ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 4, Issue 12, December 2016

Survey on data Deduplication using Similarity and Locality based Approaches

Jayashri Patil, Sunita Barve, Mayura Kulkarni,

M. E Student, MIT Academy of Engineering, Alandi, Pune

Assistant Professor, Department of Computer Engineering, MIT Academy of Engineering, Aland, Pune, India

Assistant Professor, Department of Computer Engineering, MIT Academy of Engineering, Alandi, Pune, India

ABSTRACT: in large scale storage system data deduplication has gain more popularity and attention. Deduplication is one such storage optimization technique that avoids storing duplicate Copies of data and only one occurrence of the data are stored on storage media .It is essentially a compression method for removing redundant data. As a Space efficient method data deduplicationis used, in storage system to data backup. Storage space is saved by removing redundant data and also in network storage system the transmission of duplicate data is minimize. Scalability of fingerprint-index search for centralized data deduplication is main challenge. For high throughput and performance, removing duplicate contents and balancing load by low RAM overhead SiLo scalable deduplication system is used. Similarity and locality exploit both the similarity and locality approach which are complementarily. In SiLo deduplication system, small files which are related are grouped into a segment and segmentation of large file is done. In Silo RAM usage is reduced for index lookup.

KEYWORDS: Deduplication, Similarity, Locality, Storage system.

I. INTRODUCTION

Only single instance of data is stored in data duplicationinstead of storing multiple copies of same data.It is methodof removing duplicates copies of data and duplicate copies arereplace with pointers,which points to the identical copy whichis stored in storage as a single instance of data. Data set orstream is examined at sub-file level and only identical data isstored or saved. The workflow of data deduplication consists ofInput file, Hash Computation, Computing hash with hash indextable; whether match found or not if yes set pointer to existingdata location and if no save data to memory and its hashto hash index. The duplicate data segments in deduplicationtechnology are detected with the help of fingerprint. Hashfunctions such as MD5, SHA-1 are used by fingerprint toidentify identical segment.Based on granularity deduplication can be categorized asfile level deduplication and block level deduplication. In caseof File Level deduplication the entire file is considered, thuseven small append or update make file different from previousversion and thus reducing deduplication ratio. No duplicatefile exists at file level deduplication. Where as in caseof block level deduplication data blocks are considered fordeduplication.In block level similar data segment of a file willbe detected.In offline data deduplication technique, the deduplicationprocess is done after storing the data on the storagedisk. In online data deduplication, duplicate data is deletedbefore writing to the storage disk. Data deduplication can alsobe categorized as target based deduplication and source baseddeduplication .In targert based deduplication the client does not modifie and client does not perform any deduplication whichimproves storage utilization and does not save bandwidth. Insource based deduplication client do the deduplication processonly identical data is backup, it saves bandwidth as well asstorage space, but there is extra computational load on thebackup client.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 4, Issue 12, December 2016

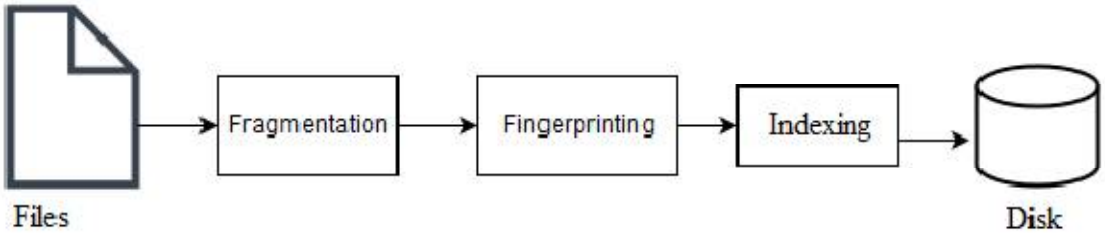

Fig. 1. Deduplication process

As the duplicate data increases the performance decrease.The volume of fingerprints grows with increase in backupdata, on disk drive more amount of space is acquired byfingerprint. Thus fingerprint indexing leads to bottleneck performance.Due to this there is a frequent disk access to locatefingerprints which blocks the process of data deduplication.The fingerprints of the same file are stored separately ondisk drivers.Whenever the fingerprints are referred there issignificant performance degradation.Many approaches havebeen proposed to address the performance bottleneck.Twoprimary approaches for data deduplication are similarity-baseddeduplication and locality-based deduplication.

A. LOCALITY APPROACH

In locality based approach the order of the backup stream issame for each backup with high probability. Due to whichthere is increases in the RAM utilization and accesses toon-disk index is reduced, which alleviate disk bottleneckproblem.Normally chunk lookups are one by one but somebackup streams have high locality between the first, second,and next backups have a very high probability that chunks arein the same order. However this approach shows low speed onbackup stream with weak locality.

