Data De-Duplication on Encrypted Data Lake in Cloud Environment

A.G.Ashmita
Sri Venkateswara College of Engineering,
Pennalur-602115, TN, India

R.Anitha
Sri Venkateswara College of Engineering,
Pennalur-602115, TN, India

Abstract— **In cloud computing one of the services is cloud storage where data is remotely maintained, managed and backed up. Due to increase in data exponentially day to day, issues related to the storage space, data confidentiality and volume of search space complexities increases. To resolve these issues, the proposed model aims to, address the demand of storing the data redundantly by means of efficient de-duplication technique and also to protect the confidentiality of sensitive data while supporting de-duplication. Monitoring the activities on top of the storage environment in datalake to provide security to the storage nodes. As the storage nodes are geographically distributed the prime focus is on Optimal data storage and retrieval storage management and data security. Improved efficiency in data storing and retrieval, optimal storage allocation and de-duplication are involved along with the increase in the data security. As the data is huge in volume, the efficient way of storing and retrieving the data in an optimal way is to be addressed. The optimal way of storage and retrieval reduces the latency and thereby increasing the throughput. As the data is stored in a remote location security is also addressed efficiently.**

Keywords— Data Lake, De-Duplication, Encryption

I. INTRODUCTION

Cloud computing a buzz word, according to Ian Foster et al. [1] is defined as “A large-scale distributed computing paradigm that is driven by economies of scale, in which a pool of abstracted, virtualized, dynamically-scalable, managed computing power, storage, platforms, and services are delivered on demand to external customers over the Internet”. Cloud can be deployed as Public Cloud, Private Cloud, Hybrid Cloud, and Community Cloud. The basic services offered by the cloud are categorized as Software as a Service, Platform as a Service and Infrastructure as a Service.

Cloud computing is a general term for anything that involves delivering hosted services over the Internet. According to NIST (National Institute of Standards and Technology) it defines “Cloud computing is a model for enabling convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction. This cloud model promotes availability and is composed of five essential characteristics, three service models, and four deployment models” (Peter Mell et al., 2011).

RELATED WORKS

In order to preserve the privacy of data holders in cloud, data are often stored in cloud in an encrypted form. However, encrypted data introduce new challenges for cloud data deduplication[1], which becomes crucial for big data storage and processing in cloud. Traditional de-duplication schemes cannot work on encrypted data. Existing solutions of encrypted data deduplication suffer from security weakness. They cannot flexibly support data access control and revocation. This paper, we propose a scheme to deduplicate encrypted data stored in cloud based on ownership challenge and proxy re-encryption. It integrates cloud data deduplication with access control. Performance evaluation is based on extensive analysis and computer simulations. The results show the superior efficiency and effectiveness of the scheme for potential practical deployment, especially for big data deduplication in cloud storage. Even though the proposed model achieved 50% access latency, It solves complex requirements of developers and cloud providers alone not the issues of user.

Deduplication has become a widely deployed technology in cloud data centers to improve IT resources efficiency.[2] Tradeoff between the conflicting goals of scalable deduplication throughput and high duplicate elimination ratio. This paper proposes AppDedupe, an application-aware scalable inline distributed deduplication framework in cloud environment. Challenges are meant by exploiting application awareness, data similarity and locality to optimize distributed deduplication with inter-node two-tiered data routing and intra-node application-aware deduplication. AppDedupe builds application-aware similarity indices with super-chunk handprints to speedup the intra-node deduplication process with high efficiency.

To improve the I/O performance of primary storage systems in the Cloud a performance-oriented I/O Deduplication (POD),[3] rather than a capacity-oriented I/O Deduplication. The design of POD aims to achieve the following three objectives.

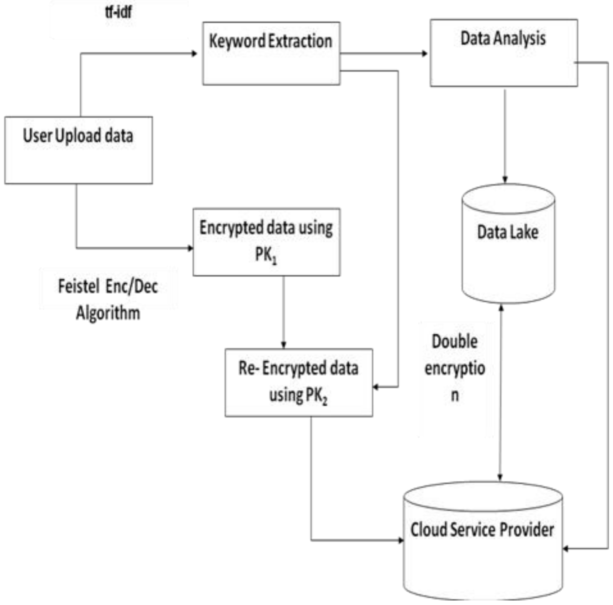
Reducing small write, Improving cache efficiency and Guaranteeing read performance. POD resides in the storage node and interacts with the File Systems via the standard read/write interface.

