ISSN: 2350-0328

International Journal of Advanced Research in Science, Engineering and Technology

Vol. 4, Issue 6 , June 2017

An Efficient Data Deduplication in Cloud Environment with Improved Reliability

Namrata Kawtikwar, Prof. M.R. Joshi

P.G. Student, Department of Computer Science & Information Technology , HVPM COET Amravati, Maharashtra,
Assistant Professor, Department of Computer Science& Information Technology , HVPM COET Amravati, Maharashtra

ABSTRACT: Day by day the use of memory is increases rapidly. The process of eliminating the repeated or duplicates copies of data is called as Data deduplication. This data deduplication process is widely used in cloud storage to decrease storage space and upload bandwidth. By using, deduplication system progress of storage utilization and reliability is increases. In addition, the dare of privacy for sensitive data also take place when they are outsourced by users to cloud. Planning to address the above security test, this paper constructs the idea of deduplication system. The paper recommends a distributed deduplication systems with two methods i.e. File level and the block level. A deduplication technique, on the other hand, can reduce the storage cost at the server side and save the upload bandwidth at the user side. Deduplication has received much attention from both academic and industry because it can more improves storage utilization and save storage space, especially for the applications with high deduplication ratio such as accession storage systems.

I. INTRODUCTION

Now days with the huge increasing of population and the using of technology, it leads to many problems. The growth in technology is increasing the amount of storage or communication and technique devices. By the unpredictable development of digital data, deduplication techniques are broadly engaged to backup data and decrease network and storage transparency by notice and eradicate redundancy among data. As an alternative of maintaining multiple data copies with the same content, deduplication reducing redundant data by maintaining only single copy and referring other redundant data to that copy. Deduplication has inward much concentration from both academic world and industry since it can really recover storage utilization and keep storage space, particularly for the applications with high deduplication ratio such as archival storage systems. A number of deduplication systems have been projected based on various deduplication scheme such as client-side or server-side deduplication, file-level or block-level deduplications. Specially, with the advent of cloud storage, data deduplication procedure grows to be more gorgeous and essential for the management of ever-increasing quantity of data in cloud storage services. Deduplication has received much attention from both academic and industry because it can more improves storage utilization and save storage space, especially for the applications with high deduplication ratio such as accession storage systems. For eliminating duplicate copies of data we use data deduplication technique. To reduce storage space and for uploading bandwidth mostly it has been used.

II. LITERATURE REVIEW

[Yinjin Fu, Hong Jiang 2014] proposes ALG-Dedupe, an application aware local-global source-deduplication scheme for cloud backup in the personal computing environment to improve deduplication efficiency. An intelligent deduplication strategy in ALG-Dedupe is designed to exploit file semantics to minimize computational overhead and maximize deduplication effectiveness using application awareness. It combines local deduplication and global deduplication to balance the effectiveness and latency of deduplication. The proposed application-aware index structure can significantly relieve the disk index lookup bottleneck by dividing a central index into many independent small indices to optimize lookup performance.

[M. Bellare, S. Keelveedhi, and T. Ristenpart 2013] introduced the idea of security and scheme for symmetric encryption in concentrate security framework. They give different idea of security and analyze the good involution of reduction among them. They provide method of encryption using a block cipher, cipher block chaining and counter

ISSN: 2350-0328
International Journal of Advanced Research in Science, Engineering and Technology
Vol. 4, Issue 6 , June 2017

mode. They had two goals .First is to study the idea of security for symmetrical encryption and second is to provide concrete security analysis of fixed symmetric encryption device.

[P. Anderson, L. Zhang et al. 2010] proposed a solution here the data which is common between users to increase the speed of backup and reduce the storage requirement namely backup algorithm. Supports client-end per user encryption is necessary for confidential personal data. This provides the potential to significantly decrease backup times and storage requirement. Storing huge amount of data in personal computer or laptops causes poor connectivity also may be theft due to hardware failure. However Network bandwidth can be a bottle-neck and Backing up directly to a cloud can be very costly are not addressed. Conventional backup solutions are not well suited to this environment. So client side deduplication necessary for confidential personal data

III. EXISTING SYSTEM

Data de-duplication techniques are very interesting techniques that are widely employed for data backup in enterprise environments to minimize network and storage overhead by detecting and eliminating redundancy among data blocks. There are many de-duplication schemes proposed by the research community. The reliability in de-duplication has also been addressed. However, all of these works have not considered and achieved the tag consistency and integrity in the construction.
There are some disadvantages of the existing system like:

- The traditional deduplication methods cannot be directly extended and applied in distributed system.
- Data reliability is actually a very critical issue in a deduplication storage system because there is only one copy for each file stored in the server shared by all the owners.
- Most of the deduplication system performs only the file level deduplication. No block level deduplication is performed there.
- Furthermore the challenges for data privacy also arise as more and more sensitive data are being outsourced by user to cloud.

