ISSN 2348 – 9928
IJAICT Volume 5, Issue 9, September 2018 Doi:01.0401/ijaict.2015.12.06 Published on 05(9) 2018

A SECURE AUTHORIZED NLP BASED DUPLICATE CHECK SCHEME FOR DATA DEDUPLICATION

Mr. S. Venkatraman
Associate Professor

Dr. D. Srinath
Associate Professor

Mr. C. Prasanth
Student

Computer Science and Engineering,
Panimalar Institute of Technology,
Chennai, Tamilnadu, India

Abstract— Data deduplication is a technique for reducing the amount of storage space an organization needs to save its data. It has been widely used in cloud computing and storage to minimize the amount of storage space and conserve bandwidth. To conceal the confidentiality of sensitive data, the convergent encryption technique has been proposed to encrypt the data before outsourcing. Several new deduplication constructions are presented to support authorized duplicate check in hybrid cloud architecture. But it would check only the contents that are same in both the file. In this paper, we propose a system that performs natural language processing to identify the elements with similar meanings and deduplicate the file, so that that deduplication is more efficient.

Keywords— Deduplication, Convergent Encryption, Hybrid Cloud.

I. INTRODUCTION

Cloud computing [1][3][16] is having an exponential popularity growth as it provides necessary storage area and parallel processing resources at relatively cheaper rates. Since the data is enormous, there is a need to avoid duplicate copies to save the wastage of memory and reduce the cost expenditure. It is also important to maintain the confidentiality of the storage data. So it is needed to encrypt the data before uploading it to the cloud [2][4]. If each user encrypts the data in their own way it becomes difficult to deduplicate the files. So a convergent encryption technique is used to encrypt the data before storing it in the cloud. Hybrid cloud is a combination of both public cloud and private cloud [6].

A private cloud is used to store the critical information of an organisation, while public cloud is used to store ordinary and easily accessible information. The privileges of the users must also be considered before performing deduplication. The files of the users with the same privileges must only be considered for deduplication process. The privilege keys are also encrypted. Now the encrypted data files are stored in the public cloud. The encrypted privilege keys are stored in the private cloud [16].

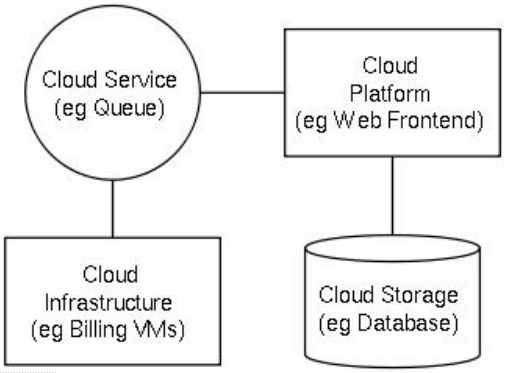

Fig 1: Architecture of Cloud Computing

In this paper, thededuplication techniques are analysed to determine which technique is effective for deduplication process. It is important to determine the best deduplication technique as it is an effective way for cost-saving. The cost-saving method may make a small difference in the expenditure of an individual. But for an organisation that deals with millions of files and billions of information, the difference in expenditure is huge [7]. It may even increase their annual profit by some percentage. Although data deduplication brings a lot of benefits, security and privacy concerns arise as users' sensitive data are susceptible to both insider and outsider attacks.

Traditional encryption, while providing data confidentiality, is incompatible with data deduplication. Specifically, traditional encryption requires different users to encrypt their data with their own keys [9][10]. Data deduplication is a specialized data compression technique for eliminating duplicate copies of repeating data in storage. The technique is used to improve storage utilization and can also be applied to network data

transfers to reduce the number of bytes that must be sent [11][12]. Instead of keeping multiple data copies with the same content, deduplication eliminates redundant data by keeping only one physical copy and referring other redundant data to that copy [13].

In this paper, the post-process deduplication and the in-line deduplication techniques are analysed. Based on the place where the deduplication occurs, the source deduplication and the target deduplication are analysed.

