Volume No:2, Issue No:6 (November-2016) ISSN No : 2454-423X (Online)
International Journal of Research in Advanced Computer Science Engineering
A Peer Reviewed Open Access International Journal
www.ijracse.com

An Advanced and Efficient Data Deduplication Detection and Elimination Scheme

A.Nagarjuna, M.Tech
Academic Consultant,
S. V. U. C. E,
S. V. University.

M.Damodhar, M.Tech
Academic Consultant,
S. V. U. C. E,
S. V. University.

ABSTRACT:

Data reduction has become increasingly important in storage systems due to the explosive growth of digital data in the world that has ushered in the big data era. One of the main challenges facing large-scale data reduction is how to maximally detect and eliminate redundancy at very low overheads. In this paper, we present DARE, a low-overhead deduplication-aware resemblance detection and elimination scheme that effectively exploits existing duplicate-adjacency information for highly efficient resemblance detection in data deduplication based backup/archiving storage systems. The main idea behind DARE is to employ a scheme, call Duplicate-Adjacency based Resemblance Detection (DupAdj), by considering any two data chunks to be similar (i.e., candidates for delta compression) if their respective adjacent data chunks are duplicate in a deduplication system, and then further enhance the resemblance detection efficiency by an improved super-feature approach. Our experimental results based on real-world and synthetic backup datasets show that DARE only consumes about 1/4 and 1/2 respectively of the computation and indexing overheads required by the traditional super-feature approaches while detecting 2-10 percent more redundancy and achieving a higher throughput, by exploiting existing duplicate-adjacency information for resemblance detection and finding the “sweet spot” for the super-feature approach.

Introduction:

Data deduplication is a dictionary based data reduction approach popular in the backup/archiving storage area due to its demonstrated ability to effectively to compress backup/archiving datasets by a factor of 4-40X [1, 2]. In general, a chunk-level data deduplication scheme splits data blocks of a data stream (e.g., backup files and databases) into multiple data chunks of average size 8K or 4K with each being uniquely identified and duplicate-detected by a secure SHA-1 or MD5 hash sig-nature (also called a fingerprint) [3, 4, 5, 1]. This secure fingerprint-based deduplication technique eliminates redundancy at the chunk or file level and thus scales better than the traditional LZ77 and Huffman coding based GZ compression [6, 4]. Delta compression, however, has been gaining increasing attention in recent years for its ability to remove redundancy among non-duplicate but very similar data files and chunks, for which the data deduplication technology often fails to identify and eliminate [7, 2].

For example, if chunk A2 is similar to chunk A1 (the base-chunk), the delta compression approach calculates and then only stores the differences (delta 1,2) and the mapping information between A2 and A1 [8]. Thus, it is considered a promising technique to effectively complement and supplement the fingerprint-based deduplication approaches by detecting and compressing similar data missed by the latter. In this paper, we propose DARE, a low-overhead Deduplication-Aware Resemblance detection and Elimination scheme that effectively combines data deduplication and delta compression to achieve high data reduction efficiency at low overhead. The main contributions include:

- A “DupAdj” approach is proposed to exploit existing duplicate-adjacency information after deduplication to detect similar data chunks for

Volume No:2, Issue No:6 (November-2016) ISSN No : 2454-423X (Online)
International Journal of Research in Advanced Computer Science Engineering
A Peer Reviewed Open Access International Journal
www.ijracse.com

delta compression. Specifically, due to locality of similar data in backup datasets, the non-duplicate chunks those are adjacent to the duplicate ones are considered good delta compression candidates for further data reduction.

- A theoretical and empirical study of the traditional super-feature approach is conducted, which suggests that improved resemblance detection for further delta compression is possible when the aforementioned existing duplicate-adjacency information is lacking or limited.

- An investigation into the restoration of deduplicated and delta compressed backup data sug-gests that delta compression has the potential to improve the data-restore performance of deduplication-only systems by further removing redundancy after deduplication and thus en-larging the logical space of the restoration cache.

