International Conference on Engineering & Technology (ICET-2020)

Available at: International Journal of Scientific Research and Engineering Development

RESEARCH ARTICLE OPEN ACCESS

Data Deduplication System using Per File Video Audio Parity and FN Interpreter

Mr.R.Meyanand[1], M.E., T.Praveena[2],S.Krishnaveni[3],M.Santhiya[4]

[1]Assistant professor, [234] UG Students

[1234]Department of Computer Science, Selvam College of Technology, Namakkal.

[1]meyanand16@gmail.com [2]praveethangavel40@gmail.com

[3]venikumar7733@gmail.com [4]santhiyasandy47@gmail.com

--************************----------------------------------

Abstract:

Cloud computing is the conveyance of various administrations through the Internet. These assets incorporate apparatuses and applications like information, servers, databases, systems administration, and programming. At present, there is a great augmentation in the proportion of data set aside in limit organizations, close by exciting advance of frameworks organization procedures to overcome the issue of de-duplication in servers so that the storage space efficiency can be improved by removing the duplicated copies. This project presents a Per File Video Audio Parity deduplication file system (PFVAP) Technique where it decouples the information square and list by composing the location of information squares from document formula and list. It will subsequently abstain from getting the record on the read activity. For each exceptional information square, it allots an all-inclusive unique ID and it just requires one plate access to get the comparing information square reference check utilizing the global ID. In order to guarantee the write performance, it employs finer granularity lock to optimize the block flushing strategy for the write buffer. The decoupled de-duplicated data chunks are combined by using chunking process which will give equality repetition assurance to all documents by intra-record recuperation and a more elevated level assurance for information pieces with high reference checks between the documents

Keywords —Deduplication,data chunk, reliability, decouple .

--************************----------------------------------

1. INTRODUCTION

Cloud computing has grown rapidly and gained considerable attention since it provides flexibility and scalability to organizations. Cloud computing is a large-scale distributed system which offers a pool of computing resources to cloud consumers through the internet. There are many cloud providers which run on cloud computing environment such as Amazon, Google Engine, IBM, and Microsoft. They provide services and resources to users on the basis of pay per use at anytime from anywhere. Cloud computing offers three main delivery models which are Software as a Service (SaaS), Platform as a Service (PaaS), and Infrastructure as a Service (IaaS). In Software as a Service (SaaS), Applications and access management tools are provided to users. Platform as a Service (PaaS) provides tools such as operating systems, databases, and network so consumers can install and develop their own software and applications. Infrastructure as a Service (IaaS) provides access to physical devices such as hardware and network so consumers can install and develop their own operating systems and applications. Software as a Service (SaaS) is a method of delivering software applications over interconnected network i.e. internet on the subscription basis. With SaaS, Cloud Service providers host and manage the software applications and underlying infrastructure with a service agreement. They assure the availability and security of both the data and the application. SaaS provides a complete software solution which you purchase on a pay-as-you-go basis from a cloud service provider. SaaS enables organizations to rise up quickly and go with the applications with minimum initial cost.

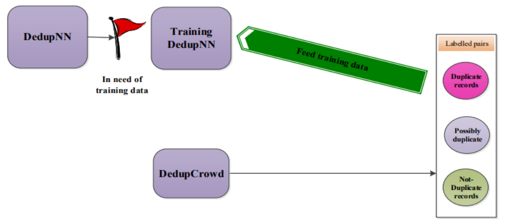


Figure 1. Working of *DedupCrowd* and how it provides training data for automated duplication detection

Shown in Figure 1, the output of DedupCrowd labels pairs of customer data as either duplicate, possibly duplicate or not duplicate. This knowledge not only provides an output to the process of deduplication by an in-house expert but also collects data that is used to train an automated model of deduplication (termed as DedupNN). This is important as in the real-world manual deduplication of a database is impractical and an automated version of it is needed based on the data that has been annotated by DedupCrowd using the process of crowd sourcing. One of the main obstacles in a human based computing approach is the poor performance of the crowd (CSRs or workers) and how it can be addressed. The researchers in the area of machine learning, and statistics and databases contributed significantly in developing techniques to approximate the error of workers. However, a practical framework for the online evaluation and eviction of poor workers from the crowd sourcing process is lacking. DedupCrowd addresses this shortage and makes use of the workers' estimated errors through a statistical quality control (SQC) approach to evict poorly performing workers from the crowd sourcing process. Statistical quality control has been used successfully in the area of manufacturing process quality control. The output of this module provides the status of the crowd sourcing workers, which determines whether a given worker can continue to participate as part of the crowd or not.

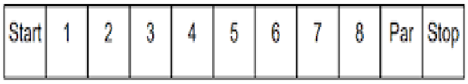
2. PROPOSED ALGORITHM

Improve the reliability of deduplication based storage systems using Per File Video Audio Parity (PFVAP) Scheme. In this techniques identify the audio video files deduplication by using inter file and intra file recovery .Inter file within single system recovery. Intra file within the two or more system recovery. The video audio files are divided into frames by using RAID scheme. The divided frames are comparing to identify the duplicated files by the help of proposed algorithms. The deduplicated data chunks are merged by using chunking process. The user can access the files through the indexer.

