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

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor & UGC Approved Journal)

Website: www.ijirset.com

Vol. 6, Issue 8, August 2017

Reliable De-Duplication Approach to Detect and Eliminate Data Redundancy and Finding the Sweet Spot through Delta Compression

Gaddala Priyanka, G.Venkata Rami Reddy, K.Balakrishna Maruthiram

Department of SE, School of Information Technology JNTUH, Hyderabad, India

ABSTRACT: As digital information is developing uncontrollably, need for data reduction has emerged as a necessary task in storage structures. For giant scale data reduction, it is very important to maximally realize and remove redundancy at low overheads. Data de-duplication could be a statistics reduction technique that reduces storage area via putting off redundant data and best one example of the records is maintained on storage media. Delta compression is an efficient methodology for eliminating redundancy among non-duplicates however terribly similar records documents and chunks. during this paper we recommend DARE (De-duplication Aware resemblance Detection and Elimination) theme to use a theme, mentioned as Duplicate-Adjacency based resemblance Detection (DupAdj), by considering concerning any statistics chunks to be comparable (i.e., candidates for delta compression) if their individual adjacent records chunks are duplicate in a de-duplication approach, once which equally smarten the resemblance detection performance by associate improved super-feature technique.

KEYWORDS: Storage System, Index Structure, Data De-duplication, Data Reduction, Delta Compression, Performance Evaluation

I. INTRODUCTION

The digital data is mounting day by day during this digital world, handling huge information within the storage system is extremely imperative. The data volume is developing; additional data has been created within the former two years than within the entire previous antiquity of the civilization. As per the surveillance in each second concerning 1.7megabytes are made for each creature. The method of eradicating the duplicate data in a data set is De-duplication of data. It is the utmost important way to abbreviate the redundant data within the huge information. Reducing of the duplicate data can at all times raise the storage management and additionally can provide a great efficiency computing. Information de-duplication is usually measured by ratio; the larger the ratio here is a shrinking come. The competition faced by the de-duplication method is the way to in the main establish redundant information with high potency. In general the information de-duplication is accomplished by rending the input information into minor fragments and configuration the fragments, victimization information de-duplication algorithm the redundant information is detected and more duplication method is continued. There are several duplicate detection algorithms and in this current work Duplicate contiguousness detection algorithm and Super feature approach is used. Unitization is that the splitting the data into minor streams. The computer file is chunked victimization window formula in content primarily based unitization. each computer memory unit slides victimization window protocol and thus the chunk limit is formed. In Rabin fingerprint formula finger prints square measure designed for the data which might have an impact on the cupboard area. The window formula has no unseen channel, so it will be less in price wise. Content based unitization is that the technique of unitization the data by their content so on the window and agency can chunk the data powerfully. The CDC (Content defined chunking) procedure could be a important module in information de-duplication that splits the inward information stream into chunks once the content window before the breakpoint gratifies a planned scenario. Since chunk is that the essential unit that finds repetitions, the agency formula has major impact on the de-duplication relation. In addition, since the total information flow should be chunked previous different de-duplication processes, the agency formula in addition includes a control on de-

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

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor & UGC Approved Journal)

Website: www.ijirset.com

Vol. 6, Issue 8, August 2017

duplication performance. Generally, there square measure four steps concerned in information de-duplication: (1) unitization that uses a agency formula to separate the data stream into chunks; (2) chunk method that calculates the fingerprint (for example, SHA-1 value) for each of the chunks; (3) fingerprint assortment and querying that provisions the fingerprints in line with an precise arrangement for assortment and queries the catalog to obtain out chunks of identical fingerprints – these chunks square measure measured as a result of the duplicated chunks; (4) information storing and management that stores new chunks and delegates reference pointers for the duplicated chunks. There square measure two necessary dimensions that assess a de-duplication system: the de-duplication relation and performance. Whereas different steps in addition play necessary roles, the unitization step has substantial impact on the de-duplication relation and performance.