IV. PROPOSED SYSTEM

Two kind's entities will be involved in this deduplication system, including the user and the storage cloud service provider (S-CSP).
User: The user is an entity that wants to outsource data storage to the S-CSP and access the data later. In a storage system supporting deduplication, the user only uploads unique data but does not upload any duplicate data to save the upload bandwidth. Furthermore, the fault tolerance is required by users in the system to provide higher reliability.

S-CSP: The S-CSP is an entity that provides the outsourcing data storage service for the users. In the deduplication system, when users own and store the same content, the S-CSP will only store a single copy of these files and retain only unique data. A deduplication technique, on the other hand, can reduce the storage cost at the server side and save the upload bandwidth at the user side. For fault tolerance and confidentiality of data storage, we consider a quorum of S-CSPs, each being an independent entity. The user data is distributed across multiple S-CSPs.

Though deduplication technique can save the storage space for the cloud storage service providers, it reduces the reliability of the system. Data reliability is actually a very critical issue in a deduplication storage system because there is only one copy for each file stored in the server shared by all the owners. If such a shared file/chunk was lost, a disproportionately large amount of data becomes inaccessible because of the unavailability of all the files that share this file/chunk. If the value of a chunk were measured in terms of the amount of file data that would be lost in case of losing a single chunk, then the amount of user data lost when a chunk in the storage system is corrupted grows with the number of the commonality of the chunk. Thus, how to guarantee high data reliability in deduplication system is a critical problem. However, as lots of deduplication systems and cloud storage systems are intended by users and applications for higher reliability, especially in archival storage systems where data are critical and should be preserved over long time periods. This requires that the deduplication storage systems provide reliability comparable to other high-available systems.
Here, the main motives of the proposed systems are-

- To authenticate a data and to make available integrity and validity assurances on the data.

ISSN: 2350-0328
International Journal of Advanced Research in Science, Engineering and Technology
Vol. 4, Issue 6 , June 2017

- To Distributed Deduplication System.
- To implement Block-level Distributed Deduplication System.
- To maintain the consistency and integrity of data within file

V. PROPOSED METHODOLOGY

There are two types of deduplication in terms of the size: (i) file-level deduplication, which discovers redundancies between different files and removes these redundancies to reduce capacity demands, and (ii) block level deduplication, which discovers and removes redundancies between data blocks. The file can be divided into smaller fixed-size or variable-size blocks. Using fixedsize blocks simplifies the computations of block boundaries, while using variable-size blocks provides better deduplication efficiency.

A. File-Level Distributed Deduplication System

A file is nothing but a data unit. While investigating duplication in file, it characteristically uses hash function and generates hash value for file which treated as file identifier. In case that more than one file has the same hash value, they are considered to have the similar contents and only one unique file of these files will be stored or saved in Storage.

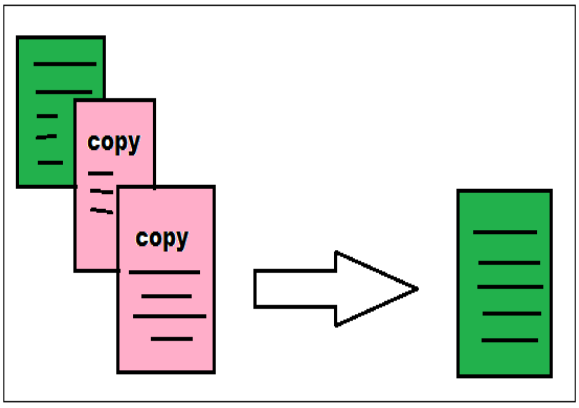


Fig 1: File Level Deduplication

It support capable duplicate check, tags for each file will be calculated and send to storage cloud service provider. To prevent alignment invasion organized by the cloud based service provider.

