ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 6, June 2017

A Survey Paper on Deduplication on Encrypted Big Data Using HDFS Framework

Sabale Nikita C[1], Prof.N.G.Pardeshi[2]

PG Student, SRES'COE, Kopargaon, SPPU, Maharashtra, India[1]

Associate Professor, SRES'COE, Kopargaon, SPPU, Maharashtra, India[2]

ABSTRACT: The biggest challenge for big data from a security point of view is the protection of user's privacy. Big data frequently contains huge amounts of personal identifiable information and therefore privacy of users is a huge concern. However, encrypted data introduce new challenges for cloud data deduplication, which becomes crucial for big data storage and processing in cloud. Traditional deduplication schemes cannot work on encrypted data. Existing solutions of encrypted data deduplication suffer from security weakness. They cannot flexibly support data access control and revocation. Therefore, few of them can be readily deployed in practice. In this paper, we propose a scheme to deduplicated encrypted data stored in cloud based ownership challenge and proxy re-encryption. It integrates cloud data deduplication with access control. We evaluate its performance based on extensive analysis and computer simulations. The results show the superior efficiency and effectiveness of the scheme for potential practical deployment, especially for big data deduplication in cloud storage.

KEYWORDS: Big data, Cloud computing, Data de-duplication,Access control, Proxy re-encryption, Data owner management

I. INTRODUCTION

Ouraimed 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 or blocks) only once. Therefore, if a user wants to upload a file (block) which is already stored, the cloud provider will add the user to the owner list of that file (block). Deduplication has proved to achieve high space and cost savings and many Big Data storage providers are currently adopting it. Deduplication can reduce storage needs by up to 90-95% for backup applications and up to 68% in standard file systems.

Cloud computing provides seemingly unlimited "virtualized" resources to users as services across the whole Internet, while hiding platform and implementation details. Today's cloud service providers offer both highly available storage and massively parallel computing resources at relatively low costs.As cloud computing becomes prevalent, an increasing amount of data is being stored in the cloud and shared by users with specified privileges, which define the access rights of the stored data. One critical challenge of cloud storage services is the management of the ever increasing volume of data.To make data management scalable in cloud computing, de-duplication has been a well-known technique and has attracted more and more attention recently.

Data de-duplication is a specialized data compression technique for eliminating duplicate copies of repeating data in storage. The technique is used to improve storage utilization and can also be applied to network data transfers to reduce the number of bytes that must be sent. Instead of keeping multiple data copies with the same content, de-duplication eliminates redundant data by keeping only one physical copy and referring other redundant data to that copy. De-duplication can take place at either the file level or the block level. For file level de-duplication, it eliminates duplicate copies of the same file. De-duplication can also take place at the block level, which eliminates duplicate blocks of data that occur in non-identical files.Cloud computing is an emerging service model that provides computation and storage resources on the Internet. One attractive functionalitythat cloud computing can offer is cloud storage. Individuals and enterprises are often required to remotely archive their data to avoid any information loss in case there are any hardware/software failures or unforeseen disasters.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 6, June 2017

Instead of purchasing the needed storage media to keep data backups, individuals and enterprises can simply outsource their data backup services to the cloud service providers, which provide the necessary storage resources to host the data backups. While cloud storage is attractive, how to provide security guarantees for outsourced data becomes a rising concern. One major security challenge is to provide the property of assured deletion, i.e., data files are permanently inaccessible upon requests of deletion. Keeping data backups permanently is undesirable,assensitive information may be exposed in the future because of data breach or erroneous management of cloud operators. Thus, to avoid liabilities, enterprises and government agencies usually keep their backups for a finite number of years and request to delete (or destroy) the backups afterwards. For example, the US Congress is formulating the Internet Data Retention legislation in asking ISPs to retain data for two years, while in United Kingdom; companies are required to retain wages and salary records for six years.

