Protected De-duplication of encrypted data in cloud supported by attribute based storage

Priyanka Madhiraju[1] & V. Vinay Kumar[2]
[1]Assistant Professor, Department of computer science and engineering, Matrusri Engineering College, Saidabad.
Phone: 9866152015, email:madhirajupriyanka@gmail.com
[2]Assistant Professor, Department of computer science and engineering, Matrusri Engineering College, Saidabad.
Phone: 8125671724, email:vijaycse2012@gmail.com

Abstract*: As there's tons of increase within the number of users and therefore the size of their data, data de-duplication comes in handy for the cloud service providers. De-duplication may be a process that eliminates redundant copies of knowledge and reduces storage overhead. Cloud providers greatly reduce their storage costs and data transfer costs, by storing a singular copy of duplicate data.*

The most important and famous cloud service is data storage. The privacy of knowledge providers is extremely important so, the data in cloud is usually stored in an encrypted format. De-duplications don't want to make the management of data scalable in cloud storage. It is raised to protect the safety of knowledge .The de-duplication of data should be authoritative. Unlike from conventional de-duplication systems, the distinct privileges of users are often considered in identical check of the data . The user is hardly granted to perform the duplicate check of data apparent with the corresponding privileges.

Attribute-based encryption (ABE) has been mostly utilized in cloud computing where a data owner deploys her/his encrypted data to a cloud service provider (CSP), and may share the information with users acquiring specific authorization. The fundamental ABE system do not support secure de-duplication. In this system, we show an attribute-based storage system with protected de-duplication during a hybrid cloud setting, where a personal cloud is accountable for deduplicaiton of data and a public cloud manages the storage. It do not often provide confidentiality, instead share data among users by enumerating access policies in place of sharing decryption keys and it achieves the quality assumption of semantic security for data confidentiality while existing systems only achieve it by defining a delicate security notion.

Keywords: Attribute-based encryption (ABE), Access policy, Access Structure, De-duplication, Hybrid cloud, Authorized duplicate check, Ciphertext policy attribute based encryption (CP-ABE).

.

II.Introduction

Cloud computing has arrived together of the fastest growing segments of data Technology industry which provides variant services like software, platform and infrastructure for internet users. the best test for huge information from a security perspective is that the assurance of client's protection. Information de-duplication may be a specific information pressure system for wiping out copy duplicates of rehashing information away instead of keeping numerous information duplicates with similar substance. Physical De-duplication adapts repetitive information by keeping out De-duplicating other excess information . De-duplication can occur at any level (document level or e piece level). For record level de-duplication, it adapts a copy of duplicates of comparable document.[4] De-duplication can likewise happen at the piece level, which takes copy squares of data that happen in non-indistinguishable documents. Distributed computing may be a rising administration display that provides calculation and capacity assets on the online .

Data de-duplication is of two types:

• **File level de-duplication:** It checks certain attributes of given files across the stipulated index, if the file is absolute, updates are wiped out it. it's commonly referred to as single instance storage, if file is not different then only a pointer like existing file that is stored is updated.

• **Block level de-duplication:** In block level data de-duplication, file is being branches into segments, blocks or chunks, those chunks of knowledge are checked for repetition or already stored data.

This project proposes to avoid de-duplicate encrypted data stored in cloud supported ownership challenge and proxy re-encryption. Propose a scheme supported data owner-ship challenge to unravel the difficulty of de-duplication within the situation where the info holder isn't available or difficult to urge involved. During this method using double encryption key for encrypted data stored in cloud. Both AP key and personal key generate the encrypted key that key using for encryption.

III. Literature survey

Cloud storage service providers like Dropbox, Google Drive, Mozy, et al. perform de-duplication to save lots of space by only storing single copy of every single file uploaded. However, if clients formally encrypt their data, storage savings by de-duplication are totally lost. This is often because the encrypted data are saved as different contents by applying different encryption keys. Present industrial solutions fail in encrypted data de-duplication. One critical challenge of cloud storage services is that the management of the ever-increasing bulk of knowledge. Another challenge is to take care of confidentiality during duplicate check which is extremely essential. The prevailing system isn't fully secured. The information is often obtained by the attackers.

