Int. J. Chem. Sci.: 14(4), 2016, 2933-2938
ISSN 0972-768X
www.sadgurupublications.com

DE-DUPLICATION OF DATA IN CLOUD

R. SHOBANA*, K. SHANTHA SHALINI, S. LEELAVATHY and V. SRIDEVI

Department of Computer Applications, Aarupadai Veedu Institute of Technology, Vinayaka Missions University, CHENNAI (T.N.) INDIA

ABSTRACT

Rendering efficient storage and security for all data is very important for cloud computing. Securing and privacy preserving of data is of high priority when it comes to cloud storage. Therefore to provide efficient storage for cloud data owners and render high security for data this paper proposes Cloud Computing Secure Framework (CCSF). Thus CCSF consists of four segments: 1) Identity Management 2) Intrusion detection and prevention system 3) Data deduplication 4) Secure Cloud Storage. Intrusion detection and prevention are performed manually by network operators in the existing system. In our proposed architecture the intrusion detection and prevention is performed automatically by defining rules for the major attacks and alert the system automatically. The major attacks/events includes vulnerabilities, cross site scripting (XSS), SQL injection, cookie poisoning, wrapping. Data deduplication technique allows the cloud users to manage their cloud storage space effectively by avoiding storage of repeated data's and save bandwidth. The data are finally stored in cloud server namely CloudMe. To ensure data confidentiality the data are stored in an encrypted type using Advanced Encryption Standard (AES) algorithm.

Key words: De-duplication, Cloud computing.

INTRODUCTION

Cloud computing is one of the emerging technology, which helped several organizations to save money and time adding convenience to the end users. Thus the scope of cloud storage is vast because the organizations can virtually store their data's without bothering the entire mechanism. Cloud Computing provides key advantage to the end users like cost savings, Able to access the data irrespective of location, performance and security.

In our proposed system we invoke a effective user authentication using fingerprint feature extraction, image based authentication during file upload/download, eliminating repetition of data in cloud server and implemented through multiple cloud storage.

*Author for correspondence; E-mail: shobanasenthil29@gmail.com, shanthashalini@gmail.com

Most of the existing authentication system has certain drawbacks, for that reason graphical passwords are most preferable authentication system where users click on images to authenticate themselves. Our proposed system states image based effective authentication. When the Admin uploads the file in the cloud, the admin will split the image into 4 parts. The admin will hold 2 parts and the user of that respective group can view the other 2 parts. The images are spilt randomly using pseudo random generator technique. When the user tries to download a file, the user can send the requisition to the respective admin along with the user side available 2 parts. The admin will verify both the parts and if the authentication is passed, the file will be sent to the user in an encrypted way.

Data deduplication is one of the techniques which used to solve the repetition of data. The deduplication techniques are generally used in the cloud server for reducing the space of the server. To prevent the unauthorized use of data accessing and create duplicate data on cloud the encryption technique to encrypt the data before stored on cloud server. Cloud Storage usually contains business-critical data and processes; hence high security is the only solution to retain strong trust relationship between the cloud users and cloud service providers. Thus to overcome the security threats, this paper proposes multiple cloud storage. Thus the common forms of data storage such as files and databases of a specific user is split and stored in the various cloud storages (e.g. Cloud A and Cloud B).

EXPERIMENTAL

Literature review

Data deduplication in cloud computing systems

Cloud computing is a paradigm shift in the Internet technology. Data deduplication can save storage space and reduce the amount of bandwidth of data transfer.

Secure and constant cost public cloud storage auditing with deduplication

Deduplication system in the cloud storage is used to reduce the storage size of the tags for integrity check.

Fingerprint verification based on minutiae features: a review

The fingerprint feature extraction and matching is performed using Minutiae Map algorithm (MM). Minutiae is the reference to bifurcation and termination values of the ridges in the fingerprint. The distribution on the fingerprint provides a unique signature for each and every individual.

Methodology

- Detect deduplication
- File encryption and file uploading
- Multiple cloud storage
- File exchange and file retrieving

Functions of data deduplication

It compares objects (usually files or blocks) and removes objects (copies) that already exist in the data set. The deduplication process removes blocks that are not unique.

1. Divide the input data into blocks or “chunks.”
2. Calculate a hash value for each block of data.
3. Use these values to determine if another block of the same data has already been stored.
4. Replace the duplicate data with a reference to the object already in the database.

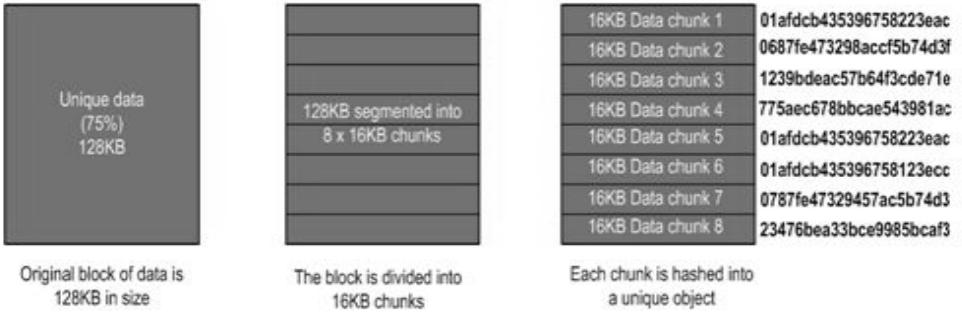


Fig. 1: working mode of data deduplication

Once the data is chunked, an index can be created from the results, and the duplicates can be found and eliminated.

Only single instance of data is stored

The actual process of data deduplication can be implemented in a number of different ways. We can eliminate duplicate data by simply comparing two files and making the decision to delete one that is older or no longer needed.