II. RELATED WORK

Bo Mao, Hong Jiang Suzhen Wu dialect, Lei Tian (2014) have projected Performance oriented I/O De-duplication approach. If data de-duplication is directly applied on primary storage then it will cause two problems, fragmentation of information on disks and area competition in memory. Because of this they planned Performance-Oriented I/O De-duplication. Two approach specifically are considered in POD specifically selective de-duplicate and iCache. Selective de-dupe is consider to remove data fragmentation drawback and iCache is consider for memory management drawback POD support options like capability saving, performance improvement, little writes elimination, massive writes elimination and cache partitioning strategy. POD achieves comparable or higher capability saving than de-dupe. Wen Xia, Hong Jiang, Dan Feng, and Lei Tian, (2015) planned two data reduction approaches resemblance and duplicate detection. Resemblance detects similar data object. Granularity at byte Level measurability is weak. It is Delta compression technique based on super feature. Duplicate detection is de-duplication technique supported secure-Fingerprint. It discovers duplicate data object. Roughness is at chunk level scalability is robust. Approaches are twofold: Memory overhead and computation overhead. Once the section is loaded into the neighborhood cache, the 2 pointers area unit associated to doubly connected list; the doubly connected list is freed once the section is far away from cache. Computation overhead is removed by confirming the similarity degree of the Duplicate adjacent detected chunks. To observe similar data chunks DARE with efficiency exploit existing duplicate-adjacency data, this achieve highest throughput, data reduction approach. T. Yang, H. Jiang, D. Feng, Z. Niu, K. Zhou and Y. Wan, (2010) planned DEBAR, a ascendable and high performance de-duplication storage design for Backup and archiving. DEBAR improved capability, output and quantifiability for de-duplication. DEBAR is compared with DDFS during this paper. Additional backup client are supported by DEBAR As compared to DDFS. Numerous application is supported DEBAR like geographic system grid, WAN, data sharing platform for scientific and engineering application. In DDFS bloom filter is used to reduce disk index access, it improve de-duplication however there is poor quantifiability. For avoiding fingerprint search disk bottleneck in data de-duplication DEBAR uses distributed index that exploits inherent neighborhood in backup stream. The main advantage of mistreatment DEBAR is that it needed half memory area for de-duplication compared to DDFS. For high throughput DEBAR can simultaneously run multiple backup servers. TDFS perform to data de-duplication scheme two de-dupe, in dedupe-1 data chunks area unit collected and dedupe-2 new data chunk is identified.

III. FRAME WORK

The system will provide an authorized de-duplication on encrypted data which might be in the sort of document. The system effectively achieves the storage space management during a secure and authorized manner additionally because the system permits to maximally discover and eliminate redundancy at very low overheads by mistreatment DARE scheme. Propose system is divided into three parts: 1. data user authentication theme is used during a private cloud server to prevent system from attackers. 2. Convergent encryption is used to encipher the data and perform duplicate check. 3. DARE maximally observe and eliminate redundancy at very low overheads. DARE, a low-overhead De-duplication-Aware resemblance detection and Elimination scheme for de-duplication based backup and archiving storage system. the most idea of DARE is to effectively exploit existing duplicate-adjacency data to observe similar data chunks (DupAdj), refine and supplement the detection by exploitation an improved super-feature approach (Low-Overhead Super-Feature) once the existing duplicate-adjacency data is lacking or restricted. Additionally, we tend to

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

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor & UGC Approved Journal)

Website: www.ijirset.com

Vol. 6, Issue 8, August 2017

present an analytical study of the existing super-feature approach with a mathematic model and conduct an empirical analysis of this approach with many real-world workloads in data de-duplication systems. In existing system DARE De-duplication technique is used only in-house computer, in our projected system you can use DARE De-duplication technique in cloud storage additionally and you can perform DARE De-duplication technique on encrypted data. Our experimental analysis results, supported real-world and artificial backup datasets, show that DARE significantly outperforms the normal Super-Feature approach.

