International Journal of
Emerging Research in Management &Technology
ISSN: 2278-9359 (Volume-6, Issue-6)

Research Article | **June 2017**

Fast and Efficient Cloud Data Utilization with Deduplication

Sunil S
MTech Scholar, Department of CSE,
School of Computing and Information Technology
Reva University, Bangalore, India

A Ananda Shankar
Associate Professor, Department of CSE,
School of Computing and Information Technology
Reva University, Bangalore, India

Abstract—

Cloud storage system is to provides facilitative file storage and sharing services for distributed clients.The cloud storage preserve the privacy of data holders by proposing a scheme to manage encrypted data storage with deduplication. This process can flexibly support data sharing with deduplication even when the data holder is offline, and it does not intrude the privacy of data holders. It is an effective approach to verify data ownership and check duplicate storage with secure challenge and big data support. We integrate cloud data deduplication with data access control in a simple way, thus reconciling data deduplication and encryption.We prove the security and assess the performance through analysis and simulation. The results show its efficiency, effectiveness and applicability.In this proposed system the upload data will be stored on the cloud based on date.This means that it has to be available to the data holder who need it when they need it. The web log record represents whether the keyword is repeated or not. Records with only repeated search data are retained in primary storage in cloud. All the other records are stored in temporary storage server. This step reduces the size of the web log thereby avoids the burden on the memory and speeds up the analysis.

Keywords- De-duplication, Proxy Re-encryption, Cloud storage, Data Ownership.

I. INTRODUCTION

Cloud computing is a rapidly growing technology where resources such as storage devices, platform and applications are shared over the internet and is widely used by multiple users in small and medium business. Cloud services can be provided and delivered remotely by vendors such as Amazon or Microsoft as “public clouds”, or the resources are designed, installed, monitored and controlled internally as “private clouds”. Cloud data retrieval is an important service to be considered as certain specific data files the users are interested during a given session must be retrieved in an efficient way and quickly.

Cloud users upload personal or confidential data to the data center of a Cloud Service Provider (CSP) and allow it to maintain these data. Since intrusions and attacks towards sensitive data at CSP are not avoidable , it is prudent to assume that CSP cannot be fully trusted by cloud users. Moreover, the loss of control over their own personal data , leads to high data security risks, especially data privacy leakages . Due to the rapid development of data mining and other analysis technologies, the privacy issue becomes serious . Hence, a good practice is to only outsource encrypted data to the cloud in order to ensure data security and user privacy . But the same or different users may upload duplicated data in encrypted form to CSP, especially for scenarios where data are shared among many users. Although cloud storage space is huge, data duplication greatly wastes network resources, consumes a lot of energy, and complicates data management. The development of numerous services further makes it urgent to deploy efficient resource management mechanisms . Consequently, deduplication becomes critical for big data storage and processing in the cloud.

Web Mining is type of data mining which is used for web. Web Mining has become an visible and an important movement because of the main reason that today storage of data has become enormous that retrieving and processing them has become an overhead. Two types of approaches were taken in initially for defining Web mining they were process-centric view and data-centric view. Process centric accounted the sequence of tasks where as data centric accounted for the types of web data that was being used in the mining process. The second type is taken into consideration widely in recent times. Web Mining have a three categories: Web Content Mining, Web Usage Mining and Web Structure Mining. The implementation of these categorises on World Wide Web have been well reviewed.

II. MAIN ARCHITECTURE

CSP that offers storage services and cannot be fully trusted since it is curious about the contents of stored data, but should perform honestly on data storage in order to gain commercial profits,data holder that uploads and saves its data at CSP. In the system, it is possible to have a number of eligible data holder that could save the same encrypted raw data in CSP. The data holder that produces or creates the file is regarded as data holder. It has higher priority than other normal data holder. An authentication provider (AP) that does not collude with CSP and is fully trusted by the data holders to verify data ownership and handle data deduplication. In this method, Authentication Provider cannot know the data stored in CSP and CSP should not know the plain user data in its storage. In this part we applied data preprocessing technique at the starting. Then we counted the frequency of the access the file, the term Frequency of file means that the numbers of visibility of that keyword in a particular file.

