ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 7, Issue 6, June 2018

Data Deduplication for Efficient Access on Cloud Storage

Juilee Dilip Mahajan[1], Prof. Rekha A. Kulkarni[2]

PG Student, Department of Computer Engineering, PICT Pune, India[1]

Asst. Professor, Department of Computer Engineering, PICT Pune, India [2]

ABSTRACT: Reduction of data redundancy is very peak issue currently in all types of storage systems. It is very necessary to manage the huge amount of data wherever it is stored. To conquer both this issues, data deduplication is an efficient approach. In data deduplication, only one original instance will be stored on comparing both data files. For its processing, there will be comparison between two or more files for the purpose of storage. All these files will be compared with each other to get the similarity between each file. If files found similar with each other, then previously stored original file will be retained only. This complete phenomenon reduces the data redundancy in storage system. Along with this, the data sent over a network will be easily transferred.

KEYWORDS: Deduplication, Fragmentation, Indexing, Performance, Redundancy, Similarity

I. INTRODUCTION

Data deduplication is the approach which reduces the redundancy by removing duplicate data as well as it reduces the rate of transmission of data in network of very low bandwidth. There are mainly two ways of applying deduplication. First is, file level deduplication, in which the similarity index will be checked only at complete file at the same time. The second way is block level deduplication, in which the files are compared at the block level. It means, blocks of particular size are made depending on user requirement and the blocks will be compared. The block level deduplication gives more accurate similarity index than file level. This is because; each byte of a file will be taken into consideration for comparison. This deduplication will be done on the basis of renowned algorithms namely SHA (Secure Hash Algorithm) or MD5 (Message Digest Algorithm). The comparison of blocks is based on the hash values of previously stored data and the data to be stored obtained by applying these algorithms. Before applying these algorithms encryption of the data to be stored is essential. After encryption, fragmentation will be in process. For data reduction purpose, fingerprint based data reduction approach was also used. But this approach has some bottlenecks. As this approach works on huge amount of data, the chunk having identical difference will be considered only. If there is any small chunk having small difference, it may get ignored. Often modification in data on small amount may also be neglected. Therefore, file level and block level deduplication are preferred.

II. EXISTING SYSTEM

Data deduplication is the technique which is day-by-day becoming well known in much intensive storage systems.
Various ways can be used to apply this technique. Previously, fingerprint based deduplication technique was used to detect the duplicate data. For this, chunking of data also known as fragmentation was the prime part. For fragmentation, content defined chunking approach was used. This approach was used to address the boundary shift problem. The data which is not duplicate can also be processed by this content defined chunking approach.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 7, Issue 6, June 2018

III. MOTIVATION

As cloud is intensively large storage system, huge amount of data is stored in it. But retrieving the exact data needed is tedious task sometimes. For this purpose, deduplication technique is much effective. As deduplication technique minimizes the redundancy in storage system, it is easy to fetch the accurate data required. Presence of a single instance of a particular file makes retrieval of data clear and easy. Because deduplication is simple and understandable technique, its motive is pretty clear.

