International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056
Volume: 04 Issue: 02 | Feb -2017 www.irjet.net p-ISSN: 2395-0072

A Survey: Enhanced Block Level Message Locked Encryption for data Deduplication

Priyanka Ahirwar[1], Dr. Jitendra Agrawal[2], Dr. Sanjiv Sharma[3]

[1] *M.Tech Scholar, School Of Information Technology, RGPV, Bhopal, Madhya Pradesh, India*

[2] *Associate professor, School Of Information Technology, RGPV, Madhya Pradesh, Bhopal India*

[3]*Professor, School Of Information Technology, RGPV, Madhya Pradesh, Bhopal India*

Abstract - *Data deduplication is one of the emerging techniques to improve the capacity of the storage media (Hard disk, Tape, CD, DVD, ROM) by removing redundant data and provide storage only for unique data. That increases the storage efficiency of the storage media. There are techniques like a large scale deduplication in that a two stage deduplication scheme is used in that scheme first a sampling technique is used to generate subsets and in second stage an active selection technique is used to reduce redundancy in the document. In this paper a survey over various techniques which used to remove redundant data from the document and provide storage only for unique data.*

Key Words **- Deduplication, storage media, Redundant data.**

1.INTRODUCTION :

The world is producing large amount of data rapidly. According to a Survey, information that is producing per year is growing by fifty seven percent annually. Thirty five percent of this information is generated by large enterprises. so it is very critical for a disaster recovery site to backup data regularly. This rapidly growing data arises many challenges to the existing storage systems. Significance fraction of information contains duplicates.

Thus deduplication techniques have been invented. Deduplication is the process to remove duplicate copies in the data, that also known as intelligent compression and single event storage. That technique put huge advantage over storage efficiency and also improves network efficiency to transfer data by reducing bits to transfer data. In that technique unique data samples, data chunks, or byte patterns are

collected and stored into the storage when a match is found with the stored data then that data remove by a small reference. Thus storage based deduplication improve the efficiency of the storage and reduce the cost overhead of storing more files of the same data. But in that process data will be stored in different forms and deduplication process stored data that arise serious concern about integrity of the data.

Mainly there are three type of deduplication is occurs in that in-line deduplication, post-process deduplication, Mutual co-operation based deduplication.

(a)In-line deduplication: In –inline deduplication, the process of reducing duplicate data conducted before it stored in to the storage media, if there any redundant data is found in the process than that data is not stored at the device.

(b)Post process deduplication: In post process deduplication first data stored at the devices and then the task of checking redundant data and then the process of removing redundancy is performed. If there any redundant data is found then that data will be removed.

(c)Mutual co-operation based deduplication:
In that both techniques inline deduplication and post process deduplication are used to provide better performance to the user. In that deduplication task performs at both ends, at client end and server end.
There are some other techniques which can use various encryption techniques to provide secure deduplication for data that preserves the security of the data during the process of deduplication.

2. LITERATURE REVIEW:

E. Manogar, S. Abirami [14] a study over various data deduplication technique is presented. Generally there are two type of techniques are used to reduce redundant data from the storage called data deduplication, data reduction. A study over various deduplication techniques is presented. in that there are three type of data deduplication techniques called location based, Time based, chunk based are used. In location based technique deduplication is performed on the basis of different positions that can be performed either at source side or target side. In time based technique data can be processed by three processes before written (inline) into the disk, after written into the disk (post) or both inline processing or post processing. In chunk based technique data files are divided into small size of bytes called chunks, and these are used to eliminate redundant data.

Pierre Meye, Philippe Raipin Fred'eric Tronel, Emmanuelle [2] a secure data deduplication scheme is presented. In that scheme encryption keys are generating on continuous and consistent basis, a single key of each chunk is generated that reduces the redundancy of the data. In that technique that key is also only known for the intended users like sender and receiver. That keeps the privacy of the data. Chunks are encrypted thus it hard to find which chunk is uses by which user. In this technique a convergent encryption scheme is used to encrypt chunks. An authentication and anonymous model is presented that helps to provide a secure and deduplicated data storage.

Rongmao Chen, Yi Mu, Guomin Yang and Fuchun Guo [15] Presents Block level Message Locked Encryption (BL-MLE) scheme which used to provide deduplication for large scale data in cloud computing. In existing technique message locked encryption (MLE) is used, but in that technique there is large amount of data need to be handled by the end used and cloud server. In BL-MLE a management scheme is used which handles small amount of metadata for deduplication purpose. That way it enhances the performance of the whole system. In that paper Dual Level Source Based (DLSB) Deduplication is used. In that scheme firstly an identifier is sent to the user to check deduplication in the server, if any deduplicated file detected then ownership request is generated by the use of POW (proof of ownership) protocols. Otherwise that identifier added to the server which pointed to the file in the server.

