Journal of Current Research in Engineering and Science
Bi-Annual Online Journal ( ISSN : 2581 - 611X )
Volume 4- Issue 1, Paper 14 January

# A SURVEY ON DEDUPLICATION TECHNIQUES HANDLING BIGDATA IN HDFS

*Dr. P. Balamurugan[1], K. H. Vani[2]*
*[1, 2] Assistant Professor,*
*[1]Department of Computer Science, [2]Department of computing*
*[1] Government arts college (Autonomous), Coimbatore*
*[2]Coimbatore institute of technology, Coimbatore*

***Abstract****— The need for data storage space is already dramatically growing. Due to increased storage demands, the machine society attracts to HDFS storage. Data security and cost considerations in HDFS storage are critical problems. In addition to wasting storage, a duplicate file often improves access time. Detecting and removing duplicate data is also an important role. Data deduplication has increased attention and popularity in large-scale computing facilities, an effective solution to data reduction. It removes redundant data at file or sub file level and detects duplicate contents through its encrypted protected hash signature. The basic concept of deduplication is the storing for just a single period of duplicate records. Thus, the HDFS provider must first connect a person able to download a saved file to the owner list for that specific file. This is why deduplication has been embraced quickly by numerous HDFS storage providers. Today, it is now a common way to minimize disc space and download capacity and helps to increase data scalability. De-duplication often reduces the apprehension of surplus data by retaining a single physical copy and applies to this copy all surplus data and is the safest solution for different copies of the same data. This literature survey has critically assessed various algorithms and techniques for safe deduplication approaches across accurate and effective methodologies. The findings indicate that a combination of encrypted deduplication approaches with improved protection functionality will provide exceptional security choices for reliable and accurate deduplication.*

*Keywords—Deduplication, HDFS, Storage, Survey*

## I. INTRODUCTION

One of Big Data's most common systems. Hadoop offers several benefits in comparison to other big-data solutions, such as open source, high parallelism, catastrophe recovery, etc [7]. The two key component parallel processing and efficient storage is carried out off-shelf by MapReduce and HDFS (Hadoop Distributed File System). In specific, MapReduce, Hadoop's programming model, fits into parallel cluster deduplication processing since it first splits and then combined its two-stage processing mode. Consequently, Hadoop-based deduplication solutions have drawn many researchers and companies [4]–[6]. [5]. the good performance index management is important for the cluster deductions as the chunk index increases linearly with the data quantity. HBase, the Hadoop NoSQL database, is also used as chunk index table to store chunk index table in the existing Hadoop-based deduplication scheme in order to eliminate the scalability constraint of index quest [10]. This approach, with Hadoop and HBase working together, is a popular approach for the Hadoop-based Big Data solution but not an optimum model for Big Data deduplication. For eg, MapReduce's regular two-phase mode waste time sorting the interim results before entering the Reduce phase, and the single table storage mode of HBase often fails to meet the parallel cluster deduplication processing requirement [12].

The same data is processed over and over again, using unwanted physical space on a disc or tape, power electricity and refreshing the disk/tape drives and backup bandwidth. This creates a chain of inefficiencies in costs and resources in the organization. Traditional encryption is incompatible with data deduplication thus providing data protection. In particular, various users need their own keys to encrypt their files. Inseparable copies of data from various users would also contribute

Journal of Current Research in Engineering and Science
Bi-Annual Online Journal ( ISSN : 2581 - 611X )
Volume 4- Issue 1, Paper 14 January

to different cipher messages, rendering deductions impossible [1].

The remainder of this article is organized accordingly. The associated work and the inspiration for the proposed scheme are presented in Section II. Section III outlines the design and functionality of the prototype. Section IV provides a comparative survey analysis. Finally, Section V presents findings and future work.

## II. BACKGROUND STUDY

Cho, E. M., & Koshiba, T. [2] the authors have built a stable deduplication scheme where several groups share data by using verified hash convergent group signature. This was an effort to try to deduct cross-group users in a true big data management. The developers took use of current systems instead of introducing a whole different one. The authors have developed a basis for a community signing scheme that can shield HDFS providers from replication and defend against unpredictable data attacks.