B. SIMILARITY APPROACH

Similarity based approaches are designed to overcome theproblem encountered by In locality-based approaches Backupstreams that either have weak locality or backup streamthat lack locality in that case locality approach have problem.Similarity based approach overcome this problem of locality.In backup stream, instead of locality they exploit datasimilarity from the backup extract similar characteristics andreduce the RAM usage. Instead of lookups per chunks or perlocal chunks (locality) the lookups are per files. Although is much faster than locality approach it can sacirfice the duplicationaccuracy. The main idea behind SiLo is that for smallfiles combine into segments to reduce number of fingerprints. For large files divide into segments to increase the similaritydetection.Group similar segments order into blocks (preservelocality).

II. LITERATURE SURVEY

Bo Mao, Hong Jiang Suzhen Wu, Lei Tian (2014)have proposed Performance Oriented I/O Deduplicationapproach [1]. If data deduplication is directly applied onprimary storage then it will cause two problems, fragmentationof data on disks and space contention in memory.Due to
This they proposed Performance-Oriented I/O Deduplication. Two approach namely have been considered in PODnamely selective dedupe and iCache.Selective dedupe isconsider to remove data fragmentation problem and iCacheis consider for memory management.POD support featureslike capacity saving, performance enhancement, small writeselimination, large writes elimination and cache partitioningstrategy. POD achieves comparable or better capacity savingthan idedupe. I/O performance of primary storage is improvedby POD.

Mazhar Ali, Kashif Bilal, Samee U. Khan, BharadwajVeeravalli, Keqin Li, Albert Y. Zomaya, (2015) haveproposed T-coloring[2] .They have consider Securityand performance.In this methodology file is divided intofragment and each

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 4, Issue 12, December 2016

fragment is replicated to different storagenode.Only single fragment of particular File is Stored on eachnode.Fragment of particular file is stored on different storagenode to increase the Surface area for attacker ,if in case theattacker access the one fragment of file he/she is unable toaccess the another fragment.Because fragment are stored ata centrality distance from each other which is difficult toguess. In DROP methodology the fragment is also replicatedto provide data availability, reliability and improve the dataretrieval time.DROP also performs a controlled replication toincrease the data availability, reliability, and improve dataretrieval time. For reconstruction of file, it provide improveretrival time for accessing particular file fragment, fragmentare placed on the node in such a way that they provides the Decreased access cost.

Wen Xia, Hong Jiang, Dan Feng,and Lei Tian, (2015)Proposed two data reduction approaches duplicate detectionand resemblance detection[3] .In resemblance detection detectssimilar data object.Granularity at byte-Level.Scalability isweak.It is Delta compression method based on super-Feature.Duplicate detection is deduplication method based onSecure-Fingerprint.It detect duplicate data object.Granularityis at chunk level.Scalability is Strong.Approaches are twofold
:Memory overhead and computation overhead.When thesegment is loaded into the locality cache,the two pointersare associated to doubly linked list ,the doubly linked list isfreed when the segment is removed from cache.Computationoverhead is removed by confirming the similarity degreeof the DupAdj-detected chunks. To detect similar datachunks DARE efficiently exploit existing duplicate-adjacencyinformation, this achieve highest throughput, data reductionapproach.

T. Yang, H. Jiang, D. Feng, Z. Niu, K. Zhou and Y.Wan, (2010) proposed DEBAR,a a scalable and highperformancededuplication storage architecture for Backupand archiving[4].DEBAR improved capacity ,throughput andscalability for dedduplication.DEBAR is compared with DDFSin this paper. More backup client are supported by DEBAR Ascompared to DDFS.Various application are supported DEBARsuch as geographic information system grid,WAN,data sharingplatform for scientific and engineering application.In DDFS bloom filter is used to reduce disk index access,it improvededuplication but there is poor scalability .For avoidingfingerprint lookup disk bottleneck in data deduplicationDEBAR uses sparse index which exploits inherent locality inbackup stream.The main advantage of using DEBAR is that itrequired half memory space for dedduplication compared toDDFS.For high throughput DEBAR can simultaneously runmultiple backup servers.TDFS perform to data deduplicationscheme two dedupe,in dedupe-1 data chunks are collectedand dedupe-2 new data chunk is identified.

M. Fu et al(2016),proposed Rducing fragmentation forinline deduplication backup storage[5].Two drawback offragmentation,first is restore performance is decreased andsecond it result in invalid chunk.For reducing fragmentationproblem two algorithm are proposed History-Aware Rewritingalgorithm and Cache-Aware Filter.Two container sparsecontainer and out of order container decreases the restoreperformance.Fragmentation is in sparse and out of ordercontainer.To identify and reduce sparse containers HAR isused which exploit histrocial information and CAF exploitcache knowledge to identify and reduce out of order container.