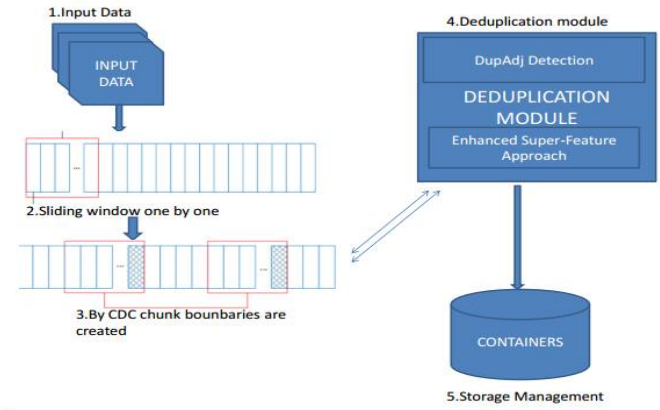


Figure 1: Proposed System Architecture

More specifically, the DupAdj approach achieves a similar data reduction efficiency to the pure super-feature approach and DARE detects 2-10 percent more redundant data whereas achieving the next throughput of data reduction than the pure super-feature approach. Meanwhile, DARE only consumes regarding 1/4 and 1/2 respectively of the computation and categorization overheads needed by the normal super feature approach for resemblance detection. It is necessary to note that our analysis additionally demonstrates the superior data-restore performance of the DARE-enhanced de-duplication system over the de-duplication-only systems via delta compression, where the previous outperforms the latter by an element of 2 (2X). The projected system will flexibly support access management on encrypted data with de-duplication. 2. Low cost of Storage. 3. Large-scale data reduction. 4. Removing redundancy among similar data chunks has gained increasing attention in storage systems. 5. Accurately discover the most similar candidates for delta compression with low overheads. By using the De-duplication module, DARE will first detect duplicate chunks for the input data stream. The DupAdj approach detects similitude by exploiting duplicate-adjacency data of a de-duplication system. Locate and will not be hold on, instead link is provided. If the fingerprint does not match with the existing fingerprint, then it is considered because the new fingerprint. For non-duplicate chunks, our approach uses DupAdj i.e. it will check for adjacent chunks and store them in separate table. For those chunks, again SHA-256 algorithm is applied. The DupAdj detection: an input phase, it will traverse all the chunks by the aforementioned doubly-linked list to find the already duplicate-detected chunks. Figure 2 shows that if chunk B2 and E2 is duplicate, chunks around them are considered potentially similar chunks (i.e. B1&E1 and B3&E3),and then migrate the similar chunks into one table so as to remove redundancy as much as attainable. If similar is detected, link are provided, else are thought-about for delta compression. Only will difference will be stored. The non-similar data and deltas are hold on separately in tables by using this projected system the subsequent advantages are Large-scale data reduction. Removing redundancy among data for likeness detection and finding the “SWEET SPOT” for the super feature approach, similar information chunks has gained increasing attention in storage systems. Accurately detect the most similar candidates for delta compression with low overheads.

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

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor & UGC Approved Journal)

Website: www.ijirset.com

Vol. 6, Issue 8, August 2017

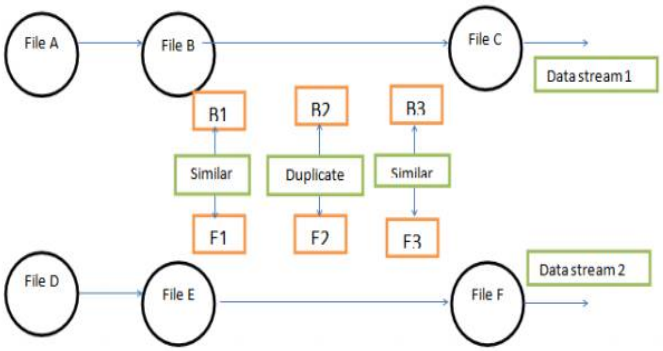
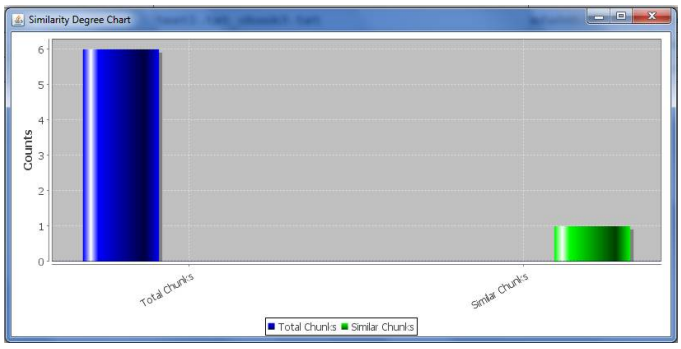


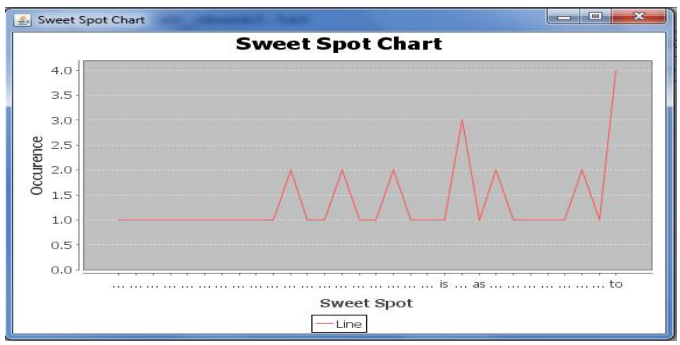
Figure 2: Architecture of DupAdj

IV. EXPERIMENTAL RESULTS

In our experiments, any number of users they can upload the files into the system those uploaded files are stored in chunk format in that system. After that to check the duplicate in that to check the similar chunk in that uploaded file after that to check on duplicate adjacency detection in that upload the different file and it will check similar chunk and it shows the similar chunks and reference will be generate for particular chunk after that to recover the file the file will be recovered and download into our system. In the below chart we can observe that difference between the length of both total chunks and similar chunks.