2.1 PROPOSED SYSTEM

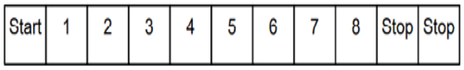
To access deduplication of audio video files remove the same content but different file name. In this proposed scheme using two algorithms such as,

Per Frame Parity- The video files are divided into the data blocks to compare the frame by frame identify the duplicated files.



Start	1	2	3	4	5	6	7	8	Par	Stop

Fig 2.1 With Parity check.



Start	1	2	3	4	5	6	7	8	Stop	Stop

Fig 2.2 Without Parity check

Per Analog Parity- The duplicated audio files are consider as a signals, the digital signal converted into analog signal to compare the same by using block level comparison. Both audio and video files recovery the deduplicated data chunks are merged by using chunking process.

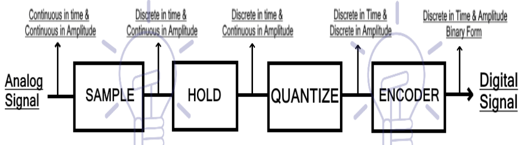


Fig 2.3.Analog to Digital converter

2.2 ADVANTAGES OF PROPOSED SYSTEM

- Faster recovery time objectives.
- Reduced tape backups.
- Enable Data Deduplication on existing file system.

3. RELATED WORKS:

3.1. DATA BLOCKS RAID SYSTEM

In traditional deduplication systems, most data blocks from previous backup have a slight modification and the modification is normally confined to some specific areas. These deduplication systems normally use RAID as storage backend and frontend. Therefore, most data deduplication systems put continuous unique data blocks and corresponding meta data into container, which is a read/write unit used to maintain locality and take the advantages of RAID bandwidth as much as possible. The restoring operations can be accelerated by parallel reading data blocks from container in RAID since most adjacent data blocks are stored in the same container. The most important point is that the container is immutable once written, and it will be recycled until there is no valid data block. We describe the advantages of this scheme used in LDFS: 1) it supports fast read any bit in any file without reading not used data blocks 2) it can save disk bandwidth 3) it benefits garbage collection

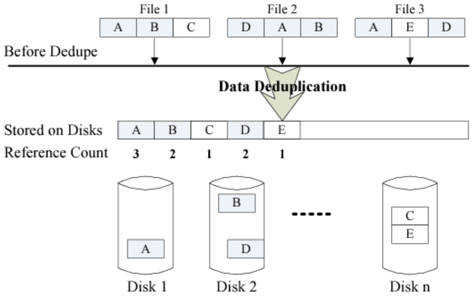


Fig 3.1.1 Data deduplication in storage system.

3.2. DEDUPLICATION THROUGHPUT

Research vendors mainly focus on alleviating disk bottle-neck to improve deduplication throughput. DDFS is the first system that exploits data duplicate locality to alleviate the disk bottleneck under limited RAM capacity. Other systems, such as Sparse Indexing and Foundation, Chunk Stash, and other literature, also take the advantage of the duplicate locality by indexing only a subset of chunks in RAM to relieve disk bottleneck. It is a deduplication approach thus store in a bin that are exploits data similarity, rather than the data duplicate locality, to improve the deduplication throughput for low-locality workloads. Silo alleviates the disk bottleneck by exploiting both data duplicate locality and similarity that have been used in DDFS and Extreme Binning under limited RAM usage.

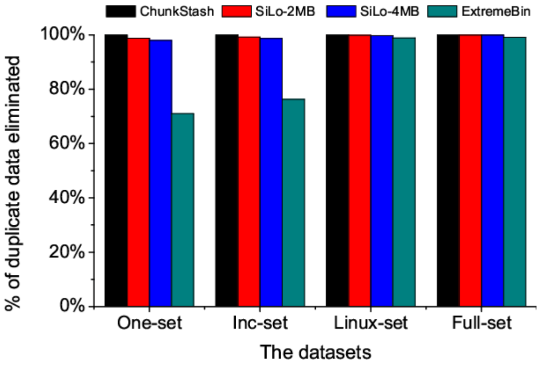


Fig 3.2.1 Comparision between dataset and duplicated data eliminated

3.3. DATA RELIABILITY

There are a few researchers that have addressed the data reliability in deduplication storage systems. It proposes to build a deduplication archival storage system that provides high data reliability with minim data Redundancy. It chooses a replication level for every chunk and the replication level is a function of the amount of data that would be lost if the chunk were lost. R-ADMAD is another effective method that packs variable-length data chunks into fixed sized objects, and exploits ECC codes to encode the objects and distributes them among storage nodes in a redundancy group, thus to improve the data reliability. It have analyzed the deduplicated storage and erasure-coded storage, and further proposes a method for system designers to determine when the data deduplication will save space and also improve the data reliability by erasure-coding.