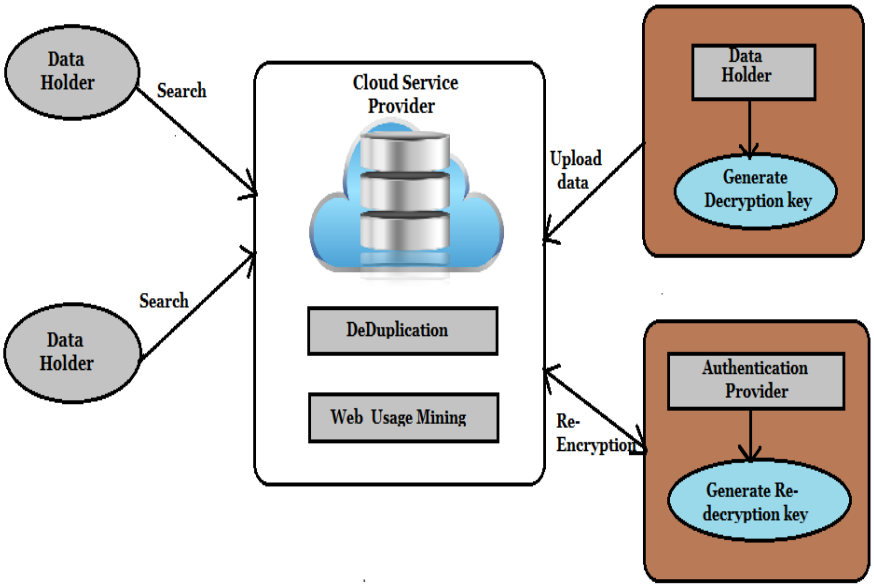

Figure 1: Main Architecture

III. ALGORITHM

Apriori is an influential algorithm presented by R. Agrawal and R. Srikant in 1994 for the purpose of mining frequent document for Boolean association rules. The name of this algorithm has been derived by the fact that the algorithm uses prior knowledge of frequent document properties. The apriori algorithm uses iterative approach that is called level-wise search. In this algorithm, n-document are used to find out (n+1)-documents. The frequent search document is acquired by inspect the database to collect the count for each document, and collecting the document that satisfy minimum support count for each document, and accumulating only those documents that have minimum support count. This algorithm is used to Finding the frequently used vital data for the encryption based on the counting of access in particular document. Then also this algorithm to identifying the sensitive dataset from the original data set.

ALGORITHMIC FRAME

procedure Apriori (M, minSupport)
{ //M is the database and minSupport is the minimum support
L1= {frequent documnets};
for (k= 2; L_{k-1} !=∅; k++)
{
C_k= data holder generated from L_{k-1}
//that is cartesian product L_{k-1} x L_{k-1} and eliminating any k-1 size document that is not
//frequent
for each search t in database do
{
#increment the count of all dataholders in Ck that are contained in t
L_k = dataholder in C_k with minSupport
}//end for each
}//end for
return $\cup_k L_k$;
}

IV. IMPLEMENTATION

A user uploads its data to cloud by sending key. The cloud service provider should first check if the same token has existed (by comparing the token with the records in cloud service provider, which is inevitable in any deduplication schemes). Then, Cloud Service Provider chooses to save the data if the token does not exist. If the data holder uploads the same data, Cloud Service Provider contacts Authentication Provider for gaining a re-encryption key if the ownership challenge is positive. In this case, Cloud Service Provider has to finish the re-encryption operation of Re-Encryption , which requires 1 pairing. Cloud Service Provider is responsible for allowing the access to the same data for all data owner by avoiding storing the same data in the cloud.

Data Deduplication: To upload file the user and the CSP perform both deduplications. The file-level deduplication operation is identical to that in the baseline approach. More precisely, the user sends the file tag to the CSP for the file duplicate check. If a file duplicate is found, the user will run the PoW protocol POWF with the CSP to prove the file ownership. If no duplicate exists, CSP stores the ciphertext with key and returns the corresponding pointers back to user for local storage. In deduplication on the other hand of keeping the multiple data copies with the same file content, deduplication eliminates repeated data by keeping only single copy and referring other redundant data to that

single copy. The file level deduplication to eliminates duplicate copies of the same file. Deduplication can also be used at the block level, which eliminates duplicate blocks of data that occur in non identical files.

