ISSN 2348-1196 (print)
International Journal of Computer Science and Information Technology Research ISSN 2348-120X (online)
Vol. 3, Issue 3, pp: (291-298), Month: July - September 2015, Available at: www.researchpublish.com

Secure and Authorized Deduplication in Hybrid Cloud

N. Venkateswara Rao[1], K.Vaddi Kasulu[2]

Final M Tech Student[1], Assoc professor[2], [1,2] Dept of Computer Science and Engineering, Eluru College of Engineering and Technology, Eluru, West Godavari dt, A.P. India

Abstract: **Many techniques are using for the elimination of duplicate copies of repeating data, from that techniques, one of the important data compression technique is data duplication. Many advantages with this data duplication, mainly it will reduce the amount of storage space and save the bandwidth when using in cloud storage. To protect confidentiality of the sensitive data while supporting de-duplication data is encrypted by the proposed convergent encryption technique before out sourcing. Problems authorized data duplication formally addressed by the first attempt of this paper for better protection of data security. This is different from the traditional duplication systems. The differential privileges of users are further considered in duplicate check besides the data itself. In hybrid cloud architecture authorized duplicate check supported by several new duplication constructions. Based on the definitions specified in the proposed security model, our scheme is secure in terms of definition and implementation.**

Keywords: **Deduplication, authorized duplicate check, confidentiality, hybrid cloud, convergent encryption, symmetric encryption.**

1. INTRODUCTION

Cloud computing was a great impact in new generation technology. Every user has large amount of data to share to store in an efficiently available secured place [11]. The concept of deduplication [9] is arrived here to efficiently utilize the bandwidth and disk usage on cloud computing. To avoid the duplication copies of the same data on cloud may cause lose of time, bandwidth utilization and storage space. Cloud computing is internet-based, a network of remote servers connected over the Internet to store, share, manipulate, retrieve and processing of data, instead of a local server or personal computer [11]. The benefit of cloud computing are enormous. It enables us to work from anywhere. The most important thing is that customer doesn't need to buy the resource for data storage. When it comes to Security, there is a possibility where a malicious user can penetrate the cloud by impersonating a legalize user, there by affecting the entire cloud thus infecting many customers who are sharing the infected cloud [11][12]. There is also big problem, where the duplicate copies may upload to the cloud, which will lead to waste of band width and disk usage. To improve this problem there should be a good degree of encryption provided, that only the customer should be able to access the data and not the legitimate User. Fig.1 shows to formally solve the problem of authorized data deduplication. Data deduplication is a data compression technique for removing duplicate copies of identical data, and it is used in cloud storage to save bandwidth and to reduce the amount storage space [1]. The technique is utilized to enhance the storage use and can likewise be applied to network data exchange to reduce the amount of bytes that must be sent. Keeping multiple data copies with the identical content, de-duplication removes redundant data by keeping only one copy and referring other identical data to that copy [1][11].

ISSN 2348-1196 (print)
International Journal of Computer Science and Information Technology Research ISSN 2348-120X (online)
Vol. 3, Issue 3, pp: (291-298), Month: July - September 2015, Available at: www.researchpublish.com