Fu, Y. et al. [3] the authors were described AppDedupe, a distributed in-line flexible device deduplication framework for Big Data management that offers a compromise between scalable output and distributed deduplication efficiency through application recognition, data similarities and locality. It uses a two-step data routing system for routing super-chunk granularity data, to reduce the cross-node data redundancy with strong load balancing and low overhead communication, and uses application-ware similarity index optimization to increase deduplication performance in each node with very low RAM consumption.

Kumar, N. et al. [4] Bucket based methodology said, various buckets are used to store data and when map data is retrieved, it reduces the data already contained in buckets such that this technique improves great data storage performance.

Kumar, N. et al. [5] the deduplication technique focused on genetic evolution has the higher deduction proportion, which makes duplicated data quicker than other methods.

Saharan, S. et al. [8] proposed a new strategy of deduplication, "QuickDedup" that exceeds standard hash-based methods to deduplication on different metrics. Rather than computing hazards for each block on the disc, QuickDedup employs a new byte comparative technique, which minimizes the entire population of identical blocks and reduces them to only a number of partially similar blocks.

Wen, M. et al. [9] the authors have created an effective safe deduplication system for the outsourcing of big data in cloud computing. Authors supported both effective protection and safe data deductibility in the cloud computing environment through searching keywords through encrypted data. Data owner may in particular download data deduplicated files to the cloud by using the convergent keys, while data users can recover the appropriate files by giving tags or keywords.

Yan, Z. et al. [11] the developers suggested a realistic scheme to deal with secured cloud Big Data with possession and PRE-based deduplication. Also with offline data holders, the suggested solution will flexibly enable data upgrade and deduplication sharing. Encrypted data may be accessed safely as the symmetrical keys required for data decryption are only available to the registered data holders. Extensive performance review and tests have shown that the protection model mentioned is safe and reliable and very suitable for deduplication of large data.

Journal of Current Research in Engineering and Science
Bi-Annual Online Journal ( ISSN : 2581 - 611X )
Volume 4- Issue 1, Paper 14 January

## III. PROPOSED MODEL

The term "large data" is added because the data development is drastic and cannot be processed by conventional data processing applications. The vast amount of data, organized and unstructured, is referred to as large data [3]. For example: For example: Millions of images are posted regularly by their users on Facebook. Data access is growing day by day. Everyone has a storage space of at least 1TB. Figure 1 shows the characteristics of large files.

Data deduplication is one of the essential strategies in data compression to remove double copies of repeated data and deduplication is used in HDFS storage to reduce the need for room and bandwidth savings. The methods such as convergent encryption techniques have been proposed and used for encrypting the data before being outsourced to secure the integrity of critical data while promoting deduplication. This paper attempts to resolve the issue of permitted deduplication of data formally. Here we want to effectively resolve the issue of processing large volumes of data by deducting chunk level of differential rights in HDFS computation. For the data replication check in the machine we perform an authenticated duplication check. The proof is then applied to the file for the file replication search at the moment the file is loaded; this proof decides the right of entry to the file. It defines who may conduct a file replication search. For sending duplicate checks, users must upload their file and evidence of their possession of the file. The duplicate search request is only allowed if there is an HDFS file and there are user rights.

**Deduplication of Client Side Backup**

The deduplication hash calculations are first created on the source, i.e. client computer, as the name suggests, and files with indistinguishable hazards are yet to be submitted to the files already on the target equipment. Target devices create references in the links to remove duplicate copies of data from directories. The benefits of this deduplication are that it prevents unwanted file and data transmission through the network thus eliminating traffic burden.