ISSN: 2350-0328
International Journal of Advanced Research in Science, Engineering and Technology
Vol. 4, Issue 6 , June 2017

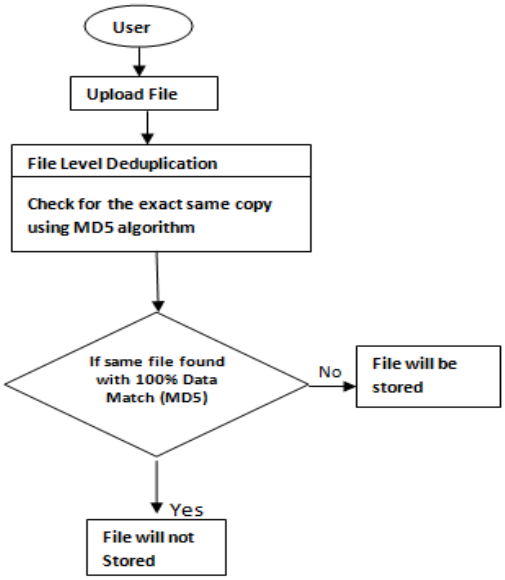


Fig 2: File Level deduplication Process

Above figure shows the File level deduplication process. Here we are checking for the exact same copy of the file. If the user uploads the new file and if that file is already present on the server then server will give the message that; the "file is already exist" and will not save the file again.

B. Block-Level Deduplication System

File level dedupliction is just not enough. Here we are doing the block level deduplication where we will check the content of the files. We will check the similar data content with the already stored file.

In this part, we appear how to derive the fine grained block level distributed deduplication. In this kind of Deduplication, it divides entire file into numerous fix size chunks or variable size chunks then it calculates a hash value for each of the chunk for investigating duplication blocks. Block-level Deduplication reduces duplicate chunks of files that may occur in non- alike files. Here, the client also demands to perform the file level deduplication before uploading file. The user partition this files into blocks, if no duplication is found and performs block-level deduplication system. The system set up is similar to file-level deduplication and also block size parameter will be defined.

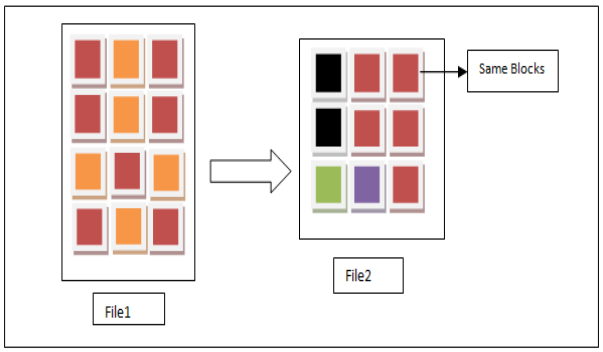


Fig 3 : Block Level Deduplication

ISSN: 2350-0328
International Journal of Advanced Research in Science, Engineering and Technology
Vol. 4, Issue 6 , June 2017

Above figure shows about the Block Level deduplication. Here file1 is the file which is already stored on the server and the file2 is the new uploaded file. File2 contains the similar blocks of data as in the file1. By performing the block level deduplication process we will come to know that how many percent of data is similar with the stored file.

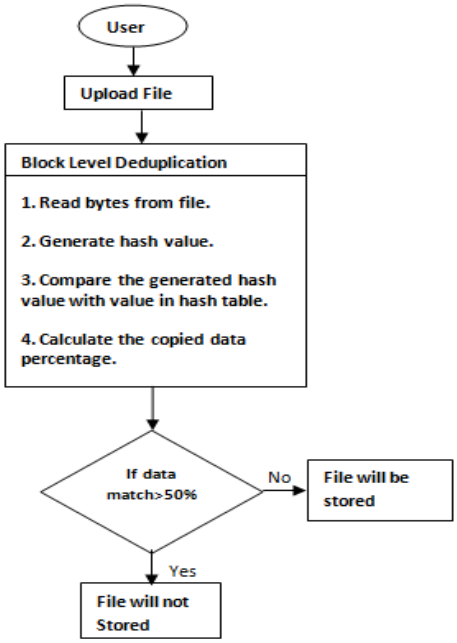


Fig 4: Block Level deduplication Process

Above figure shows the Block level deduplication process.

