

# Encrypted Data Deduplication Approach in Hadoop Environment

Nikita C. Sabale<sup>1</sup>, Prof. N. G. Pardeshi<sup>2</sup>

P.G. Student, Department of Computer Engineering, Sanjivani College of Engineering, Kopergaon, India<sup>1</sup>

Associate Professor, Department of Computer Engineering, Sanjivani College of Engineering, Kopergaon, India<sup>2</sup>

**ABSTRACT:** A big data consists of a huge amount of personal identifiable information and the biggest challenge for it from a security point of view is the protection of user's privacy. However, encrypted data introduce new challenges for data deduplication, which becomes crucial for big data storage and processing in hadoop. Traditional deduplication schemes cannot work on encrypted data technique. Existing solutions of encrypted data deduplication suffer from security problems. That means they cannot support data access control and revocation mechanisms. Therefore, few of them can be readily deployed in practice. In this system, we propose a scheme to deduplicate encrypted data stored in hadoop based on ownership challenge and proxy re-encryption. It integrates 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 hadoop.

**KEYWORDS:** Data Deduplication, Hadoop Environment , Big data , Data Owner Management , Encrypted data upload , Encrypted data deletion, Access Control.

## I. INTRODUCTION

Nowadays big data is used by many people but maintaining the security of data and the protection of users privacy is important. Encrypted data provides new challenges for data deduplication, which becomes crucial for big data storage and processing in storage [2]. Previously deduplication cannot work on an encrypted data. Proposed system prove the security and performance of the scheme through analysis and simulation. To save storage and preserve the privacy of data holders by proposing a scheme to manage encrypted data storage with deduplication. Thus, the proposed system removes the duplicate copies of data and improve the reliability. The results show the preferred efficiency and effectiveness of the scheme for potential practical deployment, exclusively for big data deduplication in cloud storage.

Recently, data deduplication provides lot many benefits. Money will get saved on disk expenditures as it requires lower storage space. The most effective use of disk space allows for longer disk retention periods, which gives better recovery time objectives (RTO) for a longer duration and decreases the need for tape backups. Data deduplication also decreases the data that must be sent across a WAN for remote backups, imitations, and disaster recovery.

In existing surveys we find some issues mentioned as below

- 1) Data redundancy issue generated in storage.
- 2) Some systems can be take very long time for data processing
- 3) No data security mechanism yet develop in any system.

After finding all these issues in existing approaches we develop a system which can work with parallel processing finding aduplication issue.

Data de-duplication is a special data compression technique which eliminates duplicate copies of repeating data in storage [1]. The technique is preferable to improve storage utilization and also applied for 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. In [6], the authors shows that de-duplication of data can take place at either the file level or at the block level. For file level deduplication, it eliminates duplicate copies of the same file. For block level, it eliminates duplicate blocks of data that occur in non-identical files. The simple idea on the side of deduplication is to store duplicate data (either files or

# International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: [www.ijirset.com](http://www.ijirset.com)

Vol. 6, Issue 6, June 2017

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).

Cloud computing and Hadoop can propose the functionality of storage. Individuals and enterprises are often required to remotely accomplish their data to avoid any data loss in case of any hardware/software failures or unexpected disasters [8]. Instead of procuring the required storage media to keep data backups, individuals and enterprises can simply outsource their data backup services to the cloud service providers (CSP), which presides the necessary storage resources to host the data backups. One of the peak security challenge is to provide the property of assured deletion, i.e., data files are permanently not accessible upon requests of deletion. Keeping data backups permanently is undesirable, as confidential information may be exposed in the future because of data fissure or erroneous management of cloud operators.

The work in this paper is divided in two stages. 1) Duplicate data elimination 2) Storing files at HDFS. Data duplication is get checked with Vector base cosine similarity algorithm (VCS), in which each file is get compared with files of database. After that storing of data is done. The above algorithm mainly works on threshold value, i.e., if value of count greater than or equal to threshold value then that file cannot stored on Hadoop otherwise it will get uploaded.