Figure 2: Upload Document

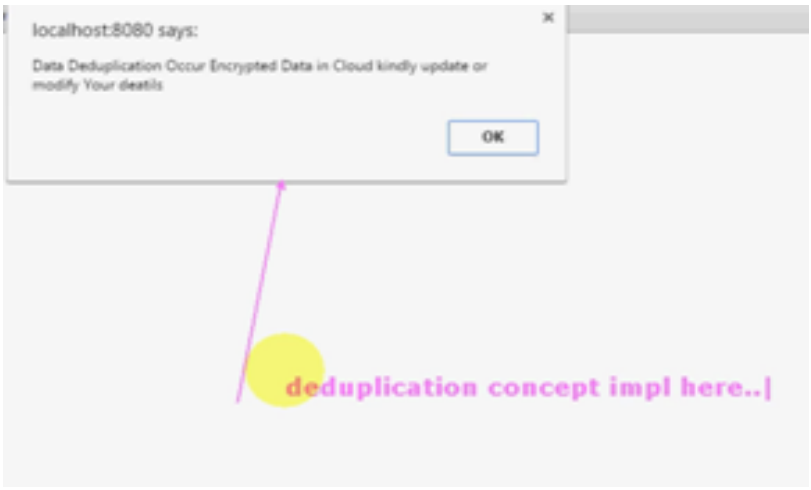

Figure 3: Deduplication

RE-ENCRYPTION: Re-encryption is a means for confidential and flexible technique for a data holder to store and share data. A data holder can encrypt the file with a public key and then store the ciphertext in a trusted server.When a user arrives, the data holder can delegate a reencryption key associated with the particular user to the trusted server.Then the re-encrypt the initial ciphertext to the desired user.The purpose of re-encryption schemes is to prevent the revelation of the keys involved in re-encryption and the plaintext that needs to be re-encrypted to the server. The re-encryption schemes are basically a version of existing encryption schemes consisting of selection of text, generation of keys, sharing or transmitting of keys between the parties, changeover from plaintext to cipher-text on one end and changeover from cipher-text to plaintext on the other end, the difference arises with the introduction of two more properties Directionality and Transitivity.

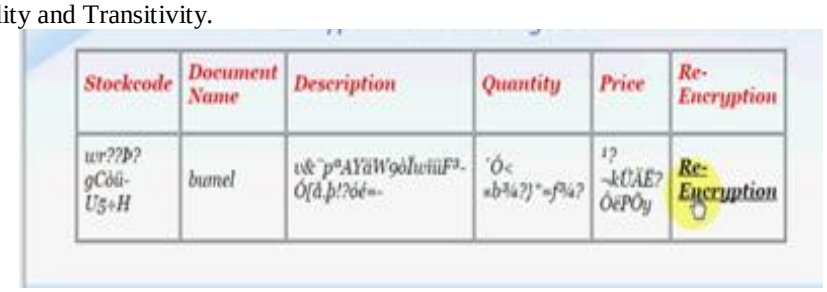

Figure 4: Reencryption

Web usage mining: Cloud computing is an emanating technology allowing data holder to perform data processing, use as storage and data admission services from around the world through internet. The Cloud service

providers charge depending on the data holder usage. Imposing confidentiality and scalability on cloud data increases the complexity of cloud computing. As sensitive information is centralized into the cloud, this information must be encrypted and uploaded to cloud for the data privacy and efficient data utilization. As the data becomes complex and number of users are increasing searching of the files must be allowed through keyword of the data holder interest. The traditional searchable encryption schemes allows data holder to search in the encrypted cloud data through keywords, which support only Boolean search, i.e., whether a keyword exists in a file or not, without any relevance of data files and the queried keyword. Searching of data in the cloud using keyword ranked search results too coarse output and the data privacy is opposed using server side ranking based on order-preserving encryption (OPE).