Zhe Sun, Jun Shena, Jianming Yong [3] a flash memory based penalty index lookup scheme for RAM is presented. In that scheme Flash memory reduces the gap between RAM and Hard-disk and provides suitable mechanism for that data is used, in that a chunk stash is used, in that chunks metadata stored a flash memory. That arranges chunks data in large log structure of chunks which provides a sequential write policy. That reduces the duplication in the RAM and improves efficiency of the ram. In that a NAND based Flash memory is used which provides better storage capacity.

Mark W. Storer Kevin Greenan Darrell D. E. Long Ethan L. Miller [4] presents a sparse indexing technique to deduce duplication in the storage. In that a sampling is used to exploits inherent duplication in the data to enhance the performance of the storage. There are two types of deduplication are there inline deduplication, out of line deduplication. In inline deduplication data is deduplicated before it hitting to the disk, and in out of line deduplication, data deduplicated when data stored at disks. In that technique sparse indexing, content based segmentation, and sampling is used divide incoming streams in to segments and deduplication is performed to remove duplicate data from that content.

Guilherme Dal Bianco, Renata Galante, Marcos Andre Gonc ¸alves, Sergio Canuto, and Carlos [1] Presented a deduplication scheme for cloud computing. In that scheme Hadoop based HDFS (Hadoop Distributed File System) system and HBase (Hadoop Database) is used to provide backend storage system. A deduplication technique is used to provide a scalable and parallel deduplicated storage system. In that technique HDFS and HBase are used to provide an enhanced functionality for that system by keeping the record of the data that stored at the storage system. In that system HDFS and HBase are not used to store data rather stores location of that data. That provides a DEDU (deduplication) system to store data.

Chun-I Fan, Shi-Yuan Huang, and Wen-Che Hsu [16] presents an encrypted data deduplication mechanism. In that scheme two steps are used in first step cipher text will be constructed and in second step deduplication of that cipher text is conducted. To encrypt data an AES encryption based technique is used that enhance the performance of deduplication. By the use of encrypted cipher text for storage preserve the integrity of the data and maintain the confidentiality of the sensitive document. In the storage only unique updated data block in the given cipher structure in entered. That reduces the redundancy of the data. that way a enhance technique is provided to provide an enhance and secure framework for deduplication task.

Pedro Neves, Paulo Ferreira, João Barreto [17] presented a web prefetching and data deduplication technique. Web prefetching is used to decrease the search latency for the user. In web prefetching user's request are preprocessed before it demanded by the used that improve the latency of the search. For deduplication purpose there are three type of techniques called classic caching, delta encoding, compare by hash, are used. In classic caching, all the resources are stored in the browsers cache. For any user's request first search is performed in browser's cache if that content is finding in cache. In delta encoding two files are compared and their deference is computed. That used for deduplication purpose. In compare by-hash, data is divided into small size of blocks called chuncks. Encrypted hash value for these chunks is used to compare and reduce redundancy in the storage system.

Qinlu He, Zhanhuai Li, Xiao Zhang [18] a description over the data deduplication technique is presented. Data deduplication is the technique which used to reduce redundancy in the storage data, there are two types of strategies are used for deduplication purpose one is file level deduplication, block level data deduplication. In file level deduplication, single instance storage is used to perform deduplication task. In block level data deduplication, data files are divided into blocks and these blocks are compared to either these blocks are contains same value or not. That way the task of data deduplication isperformed.

Table 2.1: Comparison for the various deduplication technique

Technique	Advantage	Disadvantage
Out of line deduplication	In that technique when data is stored into the server then deduplication task is performed. That reduces the possibility of loss of data.	In that a large space is required first to store data. That generates space overhead for the technique.
Inline Deduplication	In that technique a deduplication task is performed at client end and then data transmitted over the network channel. That reduces the space over head for the storage.	In that, an enhanced technique is required to provide better performance to deduplicate data. It generate network bottleneck problem.
MLE (Message Locked Encryption)	In that a message locked encryption technique is used to provide better and secure mechanism to store data.A deduplication task is performed in that encrypted data to provide better performance to deduplicate data.	In that whole data is consider at once to perform deduplication task. That degrades the performance of the whole technique and too much time to deduplicate data
BL-MLE (Block Level Message Locked Encryption)	In that technique small blocks of metadata are used to perform deduplication task. It enhance the performance of the MLE-technique	In that dual phase spruce level or client level deduplication is used to perform deduplication, which poses defect of generation of network bottleneck.