Paper is organized as follows. Section II describes automatic text detection using morphological operations, connected component analysis and set of selection or rejection criteria. The flow diagram represents the step of the algorithm. After detection of text, how text region is filled using an inpainting technique that is given in Section III. Section IV presents experimental results showing results of images tested. Finally, Section V presents conclusion.

## II. RELATED WORK

Many Existing solutions of encrypted data deduplication faces major problem of security weakness. They are unable to flexibly support data access control and revocation. Therefore, few of them can be easily implemented in practice. Jin Li, Yan Kit Li. [8] proposed scheme of deduplicate encrypted data stored in cloud based on ownership challenge and proxy reencryption.

Data de-duplication is achieved by providing the proof of data by the data owner. M. Bellare. [2], formalize a new cryptographic primitive, Message-Locked Encryption (MLE), where the key encryption and decryption are performed is it derived from the message. MLE provides a way to accomplish secure de-duplication, aim formerly targeted by numerous cloud database storage providers. They supply definitions of privacy and a form of integrity that they call tag consistency. They provide ROM security analyses of a natural family of MLE schemes that includes deployed techniques. They built connections with deterministic encryption, hash functions secure on correlated inputs.

Shweta D. Pochhi. [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 generates 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.

Backalalakshmi. [10], presented a system of giving the data de-duplication by giving identity of data by the data owner. 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 system named Sedic. The system supports the privacy aware data computing. The system is based on MapReduce function. They address the problem of authorized deduplication of public cloud data. Here the private cloud is assumed as honest but curious.

Bhushan Choudhary. [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 deduplication and secure the information secrecy demonstrated to secure the information by transforming the predictable message into unpredictable message.

James S. Plank Lihao Xu [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. SYSTEM ARCHITECTURE

# International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: [www.ijirset.com](http://www.ijirset.com)

Vol. 6, Issue 6, June 2017

In the proposed research work we 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 provide benefit access control of data management and proxy revocation of system. There are basically two types of data de-duplication as:

- i) File-level de-duplication: This de-duplication technique is commonly called as single instance storage, file-level data deduplication compares a file that has to be archived or backup that has already been stored by checking all its attributes against the index.
- ii) Block-level de-duplication: Block-level data de-duplication operates on the basis of sub-file level. As the name implies, that the file is being broken into segments blocks or chunks that will be examined for previously stored information vs redundancy.

Fig.1 shows the architecture of proposed Deduplication system based on HDFS ,which consist of data owner and storage system.