Scope:
The System can enhance the security by encryption of data and reduce the amount of storage space and save the bandwidth by deduplication. It is used in the following areas –

- In Social Network Applications
- In Online Shopping Application
- In Multimedia data storage & security

Objectives:

- To save cloud storage and preserve the privacy of data holders by proposing a scheme to manage encrypted data storage with deduplication.
- To support flexibly for data sharing with deduplication even when the data holder is offline, and it does not intrude the privacy of data holders.
- To propose an effective approach to verify data ownership and check duplicate storage with secure challenge and big data support.
- To prove the security and performance of the scheme through analysis and simulation.

II. RELATED WORK

The biggest challenge for big data from a security point of view is the protection of user's privacy. Traditional deduplication schemes cannot work on encrypted data. Existing solutions of encrypteddata deduplication suffer from security weakness. They cannot flexibly support data access control and revocation. Therefore, few ofthem can be readily deployed in practice. The propose scheme deduplicate encrypted data stored in cloud based onownership challenge and proxy re-encryption.

1. Jin Li, Yan Kit Li, Xiaofeng Chen, Patrick P. C. Lee, Wenjing Lou [8] are achieving the data de-duplication by providing the proof of data by the data owner. New de-duplication constructions supporting authorized duplicate check in hybrid cloud architecture in which the duplicate-check tokens of files are generated by the private cloud server with private keys.
2. Shweta D. Pochhi, Prof. Pradnya V. Kasture [9] presented that the data and the Private cloud where the token generation will be performed for each file. Before uploading the data or file to public cloud, the client will send the file to private cloud for token generation which is unique for each file. Private clouds then generate a hash and a token and send the token to client. A system which achieves confidentiality and enables block-level de-duplication at the same time. Before uploading the data or file to public cloud, the client will send the file to private cloud for token generation which is unique for each file.
3. Backialakshmi.NManikandan [10] presents a system of giving the data de-duplication by giving identity of data by the data owner. We refer a system include identity of data owner so it will help as handle better security issues in cloud computing.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 6, June 2017

4. BhushanChoudhary, AmitDravid [11] presented a de-duplication system in the cloud storage proposed to reduce the storage size of the tags for integrity check. To upgrade the security of de-duplication and secure the information secrecy demonstrated to secure the information by transforming the predictable message into unpredictable message.
5. James S. Plank LihaoXu [12] shown that the construction of the distribution matrix in Cauchy Reed-Solomon coding impacts the encoding performance. We note the rather counter-intuitive result that Cauchy Reed-Solomon coding can perform better for larger values of w while holding the other parameters constant.

III. EXISTING SYSTEM APPROACH

From the above literature survey we have concluded that an existing data de-duplication system, 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.

DISADVANTAGES OF EXISTING SYSTEM:

I. In existing system apply de-duplication on only without encrypted data.
II. The data owners only outsource their data storage by utilizing public cloud while the data operation is managed in private cloud.

IV. PROPOSED SYSTEM APPROACH

In the proposed research work to design and implement a system which will provide the parallel processing to detect the data de-duplication problem in big data environment. The system also provides benefit access control of data management and proxy revocation of system.

ADVANTAGES OF PROPOSED SYSTEM:

I. The proposed system can flexibly support access control on encrypted data with deduplication.
II. Low Cost of Storage.
III. The proposed system can efficiently perform big data deduplication.

The system contains three types ofentities:

1) CSP that offers storage services and cannot befully trusted since it is curious about the contents of storeddata, but should perform honestly on data storage in order to gain commercial profits.

2) Data holder (u_i) that uploadsand saves its data at CSP. In the system, it is possible tohave a number of eligible data holders (u_i, i = 1, . . . , n) that could save the same encrypted raw data in CSP. Thedata holder that produces or creates the file is regarded asdata owner. It has higher priority than other normal data holders.