1.1 TYPES OF DEDUPLICATION

A. Based on Technique

1. POST-PROCESS DEDUPLICATION

In this process, the files are first stored in the cloud. Later, the deduplication process is carried out to detect the duplicate files. The advantage of post process deduplication is that there is no need to wait for calculating hash values and deduplication process before storing it in the cloud [14]. This ensures that the storage performance has not been degraded. The disadvantage of this process is that the duplicate copies can be stored in the cloud for a short time. This may affect the storage of other files when the remaining storage is very less [8].

2. IN-LINE DEDUPLICTION

In this process, the calculation of hash values and deduplication process takes place when the data enters the cloud in real time. If there is a redundant data block that is already present in the previously stored data files, the block is replaced by a reference pointer to the data block in the original file. This technique is efficient in terms of storage as it never allows the redundant data to be stored in the cloud. But in terms of time, it may take a while to perform hash calculations for deduplication [8][9].

B. Based on Place

1. SOURCE DEDUPLICATION

In source deduplication, the deduplication process takes place near the place where the data is created. Thus the data on the data source is reduplicated periodically. The deduplication takes place by creating the hash values for newly created data and comparing it with hash values of pervious files. When a duplicate data copy is found, the duplicate data block is removed and a pointer is used to refer the old file. When a block of data is modified in the file, then a copy of the block is created using Copy-on write system.

2. TARGET DEDUPLICATION

In target deduplication, the deduplication process takes place in the secondary storage area.Generally this will be a backup store such as a data repository or a virtual tape library[6].

II. LITERATURE SURVEY

Reclaiming Space from Duplicate files in a Serverless Distributed File system uses convergent encryption technique to coalesce duplicate files into single file [2]. In our paper we use convergent encryption technique to compare the contents of encrypted files. Secure Data Deduplication uses encryption technique that creates encryption keys based on chunks [12]. In our paper we use encryption technique to create identical keys for identical chunks.

Fast and Secure Laptop Backups with Encrypted De-duplication takes the common data between files to increase speed of backups [1]. In our paper we take out the common contents to prevent wastage of memory. 'RevDedup', a deduplication system that optimizes reads to latest VM image backups using an idea called reverse deduplication.

RevDedup removes duplicates from old data, thereby shifting fragmentation to old data while keeping the layout of new data as sequential as possible in contrast with conventional deduplication that removes duplicates from new data. It achieves high deduplication efficiency high backup and read throughput [3].

Weak Leakage-Resilient Client-side Deduplication of Encrypted Data in Cloud Storage provides a secure client-side deduplication scheme that protects data confidentiality against both outside adversaries and honest-but-curious cloud storage serve. The scheme is proved secure with respect to any distribution with sufficient min-entropy [5].

Secure and Constant Cost Public Cloud Storage Auditing with Deduplication a scheme based on techniques including polynomial-based authentication tags and homomorphic linear authenticators. It allows deduplication of both files and their corresponding authentication tags. Data integrity auditing and storage deduplication are achieved simultaneously. It provides constant cost scheme that achieves secure public data integrity auditing and storage deduplication at the same time in the Secure and Constant Cost Public Cloud Storage Auditing with Deduplication [6].

An Application-aware Local-Global source deduplication scheme that improves data deduplication efficiency by exploiting application awareness, and also combines local and global duplicate detection to strike a good balance between cloud storage capacity saving and deduplication time reduction[14].

An application based deduplication approach and indexing scheme containing block that preserved caching which

maintains the locality of the fingerprint of duplicate content to achieve high hit ratio and to overcome the lookup performance, reduced cost for cloud backup services and increase dedulpication efficiency. It is shown the scheme improves the backup performance, reduce the system overhead and improve the data transfer efficiency on cloud [15].

Fast and Secure Laptop Backups with Encrypted De-duplication supports client-end per-user encryption which is necessary for confidential personal data and also supports a unique feature which allows immediate detection of common subtrees, avoiding the need to query the backup system for every file[1].

III. SYSTEM ARCHITECTURE

In our system, user tries to upload the file; the file undergoes Natural Language Processing (NLP). Thus the meaningful entities are extracted from the file. Now, a convergent key encryption technique is used to generate key based on the contents. With the help of the keys, the file is encrypted.

