A SURVEY ON MINIMIZING INFORMATION LEAKAGE IN MULTICLOUD STORAGE SERVICE

[1]Devika S. Gandhi, [2]Sachin A. Murab, [3]Parag D. Thakare, [4]Aniruddha A. Kolpyakwar

[1]Student of M.E. II year, [2] Head Of Department, [3]Assistant Professor, [4]Assistant Professor
[1,2,3] Department of Computer Engineering,
[1,2,3] Jagadambha College of Engineering & Technology, Yavatmal, India.

Abstract: Multiple clouds are used to store important data by using many advanced technology. Separated data over different cloud storage providers(CSPs) automatically provides a certain degree of data leakage control, so that no single point of attack can leak all the data.However, unplanned distribution of data chunks can lead to high data disclosure even while using multiple clouds. an important data leakage problem caused by unplanned data distribution in multicloud storage services. Then, we present **StoreSim**, an data leakage aware storage system in multicloud. StoreSim aims to store similar data on the same cloud, thus minimizing the user's data leakage across multiple clouds. We design an approximate algorithm to efficiently generate similarity preserving signatures for data chunks based on MinHash and Bloom filter, and also design a function to compute the data leakage based on these signatures. Next, an effective storage plan generation algorithm based on clustering for distributing data chunks with minimum data leakage across the multiple clouds. Finally, evaluate our scheme using two real datasets from Wikipedia and GitHub. We show that our scheme can reduce the data leakage by up to 60% compared to unplanned placement. Furthermore, our analysis on system attackability demonstrates that our scheme makes attacks on data more complex. Data security plays an important role in cloud in which lot of data is get shuffled and became unsecured while sharing to TPA or other users.
.

***IndexTerms* - Multicloud Storage, Data leakage, System Attackability.**

I. INTRODUCTION

Data integrity plays an important role in data verification and efficiency for which various steps of algorithm will be perform with the increasingly uptake of devices such as laptops, cell phones and tablets. Users require an universal and large network storage to handle their ever-growing digital lives. To meet these demands, many cloud-based storage and file sharing services such as Google Drive, dropbox and Amazon S3, have gained popularity because of easy to use interface and low storage cost. However, these centralized cloud storage services are criticized for grabbing the control of users data. which allows storage providers to run analytics for marketing and advertising. Also, the data in users' data can be leaked e.g., by means of malicious insiders, backdoors, bribe. One possible solution to reduce the risk of data leakage is to employ multicloud storage systems in which no single point of attack can leak all the data. A malicious entity, such as the one revealed in recent attacks on privacy, would be required to pressure all the different cloud storage providers on which a user might place her data, in order to get a complete picture of her data. Put simply, as the saying goes, do not put all the eggs in one basket.

II. LITERATURE REVIEW

We have studied the current requirements to manage leakage in multicloud. The below papers helped us to understand the previous work on this topic:

i. DEPSKY, a system that improves the availability, integrity and confidentiality of information stored in the cloud through the encryption, encoding and replication of the data on diverse clouds that form a cloud-of-clouds. the monetary costs of using DEPSKY on this scenario is twice the cost of using a single cloud.
DEPSKY addresses four important limitations of cloud computing for data storage in the following way:
Loss of availability, Loss and corruption of data, Loss of privacy, Vendor lock-in.
Depsky minimized the cost of data transfer from one cloud to another by storing only fraction of the total amount of data in each cloud.[1]

ii. Scalia, a cloud storage brokerage solution that continuously adapts the placement of data based on its access pattern and subject to optimization objectives, such as storage costs. Scalia efficiently considers repositioning of only selected objects that may significantly lower the storage cost.
Scalia, a system that continuously adapts the placement of data among several storage providers subject to optimization objectives, such cost minimization.[2]

iii. The Cooperative File System (CFS) is a new peer-to-peer read only storage system that provides provable guarantees for the efficiency, robustness, and load-balance of file storage and retrieval CFS delivers data to clients as fast as FTP CFS is scalable A CFS file system exists as a set of blocks distributed over the available CFS servers. CFS client software interprets the stored blocks as file system data and meta-data and presents an ordinary read-only file-system interface to applications.[3]