There are differing types of the information de-duplication, which are Inline de-duplication, Post-process de-duplication, Source based de-duplication, Target based de-duplication and Global de-duplication. In Inline de-duplication, is achieved at the time of storing data on storage system. It decreases the coarse disc space needed in system. It increases CPU overhead within the production environment but limits the entire amount of knowledge ultimately transferred to backup storage. This system allows most of the diligence to be wiped out RAM, which minimizes I/O overhead. It writes the backup data into a cache before the de-duplication process. It doesn't naturally write the complete backup to disk ahead starting the process; once the information starts to hit the disk the de-duplication process begins. The de-duplication process is escapes the backup process so we will de-duplicate the info outside the backup window without demeaning our backup performance.

Table 1: Summary of the Inline process and Post-process

Types →		Inline process		Post-Process	
Operations ↓	Process →	Redundant Block	New Block	Redundant Block	New Block
Disk Write		Not Allowed	Not Allowed	Write Block	Write Block
Disk Read		Not Allowed	Not Allowed	Read Block	Read Block
CPU		Redundancy Check	Redundancy Check	Redundancy Check	Redundancy Check

The Source side de-duplication uses client software to match new data blocks on the first memory device with already protected data previously protected data chunks. It is in contrast from all other sorts of de-duplication there in duplicate data is first only found before it's to be sent over the network. Previously stored data chunks aren't transmitted. Source-based de-duplication uses fewer bandwidth for data transmission. This may absolutely build burden on the CPU but at same time reduces the load on the network. From following fig, there are three replicates of files A and B each. So only single copy of data is shipped on to the network to cloud storage due to this, network bandwidth is rescued and for client and it requires more CPU utilization.

In Target based de-duplication [5],[6] it'll remove the repetitions from a backup transmission as and when it passes through a device that is in between the source and target. Target de-duplication appliances represent the high end of recent de-duplication. Unlike source de-duplication, the target de-duplication does not trim the entire knowledge that require to be transmitted across a WAN or LAN during the backup, but it

reduces the quantity of space for the storage required.

In Global data de-duplication [6], there is a practice of eliminating unnecessary data when backing up data to more number of de-duplication devices. This example might require backing up data to quite one target de-duplication system or within the case of source de-duplication it'd require backing up to multiple backup nodes which can themselves be backing up multiple clients. When data is sent from single node to a different, the second node recognizes that the primary node already features a copy of the information, and doesn't make a further copy.

IV.Proposed System:

In our proposed research work to style and implement a system which can provide the multiprocessing to find the information de-duplication dispute in storage at cloud. We strengthen our system in security. Particularly, we present an arduous scheme to support vigorous security by encrypting the file with differential prerogative keys.
During this way, the users without corresponding prerogative keys cannot perform the duplicate check.

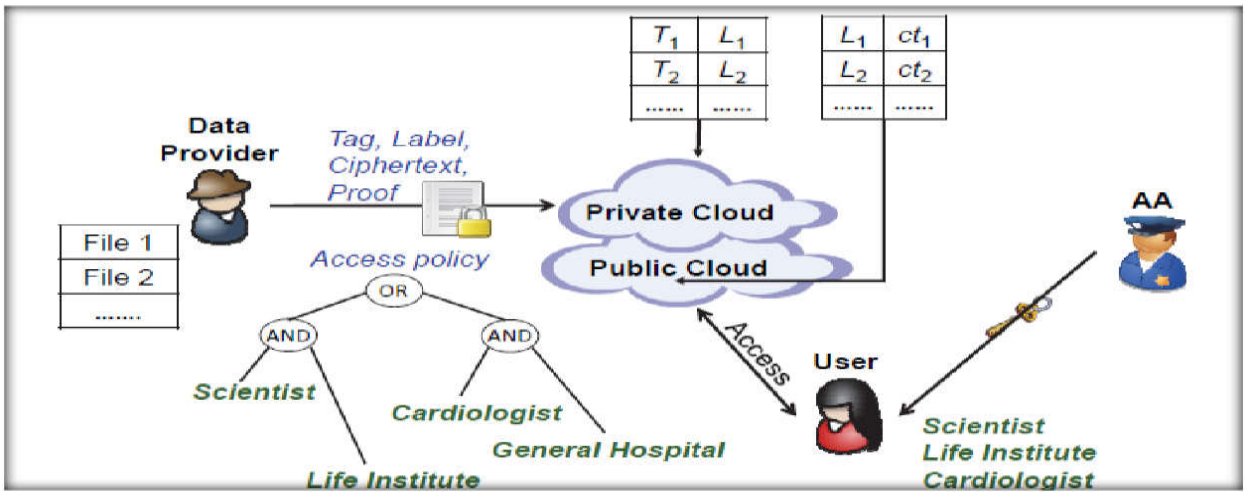

Fig 1: System Architecture

Methodology:

The proposed methodology uses the following:

1. Attribute-Based Encryption:
   Ciphertext -Policy Attribute-based encryption may be a sort of public-key encryption during which the key key of a user and therefore the ciphertext are dependent upon attributes. an important security aspect of Ciphertext-Policy Attribute-Based Encryption is collusion-resistance: An adversary that holds multiple keys should only be ready to access data if a minimum of one individual key grants access.
2. Zero-knowledge proof:
   In cryptography, a zero-knowledge proof or zero-knowledge protocol may be a method by which one party (the prover) can convince another party (the verifier) that a given statement is true, without conveying any information aside from the very fact that the statement is indeed true. differently of understanding this is able to be: Interactive zero-knowledge proofs require interaction between the individual (or computer system) proving their knowledge and therefore the individual validating the proof.
3. Commitment Scheme:
   A commitment scheme may be a cryptographic primitive that permits one to plan to a selected value (or chosen statement) while keeping it hidden to others, with the power to reveal the committed value later. Commitment schemes are designed in order that a celebration cannot change the worth or statement after they need committed to it: that's , commitment schemes are binding. . The message within the box is hidden from the receiver, who cannot open the lock themselves. Since the receiver has the box, the message inside can't be changed—merely revealed if the sender chooses to offer them the key at some later time.

4. SHA256 Algorithm:

Secure Hashing Algorithms, also referred to as SHA, are a family of cryptographic functions designed to stay data secured. It works by transforming the info employing a hash function: an algorithm that consists of bitwise operations, modular additions, and compression functions. The hash function then produces a hard and fast size string that appears nothing just like the original. SHA-256, one begins by converting the message to a singular representation of the message that's a multiple of 512 bits long , without loss of data about its exact original length in bits.

Steps Involved in CP-ABE:

Ciphertext-Policy Attribute Based Encryption consists of four fundamental algorithms:
Setup, KeyGen, Encrypt, Decrypt

- Setup: Setup algorithm does not take any inputs other than the implicit security parameter and outputs public parameters (PK) and a master key (MK) for the cloud system. This algorithm is run by Attribute Authority.

- KeyGen: KeyGen is an algorithm which generates the Private keys (sk) for the users registered on the cloud with the help of public parameters (PK), Master key (MK) and set of attributes with which the user registers. This is followed directly by the setup algorithm and is performed by Attribute Authority.

- Encrypt: The Encrypt algorithm takes the public parameters, message as well as Access structure, provided when the file is being uploaded onto cloud. It outputs a Trapdoor key and a tuple containing Label(L), Tag(T), Cipher-Text(CT) and proof(pf). Label and Tag are associated with the message, cipher-text involves the encrypted data of the message and proof is the relationship on Tag, Label and cipher-text. This algorithm is performed when the data provider is uploading some file onto the cloud.

- Decrypt: Decrypt algorithm converts the cipher-text, uploaded onto the cloud, back into normal text file. It takes public parameters, pair of label and cipher-text, private key associated with the set of attributes and outputs the message. This algorithm is run by
the user and the file can be decrypted if and only if the user satisfies the access structure.

The system works in the following way: First the users get registered on the cloud with their attributes then the Attribute Authority performs both setup and KeyGen algorithms to supply the users with their respective private keys. When the information provider (user) uploads a file onto the cloud, the file is encrypted with the assistance of encrypt algorithm and is moved onto the private cloud. In the private cloud, the file undergoes validity testing and equality testing. Validity test is performed so as to see whether the file is uploaded by a legitimate user or not and therefore the file is discarded if it fails the validity test. The file is moved to Equality testing if the validity test is successful and is completed so as to see if the file uploaded may be a duplicate file or not. If the file uploaded is duplicate then the system prompts a mistake showing that the file already exists within the cloud and therefore the duplicate file is discarded.