3) An authorizedparty (AP) that does not collude with CSP and is fullytrusted by the data holders to verify data ownership andhandle data deduplication.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 6, June 2017

Table -1: System Notation

Key	Description
(pki, ski)	The public key and secret key of user u_i for PRE
DEK_i	The symmetric key of u_i
H()	The hash function
M	The user data
E, R, D	The encryption, re-encryption and decryption algorithm of PRE

V. PROPOSED SYSTEM ARCHITECTURE

The proposed scheme deduplicate encrypted data at CSP by applying PRE to issue keys to different authorized data holders based on data ownership challenge.

The main aspects of scheme are:

i) **Encrypted Data Upload:** If data duplication check is negative, the data holder encrypts its data using a randomly selected symmetric key DEK in order to ensure the security and privacy of data, and stores the encrypted data at CSP together with the token used for data duplication check.

ii) **Data Deduplication:** Data duplication occurs at the time when data holder tries to store the same data that has been stored already at CSP. This is checked by CSP through token comparison. If the comparison is positive, CSP contacts AP for deduplication by providing the token.

iii) **Data Deletion:** When the data holder deletes data from CSP, CSP firstly manages the records of duplicated data holders by removing the duplication record of this user. If the rest records are not empty, the CSP will not delete the stored encrypted data, but block data access from the holder that requests data deletion. If the rest records are empty, the encrypted data should be removed at CSP.

iv) **Data Owner Management:** In case that a real data owner uploads the data later than the data holder, the CSP can manage to save the data encrypted by the real data owner at the cloud with the owner generated DEK and later on, AP supports re-encryption of DEK at CSP for eligible data holders.

v) **Encrypted Data Update:** In case that DEK is updated by a data owner with DEK0 and the new encrypted raw data is provided to CSP to replace old storage for the reason of achieving better security, CSP issues the new re-encrypted DEK0 to all data holders with the support of AP.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 6, June 2017