IV. REVIEW OF LITERATURE

D. Meister, et. Al [2] proposes an approach which deals with blocks of data. The locality of block can be identified on the basis of time. The block which recently stored in backup is taken into consideration. Due to this, data which is to be accessed can be obtained easily. Mainly, this prevents old age of data.
The Paper by M. Lillibridge, et. Al [3] represents about the performance of the deduplication process based on the fragmentation of chunks. Whenever deduplication is done based on chunks, it may take more time to store the data.
This will surely reduce the performance of the system. Also this issue can lead to lower the order of magnitude of the system.
The authors V. Tarasov, et. Al [4] have designed and created an environment, in which the dataset formation emulation for various aspects is taken into consideration. This environment is also reliable for controllable micro benchmarking of deduplication technique solutions. Large number of datasets can be generated with preservation of original data characteristics and properties.
“A study of practical deduplication” [5], this paper represents actual study of deduplication technique done by authors. In this study, the comparison between file level deduplication and block level deduplication is done. For comparison, they accessed data of file system of 857 desktop computers at Microsoft nearly for duration of almost 4 weeks. In this comparison they found that, block level deduplication is more efficient than file level deduplication. As byte-by-byte data is compared, block level is dominant over file level deduplication.
The paper, “Venti: a new approach to archival storage” [6] elaborates about a network storage system, called venti. It is particularly made for data archival. This system works for a unique hash value created for content which acts as the block identifier in read and write operations. This tends to indexing which avoids duplication of data. This paper have built a prototype of the system and presented some preliminary results for performance.
In another paper, El-Shimi, et. Al [7] have done analysis of primary data deduplication. They have used some information about the design of a new primary data deduplication system. Both file level and block level deduplication are taken in analysis of deduplication technique.
In Study of chunking algorithm in data deduplication [8] paper, the authors have proposed various chunking algorithms. Basically, chunk is a small part of a file used for comparison between previously stored files. For detection of duplication, comparison between two files is the prior part. For that chunking is essential. Therefore, different chunking models and algorithms can be found in this paper.
A new encoding technique is introduced by P. Kulkarni, et. Al [9] to eliminate the redundancy in this paper named as REBL. Authors have particularly implemented this technique for block level. This method is efficient for collection of huge data. It is somewhat easy to deal with small dataset to reduce its redundancy. But for large collection of files, it is useful. According to authors, REBL technique is much better than CDC i.e. content defined chunking. Generally, deduplication deals with input and output which is unique to be stored in the storage. It means it is i/o oriented.
In this paper, authors B. Mao, et. Al [10] proposed a new approach called POD, performance oriented deduplication. It works more for performance rather than reads and writes in input and output. Because of this, it saves the fragmentation of data and increases efficiency. This is beneficial to save the primary storage and hence improves the performance.
Deduplication is now widely used technique for storage systems. It reduces the redundancy. But there may be issue of data security. So, the paper “A study on authorized deduplication techniques in cloud Computing” [11] talks about the information security through its authorization. Authors of this paper have proposed various ways of authorization for technique of deduplication.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 7, Issue 6, June 2018

V. SYSTEM ARCHITECTURE

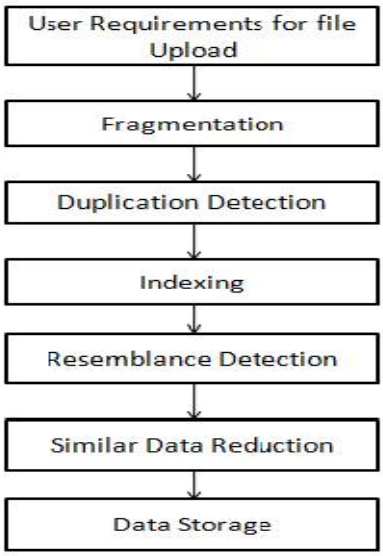


Fig. 1. System Architecture for Proposed System

VI. SYSTEM OVERVIEW

In this system, firstly the storage system will be chosen in cloud. For applying deduplication technique, user will upload the file. The uploaded file will be in encrypted format. AES algorithm will be used for the encryption of data. On the basis of type of deduplication, the chunks or fragments of a file will be created by fragmentation algorithm. The chunks of the file to be verified are mostly of same size according to the user. On completion of fragmentation, comparison between the previously stored file and the uploaded file takes place. Duplicate Adjacency Detection is the process used for comparison. This process tends to identify the duplicate data. The comparison is basically based on the hash value generated by MD5 algorithm. The identified duplicate data have to be eliminated. After duplicate adjacency detection, for further refinement, improved super feature approach can be used. Usage of this approach is able to provide exact unique data. This is surely helpful for reduction of redundancy of data. According to this architecture, duplicate data can be removed and unique data can be obtained efficiently. This unique data will be stored in cloud storage only.