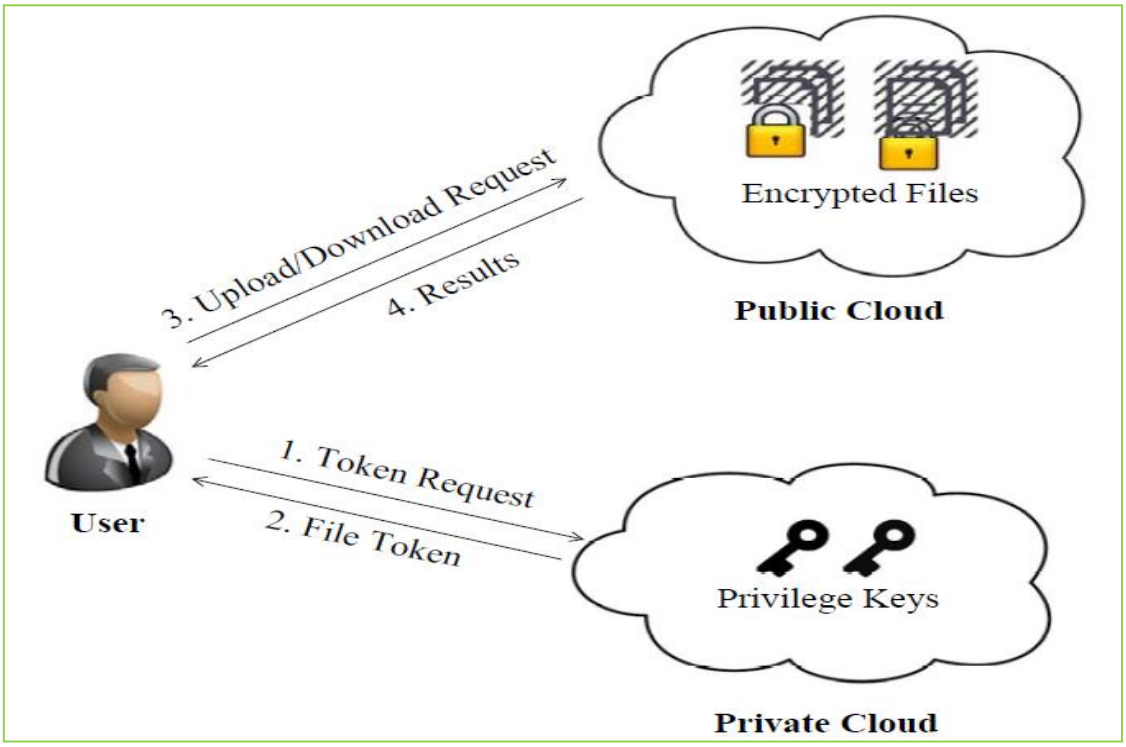

Fig. 1 Architecture of Authorized deduplication

De-duplication occurs either at block level or at file level. In file level de-duplication, it removed duplicate copies of the identical file [1]. Deduplication can also take place in the block level that eliminates duplicate blocks of data that is occurred in non identical files. Data deduplication having huge amount of advantages like providing security as well as privacy concerns arise as users sensitive or delicate data are at risk to both insider and outsider attacks. The traditional encryption requires many different customers for encrypting the data files with their own private keys. Thus, the same data copies of different customers will lead to different cipher texts, making de-duplication impossible. To secure the privacy of sensitive information while supporting deduplication, the convergent encryption [1][10] strategy has been proposed to encode the information before outsourcing. This project will work to dissolve the security issue and to evaluate the efficient utilization of cloud band width and disk usage.

2. PRELIMINARIES

In this section, we first define the notations used in this project, review some secure primitives used in our secure deduplication. The notations used in this project are listed in TABLE 1.

Figure Table 1

Acronym	Description
S-CSP	Storage-cloud service provider
PoW	Proof of Ownership
(pk_U, sk_U)	User's public and secret key pair
k_F	Convergent encryption key for file F
P_U	Privilege set of a user U
P_F	Specified privilege set of a file F
$\phi'_{F,p}$	Token of file F with privilege p

The main aim of proposed system is to efficiently solving the problem of deduplication with differential privileges in cloud computing and also provide a secured authorized deduplication using an Proof of Ownership(POW)[1]. We consider a hybrid cloud architecture consisting of a public cloud and a private cloud. Unlike existing data deduplication systems, the private cloud is involved as a proxy to allow data owner/users to securely perform duplicate check with differential privileges [1]. The data owners only outsource their data storage by utilizing public cloud while the data operation is managed in private cloud. A new deduplication system supporting differential duplicate check is proposed under this hybrid cloud architecture where the S-CSP resides in the public cloud. The user is only allowed to perform the duplicate check for files marked with the corresponding privileges. And the system uses an Proof of Ownership(POW)to support efficient authorization. OAuth acts as an intermediary on behalf of the end user, providing the service with an access token that authorizes specific account information to be shared [13]. Only the users which are authenticated in this way can undergone deduplication check.