**Primary and secondary storage:**

Primary storage facilities are built with the exception of the lowest possible cost for the finest efficiency. Furthermore, primary storage systems are less accommodating of any application or process and more cost-effective in terms of efficiency. Thus, two factors must be addressed when designing secondary storage: maximum efficiency and the lowest possible expense. Furthermore, the secondary storage facility mostly includes replicates and the secondary data versions. These versions of the data are not naturally used for real building activities such that they are forgiving of any loss of results in order to increase productivity. As data is transmitted across the network, the possible lack of data primarily concerns. And they store data differently than the way it was written for data deduplication programmes.

Journal of Current Research in Engineering and Science
Bi-Annual Online Journal ( ISSN : 2581 - 611X )
Volume 4- Issue 1, Paper 14 January

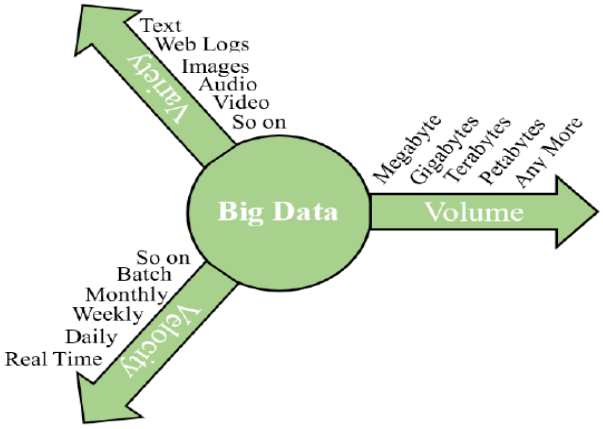


***Figure 1: Characteristics of Big Data***

Information deduplication is the redundant data removal process [5][6]. It's the method used to trace the same pieces in a storage unit and delete them. It is an easy way to store information or records. The deduplication mechanism as seen in Figure 3. In general, the deduction is split into three components, i.e. data device, location and disc positioning.

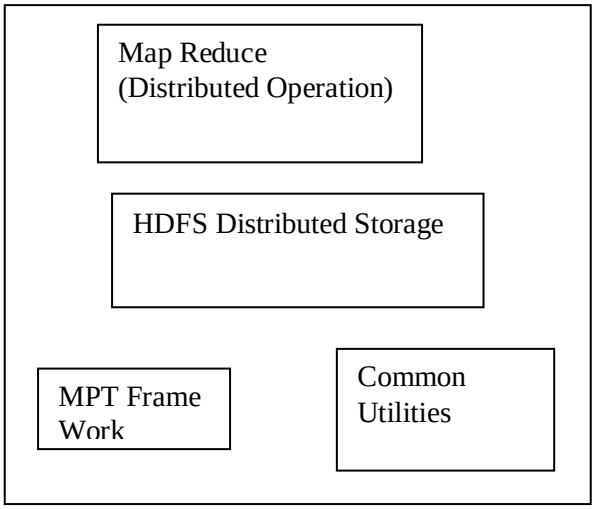


***Figure 2: Modules for Hadoop***

Data deduplication techniques are categorized in two sections, i.e. the deduplication of the file level and the deduplication of the block. When the hash value for two directories is the same, it is called the same in file level deduplication. Just a backup of a file is maintained. The same file shall be deleted with duplicate or obsolete versions. This is sometimes regarded as single-instance storage (SIS). In the deduplication level of blocks (chunk), the file is broken into blocks (chunks) and then tests and eliminates redundant blocks from the data. Just one copy is kept of each block. It eliminates room rather than SIS. It can also be separated into set duration and deduplication of variable lengths. The size of each block is consistent or set in a fixed length deduplication while the size of each block is different or not fixed in a variable length deduplication.