VII. ALGORITHMS

A. Fragmentation Algorithm

Input: File
Output: Chunks
Step1: For fragmentation go to step 2
Step2: Input original file
Step3: Size = size of original file
Step4: Fs = Size of fragments
Step5: NoF = number of fragments
Step6: Fs = Size/Nof
Step7: We get fragments with merge option

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 7, Issue 6, June 2018

Step8: End

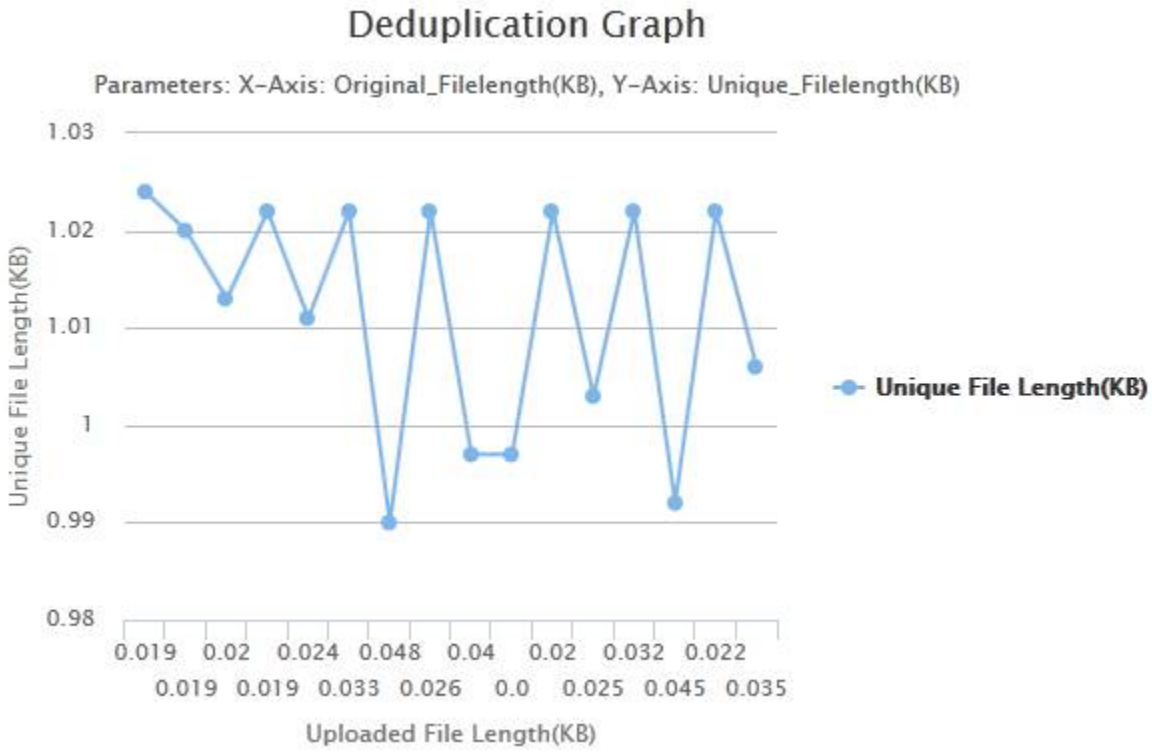
B. AES Algorithm For Encryption

AES (advanced encryption standard). It is symmetric algorithm. It is used to convert plain text into cipher text. The need for coming with this algorithm is weakness in DES. The 56 bit key of DES is no longer safe against attacks based on exhaustive key searches and 64-bit block also consider as weak. AES is to be used 128-bit block with128-bit keys.
Input:
128 bit /192 bit/256 bit input (0,1)secret key(128 bit)+plain text(128 bit).
Process:
10/12/14-rounds for-128 bit /192 bit/256 bit input Xor state block (i/p)
Final round: 10, 12, 14
Each round consists: sub byte, shift byte, mix columns, add round key.
Output:
cipher text(128 bit)

C. MD5 Message-Digest Algorithm

This algorithm is used on large scale to produce 128-bit hash value by cryptographic hash function. Generally, this function is represented in text format of a 32 digit hexadecimal number. For different cryptographic applications MD5 algorithm has been used. Data integrity verification can be done by this algorithm.