We can observe that total chunks file length is higher than similar chunks file length. The difference will be shown in the sense of counts. Removing redundancy among data for resemblance detection and finding the “Sweet Spot” for the super feature approach, similar information chunks has gained increasing attention in storage systems. Accurately detect the most similar word occurrence in de-duplication systems in terms of cost effectiveness. In the below chart we can observe that Sweet Spot word count.



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

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor & UGC Approved Journal)

Website: www.ijirset.com

Vol. 6, Issue 8, August 2017

We can observe that sweet spot means most occurrences of similar words the difference will be shown in the sense of counts. Through our implementation we can store the file in chunks format and detect the duplicate chunks as well as we can count the occurrence of each word in chunks is called Sweet Spot and also we can increase the storage space of cloud with low cost when compare to current techniques.

V.CONCLUSION

DARE uses a distinctive technique, DupAdj, that exploits the duplicate-adjacency facts for efficient resemblance detection in present de-duplication systems, and employs a stepped forward first-rate-feature approach to more detecting resemblance once the duplicate-adjacency data is lacking or restrained. Results from experiments driven through real-global and artificial backup datasets recommend that DARE could also be a robust and efficient tool for maximizing statistics reduction by means that of additional detecting corresponding to statistics with low overheads. mainly, DARE simplest consumes concerning 1/4 and half of respectively of the computation and categorization overheads needed by means of the traditional super-function techniques while detection 2-10 % a lot of redundancy and reaching a higher throughput. Moreover, the DARE-enhanced statistics reduction approaches tried to be able to improving the records-restore overall performance, speeding up the de-duplication most effective approach with the help of issue of 2(2X) by using delta compression to additional eliminate redundancy and effectively enlarge the logical house of the restoration cache. Our preliminary results on the facts-restore performance recommend that supplementing delta compression to de-duplication will effectively increase the logical area of the restoration cache; however the records fragmentation in data reduction systems remains a significant problem.

REFERENCES

[1] 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.
[2] P. Shilane, G. Wallace, M. Huang, and W. Hsu, “Delta compressed and deduplicated storage using stream-informed locality,” in Proc. 4th USENIX Conf. Hot Topics Storage File Syst., Jun. 2012, pp. 201–214.
[3] Q. Yang and J. Ren, “I-cash: Intelligently coupled array of SSD and HDD,” in Proc. 17th IEEE Int. Symp. High Perform. Comput. Archit., Feb. 2011, pp. 278–289.
[4] G. Wu and X. He, “Delta-FTL: Improving SSD lifetime via exploiting content locality,” in Proc. 7th ACM Eur. Conf. Comput. Syst., Apr. 2012, pp. 253–266.
[5] D. Gupta, S. Lee, M. Vrable, S. Savage, A. C. Snoeren, G. Varghese, G. M. Voelker, and A. Vahdat, “Difference engine: Harnessing memory redundancy in virtual machines,” in Proc. 5th Symp. Oper. Syst. Design Implementation. Dec. 2008, pp. 309–322.
[6] R. C. Burns and D. D. Long, “Efficient distributed backup with delta compression,” in Proc. 5th Workshop I/O Parallel Distrib. Syst., Nov. 1997, pp. 27–36.
[7] D. T. Meyer and W. J. Bolosky, “A study of practical deduplication,” ACM Trans. Storage, vol. 7, no. 4, p. 14, 2012.
[8] 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.
[9] A. 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.
[10] L. L. You, K. T. Pollack, and D. D. Long, “Deep store: An archival storage system architecture,” in Proc. 21st Int. Conf. Data Eng., Apr. 2005, pp. 804–815.
[11] A. Muthitacharoen, B. Chen, and D. Mazieres, “A low-bandwidth network file system,” in Proc. ACM Symp. Oper. Syst. Principles., Oct. 2001, pp. 1–14.
[12] P. Shilane, M. Huang, G. Wallace, and W. Hsu, “WAN optimized replication of backup datasets using stream-informed delta compression,” in Proc. 10th USENIX Conf. File Storage Technol, Feb. 2012, pp. 49–64.
[13] S. Al-Kiswany, D. Subhraveti, P. Sarkar, and M. Ripeanu, “Vmflock: Virtual machine co-migration for the cloud,” in Proc. 20th Int. Symp. High Perform. Distrib. Comput. Jun. 2011, pp. 159–170.
[14] X. Zhang, Z. Huo, J. Ma, and D. Meng, “Exploiting data de-duplication to accelerate live virtual machine migration,” in Proc. IEEE Int. Conf. Cluster Comput., Sep. 2010, pp. 88–96.
[15] F. Douglis and A. Iyengar, “Application-specific delta-encoding via resemblance detection,” in Proc. USENIX Annu. Tech. Conf., General Track, Jun. 2003, pp. 113–126.