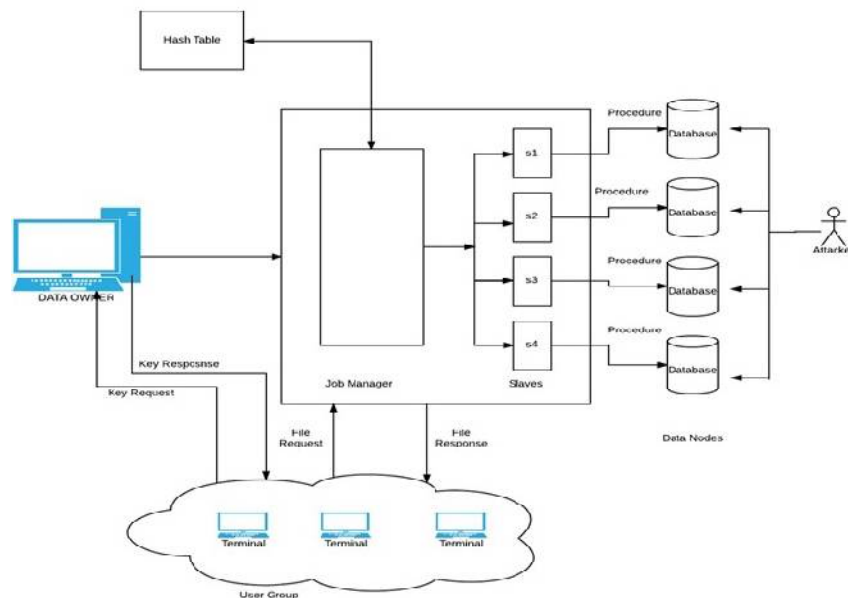


Fig 1. Architecture of Proposed System

Data owner basically gives an access permissions to respective users to allow access to a file. It performs operations as upload the file on the storage, encrypt the file, share/ revoke file, delete/update file. The utilities consists of either discard of file or stores it on server. Hadoop Distributed File System (HDFS) maintains consistency of data distributed across the large number of data nodes. Large files are distributed across the cluster and managed via a metadata server known as the primary namenode. HDFS maintains a number of high availability features including replication and rebalancing of data across nodes. It executes map-reduce and checks the similarity score with the help of VCS algorithm. In this module system will process all the execution in parallel, we used one hash table at the time of data insertion once data has insert into database it will make a history into the hash table. For the efficient retrieval we can use the hash table. User management basically uses the files by downloading from hadoop or data owner. User can request for a file and get access to a particular file. User also decrypts the file, download the file with specific key and view file contents. System Admin manages the load of server also maintains the data records and manages the security. It performs loadbalancing, generates look up tables and manages data security schemes with encryption.

Algorithm 1 performs duplication checking between two or more files and compares the score with threshold value, if score greater than the threshold value then file not stored onto the server else it will get stored .

# International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: [www.ijirset.com](http://www.ijirset.com)

Vol. 6, Issue 6, June 2017

## Algorithm 1: Vector base cosine similarity (VCS)

**Input-** Query Q, Threshold t

**Output-** duplicate if returns 1 else unique

Here we have to find similarity of two vectors:

$a = (a_1, a_2, a_3, \dots)$  and  $b = (b_1, b_2, b_3, \dots)$

Step 1: Read each row R from dataset D

Step 2: for each ( Column c from R)

Step 3: score= Formula1(R,Q)

If(score > t) Break;

Early stop;

Else step 2 continue

End for

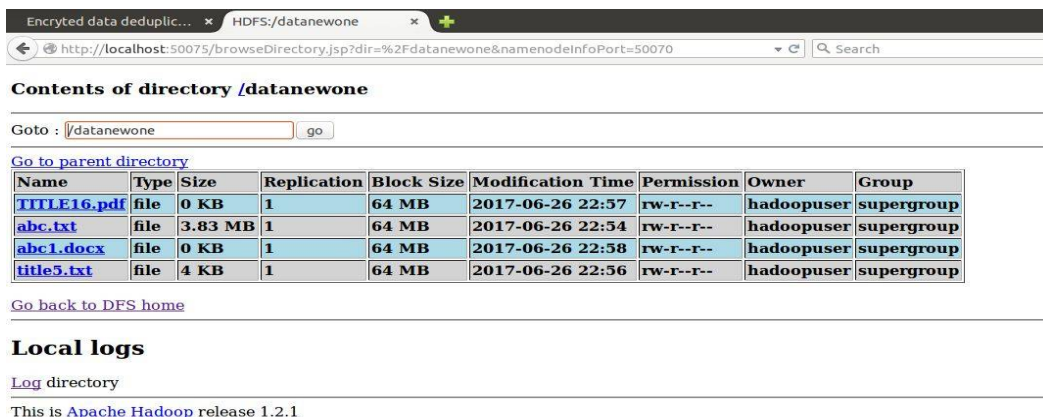
The proposed system accepts input as file, uploads the data in an encrypted form after that it checks the data deduplication and performs the operations such as modification, deletion and download of data . System also handles data owner management. The proposed system can flexibly support access control on encrypted data with deduplication. Also the proposed system can have Low Cost of Storage and efficiently perform big data deduplication.