C. Implementation Of Deduplication Algorithm

In Block-level deduplication we parse the data in terms of blocks.

1. Start
2. Declare Variable
3. Initialize variable
4. Read bytes from file in tone iteration
5. Read from file until reach EOF
   5.1 Generate Hash Value from strBuff[BLOCKSIZE]
   5.2 if (FirstBlock)
   Consider node as root element
   IncBlockCtr
   else
   search the generated Hash in Search Table
   if (Find Hash == True)
   Compute the Node
   Add the Node to a linked List
   Change the EndLink of SLL
   else

ISSN: 2350-0328
International Journal of Advanced Research in Science, Engineering and Technology
Vol. 4, Issue 6 , June 2017

Add the node in Search Table
IncTheBlockCounter
6. Calculate Deduplication Ratio
7. Display the Result for each iteration
8. END

Description:

1. As the loop runs it reads bytes of data in one iteration and generate the hash for the same.
2. Search() procedure ensures that entry is not already present in search table.
3. If the hash already exist
   - Compute the node
   - Add the node to Single linked list
4. If hash doesn't exist
   - Compute the node
   - Use hash as a key for comparison
   - Add the node in Search Table
5. Calculate deduplication percentage as a measurement also known as Deduplication ratio.

Thus we can find out the duplication of data in the uploaded file before storage.

VI. RESULT AND DISCUSSION

In this system, user can upload the file, search file, shared file, download file. The main task of the system is to check for the duplicates. In this system we are having two options for the deduplication one is File level deduplication and another one is Block level deduplication. In the following section we can see the time taken for the both deduplication types.

A. File level deduplication

Following table shows the computational time for the file level deduplication approach. This result is calculated by taking the number of files and their processing timing.

No. of Files	Computational Time for file level approach
5	2.6
7	2.9
9	3.1
11	3.4

Table 1: Computational time for file level approach

As shown in above table we have calculated the computational timing when number of files stored 5,7, 9, 11 respectively.

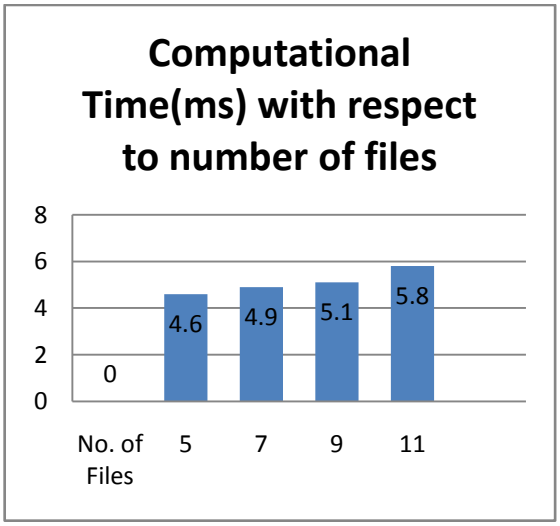
B. Block level deduplication

Following table shows the computational time for the block level deduplication approach. This result is calculated by taking the number of files and their processing timing.

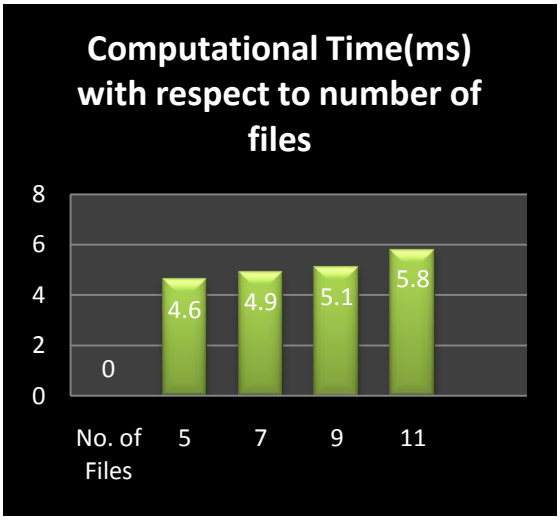
ISSN: 2350-0328
International Journal of Advanced Research in Science, Engineering and Technology
Vol. 4, Issue 6 , June 2017

No. of Files	Computational Time for Block level approach
5	4.6
7	4.9
9	5.1
11	5.8

Table 2: Computational time for Block level approach



Graph (a): File Level Analysis



Graph (b): Block Level Analysis

In the above graphs Graph (a) shows that the x axis indicates the number of files on which the file level deduplication performed. The y axis indicated the total computational time in ms to perform file level deduplication by using MD5 algorithm. And in Graph(b) the x axis indicates the number of files on which the block level deduplication performed. The y axis indicated the total computational time in ms to perform block level deduplication.

