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

International Journal of Innovative Research in Computer and Communication Engineering

*(An ISO 3297: 2007 Certified Organization)*

**Vol. 4, Issue 10, October 2016**

# Deduplication on Encrypted Big Data in Using HDFS Framework

Pritee Patil, Nitin N. Pise

PG Student, Department of Computer, Maharashtra Institute of Technology, Pune University, Pune, India

Professor, Department of Computer, Maharashtra Institute of Technology, Pune University, Pune, India

**ABSTRACT:** The greatest test for enormous information from a security perspective is the assurance of client's protection. Enormous information as often as possible contains gigantic measures of individual identifiable data and thusly security of clients is a colossal concern. Be that as it may, encoded information present new difficulties for cloud information deduplication, which gets to be significant for huge information stockpiling and preparing in cloud. Customary deduplication plans can't take a shot at encoded information. Existing arrangements of scrambled information deduplication experience the ill effects of security shortcoming. They can't adaptably bolster information get to control and renouncement. Hence, few of them can be promptly sent by and by. In this paper, we propose a plan to deduplicate scrambled information put away in cloud in light of proprietorship test and intermediary re-encryption. It incorporates cloud information deduplication with get to control. We assess its execution in light of broad investigation and PC reenactments. The outcomes demonstrate the predominant proficiency and adequacy of the plan for potential viable sending, particularly for enormous information deduplication in distributed storage.

**KEYWORDS:** Access control, Big data, cloud computing, data-deduplication.

## I. INTRODUCTION

Our meant to minimize repetitive information and augment space funds. A strategy which has been generally embraced is cross-client deduplication. The basic thought behind deduplication is to store copy information (either documents or pieces) just once. Accordingly, if a client needs to transfer a record (piece) which is now put away, the cloud supplier will add the client to the proprietor rundown of that document (square). Deduplication has demonstrated to accomplish high space and cost reserve funds and numerous Huge Information stockpiling suppliers are as of now receiving it. Deduplication can diminish capacity needs by up to 90-95% for reinforcement applications and up to 68% in standard document frameworks. Distributed computing gives apparently boundless "virtualized" assets to clients as administrations over the entire Web, while concealing stage and usage subtle elements. Today's cloud benefit suppliers offer both exceedingly accessible capacity and enormously parallel figuring assets at moderately low expenses. As distributed computing gets to be predominant, an expanding measure of information is being put away in the cloud and imparted by clients to determined benefits, which characterize the get to privileges of the put away information. One basic test of distributed storage administrations is the administration of the regularly expanding volume of information. To make information administration versatile in distributed computing, de-duplication has been an outstanding strategy and has pulled in more consideration as of late. Information de-duplication is a specific information pressure system for wiping out copy duplicates of rehashing information away. The strategy is utilized to enhance stockpiling use and can likewise be connected to network information exchanges to diminish the quantity of bytes that must be sent. Rather than keeping numerous information duplicates with similar substance, de-duplication disposes of repetitive information by keeping stand out physical duplicate and alluding other excess information to that duplicate. De-duplication can occur at either the document level or the piece level. For record level de-duplication, it disposes of copy duplicates of similar document. De-duplication can likewise happen at the piece level, which takes out copy squares of information that happen in non-indistinguishable documents. Distributed computing is a rising administration display that gives calculation and capacity assets on the Web. One appealing usefulness that distributed computing can offer is distributed storage. People and undertakings are regularly required to remotely file their information to stay away from any data misfortune in the event that there are any equipment/programming disappointments or unanticipated fiascos. Rather than buying the required stockpiling media to keep information reinforcements, people and ventures can basically outsource

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

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 10, October 2016

their information reinforcement administrations to the cloud benefit suppliers, which give the fundamental stockpiling assets to have the information reinforcements. While distributed storage is appealing, how to give security certifications to outsourced information turns into a rising concern. One noteworthy security test is to give the property of guaranteed cancellation, i.e., information records are for all time blocked heaps of erasure. Keeping information reinforcements for all time is undesirable, as delicate data might be uncovered later on in view of information break or wrong administration of cloud administrators. Subsequently, to dodge liabilities, endeavors and government organizations normally keep their reinforcements for a limited number of years and demand to erase (or crush) the reinforcements a short time later. For instance, the US Congress is figuring the Web Information Maintenance enactment in approaching ISPs to hold information for a long time, while in Joined Kingdom, organizations are required to hold wages and compensation records for a long time.

## II. GOALS

1. Improve the system performance in using parallel processing in HDFS framework.
2. In distinguish ability of file token/duplicate-check token. It requires that any user without querying the private cloud server for some file token, he cannot get any useful information from the token, which includes the file information or the privilege information.
3. Provide the access control system using proxy regeneration approach which can eliminate the data collusion as well SQL injection attacks.