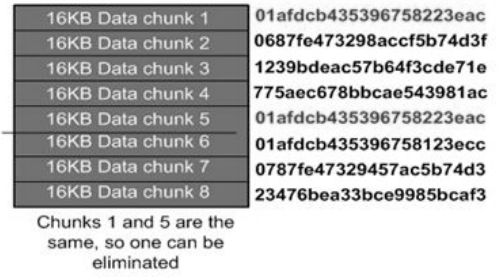


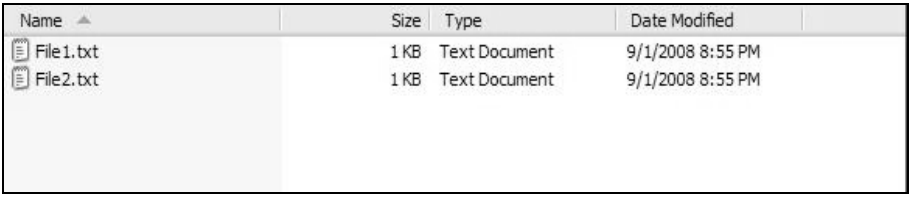
Fig. 2: Data elimination process

The most common methods of implementing deduplication are:

- File-based compare
- File-based versioning
- File-based hashing
- Block or sub-block versioning
- Block or sub-block hashing

File-based compare

File system-based deduplication is a simple method to reduce duplicate data at the file level, and usually is just a compare operation within the file system or a file system-based algorithm that eliminates duplicates. An example of this method is comparing the name, size, type and date-modified information of two files with the same name being stored in a system. If these parameters match, you can be pretty sure that the files are copies of each other and that you can delete one of them with no problems. Although this example isn't a foolproof method of proper data deduplication, it can be done with any operating system and can be scripted to automate the process, and best of all, it's free. Based on a typical enterprise environment running the usual applications, you could probably squeeze out between 10 percent to 20 percent better storage utilization by just getting rid of duplicate files.



Example: File 1. txt and File 2. txt are the same size and have the same creation time. Most likely, one is a duplicate

File-based delta versioning and hashing

More intelligent file-level deduplication methods actually look inside individual files and compare differences within the files themselves, or compare updates to a file and then just store the differences as a "delta" to the original file. File versioning associates updates to a file and just stores the deltas as other versions. File-based hashing actually creates a unique mathematical "hash" representation of files, and then compares hashes for new files to the original. If there is a hash match, you can guarantee the files are the same, and one can be removed.

A lot of backup applications have the versioning capability, and you may have heard it called incremental or differential backup. Some backup software options always use the versioning method to speed backup. Other software solutions use similar techniques to reduce wide area network (WAN) requirements for centralized backup. Intelligent software agents running on the client use file-level versioning or hashing at the client to send only delta differences to a central site. Some solutions actually send all data updates to the central site and then hash the data once it arrives, storing only the unique data elements.

Most products that use a "hashing" mechanism also require an index to store the hashes so that they can be looked up quickly to compare against new hashes to see if the new data is unique (i.e., not already stored), or there is a hash match and the new data element does not need to be stored. These indexes must be very fast or handled in such a manner that the unique data stored increases and becomes fragmented so that the solution doesn't slow down during the hash lookup and compare process.

Different solutions from various vendors use diverse hashing algorithms, but the process is basically the same. The term "hashing the data" means "creating a mathematical representation of a specific dataset that can be statistically guaranteed to be unique from any other dataset." The way this is done is to use a generally understood and approved method to encrypt each dataset, so that the metadata or resulting mathematical encryption "hash" can be used to either reproduce the original data or as a lookup within the index to see if any new data hashes compare to any stored data hashes, so the new data can be ignored.

CONCLUSION

Thus this paper compresses the data by removing the duplicate copies of identical data and it is extensively used in cloud storage to save bandwidth and minimize the storage space. To secure the confidentiality of sensitive data during deduplication, the convergent encryption technique is used to encrypt the data before outsourcing. For better data protection, this paper talks about the issue of data deduplication authorization.

REFERENCES

1. Open SSL Project (1998), www.openssl.org.
2. P. Anderson and L. Zhang, Fast and Secure Laptop Backups with Encrypted De-Duplication, in Proc. 24th Int. Conf. Large Installation Syst. Admin., 29-40 (2010).
3. M. Bellare, S. Keelveedhi and T. Ristenpart, Dupless: Serveraidedencryption for Deduplicated Storage, in Proc. 22nd USENIX Conf. Sec. Symp., 179-194 (2013).
4. H. Wang, Identity-Based Distributed Provable Data Possession in Multicloud Storage, IEEE, **8(2)**, 328-340 (2015).
5. J. M. Bohli, N. Gruschka, M. Jensen, L. L. Iacono and N. Marnau, Security and Privacy-Enhancing Multicloud Architectures, **10(4)**, 212-224 (2013).
6. N. Yager and A. Amin, Fingerprint Verification Based on Minutiae Features: A Review, Pattern Anal. Appl., **7**, 94-113 (2004); Feb 14, **7(1)**, pp. 94-113.
7. A Study on Authorized Deduplication Techniques in Cloud Computing, Int. J. Adv. Res. Computer Engg. Technol. (IJARCET), **3(12)** (2014).
8. Data Deduplication in Cloud Computing Systems, International Workshop on Cloud Computing and Information Security (CCIS) (2013).
9. Cloud Computing Security: From Single to Multi-Clouds using Digital Signature, Int. J. Engg. Technol., Manage. Appl. Sci. www.ijetmas.com, **2(6)**, (2014).

Accepted : 31.10.2016