DIFFERENTIAL AUTHORIZATION DUPLICATE CHECK IN HYBRID CLOUD ARCHITECTURE

Deepak V[1], Bhanu Pratap Y S[2], Karthik D[3], Soumya Samantaray[4], Basavaraj Jakkali[5]

[1,2,3,4] Pursuing B.E in Computer Science & Engineering, BMSCE, Bengaluru, India,

[5]Associate Professor Dept. of CSE, BMSCE, Bengaluru, India.

Abstract: Data duplication or intelligent compression is a method for eliminating duplicate copies of redundant data, and has been widely used in cloud services to decrease the amount of storage space and upload bandwidth. Sometimes we have to store duplicate data in cloud storage for confirmed data retrieval during critical situation, in other words to implement fault tolerance mechanism. This proposed system addresses Authenticated Deduplication with Security measures. In this system all users are having their own privilege and the user who is uploading the file for the first time can fix the Access Privilege and Deduplication Privilege. Based on these setting other users can access the file as well as check if the same file is already uploaded. This system uses Hybrid cloud approach to provide better security. All the files uploaded into the cloud have to undergo two processes. First is the Encryption process for data Security and next is the Tag Generation Process to identify the file presence in the cloud. In this system all the keys with privilege are stored in Private Cloud server and encrypted data is stored in Public Cloud.

***Keywords* —Deduplication, hybrid cloud, differential authorization**

I. INTRODUCTION

Cloud computing provides an infrastructure for data management and storage which is scalable, low cost and independent of location. Cloud service providers offer both highly amounts of storage and huge parallel computing at low costs. The emergence of cloud storage encourages enterprises to outsource data storage to third-party cloud storage service providers. This has led to rapid increase in volume of data being stored at remote servers. According to the International Data Corporation report, the volume of data in the world is expected to reach 40 trillion gigabytes in 2020. To tackle this problem, techniques that reduce disk space and network bandwidth are required. One such promising technique is the deduplication, where only one copy of the file is stored on the server, irrespective of any number of times the same file is sent by different clients. All clients that save their file on the server are given the link to that file. In client-side deduplication, if the server already has a copy of the file, then there is no need to upload and save the file again, thus saving bandwidth and storage space are saved.

There are two scenarios for deduplication: at file level or block level. At the file level, the duplicate copies are eliminated. It checks the whole file. But in the block level deduplication, a file is divided into blocks and blocks across non-identical files are compared for duplicates. Though data deduplication is beneficial in various aspects, there are concerns regarding the security and privacy of user’s data.

II. LITERATURE SURVEY

In this section, we have described earlier work done related to the deduplication mechanisms. Deterministic encryption, in particular convergent encryption, is a good option to achieve both confidentiality and deduplication. To prevent unauthorized access, a secure proof of ownership (PoW) protocol is also needed to provide the proof that the user indeed owns the same file when a duplicate is found.

Convergent encryption is a feasible solution to provide data confidentiality during deduplication. This encryption technique encrypts/decrypts data using convergent key. The convergent key is derived by computing the cryptographic hash value of the content of the data copy itself. After the key is generated, data is encrypted and the cipher text is sent to the public cloud for storage. The users then keep their keys. This encryption being deterministic, identical data copies will produce the same convergent key thus resulting in the same cipher text. This feature allows us to perform the duplicate checks. The cipher texts can only be decrypted by those having the keys, i.e. data owners. However, data duplication technique is vulnerable to powerful attacks. Harnik et al demonstrate that this deduplication technique can be used in a side channel attack revealing information about the contents of files of other users. Harnik et al suggested that if an attacker was able to temporarily get access to a server, he could access its internal cache, thereby getting hold of the hash values for all the recently accessed files. With this information, the attacker would get download all those files, thereby violating the confidentiality of the users’ files.