VIII. RESULTS



ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 7, Issue 6, June 2018

Above graph shows the Performance of Duplicate-Detection of data and Resemblance-Detection found in data of uploaded file. The X-axis contains original length. Y-axis contains unique file length. The graph shows the efficiency of deduplication process to identify duplicate as well as unique data which is to be uploaded.
Also shows comparison among the duplicate detection and resemblance detection approaches in terms of data reduction efficiency on uploaded data.

IX. CONCLUSION

In this paper, deduplication of data which is an efficient approach for cloud storage is elaborated. This approach is particularly referred for detecting the duplicate data. As there is huge amount of data stored in storage system, redundancy inherently occurs there. It is very crucial to remove the redundancy of data in cloud storage. Therefore, deduplication technique surely helps to overcome this issue. When duplicate data is found by this approach, it will be eliminated. On elimination of duplicate data, remaining unique data will be stored only. Proper utilization of cloud storage will be there because of this approach. Actual outcome of this system comes when file is stored in cloud storage uniquely. Also, user is able to download that uniquely stored file for further access.

REFERENCES

[1] W. Xia, D. Feng, L. Tian, "DARE: A Deduplication-Aware Resemblance Detection and Elimination Scheme for Data Reduction with Low Overheads" in IEEE Transactions on computers , Vol.65, No. 6, June 2016, pp. 1692-1705
[2] D. Meister, J. Kaiser, and A. Brinkmann, "Block locality caching for data deduplication," in Proc. 6th Int. Syst. Storage Conf., 2013, pp. 1-12
[3] M. Lillibridge, K. Eshghi, and D. Bhagwat, "Improving restore speed for backup systems that use inline chunk-based deduplication," in Proc. 11th USENIX Conf. File Storage Technol.,Feb. 2013, pp. 183-197
[4] V. Tarasov, A. Mudrankit, W. Buik, P. Shilane, G. Kuenning, and E. Zadok, "Generating realistic datasets for deduplication analysis," in Proc. USENIX Conf. Annu. Tech. Conf., Jun. 2012, pp. 261-272
[5] D. T. Meyer and W. J. Bolosky, "A study of practical deduplication," ACM Trans. Storage, vol. 7, no. 4, p. 14, 2012
[6] S. Quinlan and S. Dorward, "Venti: A new approach to archival storage", in Proc. USENIX Conf. File Storage Technol., Jan. 2002, pp. 89101.
[7] El-Shimi, R. Kalach, A. Kumar, A. Ottean, J. Li, and S. Sengupta, "Primary data deduplication-large scale study and system design," in Proc. Conf. USENIX Annu. Tech. Conf., Jun. 2012,pp.285-296
[8] A. Venish and K. Siva Sankar, "Study of Chunking Algorithm in Data Deduplication", Proceeding of the International Conference on Soft Computing System, ICSCS , Volume 2.
[9] P. Kulkarni, F. Douglis, J. D. LaVoie, and J. M. Tracey, "Redundancy elimination within large collections of files," in Proc. USENIX Annu. Tech. Conf., Jun.2012, pp.59-72
[10] B. Mao, H. Jiang, S. Wu, L. Tian, "Leveraging Data Deduplication to Improve the Performance of Primary Storage Systems in the Cloud", IEEE TRANSACTIONS ON COMPUTERS, VOL. 65, NO. 6, JUNE 2016
[11] B. Choudhary, A. 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
[12] G. Wallace, F. Douglis, H. Qian, P. Shilane, S. Smaldone, M. Chamness, and W. Hsu, "Characteristics of backup workloads in production systems," in Proc. 10th USENIX Conf. File Storage Technol., Feb.2012,pp.33-48