2.1 Convergent encryption:

Convergent encryption [4] [10] is used to encrypt and decrypt file. User can derives the convergent key [1] from each original data copy, then using that key encrypt data file. Also user derives tag for data copy to check duplicate data. If tag are same then both files are same. Both convergent key and tag are independently derives. Convergent encryption, also known as content hash keying, is used to produces identical ciphertext from identical plaintext files. The simplest implementation of convergent encryption can be defined as: Alice derives the encryption key from her file F such that K=H(F), where H is a cryptographic hash function. Convergent encryption scheme can be defined with **four primitive functions**:

- KeyGenCE(M)->K is the key generation algorithm that maps a data copy M to a convergent key K.
- EncCE(K,M)->C is the symmetric encryption algorithm that takes both the convergent key K and the data copy M as inputs and then outputs a ciphertext C;
- DecCE (K, C) ->M is the decryption algorithm that takes both the ciphertext C and the convergent key K as inputs and then outputs the original data copy M ; and
- TagGen(M)->T(M) is the tag generation algorithm that maps the original data copy M and outputs a tag T (M).

2.2Proof of ownership:

The notion of proof of ownership [8] enables users to prove their ownership of data copies to the storage server. Specifically, PoW is implemented as an interactive algorithm run by a prover and a verifier. The verifier derives a short value $\phi(M)$ from a data copy M. To prove the ownership of the data copy M, the prover needs to send ϕ' to the verifier such that $\phi' = \phi(M)$. The formal security definition for PoW roughly follows the threat model in a content distribution network, where an attacker does not know the entire file, but has accomplices who have the file. The accomplices follow the "Bounded retrieval model", such that they can help the attacker obtain the file, subject to the constraint that they must send fewer bits than the initial min-entropy of the file to the attacker[8].

2.3 Symmetric encryption:

Symmetric encryption uses a common secret key κ to encrypt and decrypt information. A symmetric encryption scheme consists of three primitive functions:

- KeyGenSE(1^{λ})->k is the key generation algorithm that generates k using security parameter 1^{λ};
- EncSE(k,M) -> C is the symmetric encryption algorithm that takes the secret k and message M and then outputs the ciphertext C; and
- DecSE(k,C) -> M is the symmetric decryption algorithm that takes the secret k and cipher text C and then outputs the original message M.

2.4 Identification Protocol:

An identification protocol [1][8] can be described with two phases: Proof and Verify. In the stage of Proof, a prover/user U can demonstrate his identity to a verifier by performing some identification proof related to his identity. The verifier performs the verification with input of public information pk_U related to sk_U. At the conclusion of the protocol, the verifier

outputs either accept or reject to denote whether the proof is passed or not. There are many efficient identification protocols in literature, including certificate-based, identity-based identification etc[2][6].

3. DESIGN GOALS

In this project, we address the problem of privacy preserving deduplication in cloud computing and propose a new deduplication system supporting for:

3.1 Differential Authorization:

Each authorized user is able to access its individual token of his file to perform duplicate check based on authority. Under this assumption, any user cannot generate a token for duplicate check out of his access or without the aid from the private cloud server.

3.2 Authorized Duplicate Check:

Authorized user is able to access his/her own token from private cloud, while the public cloud performs duplicate check directly and tells the user if there is any duplicate. The security requirements considered in this project lie in two folds, including the security of file token and security of data files [1]. For the security of file token, two aspects are defined as unforgetability and indistinguish ability of file token.

The details are given below.

3.3 Unforgeability of file token/ duplicate-check token:

User make registration in private cloud for generating file token. Using respective file token he/she upload or download files on public cloud. The users are not allowed to collude with the public cloud server to break the unforgetability of file tokens. In our system, the S-CSP is honest but curious and will honestly perform the duplicate check upon receiving the duplicate request from users. The duplicate check token of users should be issued from the private cloud server in our scheme

3.4 In distinguishability of file token/duplicate-check token:

It requires that any user without querying the private cloud server for some file token, he cannot get any useful information from the token, which includes the file information and key information.

