DEDUPLICATION IN CLOUD COMPUTING

[1]Shristi Priya, [2]Rashmi Singh, [3]Baby D. Dayana
[1]Student, [2]Student, [3]Assitant Proffesor
[1]Computer Science Department
[1]SRM IST, CHENNAI, INDIA

Abstract: The constant and endless data storing on cloud has been persuaded speedily since past few years and this leaded to deal with storage problems on cloud. To overcome this problem, a popular deduplication technique is being used in the cloud which lessens to save a same copy of data multiple times which is referred to as data redundancy. But this technique, apart from reducing burden on cloud, has many flaws for data confidentiality and cloud service users privacy. This project addresses the problems of authorized data deduplication and deals with all those flaws, introducing the convergent encryption technique which encrypts the data better using hash values of each data content for privacy purposes. The project also present several new deduplication constructions supporting authorized duplicate check in a hybrid cloud architecture. The security purposes of data is fully accomplished in deduplication with this proposal. For the demonstration purposes, a prototype is implemented for the proposed authorized duplicate check scheme and conducts an experiment which proves the encryption proposal for data security. This can be performed on a very large scale which will validate it in a better manner.

Index Terms **- Deduplication, Convergent encryption/decryption, Cloud computing, Cipher text, Differential authorization.**

I. Introduction

The idea of this project is associated with cloud computing, a virtual storage space for the data which is preserved across the whole internet. These clouds have a huge storage available in low level costs and the data or files stored on the clouds are platform independent for accessing. All these functions of cloud has been attracting more and more cloud service users since past few years leading to incalculable amount of data to be stored. This has resulted as a big challenge for cloud service providers and to handle this efficiently, data deduplication technique has come into practise. This technique is used to minimize data redundancy i.e to eliminate duplicate copies of same data stored on the cloud and through which more and more space utilization can be pursued. Only one physical copy of the repeating data will be kept on the cloud rather than holding it multiple times and other redundant data is referred to that copy. The technique takes place at file level which removes duplicate copies of same file and block level removing duplicate blocks of data of non-identical files.

Apart from all the features and security functions of cloud, a user's private data are still endangered of insiders and outsiders attacks. To avoid this or for data confidentiality, the existing encryption technique is used by the users specifically having their own private keys which gives rise to different cipher texts and hence making deduplication unworkable. To overcome this, convergent encryption are used for data privacy which uses convergent key for encoding/decoding of data and that key is obtained by evaluating cryptographic hash values of the content of each data copy. Thereafter key generation, users retain the keys and send the cipher text to the cloud and through which corresponding data copies will produce the same convergent key and hence same cipher texts. Also for prohibited access of data, a validation protocol of authorization is necessary to provide the proof that the user indeed owns the same file when a duplicate is found. After providing proof the upcoming users with the same file will be provided a pointer from the server without the requirement to upload the same file. A user can download the encoded file with the pointer from the server, which can only be decoded by the corresponding data owners with their convergent keys. Hence this convergent technique enables the data deduplication on the cloud giving all the needed securities for authorized users.

II. Existing System

In the current scenario of data storage in cloud is a well used format for placing one's confidential information on a virtual space which could be feasible to access from anywhere and at anytime. This system of collecting, preserving and sharing data among users has been widely prevailed through out the world which leads to expansion of data on a daily basis on the cloud. The management of massive amount of data turns out to be a tough confrontation for the cloud service providers and currently this is handled through a well known data deduplication technique which accepts and stores only a single copy of repeated data in the cloud reducing data redundancy and utilizing more capacity of the cloud to store data. With all those functionalities in data deduplication, it also consists of many of the drawbacks which includes:

* Existing encoding of data is used for data privacy which is not feasible with data deduplication technique.

* Creates different cipher texts for similar data which leads to impossible execution of data duplication technique.

* Confidentiality of data is not fully assured in this technique

III. Implementation and Analysis

The concept of shared storage over the internet as Cloud computing has been the ice-breaker for the current era with the forever increasing data. Deduplication is a proven technique to resolve the problem of exhausting the seemingly unlimited "virtual" cloud storage. It helps in preventing the storage of repetitive data again and again by an early data scan. With this aid, there are some constraints which shall be dealt with, such as in the case of encrypted data, deduplication is impossible. A solution to this was found in convergent encryption with proven ownership of data. Moreover, an addition of differential authorization for duplicate checks will result in an even more security-optimized system. Hence, the system is fractioned with privilege hierarchy.