Data consistency in POD mandates that the referenced data be reliably stored on disks and the key data structures not be lost in case of a power failure. It takes a request based selective Deduplication approach (Select Dedupe) to deduplicating the I/O redundancy on the critical I/O path in such a way that it minimizes the data fragmentation problem.

On conceding two factors such as data reliability data redundancy this paper provides solution to the storage allocation scheme. Against failures, it distributes copies of data objects to a set of storage nodes. Redundancy scheme can be classified into two types: replication and erasure code.[4] Replication scheme is simple and intuitional, which replicates each data block into n copies and then distributes them to different network nodes. By contrast, erasure code encodes k data blocks into (n-k) coded blocks, resulting in n blocks in total. These n blocks are then distributed into different nodes. Storage allocation scheme paves the way for redundant data blocks are distributed and stored over a set of network nodes. The storage allocation scheme allows to optimize the phases of partition, assignment and searching process. The optimal storage allocation scheme can significantly reduce the search space.

PROPOSED SYSTEM

A scheme to de-duplicate encrypted data at the cloud service provider's location by applying the machine learning in an unsupervised approach. It is applicable in scenarios where data holders are not available for de-duplication control. The proposed system contains three types of entities: 1) 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; 2) data holder that uploads and saves its data at CSP in an encrypted form. In the proposed system, it is possible to have a number of eligible data holders that could save the same encrypted raw data in CSP. The data holder that produces or creates the file is regarded as data owner. It has higher priority than other normal data holders, which will be presented in 3) an authorized party (AP) that does not collude with CSP and is fully trusted by the data holders to verify data ownership and handle data de-duplication. As it is possible that CSP and its users (e.g., data holders) can collude such collusion could make the CSP lose reputation due to data leakage.



A negative impact of bad reputation is the CSP will lose its users and finally make it lose profits. On the other hand, the CSP users (e.g., data holders) could lose their convenience and benefits of storing data in CSP due to bad reputation of cloud storage services. Thus, the collision between CSP and its users is not profitable for both of them. Hence the concrete analysis based on the machine learning approach is provided in. Therefore, the project holds an assumption such as: CSP does not collude with its users, e.g., performing re-encryption for unauthorized users to allow them to access data. Additional assumptions include: data holders honestly provide the encrypted hash codes of data for ownership verification. The data owner has the highest priority. A data holder should provide a valid certificate in order to request a special treatment. Users, CSP and AP communicate through a secure channel (e.g., SSL) with each other. CSP can authenticate its users in the process of cloud data storage. We further assume that the user policy for data storage, sharing and de-duplication is provided to CSP during user registration.

CREATION OF DATA LAKE:

Data Lake is a massive, easily accessible, flexible and scalable large data repository or large storage. DataLake is a place to store practically unlimited amounts of data of any type of same frequency, schema and format that is relatively inexpensive and massively scalable. Hadoop implements a scalable and parallel processing framework that will process exceedingly large amounts of data in a smooth way, and makes it almost impossible to lose any kind of data, as it is replicated across the cluster. As organizations rush to take advantage of huge and diverse data sets, it's difficult to manage increase in the volume, velocity and variety of information today. The data lake should serve data efficiently to business users and applications ultimately helping the users to meeting SLAs. A Hadoop-based data lake should have a strong integrated toolset that supports self-service with data discovery steps: data access, preparation and analysis. There are many aspects enabling a cohesive machine learning process for creating a data lake. This brings us to the next step in trusted data lake for providing security. With the data lake, the mantra becomes "extract and load". The approach imports and manages data of varied size, provenance, and frequency in the data lake using the predicating modeling using the keyword. The data lake is created based on the keyword of the file uploaded by the user. When a file is uploaded by the user the document is preprocessed. After preprocessing the keywords are extracted using tf-idf algorithm and is analysed using multi label way of unsupervised approach.

DATA DE-DUPLICATION:

Whenever data is uploaded to the server it will be compared with the existing resources if the resource already exists, a additional pointer is used to point the respective data using the data structure called RBF (Replica Bloom Filter). Attempts to avoid replicas in the servers in order to reduce duplication data. (De-duplication) To provide guaranteed content delivery. Hash based De-duplication Replication Bloom Filter defines the protocol for placing the metadata file in the metadata server and its replication locations. The RBF maintains the information about where the replica is available and from which location the metadata file can be accessed.