- Our experimental evaluation results, based on real-world and synthetic backup datasets, show that DARE only consumes about 1/4 and 1/2 respectively of the computation and indexing overhead required by the traditional super-feature approach for resemblance detection while achieving a superior data reduction performance.

EXISTING SYSTEM:

The existing solutions to the indexing issue of delta compression either record the resemblance information for files, instead of data chunks, so that similarity index entries can fit in the memory or exploit the locality of backup data streams in deduplication-based backup/archiving systems, which avoid the global indexing on the disk. The first approach faces an implementation difficulty in large-scale data deduplication systems since it is hard to record all the resemblance or version information of files in such systems. The second approach often fails to detect a significant amount of redundant data when the workloads lack locality.

Another challenge facing the super-feature method is the high overhead in computing the super-features. According to a recent study of delta compression and our experimental observation, the throughput of computing super- features is about 30 MB/s, which may become a potential bottleneck for deduplication-based storage systems, particularly if most index entries are fit in memory or partially on SSD-based storage for which the throughput can be hundreds of MB per second or higher.

Disadvantages of Existing System:

- Existing fingerprint-based deduplication approaches often fail to detect the similar chunks that are largely identical except for a few modified bytes, because their secure hash digest will be totally different even only one byte of a data chunk was changed.

- One of the main challenges facing the application of delta compression in deduplication systems is how to accurately detect the most similar candidates for delta compression with low overheads.

PROPOSED SYSTEM:

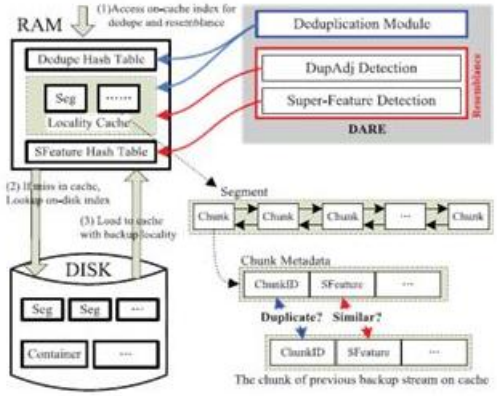
In this paper, we propose DARE, a low-overhead Deduplication- Aware Resemblance detection and Elimination scheme for deduplication based backup and archiving storage system. The main idea of DARE is to effectively exploit existing duplicate-adjacency information to detect similar data chunks (DupAdj), refine and supplement the detection by using an improved super-feature approach (Low-Overhead Super-Feature) when the existing duplicate-adjacency information is lacking or limited. In addition, we present an analytical study of the existing super-feature approach with a mathematic model and conduct an empirical evaluation of this approach with several real-world workloads in data deduplication systems.

Volume No:2, Issue No:6 (November-2016) ISSN No : 2454-423X (Online)
International Journal of Research in Advanced Computer Science Engineering
A Peer Reviewed Open Access International Journal
www.ijracse.com

Advantages of Proposed System:

1. DARE significantly outperforms the traditional Super-Feature approach.

SYSTEM ARCHITECTURE:



MODULES:

We have 3 modules,

3. Deduplication Module
4. DupAdj Detection Module
5. Improved Super-Feature Module

Module Description:

Deduplication:

By using the Deduplication module, DARE will first detect duplicate chunks for the input data stream.

DupAdj Detection:

The DupAdj approach detects resemblance by exploiting existing duplicate-adjacency information of a deduplication system.

Improved Super-Feature:

In this module, for each non-duplicate chunk, DARE will first use its DupAdj Detection module to quickly determine whether it is a delta compression candidate; If it is not a candidate, DARE will then compute its features and super-features, using its improved Super-Feature Detection module, to further detect resemblance for data reduction.