To overcome such attacks, Halevi et al. proposed the notion of "proofs of ownership" (PoW) for deduplication systems, so that a client can convincingly prove to the cloud server that he/she holds the complete data file and not just some short summary string. The proof of ownership (the HPPS protocol) works by encoding the file F using an erasure code E and then building a Merkle-tree over the encoded file. The verifier computes the encoding X = E(F) and the Merkle-tree MT(X) and stores the root of the tree in a summary string v. During an ownership check, the verifier randomly chooses 'u' leaf indexes and asks the prover for the sibling-paths of all the leaves, and verifies only if all the sibling paths are valid with respect to MT(X).

Bellare et al. characterized a new cryptographic primitive, message-locked encryption (MLE) [5], and examined its uses in space-economical secure outsourced public storage. In MLE the key used to encrypt/decrypt is derived from the message itself.

In the paper "Secure deduplication with efficient and reliable convergent key management", an attempt has been made to tackle the problem of key management efficiently and reliably in secure deduplication. A new construction called Dekey is proposed which doesn't require the key management by the users instead is managed by securely distributing the user keys among multiple servers. Here the first user uploading the file is responsible for generating the key and distributing it which eliminates the need for other users to repeat the process. To access his files, any user has to first authenticate to a minimum number of key servers, thereby obtaining the key shares which can then be used to reconstruct the secret key. This construction results in huge decrease overhead storage required to store the keys thus ensuring reliable and fail-safe key management.

In the paper "A Hybrid Cloud Approach for Secure Authorized Deduplication", differential privileges of users are considered which was not considered in previous deduplication models. The authors detail various constructions which support authorized duplicate check in hybrid cloud architecture. The private keys which include the privileges are managed in the private server and not distributed to the individual users. In order to perform a duplicate check, a user first needs to obtain the file token from the private server. This server checks the identity and privileges of the user before issuing the token. The user can then use this token to perform the duplicate check with the public cloud to see if the file is already present. If not, the user can upload his file to the server.

III. PROPOSED SYSTEM

Previous deduplication systems though providing duplicate check has no means to provide differential access. This is required in numerous present day applications. There is only a single cloud in the existing system which makes deduplication process hard. The sharing of privilege private keys among users is an issue with the existing system. It seems to be conflicting if we want to provide both deduplication and differential authorization in the same system at the same time.

In the proposed system we provide a solution that solves the problem of both deduplication and differential privileges in cloud storage. A hybrid cloud architecture is proposed which is a mixture of public cloud and private cloud. The private cloud is used as a stand-in to check for duplicate with privileges. The data storage is outsourced to the public cloud. The private cloud also stores the privileges of the users as keys. The cloud storage service providers residing in the public cloud space store the data but the checks are done in the private cloud. Any user can perform the duplicate checks only for the files corresponding to his privileges.