DATA SECURITY

The DUDE scheme provides a secure approach to protect and de-duplicate the data stored in cloud by concealing the keyword in the plaintext from both CSP and the user. The security of the proposed scheme is ensured by the key exchange functionalities. New Key generation technique is followed by making use of the CBC (Cipher Block Chain) algorithm.

The proposed security model describes the below listed modules and the parameters are also described below. The parameters, Filename and User id (Filename, User id) remains unchanged in the proposed scheme.

User level Cipher Key Generation (KG). A user A's key pair is of the form U_1= CTC(KW_1, .. KW_n) and E_1 = (F, U_1).

Cloud level Cipher-Key Generation: Cloud level Cipher key is of the form CCK_{mxn}= EFN(A_1, .. A_n) from DS.

First-Level Encryption (E_1): To encrypt a file F using U_1 in such a way that it can only be decrypted by the holder of U_1, Encrypt(F, U_1) = E_1 where CTC(KW_i, MK_1) = U_1.

Second-Level Encryption or Re-Encryption(E_2): To encrypt a message F_1 using U_2 in such a way ENCRYPT(E_1, U_2) = E_2 where ENCRYPT(U_1, CCK_{mxn}) → U_2 .

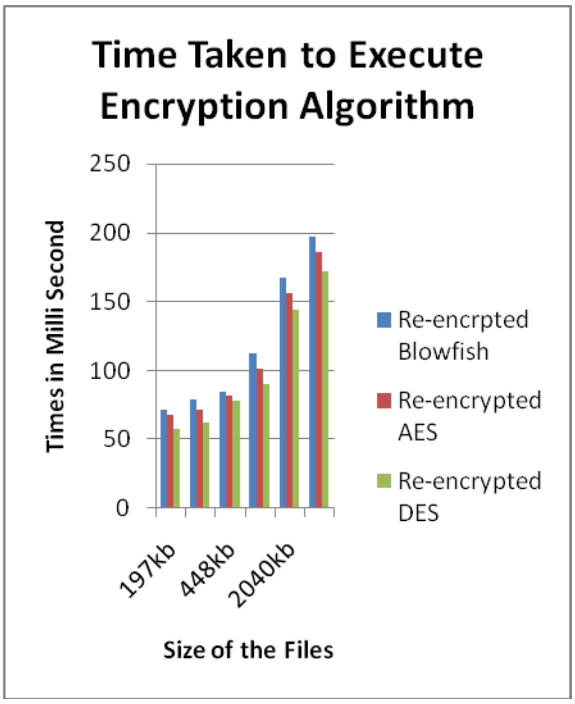
In encryption process, a key specifies the transformation of plain text to cipher text and vice-versa during decryption. The security model proposes two novel key generation mechanisms. 1. Key generation at user level using Cipher Tree Chaining network. 2. Key generation at MDS level using Enhanced modified feistel network(EFN) with a new Feistel function F, which exploits the matrix operations like transpose, shuffle, addition and multiplication along with the key matrix. The avalanche effect determines the strength of the cipher key and is defined as a change in the single input bit results in the change of output bits with a probability of 0.5.

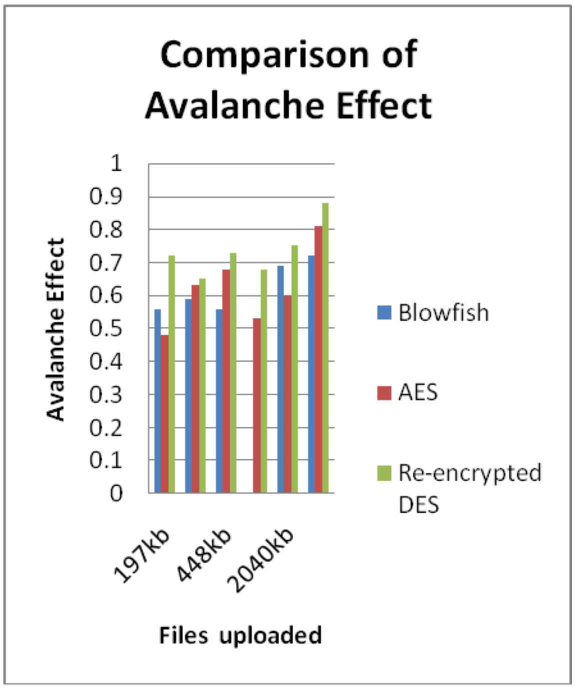
ENCRYPTED DATA UPLOAD

If data duplication check is negative, the data holder encrypts its data using a randomly selected symmetric key in order to ensure the security and privacy of data, and stores the encrypted data at CSP together with the token used for data duplication check. The data holder encrypts with and passes the encrypted key to CSP.

EXPERIMENTAL AND RESULT

The experiments have been carried out in a cloud setup using eucalyptus which contains cloud controller and walrus as storage controller. These tests were done on 5 node cluster. Each node has two 3.06 GHz Intel (R) Core TM Processors, i-7 2600, CPU @ 3.40GHZ, 4 GB of memory and four 512 GB hard disks, running Eucalyptus. The tests used 500 files of real data set.