Journal of Current Research in Engineering and Science
Bi-Annual Online Journal ( ISSN : 2581 - 611X )
Volume 4- Issue 1, Paper 14 January

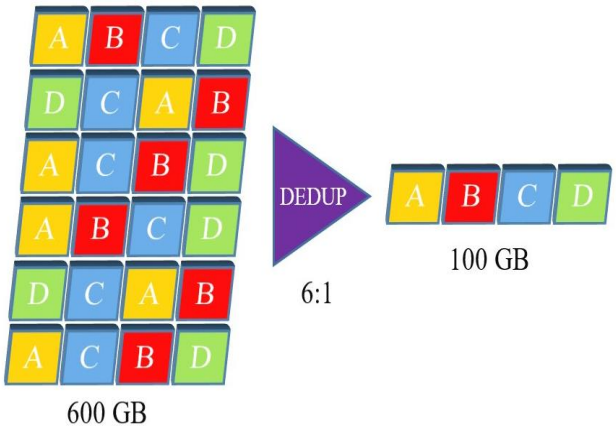


***Figure 3: Deduplication Process***

Deduplication based on disc positioning is based on how data can be put on the disc. Advance comparison or retroactive reference is methods utilized in this procedure. Current data chunks are maintained in the future and all older data chunks are linked to the recent chunks through pointers. The reverse reference adds further fragmentation in the previous chunks of data.

## IV. COMPARATIVE ANALYSIS OF SURVEY

| Author Name | Methodology | Limitations |
|---|---|---|
| Kumar, N.et al. [4] | Bucket based data deduplication technique. | Fixed size chunking splits the file into fixed-size bits and then generates hashing to distinguish duplicate fixed-size chunks. Fixed size chunks generate fixed size chunks, but if data changes, boundary shifts can cause a problem. |
| Kumar, N.et al. [5] | Optimizing the deduplication scheme by changing the relevant factors for content-defined chunking (CDCs) to recognize chunk cutting and effective fingerprinting by bucket-based index partitioning as a main ingredient. To make chunking effective. | Data deduplication consists of three stages: chunking, creation of hash values, redundancy and removal. Our third finding shows that the comparison of hash values takes time with huge databases when matching fresh hash values with a very large number of hash values already stored. |

| | | |
|---|---|---|
| Liu, Q.et al. [6] | Scalable and stable cluster deduplication framework via Hadoop cloud computing platform | MapReduce regular mechanism is not optimal. MapReduce initiates several concurrent map operations to speed up the workflow. But after this step, the dispersed intermediate findings must be obtained and sorted before entering the reduction phase. |
| Shrivastava, A., & Tiwary, A.et al. [7] | Proposed modified hash value concept. | There are several cloud users that upload the same file from various users who use bandwidth and storage. |
| Yan, Z.et al. [11] | Deduplicated encoded data processed in the cloud depending on the challenge of possession and proxy encryption. It combines the deduplication of cloud data with access control. We assess its success on the basis of comprehensive research and machine simulation. | Data holders are difficult to handle deduplication for a variety of purposes. First, data holders can not often be accessible online or for other maintenance, which can lead to storage delays. Secondly, the deduplication of communications and computations may become too difficult to include data holders in the deduplication phase. |

***Table 1: Evaluation on various authors views.***

## IV. DISCUSSION

In the proposed scheme only one part, the server has been trusted for a small number of operations, which is why we term it semi-trustful. Since extra encryption has been implemented, the data is no longer susceptible to HDFS vulnerabilities. In fact, no variable can conduct dictionary attacks on data saved on the HDFS storage provider without the key material required for additional encryption. The server is a basic half-trusted component installed at the user's office and responsible for user authentication, access control and additional symmetric encryption. The main function of the server is to keep the hidden key used for more encryption safely. In a real case, this objective can be achieved easily utilizing a hardware security module. The server plays a more essential part as data is retrieved by a customer. Until sending data to a certain client, the server must check if block signatures match the recipient's public key. The Meta Data Manager (MM) and the HDFS storage provider have no trust in the security and cannot decode data contained in the cloud storage provider. We may not take account of components that may misbehave randomly but do not perform the tasks assigned to them.