3.4. COMPRESSION RATIO

The compression ratio varies with different deduplication approaches. The early storage systems remove the redundant data at file level. It is the first archival storage system that removes redundant data at chunk level. firstly explores the variable sized chunks instead of fixed-size ones to find much more redundant data. Other methods, such as finger diff can provide more flexibility on the variability of the chunk sizes to get higher compression ratios.

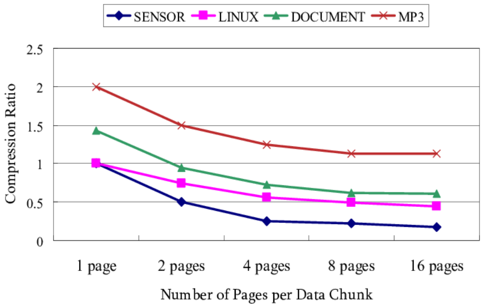


Fig 3.3.1.Data chunks compression ratio

4. CONCLUSION

Data deduplication can be divided into two parts: deduplication process used for removing redundant data and the data layout that represents an overview of data placement on the back-end disk storage. So far, most of researchers prefer focusing on deduplication process since it is closely related to compression ratio and deduplication throughput performances, while paying less attention to data layout. In this project, we propose a per-file-video-audio-parity (PFVAP)scheme to improve the reliability of deduplication-based storage systems. PFP computes the parity for each parity group of N chunks (N−1 data chunks and 1 parity chunk, where N is configurable) within each file before the file is deduplicated. Therefore, PFVAP can provide redundancy protection for all files by intra-file recovery as well as a higher level of protection for data chunks with high reference counts, critical data chunks, by inter-file recovery. The reliability analysis and evaluations show that PFP can provide better reliability.

FUTURE ENHANCEMENT

Later on we intend to convey and test the proposed arrangement and assess the common sense of the thought of notoriety and whether the severe well known/disagreeable grouping can be made all the more fine-grained. Additionally, we intend to expel the supposition of a confided in ordering support and investigate various methods for making sure about the lists of disagreeable documents

REFERENCES

[1] Suzhen Wu, Huagao Luan, Bo Mao, Hong Jiang, Gen Niu, Hui Rao, Fang Yu, Jindong Zhou"Improving reliability based storage system using per file parity",vol 31.no.3 March 2019.

[2] Jibin Wang, Zhigang Zhao, Zhaogang Xu, Hu Zhang, Liang Li, and Ying Guo" I-sieve: An Inline High

Performance Deduplication System Used in Cloud Storage"Vol 20 .no.3 Feb 2015.

[3] Jiansheng Wei, Member, IEEE, Hong Jiang, Senior,Ke Zhou, Member, IEEE, and Dan Feng, Member,IEEE"Efficiently Representing Membership for Variable Large Data Sets"Vol 25 no.4 Apr 2015.

[4] Li, Jin, Xiaofeng Chen, Mingqiang Li, Jingwei Li, Patrick PC Lee, and Wenjing Lou."Secure deduplication with efficient and reliable convergent key management" IEEE transactions on parallel and distributed systems 25, no. 6 (2014): 1615- 1625.

[5] Li, Jin, Yan Kit Li, Xiaofeng Chen, Patrick PC Lee, and Wenjing Lou. "A hybrid cloud approach for secure authorized deduplication" IEEE Transactions on Parallel and Distributed Systems 26, no. 5 (2015): 1206-1216.

[6]Puzio, Pasquale, RefikMolva, MelekOnen, and Sergio Loureiro. "ClouDedup: secure deduplication with encrypted data for cloud storage" In Cloud Computing Technology and Science (CloudCom), 2013 IEEE 5th International Conference on, vol. 1, pp. 363-370. IEEE, 2013.

[7]Liu, Jian, N. Asokan, and Benny Pinkas."Secure deduplication of encrypted data without additional independent servers" In Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security, pp. 874-885. ACM, 2015.

[8]Li, Jin, Xiaofeng Chen, Xinyi Huang, Shaohua Tang, Yang Xiang, Mohammad Mehedi Hassan, and Abdulhameed Alelaiwi."Secure distributed deduplication systems with improved reliability" IEEE Transactions on Computers 64, no. 12 (2015): 3569-3579.

[9]L.N.Bairavasundaram,G.R.Goodson,S. Pasupathy, and J. Schindler.An Analysis of Latent Sector Errors in Disk Drives. In SIGMETRICS'07,Jun. 2007.

[10]D. Bhagwat, K. Pollack, D. Long, T. Schwarz, E. Miller, and J.-F. P ˆaris.Providing High Reliability in a Minimum Redundancy Archival Storage System. In MASCOTS'06, Sept. 2