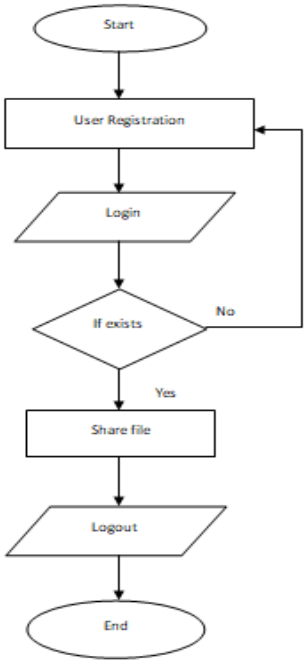

Fig. 1 File upload flowchart

The paper proposes a privacy preserving deduplication for cloud computing. An authorised user will undergo the duplicate check process of his/her file on uploading the file , on the presence of a duplicate, the user will be bound to upload the proof of ownership of the provided data. Only after the proof is validated, the user will be provided the convergent key based on data or a token based on the set of data and privileges which the user got at the time of registration.

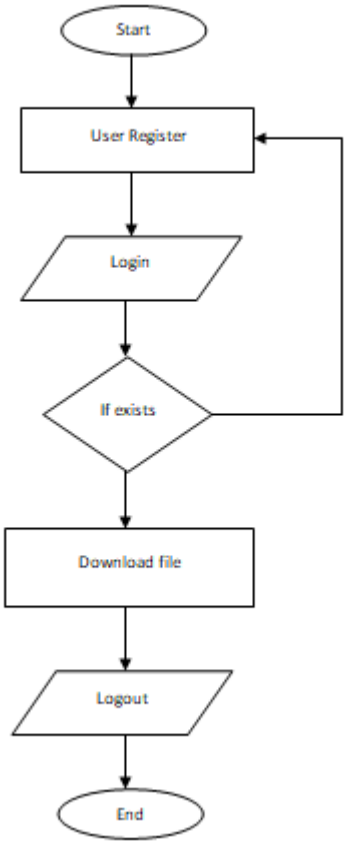

Fig. 2 File Download Flowchart

The project proposes a cloud platform such that it will implement the following features. The cloud will be allowing its users or the new registrations to access its facilities. As a standard cloud, it will provide the option to upload a new file or read or download an already uploaded file. The newer technicalities will come in picture now as if it's the case of upload of a document, it's duplication check would verify if the similar file or data is already present in the cloud storage or not. If it's a negative, file will be uploaded in ciphered form with the proof of ownership. The approval of ownership from the cloud administration will then result into generation of a convergent key based on the data which was provided in the file. The key is based on the generation of hash values based on data. Hence, similar data will lead to similar key. Also, to implement the differential authorization, a token will now be generated with two inputs, firstly convergent key and secondly the privileges associated with the user which uploaded the file. This token will be provided to user only on the approval of proof of ownership. In a different scenario, if the user uploads a file and its duplicate is present, then also the user is supposed to provide the proof of ownership of the uploaded document. If the proof is verified, then the second user will also be provided a token to access their data based on their set of privileges. This system differentiates between every single access to data all over the cloud, maintaining the privacy as well as differential access. The entire convergent key, token system and data storage details are managed by an efficient database management system.

3.1 Encryption Algorithms

It uses a key (K) to encrypt or decrypt the uploaded data.The important functions in the overall generation of encrypted data will be

1. KeyGenerationSE(1 λ)=K This is the algorithm for key (K) generation using parameter of security 1 λ .

2. EncryptionSE(K,M)=C This is the symmetric encryption algorithm which will take key K and message M and results in ciphertext C.

3. DecryptionSE(K,C)=M This is the symmetric decryption algorithm which will take key K and ciphertext C and results in

message M.

3.2 Convergent Encryption

The generation of a convergent key based on data values refers to convergent encryption. The system when it will initially accept the document, will encrypt data based on hash values generated by the data present in the document. Hence, same data will produce same hash values and ultimately same encryption sequence. This data when saved in the cloud storage, cannot be made use of, as it is encrypted. Further, the generation of token with the aid of this convergent key and the set of privileges associated with an individual's account. Hence, every user will have a unique token for their data.

3.3 Proof of ownership

This corresponds to the documentation which proves that the account associated is true owner of the provided data. This helps in authenticating only the genuine users and original documentation on the cloud.

Considering a scenario of a corporate institution, different employees will access the company cloud based on their privileges. Several documents will be present on the cloud in ciphered form. The employees can access the documents based on their privilege set, if they are qualified for the particular document, they can access it. Their limit of access can also be set based on the set of privileges. The case in which a major company project is uploaded and accessible from cloud to everyone, employees with higher privileges can control all access rights to the file. The technical team may only works on different portions of it. And the paralegal may only access the project in read only mode for documenting it.