## III. OBJECTIVE

- To develop the system in HDFS framework with 4 to 16 node cluster.
- Motivate to save cloud storage and preserve the privacy of data holders by proposing a scheme to manage encrypted data storage with deduplication.
- Flexibly support data sharing with deduplication even when the data holder is offline, and it does not intrude the privacy of data holders.
- Propose an effective approach to verify data ownership and check duplicate storage with secure challenge and big data support.
- Integrate cloud data deduplication with data access control in a simple way, thus reconciling data deduplication and encryption.
- Prove the security and assess the performance of the proposed scheme through analysis and simulation. The results show its efficiency, effectiveness and applicability.

## IV. LITERATURE SURVEY

### *A. HYBRID CLOUD APPROACH FOR SECURE AUTHORIZED DE-DUPLICATION*

De-duplication of data has many forms. Typically, there is no one best way to implement data de-duplication across an whole an organization. Instead, to maximize the benefits, organizations may deploy more than one deduplication strategy. Cloud data storage services mostly refer de-duplication, which removing redundant data by storing only single copy of every file or block [1]. It is very essential to know the backup and backup challenges, when selecting de-duplication as a solution.

Advantages: This De-duplication technique reduces the space and bandwidth requirements of data storage services, and is most effective when applied with multiple users, a common practice by cloud storage offerings.

Limitations: Data deduplication does not work with traditional encryption techniques. While using data deduplication technique it should not reduce fault tolerance mechanism. Types of data de-duplication are described below: File-level de-duplication: This de-duplication technique is commonly called as single-instance storage, file-level data de-duplication compares a file that has to be archived or backup that has already been stored by checking all its attributes against the index. The index is updated and stored only if the file is unique, if not than only a pointer to the existing file that is stored references. Only the single instance of file is saved in the result and relevant copies are replaced by"stub" which points to the original file [1].

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

International Journal of Innovative Research in Computer and Communication Engineering

*(An ISO 3297: 2007 Certified Organization)*

**Vol. 4, Issue 10, October 2016**

***B. CONTENT ADDRESSABLE STORAGE***

Eliminating multiple copies of any file is a form of the de-duplication. Single instance storage (SIS) environments can detect and eliminate redundant copies of identical files. After a file is stored in a single-instance storage system than, all the other references to same file, will refer to the original, single copy. Single instance storage systems compare the content of files to detect if the incoming file is identical to an existing file in the storage system. Content-addressed storage is typically combined with single-instance storage functionality [5]. While filelevel de-duplication avoids storing files that are a duplicate of another file, many files that are considered unique by single-instance storage measurement may have a huge amount of redundancy within the files or between files. For example, it would take only one small element (e.g., a new date inserted into the title slide of a presentation) for single-instance storage to through two large files as being different and requiring them to be stored without further de-duplication [7].

Advantages: CAS system provides higher searching speed for documents.

Limitations: This system only provides performance benefits when there are more read operations than update operations.

Sven B. et al. [10] proposed twin cloud architecture for secure deduplication in cloud storage. As the name suggest their approach uses one public cloud and one private cloud, User communicates with a private cloud (organization maintained cloud) which encrypts data before outsourcing to public cloud. This private cloud is also responsible for verification of stored data in public cloud. Their architecture uses private cloud for operations requiring security whereas other kind of queries is processed by public cloud. Their technique allows maximum utilization of resources of private cloud, and only high load queries are processed on-demand by the public cloud.

Trusted Cloud requires constant amount of storage and is used constantly in the Setup Phase for pre-computing encryption. The public cloud provides large amount of storage and is used for time-critical Query operations. Zhang et al. also proposed a hybrid cloud [1] [7] system named Sedic [7]. The system supports the privacy aware data computing. The system is based on MapReduce fuction. They address the problem of authorized deduplication of public cloud data. Here the private cloud is assumed as honest but curious. Advantages: Convergent encryption provides security while deduplication process. Security in deduplication process can be increased using twin cloud approach or hybrid cloud approach and using random encryption keys. Limitations: Increase in complexity in deduplication process is main limitation in above approaches.

## V. 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.

## VI. 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 bigdata environment. The system also provide benefit access control of data management and proxy revocation of system.

## VII. SYSTEM ARCHITECTURE

Proposed scheme contain following main aspects

Encrypted Data Upload:

If data duplication check is negative, the data holder encrypts its data utilizing an arbitrarily culled symmetric key DEK in order to ascertain the security and privacy of data, and stores the encrypted data at CSP together with the token utilized for data duplication check. The data holder encrypts DEK with pkAP and passes the encrypted key to CSP.

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

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 10, October 2016

Data Deduplication:

Data duplication occurs at the time when data holder u endeavors 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 and the data holder's PRE public key. The AP challenges data ownership, checks the eligibility of the data holder, and then issues a re-encryption key that can convert the encrypted DEK to a form that can only be decrypted by the eligible data holder.

