the encrypted data copy and its corresponding tag will be stored on the server side. Formally, a convergent encryption scheme can be defined with four primitive functions.

Algorithm

Step 1: $KeyGen_{CE}(M)$ -- *K* is the key generation algorithm that maps a data copy *M* to a convergent key *K*;

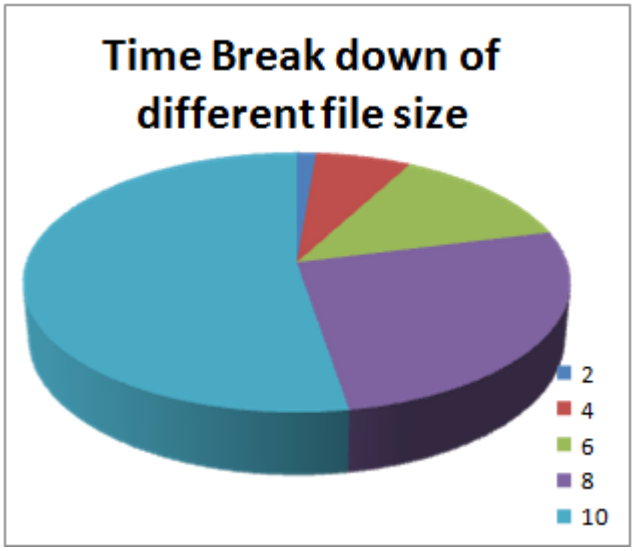
Step 2: $Enc_{CE}(K,M)$ -- *C* is the symmetric encryption algorithm that takes both the convergent key *K* and the data copy *M* as inputs and then outputs a cipher text *C*;

Step 3: Dec_{CE} (*K*,*C*) -- *M* is the decryption algorithm that takes both the cipher text *C* and the convergent key *K* as inputs and then outputs the original data copy*M*; and

*Step 4:*TagGen(*M*) -- *T*(*M*) is the tag generation algorithm that maps the original data copy *M* and outputs a tag *T*(*M*).

5. Results and Discussion

We measured the increase in transmission bandwidth due to deduplication during storage. To do so, we used TCP dump and filtered out all traffic unrelated to Drop box and deduplication. We took from this the total number of bytes (in either direction). For even very small files, the Drop box API incurs a cost of about 7 KB per upload. The ratio of bandwidth used by deduplication to that used by plain Drop box as file size increases. Given the small constant size of the extra file sent by deduplication, overhead quickly diminishes as files get larger.



We measure the baseline performance of deduplication using unique data. The server initially contains no data. The client processes submit 128GB of unique data (i.e., all blocks are globally unique) to the server. Then a client process retrieves the data using the Linux command we get. We also measure the raw disk throughput by reading/writing data directly via the native file system of our testbed. The write throughput of deduplication is 13-19% less than the raw write throughput. When the segment size is larger, fewer segments are involved and deduplication has higher unique write throughput. On the other hand, the read throughput of deduplication is very close to the raw read throughput.

Data confidentiality – Data confidentiality is a very important concept where data needs to be secured from the unauthorized users. This can be done by providing the secured authorization scheme to each and every user when they register for the data ownership.

6. Conclusion and Future Scope

In order to demonstrate that the proposed method is very best technique to reduce the duplicate information on the cloud we have tested the application and found out that the proposed application is best. we implement a prototype of our proposed authorized duplicate check scheme and conduct testbed experiments using our prototype. The proposed authorized duplicate check scheme sustains minimal overhead when compared to normal operations.

In future work we can extend this to include image processing files. Where similar files can be deduplicated and also to include audio and video deduplication.

7. Acknowledgment

The authors would like to thank a great support of VTU University Belagavi and VTU Regional Office Postgraduate Centre, Kalaburagi, Karnataka, India. Under taking this work successfully.

References

[1] P. Anderson and L. Zhang. Fast and secure laptop backups with encrypted de-duplication.
[2] M. Bellare, S. Keelveedhi, and T. Ristenpart. Dupless : Server aided encryption for deduplicated storage. In *USENIX Security Symposium*, 2013.
[3] Java The complete Reference.
[4] My sql Reference Books.