Conclusion and Future Work:

In this paper, we present DARE, a deduplication-aware, low-overhead resemblance detection and elimination scheme for delta compression on the top of deduplication on backup datasets. DARE uses a novel resemblance detection approach, DupAdj, which exploits the duplicate-adjacency in-formation for efficient resemblance detection in existing deduplication systems, and employs an improved super-feature approach to further detecting resemblance when the duplicate-adjacency information is lacking or limited. Our preliminary results on the data-restore performance suggest that supplementing delta compression to deduplication can effectively enlarge the logical space of the restoration cache, but the data fragmentation in data reduction systems remains a serious problem [19]. We plan to further study and improve the data-restore performance of storage systems based on deduplication and delta compression in our future work.

References:

[1]Wen Xia, Member, IEEE, Hong Jiang, Fellow, IEEE, Dan Feng, Member, IEEE, and Lei Tian, Senior Member, IEEE "DARE: A Deduplication-Aware Resemblance Detection and Elimination Scheme for Data Reduction with Low Overheads" IEEE TRANSACTIONS ON COMPUTERS, VOL. 65, NO. 6, JUNE 2016

[2]P. Shilane, M. Huang, G. Wallace, and W. Hsu, "WAN optimized replication of backup datasets using stream-informed delta compression," in Proc. USENIX FAST, 2012.

[3] A. Muthitacharoen, B. Chen, and D. Mazieres, "A low-bandwidth network file system," in Proc. ACM SOSP, 2001.

[4]C. Constantinescu, J. Glider, and D. Chambliss, "Mixing deduplication and compression on active data sets," in Data Compression Conference (DCC), 2011. IEEE, 2011, pp. 393–402.

Volume No:2, Issue No:6 (November-2016) ISSN No : 2454-423X (Online)
International Journal of Research in Advanced Computer Science Engineering
A Peer Reviewed Open Access International Journal
www.ijracse.com

[5]B. Zhu, K. Li, and H. Patterson, "Avoiding the disk bottleneck in the data domain deduplication file system," in Proc. USENIX FAST. USENIX Association, 2003.

[6]J. Gailly and M. Adler, "The gzip compressor," http://www.gzip.org/, 1991.

[7]P. Kulkarni, F. Douglis, J. LaVoie, and J. Tracey, "Redundancy elimination within large collections of files," in USENIX Annual Technical Conference. USENIX Association, 2004.

[8]J. MacDonald, "File system support for delta compression." Masters thesis. Department of Electrical Engineering and Computer Science, University of California at Berkeley., 2000.

[9]S. Quinlan and S. Dorward, "Venti: a new approach to archival storage," in Proc. USENIX FAST, 2002.

[10]F. Douglis and A. Iyengar, "Application-specific delta-encoding via resemblance detection," in Proc. USENIX FAST. USENIX Association, 2003.

[11]L. Aronovich, R. Asher, E. Bachmat, H. Bitner, M. Hirsch, and S. Klein, "The design of a similarity based dedupli-cation system," in Proceedings of SYSTOR 2009: The Israeli Experimental Systems Conference. ACM, 2009.

[12]M. Rabin, Fingerprinting by random polynomials. Center for Research in Computing Techn., Aiken Computation Laboratory, Univ., 1981.

[13]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. USENIX OSDI, 2008.

[14]Q. Yang and J. Ren, "I-CASH: Intelligently coupled array of ssd and hdd," in Proc. IEEE HPCA, 2011.

[15]A. Broder, "Identifying and filtering near-duplicate documents," in Combinatorial Pattern Matching, 2000.

[16]"Some applications of Rabins fingerprinting method," in Sequences II: Methods in Communications, Security, and Computer Science, 1993.

[17] "On the resemblance and containment of documents," in Compression and Complexity of Sequences 1997.

[18] V. Tarasov, A. Mudrankit, W. Buik, P. Shilane, G. Kuenning, and E. Zadok, "Generating realistic datasets for deduplication analysis," in USENIX Annual Technical Conference, 2012.

[19] M. Lillibridge, K. Eshghi, and D. Bhagwat, "Improving restore speed for backup systems that use inline chunk-based deduplication," in Proc. USENIX FAST, 2013.