## V. CONCLUSION

Replication is the safest option to optimize disc space and upload bandwidth by HDFS. The deduplication method aims to achieve efficiency, confidentiality and data safety. A variation of the two

approaches contributes to a higher deduplication and data efficiency. A successful deduplication ratio can also be obtained by more modifications of the deduplication algorithm. Some authors emphasized encrypted deduplication with safe and quick notebook backups. Today, vast amounts of business and confidential information have been stored on their personal computers and laptops. The problem is discontinuous and bad networking which often leaves your data vulnerable to robbery and often hardware failure. This paper has analyzed numerous methods and algorithms that are used to use different users' similar data to improve backup speed and reduce storage requirements. This algorithms offer excellent help for customer-end per-user encryption and keeping sensitive details private. The fundamental basis is that reliable deduplication services may be deployed with additional protection capabilities for both external and internal attackers through detecting masquerade activities. The dissuasive influence, the confusion of the intruder, and additional expenditures in preventing masquerade activity by reventing attackers play a very important role. A combination of these security features is therefore assumed to provide exceptional security choices for deduplication.

## REFERENCES

[1] Bhalerao, A., & Pawar, A. (2017). A survey: On data deduplication for efficiently utilizing cloud storage for big data backups. 2017 International Conference on Trends in Electronics and Informatics (ICEI). doi:10.1109/icoei.2017.8300844
[2] Cho, E. M., & Koshiba, T. (2017). Big Data Cloud Deduplication Based on Verifiable Hash Convergent Group Signcryption. 2017 IEEE Third International Conference on Big Data Computing Service and Applications (BigDataService). doi:10.1109/bigdataservice.2017.37
[3] Fu, Y., Xiao, N., Jiang, H., Hu, G., & Chen, W. (2017). Application-Aware Big Data Deduplication in Cloud Environment. IEEE Transactions on Cloud Computing, 1–1. doi:10.1109/tcc.2017.2710043
[4] Kumar, N., Rawat, R., & Jain, S. C. (2016). Bucket based data deduplication technique for big data storage system. 2016 5th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO). doi:10.1109/icrito.2016.7784963
[5] Kumar, N., Antwal, S., Samarthyam, G., & Jain, S. . (2017). Genetic optimized data deduplication for distributed big data storage systems. 2017 4th International Conference on Signal Processing, Computing and Control (ISPCC). doi:10.1109/ispcc.2017.8269581 [6] Liu, Q., Fu, Y., Ni, G., & Hou, R. (2016). Hadoop Based Scalable Cluster Deduplication for Big Data. 2016 IEEE 36th International Conference on Distributed Computing Systems Workshops (ICDCSW). doi:10.1109/icdcsw.2016.17
[7] Shrivastava, A., & Tiwary, A. (2018). A Big Data Deduplication Using HECC Based Encryption with Modified Hash Value in Cloud. 2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS). doi:10.1109/iccons.2018.8662984
[8] Saharan, S., Somani, G., Gupta, G., Verma, R., Gaur, M. S., & Buyya, R. (2020). QuickDedup: Efficient VM deduplication in cloud computing environments. Journal of Parallel and Distributed Computing. doi:10.1016/j.jpdc.2020.01.002
[9] Wen, M., Lu, K., Lei, J., Li, F., & Li, J. (2015). BDO-SD: An efficient scheme for big data outsourcing with secure deduplication. 2015 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS). doi:10.1109/infcomw.2015.7179387
[10] Xia, W., Feng, D., Jiang, H., Zhang, Y., Chang, V., & Zou, X. (2019). Accelerating content-defined-chunking based data deduplication by exploiting parallelism. Future Generation Computer Systems. doi:10.1016/j.future.2019.02.008
[11] Yan, Z., Ding, W., Yu, X., Zhu, H., & Deng, R. H. (2016). Deduplication on Encrypted Big Data in Cloud. IEEE Transactions on Big Data, 2(2), 138–150. doi:10.1109/tbdata.2016.2587659
[12] Zhenhua, L., Yaqian, K., Chen, L., & Yaqing, F. (2017). Hybrid cloud approach for block-level deduplication and searchable encryption in large universe. The Journal of China Universities of Posts and Telecommunications, 24(5), 23–34. doi:10.1016/s1005-8885(17)60230-9