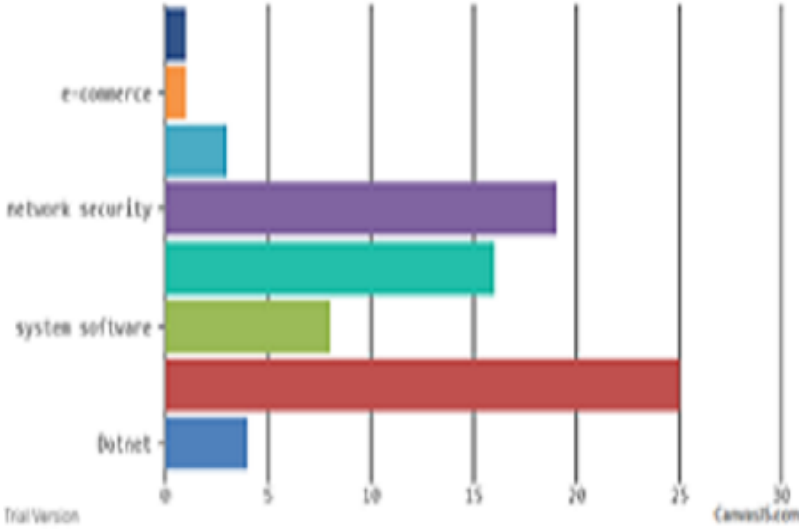

Figure 5 : Web Log details

V. RESULT

We implement this process for which we search the file and checked it whether the file is there or not. If the file search matched again and again then we increment its counting by one each time and this counting only shows its frequency.

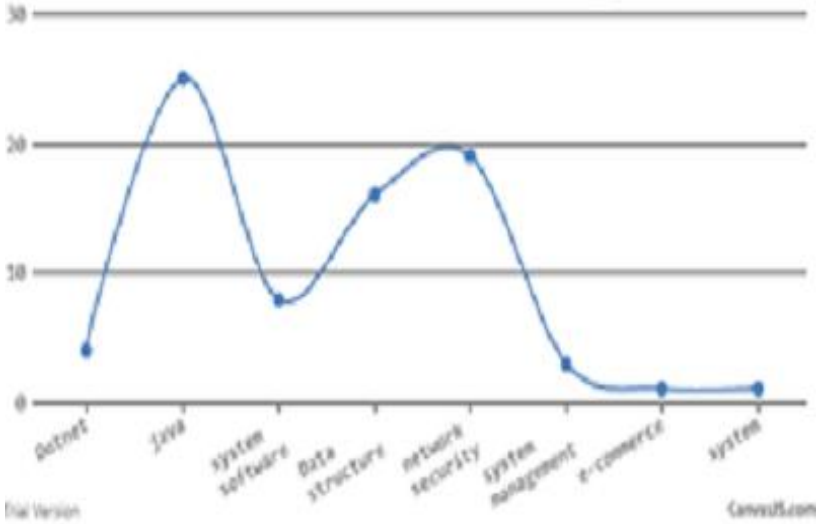

Figure 6 : Result

VI. CONCLUSION

The important task of Web Usage Mining is data preprocessing. For applying data mining techniques our data must be preprocessed, after the can be use data mining techniques like classification, clustering and association rule mining etc. Generally data preprocessing is a time consuming process. When our data would be preprocessed then we can easily detect the user's behavior that used the document. The term user behavior means that how much time user use a particular document. According to this kind of information we can judge the interesting information applying some data mining techniques. Here we counted the document access frequency. So that frequent access document and infrequent access document can be evaluated and stored in the cloud.

REFERENCES

[1] M. Bellare, S. Keelveedhi, and T. Ristenpart, "DupLESS: Server aided encryption for deduplicated storage," in Proc. 22nd USENIX Conf. Secur., 2013, pp. 179–194.