Architecture of the Proposed System

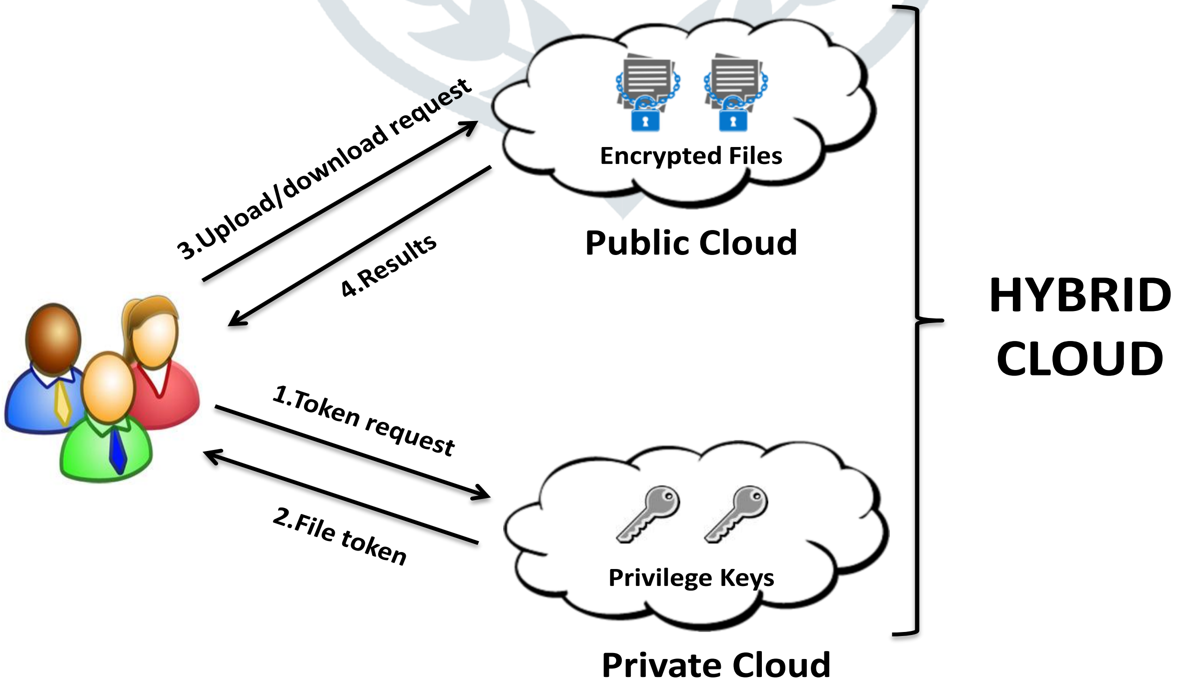

Figure 1: Architecture of the Proposed System

Sequence Diagram of Proposed System

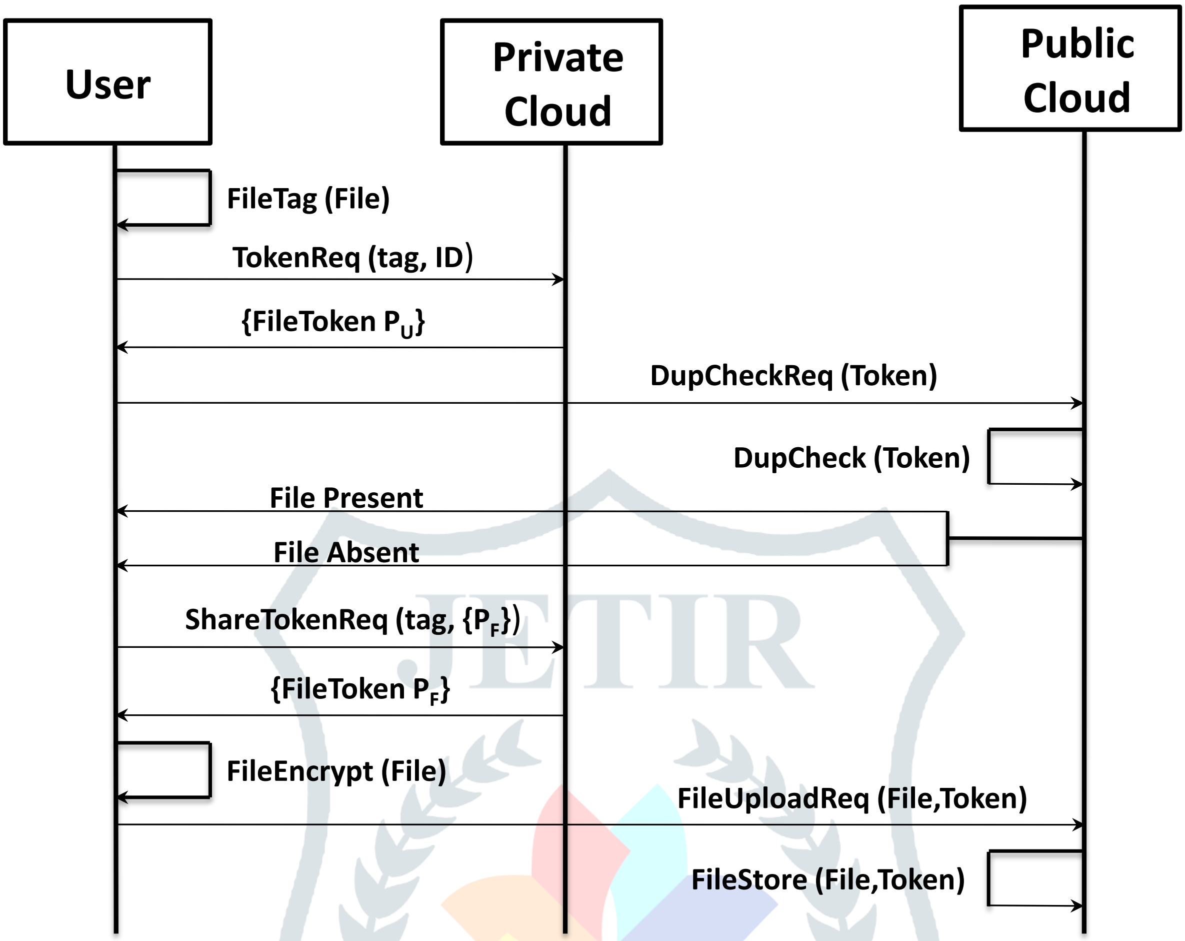

Figure 2: Sequence Diagram of the Proposed System

IV. Conclusion

The notion of authorized data deduplication is being proposed to protect the data security by including differential privileges of users in the duplicate check. As a proof of concept, we are going to simulate a prototype of an authorized duplicate check scheme on an available public cloud service provider. Security analysis is also going to be done to demonstrate that the schemes are secure. We are also going to conduct several experiments on the prototype and compare the reduction in storage space and bandwidth to the existing schemes.

References

[1] Amazon Case Studies: http://aws.amazon.com/solutions/case-studies/#backup
[2] J.Gantz and D.Reinsel, *The Digital Universe in 2020: Big Data, Bigger Digital Shadows, Biggest Growth in the Far East*, Dec 2012
[3] J.R. Douceur, A. Adya, W.J. Bolosky, D. Simon and M. Theimer, "Reclaiming Space from Duplicate Files in a Serverless Distributed File System", *Proc. ICDCS*, pp.617 -624
[4] S. Halevi, D. Harnik, B. Pinkas and A. Shulman-Peleg, "Proofs of Ownership in Remote Storage Systems", Proc. *ACM Conf. Computer and Communication Security*, pp.491 -500