3.5 Data Confidentiality:

Unauthorized users without appropriate token, including the S-CSP and the private cloud server, should be prevented from access to the underlying plaintext stored at S-CSP. In another word, the goal of the adversary is to retrieve and recover the files that do not belong to them. In our system, compared to the previous definition of data confidentiality based on convergent encryption [8], a higher level confidentiality is defined and achieved.

4. EXISTING SYSTEM

Data deduplication is one of important data compression techniques for eliminating duplicate copies of repeating data, and has been widely used in cloud storage to reduce the amount of storage space and save bandwidth. To protect the confidentiality of sensitive data while supporting deduplication, Cloud computing provides seemingly unlimited "virtualized" resources to users as services across the whole Internet, while hiding platform and implementation details. Today's cloud service providers offer both highly available storage and massively parallel computing resources at relatively low costs. As cloud computing becomes prevalent, an increasing amount of data is being stored in the cloud and shared by users with specified privileges, which define the access rights of the stored data.

Problems on existing system:

1. One critical challenge of cloud storage services is the management of the ever-increasing volume of data.
2. Traditional encryption, while providing data confidentiality, is incompatible with data deduplication.
3. Identical data copies of different users will lead to different ciphertexts, making deduplication impossible.

ISSN 2348-1196 (print)
International Journal of Computer Science and Information Technology Research ISSN 2348-120X (online)
Vol. 3, Issue 3, pp: (291-298), Month: July - September 2015, Available at: www.researchpublish.com

5. PROPOSED SYSTEM

In our system we implement a project that includes the public cloud and the private cloud and also the hybrid cloud which is a combination of the both public cloud and private cloud. In general by if we used the public cloud we can't provide the security to our private data and hence our private data will be loss. So that we have to provide the security to our data for that we make a use of private cloud also. When we use a private clouds the greater security can be provided. In this system we also provides the data deduplication which is used to avoid the duplicate copies of data. User can upload and download the files from public cloud but private cloud provides the security for that data. That means only the authorized person can upload and download the files from the public cloud. For that user generates the key and stored that key onto the private cloud at the time of downloading user request to the private cloud for key and then access that Particular file.