IV. DES Algorithm

The DES algorithmic rule could be a basic building block for providing information security. It is a symmetric encryption system that uses 64-bit blocks, 8 bits (one octet) of that are used for parity checks (to verify the key's integrity). every of the key's parity bits (1 each eight bits) is employed to visualize one in every of the key's octets by odd parity, that is, every of the parity bits is adjusted to possess an odd variety of '1's within the octet it belongs to. The key thus incorporates a "useful" length of 56 bits, which suggests that solely 56 bits are literally employed in the algorithmic rule. The algorithmic rule involves finishing up combinations, substitutions and permutations between the text to be encrypted and also the key, whereas ensuring the operations is performed in each directions (for decryption). the combinations of substitutions and permutations is named a product cipher.

V. Generation of Keys

Given that the DES algorithmic rule given higher than is public, security is predicated on the quality of encryption keys. The algorithmic rule below shows a way to get, from a 64-bit key (made of any 64 alphanumerical characters), 8 totally different 48-bit keys every employed in the DES algorithm. Firstly, the key's parity bits are eliminated therefore on get a key with a helpful length of 56-bits.

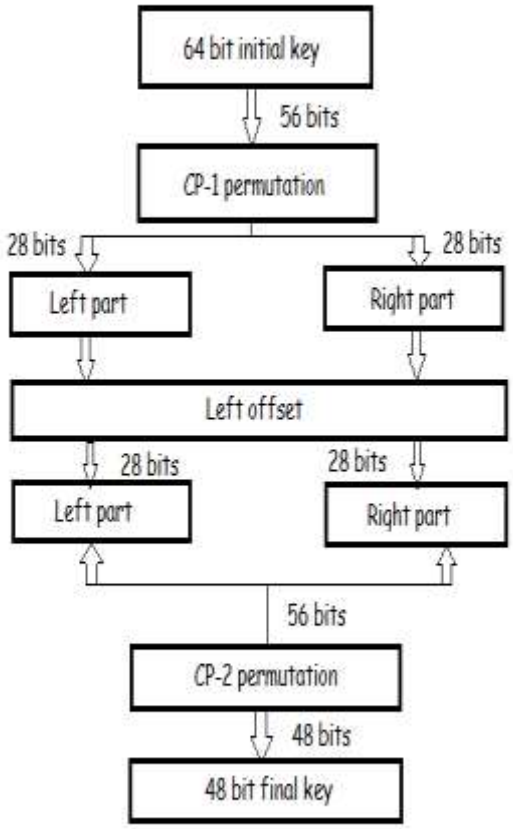

Fig. 3 Key Generation Flowchart

VI. Conclusion

The problems which arose in the current existing system was the confidentiality of data in the data deduplication technique as the present encryption technique is incompatible with data deduplication. This projects presents an idea for protecting the cloud services user's data in a more encrypted form using convergent keys technique in which each user will have its own convergent key for the encryption or decryption of personal information as this convergent keys will be retained by the user and the cipher texts will be send to the cloud.This technique gives full assurity for data security of the user as a secured ownership protocol is also followed. And for the implementation purpose, a prototype is shown as a proof for the execution of this project. This will give a far better result on a large scale basis.

VII. Future Scope

Deduplication is currently an emerging technology which can be developed with aid of efficient algorithms. The difference of a single character in a string has a minimal effect on the data whereas it's not the case with digits. The analysing power of the cloud needs to be advanced in a way to check for different problems and corresponding solutions. With an ever expanding dataset of a worldwide cloud, the need of an optimized algorithm is vital. One of the areas that can be focused on is to have a deduplication procedure for a multilingual dataset of a cloud. The comparison should be such that it is able to detect different scripts in accordance to same data.

VIII. Acknowledgement

This is to express our gratitude to everyone who contributed in effortless completion of the paper. Firstly, we would like to thank our guide Baby. D. Dayana for her valuable guidance throughout the process and provide us with all the necessary support

that was required to successfully complete our paper. Also, it would not have been possible without the guidance and support of our esteemed computer science department. Lastly, we appreciate the worthy suggestions of our colleagues.

References

[1] Fast and secure laptop backups with encrypted de-duplication, P. Anderson, L. Zhang, 2010

[2] Dupless: Server-aided encryption for deduplicated storage, M. Bellare, S. Keelveedhi, T. Ristenpart, 2013

[3] Message-locked encryption and secure deduplication.M. Bellare, S. Keelveedhi, and T. Ristenpart, 2013.

[4] Twin clouds: An architecture for secure cloud computing, S. Bugiel, S. Nurnberger, A. Sadeghi,2011.

[5] Role-based access controls, D. Ferraiolo and R. Kuhn, 1992.

[6] Proofs of ownership in remote storage systems, S. Halevi, D. Harnik, B. Pinkas, A. Shulman-Peleg, 2011