## IV. EXPERIMENTAL RESULTS

Figures shows the results of how files are stored on HDFS after uploaded by data Owner . In Figure 2 (a), it shows that the data node and name node must be active while performing upload operation. And in Figure 2 (b), it represents the actual storage of files to hadoop environment.

```
hadoopuser@ubuntu:~$ jps
3504 NameNode
4084 Jps
4011 TaskTracker
3644 DataNode
2748
3788 SecondaryNameNode
3871 JobTracker
```

Fig 2 (a) – Working of hadoop at terminal



Contents of directory /datanewone

Goto : /datanewone go

Name	Type	Size	Replication	Block Size	Modification Time	Permission	Owner	Group
<a href="#">TITLE16.pdf</a>	file	0 KB	1	64 MB	2017-06-26 22:57	rw-r--r--	hadoopuser	supergroup
<a href="#">abc.txt</a>	file	3.83 MB	1	64 MB	2017-06-26 22:54	rw-r--r--	hadoopuser	supergroup
<a href="#">abc1.docx</a>	file	0 KB	1	64 MB	2017-06-26 22:58	rw-r--r--	hadoopuser	supergroup
<a href="#">title5.txt</a>	file	4 KB	1	64 MB	2017-06-26 22:56	rw-r--r--	hadoopuser	supergroup

Go back to DFS home

**Local logs**

Log directory

This is [Apache Hadoop](#) release 1.2.1

Fig 2 (b) – Storage of files at HDFS

# International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: [www.ijirset.com](http://www.ijirset.com)

Vol. 6, Issue 6, June 2017

Name of uploaded file	Name of User	Size of file (KB/MB)	Duplication Present or not	How much duplicate and with which file	File uploaded or not
title5.txt	niki	4.2 KB	No	0.35	Uploaded
nn.txt	rahul	4.2 KB	Yes	1.0	Not uploaded to server
abc1.txt	Ram	1.5 MB	No	0.48	Uploaded
abc.docx	niki	4.0 MB	No		Uploaded
TITLE3.docx	Om	14.9 KB	No		Uploaded
TITLE3.docx	Om	14.9 KB	File already exists	0.28	Not Uploaded
TITLE17.pdf	jitu	121.4 KB	No	0.0	Uploaded
TITLE25.pdf	Om	135.7 KB	No	0.39	Uploaded
TITLE21.pdf	Nikita	139.8 KB	No	0.18	Uploaded

**Table 1: Result of duplication checking with VCS algorithm while uploading the file**

From above table it represents the files with different format by different users will get uploaded to storage system. With the help of vector base cosine similarity (VCS) algorithm the duplicate files are get identified and if data duplication present then file not saved to server i.e., it is mainly based on threshold value.

## V. CONCLUSION

In this system, we protect the data confidentiality alongwith secure de-duplication, an authorized deduplication is proposed in HDFS framework which can provide parallelprocessing with minimum time complexity. Here, problem of privacy preserving in de-duplication in cloud environment is considered and advanced scheme supporting differential authorization and authorized duplicate check is proposed. The system addresses the issue in authorized de-duplication to achieve better security. System shows that the authorized duplicate check scheme incurs minimal overhead compared to convergent encryption and network transfer.

## 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] Mozy, Mozy: A File-storage and Sharing Service. (2016). [Online]. Available: <http://mozy.com/>
- [5] A. J. R. Douceur, A. Adya, W. J. Bolosky, D. Simon, and M. Theimer, "Reclaiming space from duplicate files in a serverless distributed filesystem", in Proc. IEEE Int. Conf. Distrib. Comput. Syst., 2002, pp. 617-624, doi:10.1109/ICDCS.2002.1022312.
- [6] G. Wallace, et al., "Characteristics of backup workloads in production systems", in Proc. USENIX Conf. File Storage Technol., 2012, pp. 116.
- [7] 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.
- [8] 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.
- [9] Backialakshmi.N.Manikandan, "M SECURED AUTHORIZED DEDUPLICATION IN DISTRIBUTED SYSTEM", IJRST International Journal for Innovative Research in Science and Technology Volume 1 Issue 9 February 2015.
- [10] Bhushan Choudhary, Amit Dravid, "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.
- [11] James S. Plank Lihao Xu, "Optimizing Cauchy Reed-Solomon Codes for Fault-Tolerant Network Storage Applications", The 5th IEEE International Symposium on Network Computing and Applications (IEEE NCA06), Cambridge, MA, July, 2006.
- [12] Nikita Sabale, Prof. N.G. Pardeshi, "A Survey on Deduplication in Encrypted Big Data Using HDFS Framework", International Journal of Innovative Research in Computer and Communication Engineering, Vol. 4, Issue 9, September 2016.