[2] Dropbox, A file-storage and sharing service. (2016). [Online].Available: http://www.dropbox.com
[3] Google Drive. (2016). [Online]. Available: http://drive.google.com
[4] Mozy, Mozy: A File-storage and Sharing Service. (2016). [Online].Available: http://mozy.com/
[5] 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 Proc. IEEE Int. Conf. Distrib. Comput. Syst., 2002,pp. 617–624, doi:10.1109/ICDCS.2002.1022312.
[6] G. Wallace, et al., "Characteristics of backup workloads in production systems," in Proc. USENIX Conf. File Storage Technol., 2012,pp. 1–16.
[7] Z. O. Wilcox, "Convergent encryption reconsidered," 2011.[Online].Available:http://www.mailarchive.com/cryptography@ metzdowd.com/msg08949.html
[8] G. Ateniese, K. Fu, M. Green, and S. Hohenberger, "Improved proxy re-encryption schemes with applications to secure distributed storage," ACM Trans. Inform. Syst. Secur., vol. 9, no. 1, pp. 1–30, 2006, doi:10.1145/1127345.1127346.
[9] Opendedup. (2016). [Online]. Available: http://opendedup.org/
[10] D. T. Meyer and W. J Bolosky, "A study of practical deduplication," ACM Trans. Storage, vol. 7, no. 4, pp. 1–20, 2012,doi:10.1145/2078861.2078864.
[11] J. Pettitt, "Hash of plaintext as key?" (2016). [Online]. Available:http://cypherpunks.venona.com/date/1996/02/msg02013.html
[12] The Freenet Project, Freenet. (2016). [Online]. Available: https://freenetproject.org/
[13] M. Bellare, S. Keelveedhi, and T. Ristenpart, "Message-locked encryption and secure deduplication," in Proc. Cryptology—EUROCRYPT,2013, pp. 296–312, doi:10.1007/978-3-642-38348-9_18.
[14] D. Perttula, B. Warner, and Z. Wilcox-O'Hearn, "Attacks on convergent encryption." (2016). [Online]. Available: http://bit.ly/yQxyvl
[15] C. Y. Liu, X. J. Liu, and L. Wan, "Policy-based deduplication in secure cloud storage," in Proc. Trustworthy Comput. Serv., 2013,pp. 250–262, doi:10.1007/978-3-642-35795-4_32.
[16] P. Puzio, R. Molva, M. Onen, and S. Loureiro, "ClouDedup:Secure deduplication with encrypted data for cloud storage," in Proc. IEEE Int. Cof. Cloud Comput. Technol. Sci., 2013, pp. 363–370,doi:10.1109/CloudCom.2013.54.
[17] Z. Sun, J. Shen, and J. M. Yong, "DeDu: Building a deduplication storage system over cloud computing," in Proc. IEEE Int. Conf.Comput. Supported Cooperative Work Des., 2011, pp. 348–355,doi:10.1109/CSCWD.2011.5960097.
[18] Z. C. Wen, J. M. Luo, H. J. Chen, J. X. Meng, X. Li, and J. Li, "A verifiable data deduplication scheme in cloud computing," in Proc.Int. Conf. Intell. Netw. Collaborative Syst., 2014, pp. 85–90,doi:10.1109/INCoS.2014.111.
[19] J. Li, Y. K. Li, X. F. Chen, P. P. C. Lee, and W. J. Lou, "A hybrid cloud approach for secure authorized deduplication," IEEE Trans.Parallel Distrib. Syst., vol. 26, no. 5, pp. 1206–1216, May 2015,doi:10.1109/TPDS.2014.2318320.
[20] P. Meye, P. Raipin, F. Tronel, and E. Anceaume, "A secure twophase data deduplication scheme," in Proc. HPCC/CSS/ICESS,2014, pp. 802–809, doi:10.1109/HPCC.2014.134.
[21] J. Paulo and J. Pereira, "A survey and classification of storage deduplication systems," ACM Comput. Surveys, vol. 47, no. 1,pp. 1–30, 2014, doi:10.1109/HPCC.2014.134.
[22] Y.-K. Li, M. Xu, C.-H. Ng, and P. P. C. Lee, "Efficient hybrid inline and out-of-line deduplication for backup storage," ACM Trans.Storage, vol. 11, no. 1, pp. 2:1-2:21, 2014, doi:10.1145/2641572.
[23] M. Fu, et al., "Accelerating restore and garbage collection in deduplication-based backup systems via exploiting historical information,"in Proc. USENIX Annu. Tech. Conf., 2014, pp. 181–192.
[24] M. Kaczmarczyk, M. Barczynski, W. Kilian, and C. Dubnicki,"Reducing impact of data fragmentation caused by in-line deduplication," in Proc. 5th Annu. Int. Syst. Storage Conf., 2012,pp. 15:1–15:12, doi:10.1145/2367589.2367600.
[25] M. Lillibridge, K. Eshghi, and D. Bhagwat, "Improving restore speed for backup systems that use inline chunk-based deduplication," in Proc.USENIX Conf. File Storage Technol., 2013, pp. 183–198.
[26] L. J. Gao, "Game theoretic analysis on acceptance of a cloud data access control scheme based on reputation," M.S. thesis, Xidian University, State Key Lab of ISN, School of Telecommunications Engineering, Xi'an, China, 2015.
[27] Z. Yan, X. Y. Li, M. J. Wang, and A. V. Vasilakos, "Flexible data access control based on trust and reputation in cloud computing,"IEEE Trans. Cloud Comput., vol. PP, no. 99, Aug. 2015, doi:10.1109/TCC.2015.2469662, Art. no. 1.