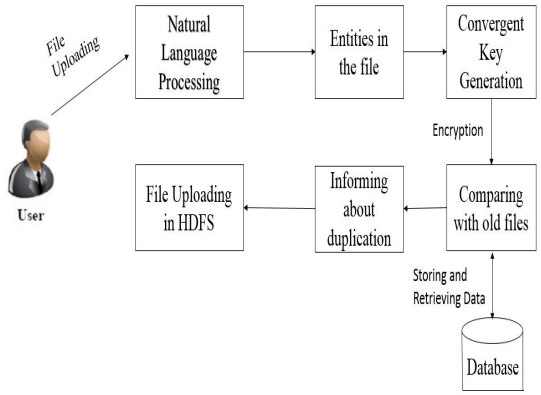

Fig 2 : System Architecture

The encrypted files are then compared with the files in the cloud and the amount of duplicate contents in the file if reported to the user. The user is then allowed to take a decision on storing the file. Based on the user's option the file is stored or deleted. This system allows deduplicating the files with different words but with similar meanings.

3.1 Natural Language Processing Of Quries

Three techniques are used to translate NLQs to NEXI in INEX 2004 and 2005. The three approaches are called Hassler, Tannier (Tannier, 2005) and Woodley (Woodley and Geva, 2005) after their authors. While each of the approaches is different, they all contain four main stages.

Detecting Structural and Content Constraints: The first stage is to detect a query's structural and content constraints. Hassler uses template matching based on words and parts-of speech. Links between structure and content are not linguistically motivated, and it is assumed that content is the last element. Woodley adds shallow syntactic parsing before applying the same kind of template matching. Tannier uses deep syntactic analysis, complemented by some specific semantic rules concerning query structure.

Structure Analysis: The second stage is to map structural constraints to corresponding XML tags. This requires lexical knowledge about the documents' structure, since the tags in the XML documents are rarely "real" words or phrases, but abbreviations, acronyms or an amalgamation of two. Furthermore, a single tag can be referred to by different names. Tannier uses grammatical knowledge to recognise some frequent linguistic constructions that imply structure.

Content Analysis: The third stage is to derive users' content requirements, as either terms or phrases. Noun phrases are particularly useful in information retrieval. They are identified as specific sequences of parts-of-speech. Tannier is also able to use content terms to set up a contextual search along the entire structure of the documents.

NEXI Query Formulation: The final stage of translation is the formulation of NEXI queries. Following NEXI format, content terms are delimitated by spaces, with phrases surrounded by quotation marks

IV. SYSTEM IMPLEMENTATION

Initially, the user is given two options. Either to register as a new user or to login as an existing user. Once the user has registered, his details including email id and password are stored in the database. Now he can enter into the cloud to store or retrieve data files. When the user wants to upload a file, he can select the file from the system. When the file is ready to be uploaded, it undergoes Natural Language Processing (NLP). The NLP is performed using an online tool called Alchemy API.

This tool removes the prepositions from the file and selects only the meaningful words. These words are then stored as an XML file. Then the convergent key is generated based on the words obtained. The advantage of storing as a XML file is that the new file's concept is enough to be compared with the concept of previously stored files, instead of checking the entire file. This reduces the wastage of resources. The files whose concept matches are only taken for content matching.

In the next step, the contents of the chosen file are compared to find how much percentage of the file contents matches with the contents of previously stored files. The amount of duplicate data is informed to the user and the option is left to the user to store the new file, replace the existing file or discard the new file. If the user decides to store the file, the file is encrypted based on the convergent key and stored in the cloud. If there is no file that matches with the concept of the file to be uploaded the content match is not performed. The file is directly encrypted and gets stored in the cloud.