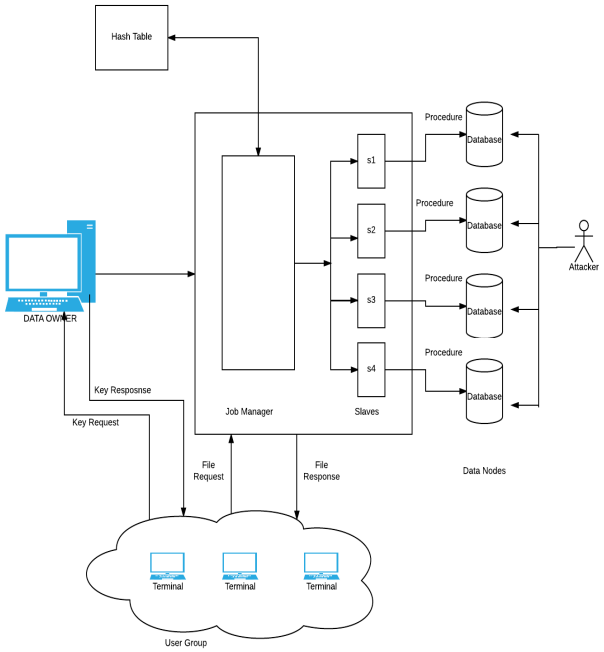


***Figure 01: System architecture***

Data Deletion:

When the data holder effaces data from CSP, CSP firstly manages the records of duplicated data holders by abstracting the duplication record of this utilizer. If the rest records are not vacuous, the CSP will not efface the stored encrypted data, but block data access from the holder that requests data effacement. If the rest records are vacuous, the encrypted data should be abstracted at CSP.

Data Owner Management:

In case that an authentic data owner uploads the data later than the data holder, the CSP can manage to preserve the data encrypted by the authentic data owner at the cloud with the owner engendered DEK and later on, AP fortifies re-encryption ofDEK at CSP for eligible data holders.

Encrypted Data Update:

In case that DEK is updated by a data owner with DEK0 and the incipient encrypted raw data is provided to CSP to supersede old storage for the reason of achieving better security, CSP issues the incipient re-encrypted DEK0 to all data holders with the fortification 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)*

**Vol. 4, Issue 10, October 2016**

## VIII. CONCLUSION

Managing encrypted data with deduplication is consequential and consequential in practice for achieving a prosperous cloud storage accommodation, especially for astronomically immense data storage. In this paper, we proposed a practical scheme to manage the encrypted sizably voluminous data in cloud with deduplication predicated 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 sanctioned data holders can obtain the symmetric keys utilized for data decryption. Extensive performance analysis and test showed that our scheme is secure and efficient under the described security model and very opportune for sizably voluminous 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 ascertain that CSP deports as expected in deduplication management.

## REFERENCES

[1]Jin Li, Yan Kit Li, Xiaofeng Chen,Patrick P. C. Lee and Wenjing Lou, "A Hybrid Cloud Approach for Secure Authorized Deduplication", IEEE Transaction On Parallel And Distributed System,Vol.PP,No.99, 2014.
[2]. Maneesha Sharma, Himani Bansal and Amit Kumar Sharma, " Cloud Computing: Different Approach & Security Challenge", IJSCE, Volume-2, Issue-1,March 2012.
[3]. Kangchan Lee, "Security Threats in Cloud Computing Environments", International Journal of Security and Its Applications, Vol. 6, No. 4, October, 2012.
[4]. Sashank Dara, "Cryptography Challenges for Computational Privacy in Public Clouds", International Journal of Security and Its Applications, Volume 4, 2002.
[5]. David Pointcheval, "Asymmetric Cryptography and Practical Security", International Journal of Security and Its Applications, Volume 4,2002.
[6]. Yogesh Kumar, Rajiv Munjal and Harsh Sharma, "Comparison of Symmetric and Asymmetric Cryptography with Existing Vulnerabilities and Counter-measures", International Journal of Computer Science and Management Studies, Vol. 11, Issue 03, Oct 2011.
[7]. Jan Stanek, Alessandro Sorniottiy, Elli Androulakiy, and Lukas Kencl, "A Secure Data De-duplication Scheme for Cloud Storage", IBM Research, Zurich, May 1994.
[8]. Jin Li, Xiaofeng Chen, Mingqiang Li, Jingwei Li, Patrick P.C. Lee, and Wenjing Lou, "Secure Auditing and De-duplicating Data in Cloud", IEEE Transactions on Computers,unpublished,2015.
[9]. Deepak Mishra and Sanjeev Sharma, "Comprehensive study of data de-duplication", International Conference on Cloud, Big Data and Trust,Vol.13.No.15,NOV 2013
[10]. Paul Anderson and Le Zhang, "Fast and Secure Laptop Backups with Encrypted De-duplication", Proceedings of Eurocrypt, Vol. 6,March 2013.
[11]Mihir Bellare,Sriram Keelveedhi and Thomas Ristenpart, "Message-Locked Encryption and Secure Deduplication", Proceedings of Eurocrypt, Vol. 6,March 2013.