iv. Samsara technique force on peer-to-peer storage systems, it doesn't required trusted third parties, symmetric storage relationships, cash payment, or certified identities. Samsara designed their own storage system with a peer-to-peer network comprised with untrusted nodes. Our work targets to use storage cloud without using decentralized Peer to Peer protocol and optimizes data placement in a centralized way.[4]

v. It provided a survey for four different multicloud architectures with various security and privacy-improved designs. an security merits by making use of multiple different clouds simultaneously. Various distinct architectures are introduced and discussed according to their security and privacy capabilities and prospects.[5]

vi. Due to the increase of attractive services available on the cloud, people are placing an large amount of their data online on different cloud platforms. However, because of unauthorised attacked, privacy is a important issue. Ordinary users cannot be recognized which of their data is sensitive, or to take sufficient measures to protect such data.
A novel conceptual frame work in which privacy risk is automatically calculated using the sharing context of data items. it is a semantic frame work based on crowed sourcing to determine the sensitivity of item and diverse attitudes of users towards privacy.[6]

vii. Analyze the fundamental characteristics of privacy and utility, and show that it is not suitable to directly compare privacy with utility. A privacy loss measure based on the JS-divergence distance which is a method of measuring the similarity between two probability distributions.[7]

viii. Near-duplicate web documents are plentiful. Two documents differ from each other in a very small portion, for example. Such differences are irrelevant for web search. So the quality of a web crowd increases if it can assess new crawled web page it check weather is a near-duplicate of a previous web page or not.
Elimination of near-duplicates saves network bandwidth, reduces storage costs and improves the quality of search. It also reduces the load on the remote host that is serving such web pages.[8]

III. PROPOSED WORK

In the proposed we are planned an activity in which the system can be integrated with the various data compression algorithm for efficient data storage in which the data storing capacity of connected cloud can be improve to much more extend . In proposed system the data integrity will updated with encryption and decryption algorithms. So that data leakage problem is minimized from 80% to 60%.

IV. PROPOSED MODEL

In this section, we describe the architecture of storesim.