Validity test is performed by verifying the proofs and Equality test is completed by verifying the tags-labels of the file with the files that exist already within the cloud. If the file uploaded may be a duplicate, then necessary modifications are made to the access structure of the first file in order that it's accessible by the acceptable users. These changes to the access structure are made possible with the assistance of trapdoor key and re-encryption algorithm. After these tests, the file is moved to public cloud, where the file is stored in an encrypted format and is visible to the users. The users who want to access the files in cloud got to login and check if their attributes satisfy the access structure of the files. they will download the file if and as long as they satisfy the corresponding access structure and Decrypt algorithm is performed which converts the encrypted file in order that the user can read and download the file.

V. Conclusion

Here, we've provided how to de-duplicate data in an attribute-based storage system with secure de-duplication in a hybrid cloud setting, where a personal cloud is liable for duplicate data disclosure and a public cloud supervises the storage. It also can be more confidential to share data with users by enumerating access policies rather than sharing decryption keys and it achieves the quality approach of semantic security

for data confidentiality. This technique leads to reduced redundancy which is vital for removing same copies of same data so as to save lot of space for storing and network bandwidth and improves efficiency of CSPs.

VI. Future Scope:

Future work includes efficient verification of knowledge ownership, scheme development with hardware acceleration at IoT devices and development of a versatile result to support de-duplication and data access controlled by either the information owner or its representative agent.

VII. References

[1] X.Alphonseinbaraj, "*Authorized data deduplication check in hybrid cloud with cluster as a service*", November-2014.

[2] Jin. Li, Yan Kit Li, Xiaofeng, P. Lee, and W. Lou., "*A Hybrid Cloud Approach for Secure Authorised Deduplication*" ,In IEEE Transactions on Parallel and Distributed Systems, 2014.

[3] Boga Venkatesh, Anamika Sharma, Gaurav Desai, Dadaram Jadhav, "*Secure Authorised Deduplication by Using Hybrid Cloud Approach*", November 2014.

[4] 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 Conference on Computer and Communications Security, pages 491–500. ACM, 2011.

[5] T.Y.J. NagaMalleswari, D.Malathi, "*Deduplication Techniques: A Technical Survey*", International Journal for Innovative Research in Science & Technology, December 2014.

[6] Pooja S Dodamani, Pradeep Nazareth,"*A Survey on Hybrid Cloud with De-Duplication*",

International Journal of Innovative Research in Computer and Communication Engineering, December 2014.

[7] Usha Dalvi, Sonali Kakade, Priyanka Mahadik, Arati Chavan, "*A Secured and Authenticated Mechanism for Cloud Data Deduplication Using Hybrid Clouds*", International Journal ofEngineering Research and Review, December 2014.

[8] Kim, Sejun Song, Baek-Young Choi, "*SAFE: Structure-Aware File and Email Deduplication for Cloud-based Storage Systems*".

[9] Deepak Mishra, Dr. Sanjeev Sharma, "*Comprehensive study of data de-duplication*", International Conference on Cloud, Big Data and Trust 2013, Nov 13-15.

[10] Jin. Li, Xiaofeng Chen, M. Li, J. Li, P. Lee, and W. Lou., "*Secure Deduplication with Efficient and Reliable Convergent Key Managemen*t", In IEEE Transactions on Parallel and Distributed Systems,June- 2014.

[11] P.Gokulraj, K.Kiruthika Devi, "*Deduplication Avoidance Using Convergent Key Management in Cloud*", International Journal On Engineering Technology and Sciences, Volume I, Issue III, July-2014.

[12] Attribute-Based Storage Supporting Secure Deduplication of Encrypted Data in Cloud
Hui Cui , Robert H. Deng , Fellow, IEEE, Yingjiu Li , Member, IEEE, and Guowei Wu
IEEE TRANSACTIONS ON BIG DATA, VOL. 5, NO. 3, JULY-SEPTEMBER 2019

AUTHORS PROFILE

Mrs. Priyanka Madhiraju is working as an Assistant Professor at Matrusri Engineering College. She has done her M.Tech and has 10 years of teaching experience. She has published 6 Research Papers in various journals .

Mr. V.Vinay Kumar is working as an Assistant Professor at Matrusri Engineering College. He has done his M.Tech and pursuing Ph.d has 8 years of teaching experience. He has published 7 Research Papers in various journals .