Avalanche effect is that by changing only one bit in a matrix, leads to a large change in the existing key, hence it is hard to perform an analysis of cipher text, when trying to come up with an attack.

$$\text{Avalanche Effect} = \frac{\text{Number of values changed in the cipher Key } C_{mxn}}{\text{Total Number of values in the cipher key } C_{mxn}}$$

Higher the avalanche effect, higher the strength of the cipher key. The avalanche effect is calculated by the formula,

CONCLUSION

Managing encrypted data with de-duplication along with the reduced search space is important for achieving a successful cloud storage service, especially for big data storage. This project proposes a DUDE (De-duplication in Encrypted data lake) scheme to manage the encrypted big data in cloud data lake with de-duplication Our scheme can flexibility support data update and sharing with de-duplication even when the data holders are offline. Encrypted data can be securely accessed because only authorized data holders can obtain the symmetric keys used for data decryption.

REFERENCES

[1] Zheng Yan, Wenxiu Ding, Xixun Yu, Haiqi Zhu, and Robert H. Deng, "De-duplication on Encrypted Big Data in Cloud", IEEE Transactions on Big Data, Vol. 2, No. 2, pp. 138-151, 2017.

[2] Yinjin Fu, Nong Xiao, Hong Jiang, Fellow, IEEE, Guyu Hu, and Weiwei Chen, "Application-Aware Big Data Deduplication in Cloud Environment", IEEE Transactions on Cloud Computing, 2017.

[3] Bo Mao,Hong Jiang suzhan Wu &Lei Tian, "Leveraging Data Deduplication To improve the performance storage system In the cloud", IEEE transaction on computer Vol. 65, No.6, pp. 1775-1788, 2016.

[4] Zhen Huang et al,Jinbang chen Yisong Lin,Pengfei You & Yuxing Peng,"Minimizing data redundancy for high reliable cloud storage", Journal of computer Networks,Elsevier, Vol.81, pp. 164-177, 2016.

[5] Fraunhafer christoph quix,Riahan Hai,Ivan vahai,"Generic and Extensible Metadata Management system for Data Lakes",Data bases and Information System in Proc. International Conference on Advanced Information Systems Engineering, pp.129-136, 2016.

[6] Isure surarachchi and Beth Plate,"A case for Integrated Provenance in data lakes", International conference on e-science, pp.349-353, 2016.

[7] Surabhi D Hedge,Ravinarayana," Survey paper on Data Lake", International Journal of Science and Research, pp. 1718-1719, Vol. 5, No. 7, 2016.

[8] Jin Li Yan Kit Li, Liaofeng chan, Patrick P.L, Lee Wenjing Lou,"A Hybrid cloud Approach for secure authorized Depulication" IEEE Transaction on parallel and Distributed systems, Vol. 26 , No. 5 , pp.1- 11, 2014.

[9] C. Yang, J. Ren, and J. F. Ma, "Provable ownership of file in deduplication cloud storage," in Proc. IEEE Global Commun. Conf., 2013, pp. 695–700.

[10] T. Y. Wu, J. S. Pan, and C. F. Lin, "Improving accessing efficiency of cloud storage using de-duplication and feedback schemes,"IEEE Syst. J., vol. 8, no. 1, pp. 208–218, Mar. 2014.

[11] C. Fan, S. Y. Huang, and W. C. Hsu, "Hybrid data deduplication in cloud environment," in Proc. Int. Conf. Inf. Secur. Intell. Control, 2012, pp. 174–177.

[12] J. W. Yuan and S. C. Yu, "Secure and constant cost public cloud storage auditing with deduplication," in Proc. IEEE Int. Conf. Communic. Netw. Secur., 2013, pp. 145–153.

[13] N. Kaaniche and M. Laurent, "A secure client side deduplication scheme in cloud storage environments," in Proc. 6th Int. Conf.New Technol. Mobility Secur., 2014, pp. 1–7.

[14] Z. Yan, W. X. Ding, and H. Q. Zhu, "A scheme to manage encrypted data storage with deduplication in cloud," in Proc.ICA3PP2015, Zhangjiajie, China, Nov. 2015, pp. 547–561.

[15] Z. Yan, X. Y. Li, and R. Kantola, "Controlling cloud data access based on reputation," Mobile Netw. Appl., vol. 20, no. 6, 2015, pp. 828–839.

[16] T. T. Wu, W. C. Dou, C. H. Hu, and J. J. Chen, "Service mining for trusted service composition in cross-cloud environment," IEEE Systems Syst. J., vol. PP, no. 99, pp. 1–12, 2014.