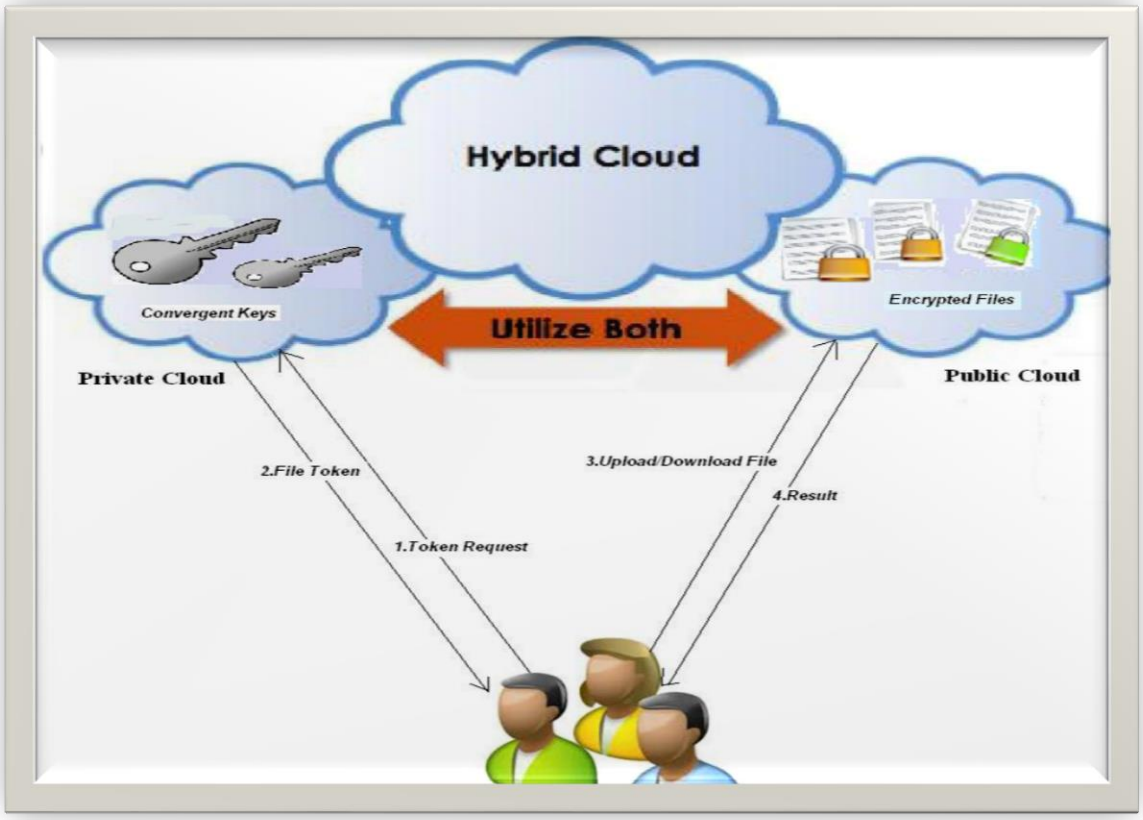

Fig. 2: Data sharing in cloud

The convergent encryption technique has been proposed to encrypt the data before outsourcing. To better protect data security, this project makes the first attempt to formally address the problem of authorized data deduplication. Different from traditional deduplication systems, the differential privileges of users are further considered in duplicate check besides the data itself as shown in figure 2. We also present several new deduplication constructions supporting authorized duplicate check in a hybrid cloud architecture. Security analysis demonstrates that our scheme is secure in terms of the definitions specified in the proposed security model. As a proof of concept, we implement a prototype of our proposed authorized duplicate check scheme and conduct testbed experiments using our prototype. We show that our proposed authorized duplicate check scheme incurs minimal overhead compared to normal operations.

ADVANTAGES OF PROPOSED SYSTEM:

- The user is only allowed to perform the duplicate check for files marked with the corresponding privileges.
- We present an advanced scheme to support stronger security by encrypting the file with differential privilege keys.

- Reduce the storage size of the tags for integrity check. To enhance the security of deduplication and protect the data confidentiality

6. IMPLEMENTATION

Implementation is the stage of the project when the theoretical design is turned out into a working system. Thus it can be considered to be the most critical stage in achieving a successful new system and in giving the user, confidence that the new system will work and be effective.

The implementation stage involves careful planning, investigation of the existing system and it's constraints on implementation, designing of methods to achieve changeover and evaluation of changeover methods.

Main Modules:

6.1 User Module:

In this module, Users are having authentication and security to access the detail which is presented in the ontology system. Before accessing or searching the details user should have the account in that otherwise they should register first.

6.2 Secure DeDuplication System:

To support authorized deduplication, the tag of a file F will be determined by the file F and the privilege. To show the difference with traditional notation of tag, we call it file token instead. To support authorized access, a secret key k_p will be bounded with a privilege p to generate a file token. Let $\phi'_{F,p}$= TagGen(F, k_p) denote the token of F that is only allowed to access by user with privilege p. In another word, the token $\phi'_{F,p}$ could only be computed by the users with privilege p. As a result, if a file has been uploaded by a user with a duplicate token $\phi'_{F,p,}$ then a duplicate check sent from another user will be successful if and only if he also has the file F and privilege p. Such a token generation function could be easily implemented as H(F,k_p), where H(.) denotes a cryptographic hash function.

6.3 Security Of Duplicate Check Token :

We consider several types of privacy we need protect, that is, i) unforge ability of duplicate-check token: There are two types of adversaries, that is, external adversary and internal adversary. As shown below, the external adversary can be viewed as an internal adversary without any privilege. If a user has privilege p, it requires that the adversary cannot forge and output a valid duplicate token with any other privilege p′ on any file F, where p does not match p′. Furthermore, it also requires that if the adversary does not make a request of token with its own privilege from private cloud server, it cannot forge and output a valid duplicate token with p on any F that has been queried.

6.4 Send Key:

Once the key request was received, the sender can send the key or he can decline it. With this key and request id which was generated at the time of sending key request the receiver can decrypt the message.