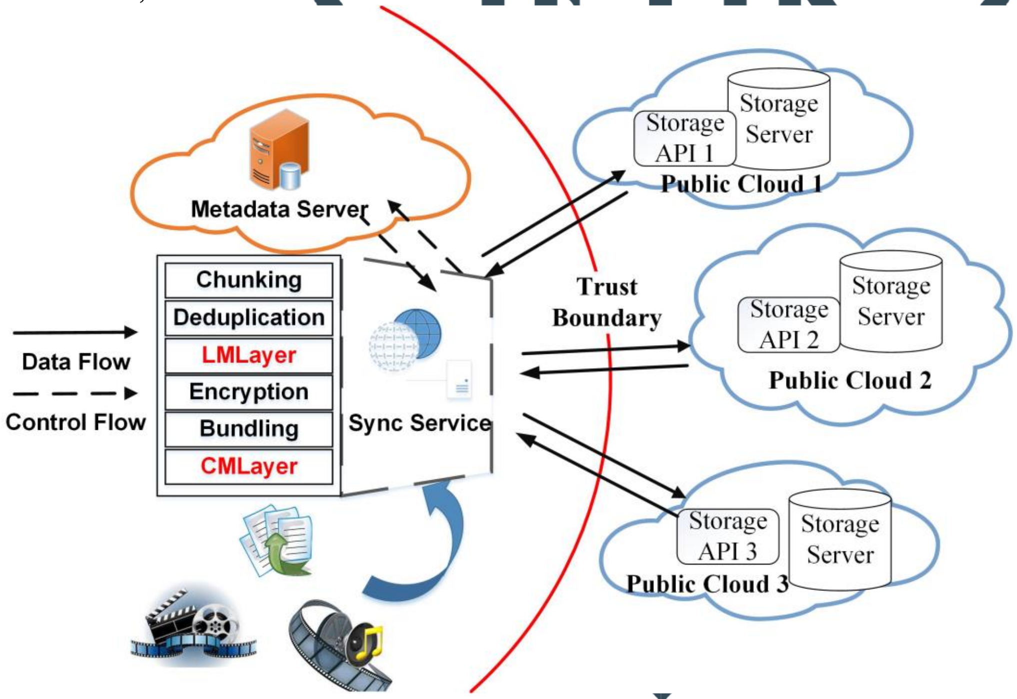

Fig.1 Architecture

The architecture of StoreSim is shown in Figure. It shows that there is a trust boundary between the metadata and storage servers. Here clients and metadata servers, which are situated inside the trust boundary, are trustable by users while servers outside the boundary are untrustworthy. For example, the metadata can be stored in private database servers while storage servers can be located in public CSPs such as Google drive, Amazon S3 And Dropbox. Storage servers are accessed through standard Application Programming Interfaces. The fig shows. To minimize the data leakage on multicloude storage, we design two components. The first component is the Leakage Measure layer (LMLayer) is used to evaluate the data leakage and then generate the storage plan which store data chunks to different clouds. The other component is the Cloud Manager Layer (CMLayer) is provides cloud interoperability in a syntactic way.

V. Work flow of System:

The following figure demonstrates the flow of execution of the system. When user give a some kind of input the it check for inputted data i.e. how much size it has, then it check the size required for storage is available or not on cloud then it apply divisional chunk for these storage. Apply data compression algorithm and then stored this compress chunk as a backup storage so that whenever data leakage occurs it help in to recovered data.

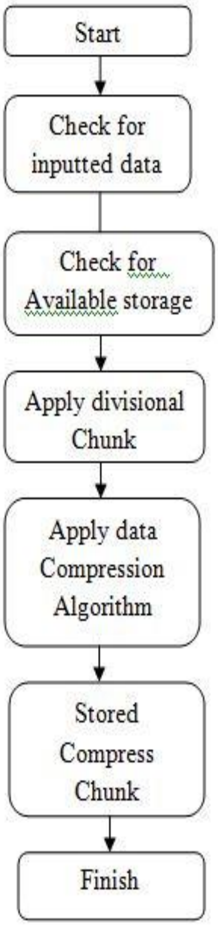

Fig 2. Flow Graph

V. CONCLUSION

We are planned an activity in which the data integrity will updated with encryption and decryption algorithm. The storage accuracy will increased and data leakage can be controlled. So the data storing capacity of connected multicloud storage services improved.

Storesim can achieve near optimal performance and reduce data leakage up to 60%

REFERENCE

[1]. A. BESSANI, M. CORREIA, B. QUARESMA, F. ANDR´E, AND P. SOUSA, "DEPSKY: DEPENDABLE AND SECURE STORAGE IN A CLOUD-OF-CLOUDS," ACM TRANSACTIONS ON STORAGE (TOS), VOL. 9, NO. 4, P. 12, 2013.

[2]. T. G. Papaioannou, N. Bonvin, and K. Aberer, "Scalia: an adaptive scheme for efficient multi-cloud storage," in Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis. IEEE Computer Society Press, 2012, p. 20.

[3]. F. Dabek, M. F. Kaashoek, D. Karger, R. Morris, and I. Stoica, "Wide-area cooperative storage with cfs," in ACM SIGOPS Operating Systems Review, vol. 35, no. 5. ACM, 2001, pp. 202–215.

[4]. L. P. Cox and B. D. Noble, "Samsara: Honor among thieves in peer-to-peer storage," ACM SIGOPS Operating Systems Review, vol. 37, no. 5, pp. 120–132, 2003.

[5]. J.-M. Bohli, N. Gruschka, M. Jensen, L. L. Iacono, and N. Marnau, "Security and privacy-enhancing multicloud architectures," Dependable and Secure Computing, IEEE Transactions on, vol. 10, no. 4, pp. 212–224, 2013

[6]. H. Harkous, R. Rahman, and K. Aberer, "C3p: Context-aware crowdsourced cloud privacy," in 14th Privacy Enhancing Technologies Symposium (PETS 2014), 2014.

[7]. T. Li and N. Li, "On the tradeoff between privacy and utility in data publishing," in Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2009, pp. 517–526

[8]. G. S. Manku, A. Jain, and A. Das Sarma, "Detecting near-duplicates for web crawling," in Proceedings of the 16th international conference on World Wide Web. ACM, 2007, pp. 141–150.