C. Li, S. Wang, Xiaochunyun, X. Zhou and G. Wu(2014),have proposed MMD[6], Multiple disk are used to boastthe reading performance ,each disk is used independentlyas logical device.Due to fragmentation in data deduplicationsystem,reading performace is decreased. For this reasonMMD storage approach is used which increases readperformance and it is diffirent from RAID.Two algorithm areused,algorithm are used to assign the container to disk.MMDperformance is higher compared to RAID0.

J. Liu, Y. Chai, C. Yan and X. Wang (2016), propose a newDelayed Container Organization [7], to increase the restoreperformance in data deduplication system.The constructionof containers is delayed after assigning data chunck in nonvolatile memory. DCO have higher restore speed, Betteroptimization based on a large amount of information, spacesaving is medium. DCO has three advantages Higher UDRsContainers are produced, More data is duplicated, Restore isspeedup.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 4, Issue 12, December 2016

III. CONCLUSION

In this paper we studied different deduplication approaches.Similarity and locality approachs overcomes the shortcomingsof various approaches in this paper. This combine approachreduce RAM usage,keep duplication accuracy,and it also increasesthroughput.Silo approach can effectively improve thedisk bottleneck with adequate overhead of CPU, memory, andstorage when performing fingerprint lookup, thus improvingthe throughput of data deduplication. There are multiple approachesand method been suggested by different authors fordata deduplication in large storage system.Various method fordata reduction, datacompression, dataencoding, data deduplicationhave been examined to improve performance.Restoreperformace is also increased.

ACKNOWLEDGMENT

I take immense pleasure in expressing my humble noteof gratitude to my project guide Mrs. Mayura kulkarni andDr.Sunita Barve, Assistant Professor, Department of ComputerEngineering, MIT Academy of Engineering, Alandi, Pune,for their remarkable guidance and useful suggestions, whichhelped me in completion the paper before deadline.

REFERENCES

[1] B. Mao, H. Jiang, S. Wu and L. Tian, POD: Performance Oriented I/ODeduplication for Primary Storage Systems in the Cloud, (2014) pp. 767-776.
[2] M. Ali, K. Bilal,S. Khan, B. Veeravalli,K. Li, A. Zomaya, DROPS:Division and Replication of Data in the Cloud for Optimal Performanceand Security, IEEE Transactions on Cloud Computing , vol.PP, no.99,pp.1-1.
[3] W. Xia, H. Jiang, D. Feng and L. Tian, DARE: A Deduplication-AwareResemblance Detection and Elimination Scheme for Data Reduction withLow Overheads, IEEE Transactions on Computers, vol. 65, no. 6, pp.1692-1705, June 1 2016.
[4] T. Yang, H. Jiang, D. Feng, Z. Niu, K. Zhou and Y. Wan, DEBAR: Ascalable high-performance de-duplication storage system for backup andarchiving,(2010) pp. 1-12.
[5] M. Fu et al., Reducing Fragmentation for In-line Deduplication BackupStorage via Exploiting Backup History and Cache Knowledge, IEEETransactions on Parallel and Distributed Systems, vol. 27, no. 3, pp.855-868, March 1 2016.
[6] C. Li, S. Wang, Xiaochunyun, X. Zhou and G. Wu, MMD: An Approachto Improve Reading Performance in Deduplication Systems, (2014) pp.93-97.
[7] J. Liu, Y. Chai, C. Yan and X. Wang, A Delayed Container OrganizationApproach to Improve Restore Speed for Deduplication Systems, IEEETransactions on Parallel and Distributed Systems, vol. 27, no. 9, pp.2477-2491, Sept. 1 2016.
[8] Jingwei Li, Jin Li, Dongqing Xie and Zhang Caiu, Secure Auditing andDeduplicating Data in Cloud. IEEE Transactions on Computers (2015).
[9] D. Meyer and W. Bolosky, A study of practical deduplication, Feb. 2011,pp. 229241.
[10] F. Guo and P. Efstathopoulos, Building a High-Performance DeduplicationSystem (2011).
[11] K. Srinivasan, T. Bisson, G. Goodson, and K. Voruganti, iDedup:Latency-Aware, Inline Data Deduplication for Primary Storage, (2012).
[12] W. Xia, H. Jiang, D. Feng, L. Tian, M. Fu, and Z. Wang, PDedupe:Exploiting Parallelism in Data Deduplication System, pp. 338-347(2012).
[13] M. Lillibridge, K. Eshghi, and D. Bhagwat, Improving restore speed forbackup systems that use inline chunk-based deduplication, Feb. 2013, pp.183197.
[14] A. F. Banu and C. Chandrasekar, A survey on deduplication methods, vol.3, no. 3, pp. 364368, 2012.
[15] G. Lu, Y. J. Nam, and D. H. Du, BloomStore: Bloom-filter basedmemory-efficient key-value store for indexing of data deduplication onflash, Apr. 2012, pp. 111.
[16] J. Li et al, Secure deduplication with efficient and reliable convergentkey management, vol. 25, no. 6, pp. 111, 2013.