International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

COPY RIGHT

ELSEVIER SSRN

IJIEMR Transactions, online available on 27th Dec2020. Link
:http://www.ijiemr.org/downloads.php?vol=Volume-09&issue=ISSUE-12

DOI: 10.48047/IJIEMR/V09/I12/114

Title: **STRENGTHNING THE PRODUCTIVITY OF STORAGE FOR BIG DATA STORAGE SYSTEMS USING DISTRIBUTED DEDUPLICATION.**
Volume 09, Issue 12, Pages: 691-694
Paper Authors

Mohd Akbar, Dr. Thirupathi Regula, Irshad Ahmad

USE THIS BARCODE TO ACCESS YOUR ONLINE PAPER

To Secure Your Paper As Per UGC Guidelines We Are Providing A Electronic Bar Code

International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

STRENGTHNING THE PRODUCTIVITY OF STORAGE FOR BIG DATA STORAGE SYSTEMS USING DISTRIBUTED DEDUPLICATION

Mohd Akbar*, Dr. Thirupathi Regula*, Irshad Ahmad*
*Lecturer, University of Technology and Applied Sciences, Muscat, Sultanate of Oman.
akb.mtech@gmail.com, regulathirupathi@gmail.com

Abstract— Cloud storage is one of the key features of cloud computing, which helps cloud users outsource large numbers of data without upgrading their devices. However, Cloud Service Providers (CSPs) data storage faces problems with data redundancy. The data deduplication technique aims at eliminating redundant information segments and maintains one single instance of the data set, even if any number of users own similar data set. Since blocks of data are spread on many servers, each block of the file has to be downloaded before restoring the file to decrease system output.
We suggest a cloud storage server data recovery module to improve file access efficiency and reduce time spent on network bandwidth. Device coding is used in the suggested method to store blocks in distributed cloud storage, and for data integrity, MD5 (Message Digest 5) is used. Running recovery algorithm helps the user to retrieve a file directly from the cloud servers without downloading every block. The scheme proposed improves system time efficiency and the ability to access the stored data quickly. This reduces bandwidth consumption and reduces overhead user processing while downloading the data file.
Keywords: Access Efficiency, Data Duplication, Cloud Storage, Bandwidth Network, Recovery Modulus.

INTRODUCTION

Compression of data reduces data sizes and minimizes data intrusion. Data compression is a classical control field that has made its mark because it requires large storage space. The compression of data ensures less storage space usage but also improves performance. File compression for various file types and sizes can be implemented. The ability to provide ongoing services (portability) and the high degree of scalability for user demands has become more popular with cloud computing. Cloud services often support all the data (user / system data) available in order to be reliable in operation. However, this causes a problem that storage space runs out. Data files of different sizes and multimedia files are often large (few megabytes to more than one gigabyte). To the cloud service provider, this becomes a challenge.

Data Duplication can reduce the overhead of data logistics due to large files. Data Duplication The main focus of Cloud Computing is always the use of most resources, but it always has a cost that grows very quickly every day. User demand can not be lowered and/or the cost of buying additional services can be reduced. Data compression will allow large-scale storage files to be uploaded to a smaller size, ensuring fewer resources can be purchased and maintained during that time. Redundancies in data across files / users / data blocks are being removed by deduction. It is only possible to remove the redundancy and store that piece of data once instead of "n" times within the data. The original file can be restored at any time by restoring the redundant data to its original position.

This means less space is needed for storing data and maximizing bandwidth. The quality and accuracy of data would also be improved across various geographic locations. Divided files into pieces can lead to many security problems such as the determination of master file owner, misrepresentation of file chunk etc. In a distributed environment , data deduction can be implemented. The distributed system may also give way with certain constraints for user-level data deduplication. Documents, file forms and essential application / users used for documents can be included in the restrictions.

LITERATURE REVIEW

Biggar H (2012) Through the regional availability of data, the cloud services provide consumers with an impeccable service. Increasing data availability leads to significant redundancies and resources required to store

these data. The techniques of compression of data will reduce the space required to store the data at various locations. Compression of data ensures that the accuracy and quality at every location is not compromised. As there is an overwhelming demand for cloud services and storage, the investment is also growing. By using data compression we can reduce the amount of investment needed and reduce the amount of data storage available in the physical space and data centers. A variety of safety protocols may be used to ensure that these compressed files are protected at different locations. We provide a trustworthy system for safe storage of deduplicates and their management to achieve high coherence and availability.

Castiglione A., Pizzolante R., De Santis A., Carpentieri B., Castiglione A., and Palmieri F. (2014) Data deduction, an effective way to reduce data, has become increasingly popular and attracted attention by the exponential growth of digital data in large-scale storage systems. It eliminates redundant information on file or sub file level and identifies duplicate content using its crippling secure hash signature, which shows that it is much more efficient computing than the conventional approach to compression in larger storage systems. It is a fingerprint-resistant fingerprint. In this work, we first review the context and important features of data deduplication, then compile and classify state-of-the-art data deduplication research according to the key data deduplication process workflow. The summary and taxonomy of the deduplication state of the art help to identify and understand the main concepts for the data deduplication systems.

Chu X., Ilyas I., and Koutris P (2016) This work shows that the cloud computing platforms provide a duplicate-free storage system. Our cloud-based deduplication storage system manages data and replication consists of two main elements, a first-end deduplication program and a back-end mass storage system. Hadoop distributed file system (HDFS) is a typical cloud storage file system used for Hadoop (HBase) databases. We use HDFS as a method for mass storage and use HBase to set up a fast framework for indexing.