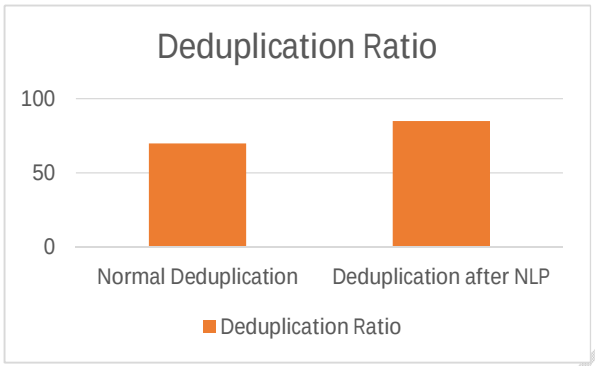

Fig 3 : Comparison of Normal and Deduplication after NLP

V. CONCLUSION

Since the storage in cloud is increased, we must try to reduce the cost expenditure by minimizing the amount of duplicate data in the cloud. So we have proposed a system that will find 15% more duplicate data by performing NLP and helps reduce the amount of duplicate contents in an effective manner, thereby saving the resources than the existing system.

References

[1] P. Anderson and L. Zhang. Fast and secure laptop backups with encrypted de-duplication. In Proc. of USENIX LISA, 2010.

[2] J. R. Douceur, A. Adya, W. J. Bolosky, D. Simon, and M. Theimer. Reclaiming space from duplicate files in a serverless distributed file system. In ICDCS, pages 617–624, 2002.

[3] C. Ng and P. Lee. Revdedup: A reverse deduplication storage system optimized for reads to latest backups. In Proc. of APSYS, Apr 2013.

[4] A. Rahumed, H. C. H. Chen, Y. Tang, P. P. C. Lee, and J. C. S.Lui. A secure cloud backup system with assured deletion and version control. In 3rd International Workshop on Security in Cloud Computing, 2011.

[5] J. Xu, E.-C. Chang, and J. Zhou. Weak leakage-resilient client-side deduplication of encrypted data in cloud storage. In ASIACCS, pages 195–206, 2013.

[6] J. Yuan and S. Yu. Secure and constant cost public cloud storage auditing with deduplication. IACR Cryptology ePrint Archive, 2013:149, 2013.

[7] K. Zhang, X. Zhou, Y. Chen, X.Wang, and Y. Ruan. Sedic: privacyaware data intensive computing on hybrid clouds. In Proceedings of the 18th ACM conference on Computer and communications security, CCS'11, pages 515–526, New York, NY, USA, 2011. ACM.

[8] G. Kakariya and Prof. S. Rangdale. A Hybrid Cloud Approach for Secure Authorized Deduplication. International Journal of Computer Engineering and Applications, Volume VIII, Issue I, October 14

[9] S. Bugiel, S. Nurnberger, A. Sadeghi, and T. Schneider. Twin clouds: An architecture for secure cloud computing. In Workshop on Cryptography and Security in Clouds (WCSC 2011), 2011.

[10] W. K. Ng, Y. Wen, and H. Zhu. Private data deduplication protocols in cloud storage. In S. Ossowski and P. Lecca, editors, Proceedings of the 27th Annual ACM Symposium on Applied Computing, pages 441–446. ACM, 2012.

[11] J. Stanek, A. Sorniotti, E. Androulaki, and L. Kencl. A secure data deduplication scheme for cloud storage. In Technical Report, 2013.

[12] M. W. Storer, K. Greenan, D. D. E. Long, and E. L. Miller. Secure data deduplication. In Proc. of StorageSS, 2008.

[13] Z. Wilcox-O'Hearn and B. Warner. Tahoe: the least-authorityfilesystem. In Proc. of ACM StorageSS, 2008.

[14] Y. Fu, H. Jiang, N. Xiao, L. Tian, F. Liu and L. Xu. Application-Aware Local-Global Source Deduplication for Cloud Backup Services of Personal Storage. In IEEE Transactions On Parallel And Distributed Systems, Vol. 25, No. 5, May 2014.

[15] J. Malhotra and P. Ghyare. A Novel Way of Deduplication Approach for Cloud Backup Services Using Block Index Caching Technique. In IJAREEIE, Vol. 3, Issue 7, July-2014.

[16] J. Li, Y.K. Li, X.Chen, P.P.C. Lee and W. Lou. A Hybrid Cloud Approach for Secure Authorized Deduplication. In IEEE Transactions on Parallel and Distributed Systems, DOP: 18, April 2014.