6.5 File Uploading:

When user want to upload the file to the public cloud then user first encrypt the file which is to be upload by make a use of the symmetric key, and send it to the Public cloud. At the same time user generates the key for that file and sends it to the private cloud. in this way user can upload the file in to the public cloud.

6.6 File Downloading:

When user wants to download the file that he/she has upload on the public cloud. He/she make a request to the public cloud. then public cloud provide a list of files that many users are upload on it. Among that user select one of the file form the list of files and enter the download option. At that time private cloud sends a message that enter the key for the file generated by the user. then user enters the key for the file that he/she is generated. then private cloud checks the key for that file and if the key is correct that means the user is valid. only then and then the user can download the file from the public cloud otherwise user can't download the file. When user download the file from the public cloud it is in the encrypted format then user decrypt that file by using the same symmetric key.

ISSN 2348-1196 (print)
International Journal of Computer Science and Information Technology Research ISSN 2348-120X (online)
Vol. 3, Issue 3, pp: (291-298), Month: July - September 2015, Available at: www.researchpublish.com

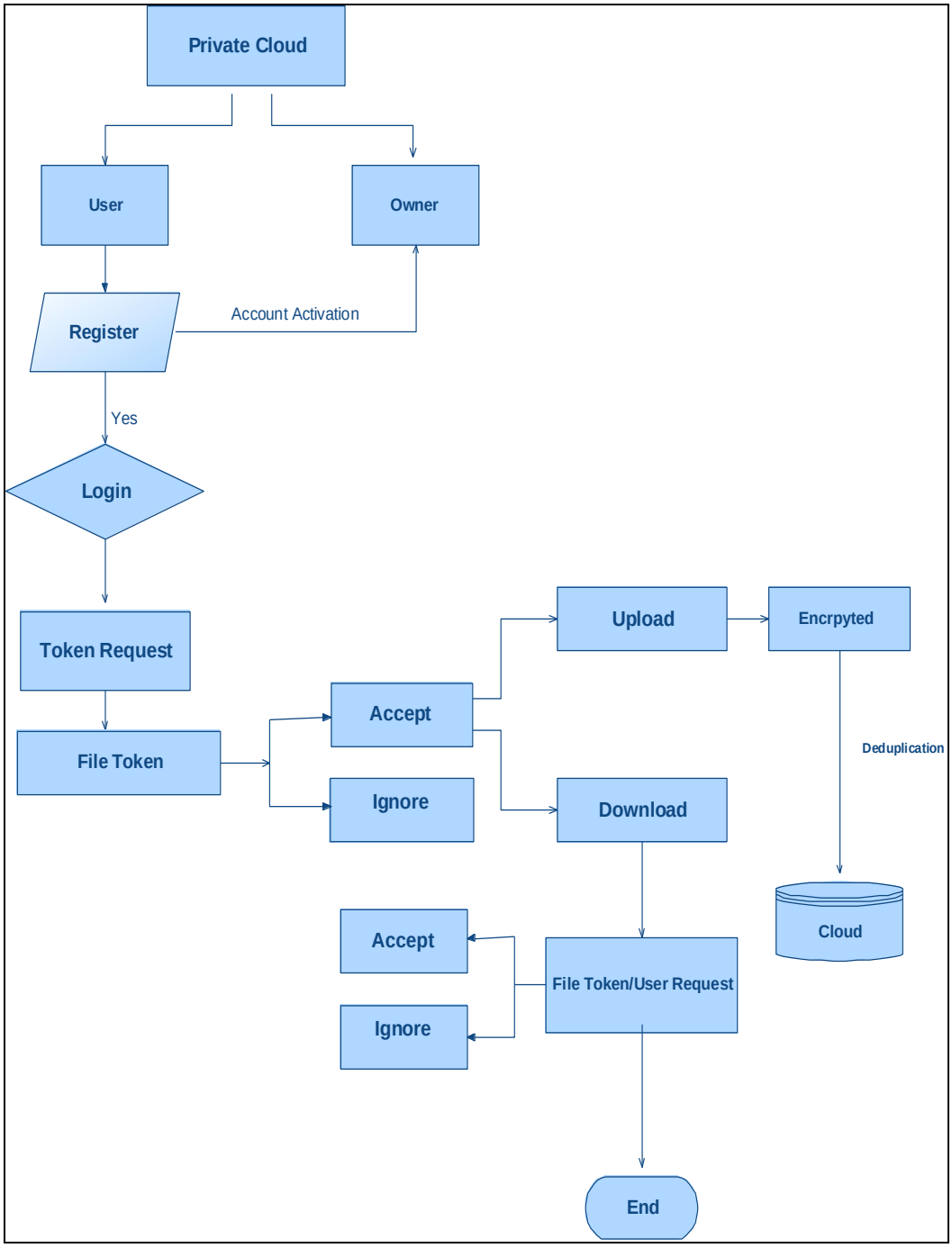

Fig. 3: Data flow diagram

7. CONCLUSION

In this project, the notion of authorized data deduplication was proposed to protect the data security by including differential privileges of users in the duplicate check. We also presented several new deduplication constructions supporting authorized duplicate check in hybrid cloud architecture, in which the duplicate-check tokens of files are generated by the private cloud server with private keys. Security analysis demonstrates that our schemes are secure in terms of insider and outsider attacks specified in the proposed security model. As a proof of concept, we implemented a prototype of our proposed authorized duplicate check scheme and conduct testbed experiments on our prototype. We showed that our authorized duplicate check scheme incurs minimal overhead compared to convergent encryption and network transfer.

REFERENCES