METHODOLOGY

Because it takes longer to decide the same work, the performance rate is also lower. The second method of deducing data in a fixed block size will provide a better success rate and take less time than file level, but will not save high data. The third method for chunking a variable block size will save you the most but take a longer time than chunking a defined block size. For variable size chunking, we used the Rabin-Karp algorithm. For small files such as audio files , images and small-size documents of up to 10-15 MB we use file-level data deduplication. File larger than films, documents or executable files that are seldom duplicated.

1. Check for deductibility of file.
2. Search for the deduplication of a fixed size block (file included).
3. Search the Rabin-Karp algorithm for duplicate chunks (across files).
4. To isolate duplicate elements, choose the best of three algorithms above.
5. Filter all duplicated elements separately from the unique instances.
6. Store the duplicated files and blocks in these unique instances.

Recovery Step of Record

In general, the user gives a command to get the file via the application interface. This allows the deduplication engine to evaluate the file and verify that it has been separated into parts. In relation to their file type and suffix the user data is separated into various folders. For example if the user wants to retrieve a file names “Test.mp3” with the user ID “U123”, the lookup would be directory U123 inside which should contain a directory mp3 and inside which should contain the master file “Test.mp3” or should contain chunks of the files with the mapping information.

RESULTS

1. File Downloading Time

File downloading time is defined as the amount of time taken to download a file.

International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

2. Bandwidth Utilization Time

It's the time of the transaction with the network bandwidth. The user sends a request with the tag value to all servers, and all blocks from various servers are received by the user, since the transfer time in the current system is high. Whereas only the tag value of the file to the metadata server is sent by the system user proposed. For the correct tag value of the file, the metadata server will retrieve a number of blocks from the servers, run the recover algorithm and return the file to the user. The user receives the file in the single request when he runs a recovery algorithm on the metadata server for creating the original file.

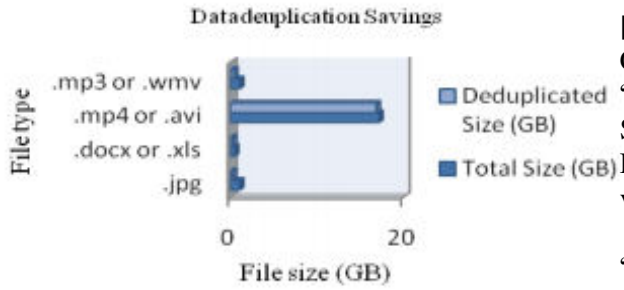

Figure: Results graph.

The time of download is reduced and the bandwidth is used for a less time by decreasing the bandwidth use. Downloading files by DDS follows methods for downloading block files and the download technique for DDEAS uses a file level approach that helps to transfer the file in less time. The existing process takes more time to download the file compared with the proposed method, consuming more bandwidth for the Network. In order to address this drawback, the metadata serer is used in the proposed framework.

CONCLUSION

The findings show that we can save ~11% of the personal details on the screen. Unlike other compression algorithms, the deduction does not perform complicated mathematics. Over a large amount of data the method is fast and reliable. The reliability of the system is only improved by increasing the size of the data. Data deduplication is also secure because the underlying protection algorithm / protocol is not interfered with. If you can properly fit into and stick to the security protocol. No authorization requirements are violated, since only the personal information of the user is used for the deduction. Overtime functions and not because the data is transferred to the storage unit. Because we have used the single user system and the data deduction process overtime. This ensures a smooth and stable data recovery process.

REFERENCES

[1] Biggar H., "Experiencing Data De-Duplication: Improving Efficiency and Reducing Capacity Requirements," The Enterprise Strategy Group, pp. 902-906, 2012.

[2] Castiglione A., Pizzolante R., De Santis A., Carpentieri B., Castiglione A., and Palmieri F., "Cloud-Based Adaptive Compression and Secure Management Services for 3D Healthcare Data," Future Generation Computer Systems, vol. 43-44, pp. 120-134, 2014.

[3] Chu X., Ilyas I., and Koutris P., "Distributed Data Deduplication," Proceedings of the VLDB Endowment, vol. 9, no. 11, pp. 864-875, 2016.

[4] Dolan M., Kochan L., Ram T., Rohr S., Tu K., and Miller S., Patent No. US20160292048, Retrieved from https://www.google.com/patents/US20160292048, Data Deduplication Using Chunk Files, Google Patent, Last Visited, 2016.

[5] Douceur J., Adya A., Bolosky W., Simon D., and Theimer M., "Reclaiming Space from Duplicate _Les in A Serverless Distributed _Le System," in Proceedings of 22nd International Conference on Distributed Computing Systems, Vienna, pp. 617-624, 2002.

[6] Demystifying Data Reduplication: Choosing the Best Solution, FalconStor Software, White Work Dynamic Solutions International, https://www.varinsights.com/doc/demystifyingdata-deduplication-choosing-0002, Last Visited, 2017.

[7] Eastlake D. Jones P., White work: Description of SHA-1, http://tools.ietf.org/html/rfc3174, Last Visited, 2017.

[8] Estes J., Patent No. US20140258245, Retrieved from

https://www.google.ch/patents/US20140258245 , Efficient Data Deduplication, Last Visited, 2014.

[9] Harnik D., Pinkas B., and Shulman-Peleg A., "Side Channels in Cloud Services, the Case of Deduplication in Cloud Storage," IEEE Security and Privacy Magazine, vol. 8, no. 6, pp. 40-47, 2010.