3. CONCLUSION:

In this paper a survey over different techniques which used for reduce duplicate data or which used for deduplication purpose is presented. Deduplication is one of the emerging technique to reduce the redundancy of the data and provide storage for the unique data. There is an overview over the techniques like two stage sampling strategy, a technique which use intra and inter domain deduplication to reduce the data, an encryption based deduplication technique is also presented and some other technique which usesd to prevent
integrity of the data is presented. For future work a technique is presented which provide an advanced deduplication comparing to the other techniques.

REFERENCES:

[1] Guilherme Dal Bianco, Renata Galante, Marcos Andre Gonc ¸alves, Sergio Canuto, and Carlos A. Heuser "A Practical and Effective Sampling Selection Strategy for Large Scale Deduplication" IEEE, September 2015.
[2]Pierre Meye*, Philippe Raipin* Fred'eric Tronel, Emmanuelle anceaume "A secure two-phase data deduplication scheme" IDC information, May 2010.
[3] Zhe SUN, Jun SHENA, Jianming YONG "A novel approach to data deduplication over the engineering-oriented cloud systems" Integrated Computer Aided Engineering, 2013.
[4] Mark W. Storer Kevin Greenan Darrell D. E. Long Ethan L. Miller "Secure Data Deduplication" ACM, October 2008.
[5] Biplob Debnath, Sudipta Sengupta, Jin Li "ChunkStash: Speeding up Inline Storage Deduplication using Flash Memory" 2010.
[6] Mark Lillibridge, Kave Eshghi, Deepavali Bhagwat, Vinay Deolalikar, Greg Trezise and Peter Camble "Sparse Indexing: Large Scale, Inline Deduplication Using Sampling and Locality" 2009.
[7] S. Chaudhuri, V. Ganti, and R. Kaushik, "A primitive operator for similarity joins in data cleaning," in Proc. 22nd Int. Conf. Data Eng., p. 5, Apr. 2006.
[8] P. Christen, "Automatic record linkage using seeded nearest neighbor and support vector machine classification," in Proc. 14th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2008, pp. 151–159.
[9] P. Christen, "A survey of indexing techniques for scalable record linkage and deduplication," IEEE Trans. Knowl. Data Eng., vol. 24, no. 9, pp. 1537–1555, Sep. 2012.
[10] P. Christen and T. Churches, "Febrl-freely extensible biomedical record linkage," Computer Science, Australian National University, Tech. Rep. TR-CS-02-05, 2002.
[11] Cohn, L. Atlas, and R. Ladner, "Improving generalization with active learning," Mach. Learn., vol. 15, no. 2, pp. 201–221, 1994.
[12] G. Dal Bianco, R. Galante, C. A. Heuser, and M. A. Gonalves, "Tuning large scale deduplication with reduced effort," in Proc. 25th Int. Conf. Scientific Statist. Database Management. 2013, pp. 1–12.
[13] M. G. de Carvalho, A. H. Laender, M. A. Goncalves, and A. S. da Silva, "A genetic programming approach to record deduplication," IEEE Trans. Knowl. Data Eng., vol. 24, no. 3, pp. 399–412, Mar. 2012.
[14] E. Manogar, S. Abirami "A Study on Data Deduplication Techniques for Optimized Storage" IEEE, 2014.
[15] Rongmao Chen, Yi Mu, Senior Member, Guomin Yang, Member and Fuchun Guo "BL-MLE: Block-Level Message-Locked Encryption for Secure Large File Deduplication" IEEE, 2015.
[16] Chun-I Fan, Shi-Yuan Huang, Wen-Che Hsu "Encrypted Data Deduplication in Cloud Storage" IEEE, 2015.
[17] Pedro Neves, Paulo Ferreira, João Barreto "Leveraging Web prefetching systems with data deduplication" IEEE, 2015.
[18] Qinlu He, Zhanhuai Li, Xiao Zhang "Data Deduplication Techniques" IEEE, 2010.
[19] A.K. Elmagarmid, P.G. Ipeirotis, and V.S. Verykios, "Duplicate Record Detection: A Survey," IEEE Trans. Knowledge and Data Eng., vol. 19, no. 1, pp. 1-16, Jan. 2007.
[20] M. W. Storer, K. Greenan, D. D. E. Long, and E. L. Miller. Secure data deduplication. n. In Proc. of StorageSS, 2008 .
[21] E. Rahm and H.H. Do, "Data Cleaning: Problems and Current Approaches," IEEE Technical Committee Data Eng. Bull., vol. 23, no. 4, pp. 3-13, Dec. 2000.