[5] M. Bellare, S. Keelveedhi, and T. Ristenpart, "Message-locked encryption and secure deduplication" in *Proc. 32nd Annual International Conference on the Theory and Applications of Cryptographic Techniques*, 2013, pp. 296-312.
[6] J. Li, X. Chen, M. Li, J. Li, P. Lee, and W. Lou, "Secure deduplication with efficient and reliable convergent key management", *IEEE Transactions on Parallel and Distributed Systems*, June 2014.
[7] Jin Li, Yan Kit Li, Xiaofeng Chen, P.C. Lee, and Wenjing Lou, "A Hybrid Cloud Approach for Secure Authorized Deduplication", *IEEE Transactions on Parallel and Distributed Systems*, May 2015.

[8] D. Harnik, B. Pinkas and A. Shulman-Peleg, “Side Channels in Cloud Services: Deduplication in Cloud Storage”, *IEEE Security & Privacy*, 8(6)40-47, 2010.

Author Profile:

Basavaraj Jakkali is an Associate Professor in the Department of Computer Science & Engineering at BMS College of Engineering. He has completed his Post Graduation from Visvesvaraya Technological University (VTU), Belgaum, in 2002. His areas of interest include Computer Organization, Microprocessors, Theory of Computation, Operating Systems, System Software and Network Security.

Deepak V is Pursuing B.E in Computer Science & Engineering, BMSCE, Bengaluru.

Bhanu Pratap Y S is Pursuing B.E in Computer Science & Engineering, BMSCE, Bengaluru.

Karthik D is Pursuing B.E in Computer Science & Engineering, BMSCE, Bengaluru.

Soumya Samantaray is Pursuing B.E in Computer Science & Engineering, BMSCE, Bengaluru.