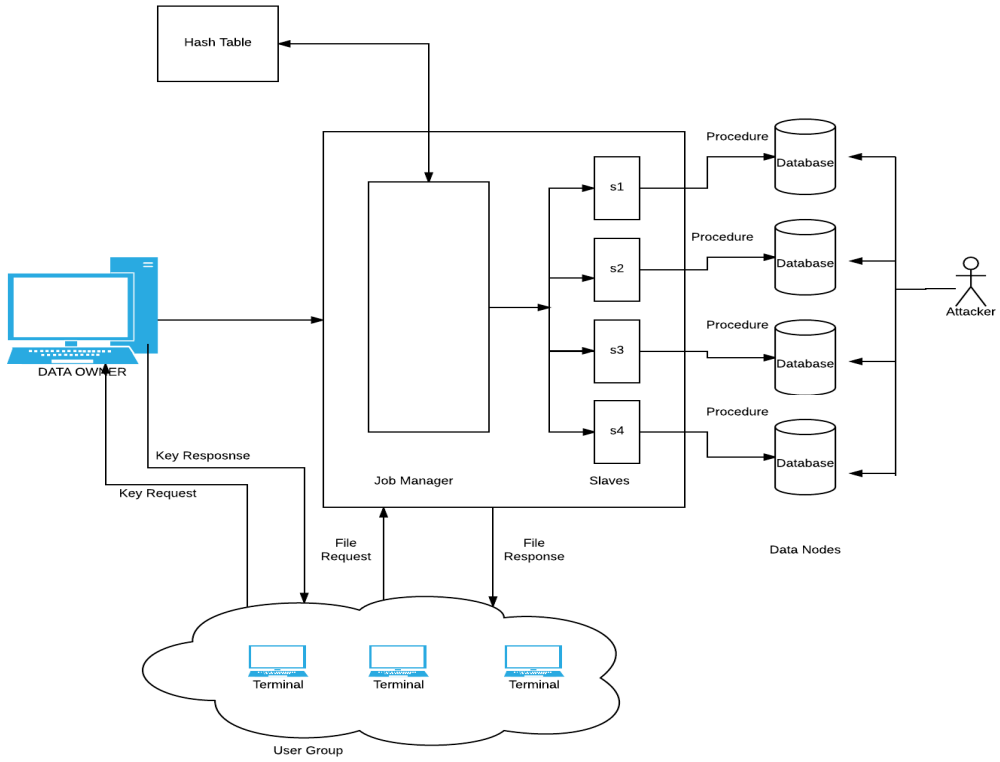

Fig.1. Proposed System Architecture

VI. CONCLUSION AND FUTURE WORK

Managing encrypted data with deduplication is important and significant in practice for achieving a successful cloud storage service, especially for big data storage. In this paper, we proposed a practical scheme to manage the encrypted big data in cloud with deduplication based on ownership challenge and PRE. Our scheme can flexibly support data update and sharing with deduplication even when the data holders are offline. Encrypted data can be securely accessed because only authorized data holders can obtain the symmetric keys used for data decryption. Extensive performance analysis and test showed that our scheme is secure and efficient under the described security model and very suitable for big data deduplication. The results of our computer simulations further showed the practicability of our scheme. Future work includes optimizing our design and implementation for practical deployment and studying verifiable computation to ensure that CSP behaves as expected in deduplication management.

REFERENCES

1. Zheng Yan, Senior Member, IEEE, Wenxiu Ding, Xixun Yu, Haiqi Zhu, and Robert H. Deng, Fellow, IEEE , "Deduplication on Encrypted Big Data in Cloud" IEEE TRANSACTIONS ON BIG DATA, VOL. 2, NO. 2, APRIL-JUNE 2016.
2. M. Bellare, S. Keelveedhi, and T. Ristenpart, "DupLESS: Server aided encryption for deduplicated storage," in Proc. 22nd USENIX Conf. Secur., 2013, pp. 179–194.
3. Dropbox, A file-storage and sharing service. (2016). [Online]. Available: http://www.dropbox.com
4. Google Drive. (2016). [Online]. Available: http://drive.google.com
5. Mozy, Mozy: A File-storage and Sharing Service. (2016). [Online]. Available:http://mozy.com/

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 6, June 2017

6. J. R. Douceur, A. Adya, W. J. Bolosky, D. Simon, and M. Theimer, "Reclaiming space from duplicate files in a serverless distributed file system," in Proc. IEEE Int. Conf. Distrib. Comput. Syst., 2002, pp. 617–624, doi:10.1109/ICDCS.2002.1022312.
7. G. Wallace, et al., "Characteristics of backup workloads in production systems," in Proc. USENIX Conf. File Storage Technol., 2012, pp. 1–16.
8. Jin Li, Yan Kit Li, Xiaofeng Chen, Patrick P. C. Lee, Wenjing Lou "A Hybrid Cloud Approach for Secure Authorized De-duplication" IEEE Transactions on Parallel and Distributed Systems: PP Year 2014.
9. Shweta D. Pochhi, Prof. Pradnya V. Kasture, "Encrypted Data Storage with De-duplication Approach on Twin Cloud" International Journal of Innovative Research in Computer and Communication Engineering.
10. Backialakshmi.NManikandan. "M SECURED AUTHORIZED DE-DUPLICATION IN DISTRIBUTED SYSTEM" , IJIRST International Journal for Innovative Research in Science and Technology— Volume 1 — Issue 9 — February 2015.
11. BhushanChoudhary, AmitDravid, "A Study On Secure Deduplication Techniques In Cloud Computing" International Journal of Advanced Research in Computer Engineering and Technology (IJARCET) Volume 3, Issue 12, April 2014.
12. James S. Plank LihaoXu "Optimizing Cauchy Reed-Solomo, Codes for Fault-Tolerant Network Storage Applications" , The 5th IEEE International Symposium on Network Computing and Applications (IEEE NCA06), Cambridge, MA, July, 2006.