[1] Hybrid Cloud Approach for Secure Authorized Deduplication. Jin Li, Yan Kit Li, Xiaofeng Chen, Patrick P. C. Lee, Wenjing Lou –DOI 10.1109/TPDS.2014. 2318320, IEEE Transactions on Parallel and Distributed Systems.

[2] M. Bellare and A. Palacio. Gq and schnorr identification schemes: Proofs of security against impersonation under active and concurrent attacks. In CRYPTO, pages 162-177, 2002.

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

[4] M. Bellare, S. Keelveedhi, and T. Ristenpart. Dupless: Serveraided encryption for deduplicated storage. In USENIX Security Symposium, 2013.

[5] M. Bellare, S. Keelveedhi, and T. Ristenpart. Message-locked encryption and secure deduplication. In EUROCRYPT, pages 296– 312, 2013.

[6] D. Ferraiolo and R. Kuhn. Role-based access controls. In 15th NIST-NCSC National Computer Security Conf., 1992.

[7] GNULibmicrohttpd. http://www.gnu.org/ software/libmicrohttpd/.

[8] S. Halevi, D. Harnik, B. Pinkas, and A. Shulman-Peleg. Proofs of ownership in remote storage systems. In Y. Chen, G. Danezis, and V. Shmatikov, editors, ACM Conferenceon Computer & Communications Security, pages 491–500. ACM, 2011.

[9] S. Quinlan and S. Dorward. Venti: a new approach to archival storage. In Proc. USENIX FAST, Jan 2002.

[10] J.R.Douceur, A. Adya, W. J. Bolo sky, D. Simon & M. Theimer. Reclaiming space from duplicate files in a server less distributed file system. In ICDCS, pages 617-624, 2002.

[11] http://ijarcet.org/wp-content/uploads/ IJARCET-VOL-3-ISSUE-12-4191-4194

[12] https://dl.packetstormsecurity.net/papers/general/ProblemsFacedbyCloudComputing.pdf.

[13] http://worldconferences.org/uplo -ads /24315IISRC40.pdf.

AUTHORS BIBLIOGRAPHY

Mr. Namburi. Venkateswara Rao is Pursuing M.Tech in Computer Science & Engineering at Eluru College of engineering and technology, Eluru, W.G. Dt, A.P, India. He has received his B. Tech IT from Sri Vasavi Engineering College Tadepalligudem affiliated to JNTUK, AP. Research interests include cloud computing, Computer Networks, Data Mining, Network security.

Mr. Kasani Vaddi Kasulu working as assistant professor at Eluru College of Engineering and Technology, Eluru, W.G.Dt, A.P, India.He has received his M.Tech Degree(CST) from Andra university,Vizag,AP. Research interests include cloud computing, Computer Networks, Data Mining, Network security .