VII. ADVANTAGES

- This application helps in easy maintenance of data on the cloud platform so that no duplicate files are saved in the cloud which helps to save the storage space.
- Proposed system provides authentication and integrity and validity assurances on the data.
- The proposed constructions support both file-level and block-level deduplication.

ISSN: 2350-0328

International Journal of Advanced Research in Science, Engineering and Technology

Vol. 4, Issue 6 , June 2017

VIII. CONCLUSION AND FUTURE SCOPE

A. Conclusion

Design of an improved technique for storage in Cloud is deduplication technique. Deduplication aids in saving the storage space. This application helps in easy maintenance of data on the cloud platform so that no duplicate files are saved in the Cloud. With the evolution of Cloud computing, storage resources of commodity machines can be efficiently utilized. This allows every organization to build its own private cloud for a variety of purposes. In order to better utilize the limited storage available in a private cloud, a suitable approach for optimization has to be used.

B. Future Scope

With the evolution of Cloud computing, storage resources of commodity machines can be efficiently utilized. This allows every organization to build its own private cloud for a variety of purposes, but there are some features that are not implemented in this system but that can be implemented in future. These type of features are mention below:

- Currently, improved technique for storage has been tested only for text files. In future, it can be further extended to support files of other types such as audio files and video files.
- This system is work on plane text files it means that this system will not work on the encrypted files. If user uploads the encrypted file then this system will not work for it. But it can be implemented in future as complex logic is required for it.

REFERENCES

[1] M. Bellare, S. Keelveedhi, and T. Ristenpart, "Dupless: Server aided encryption for deduplicated storage," in USENIX Security Symposium, 2013

[2] M. W. Storer, K. Greenan, D. D. E. Long, and E. L. Miller, "Secure data deduplication," in Proc. Of StorageSS, 2008.

[3] P. Anderson and L. Zhang, "Fast and secure laptop backups with encrypted de- duplication," inProc. of USENIX LISA, 2010

[4] J. S. Plank, S. Simmerman, and C. D. Schuman, "Jerasure: Alibrary in C/C++ facilitating erasure coding for storage applications- Version 1.2," University of Tennessee, Tech. Rep. CS-08-627,August 2008.

[5] J. S. Plank and L. Xu, "Optimizing Cauchy Reed-solomon Codesfor fault-tolerant network storage applications," in NCA-06: 5th IEEE International Symposium on Network Computing Applications,Cambridge, MA, July 2006.

[6] C. Liu, Y. Gu, L. Sun, B. Yan, and D. Wang, "R-admad: High reliability provision for large-scale de-duplication archival storage systems," in Proceedings of the 23rd international conference on Supercomputing, pp. 370–379.

[7] M. Li, C. Qin, P. P. C. Lee, and J. Li, "Convergent dispersal:Toward storage-efficient security in a cloud-of-clouds," in The 6thUSENIX Workshop on Hot Topics in Storage and File Systems, 2014.

[8] P. Anderson and L. Zhang, "Fast and secure laptop backups with encrypted de-duplication," in Proc. of USENIX LISA, 2010.

[9] Z. Wilcox-O'Hearn and B. Warner, "Tahoe: the least-authorityfilesystem," in Proc. of ACM StorageSS, 2008.

[10] A. Rahumed, H. C. H. Chen, Y. Tang, P. P. C. Lee, and J. C. S.Lui, "A secure cloud backup system with assured deletion and version control," in 3rd International Workshop on Security in CloudComputing, 2011.

[11] . S. Plank and L. Xu, "Optimizing Cauchy Reed-solomon Codesfor fault-tolerant network storage applications," in NCA-06: 5th IEEE International Symposium on Network Computing Applications,Cambridge, MA, July 2006.

[12] J. Stanek, A. Sorniotti, E. Androulaki, and L. Kencl, "A securedata deduplication scheme for cloud storage," in Technical Report,2013.