Data Compression Technique to Eliminate Duplicates in Cloud Computing

R. Saranya[1] G. Indra[2] Dr. N. Sankar Ram[3]
[1,2]Assistant Professor [3]Professor & Head
[1,2,3]Department of Computer Science & Engineering
[1,2,3]R.M.K. College of Engineering and Technology, India

Abstract— Cloud computing promises to increase the velocity with which applications are deployed. Data Compression in cloud computing deals with reducing the storage space and providing privacy for users . Each authorized user is able to get an individual token of their file from duplicate check based on the privileges. Authorized user is able to use his/her individual private keys to generate query and hence attributes are attached along with the file. Attributes are found in the private cloud and hence control immediately passes to the private cloud, where duplicate check is performed. Data stored in the public cloud is accessed only by the authorized users by providing different encryption privilege keys. Convergent and symmetric encryption techniques produce identical cipher text that results in minimum overhead. Proof of reliability assures a verifier via a proof that a user's file is available.

Key words: Cloud Computing, Data Compression, Encryption Techniques

I. INTRODUCTION

Cloud computing increases the speed and dexterity which alludes to accessing the internet in a specific data center of different hardware and software. It is a used to describe a class of network based computing that takes place over the internet. It comprises the procurement of dynamically adaptable and virtualized reserves as a indulgence over the internet. This technology allows more efficient computation by centralizing storage memory processing and bandwidth. A censorious confrontation for cloud storage is the management of aggregate volume of accumulating data. In order to manipulate the data management, data compression or data deduplication technique has been proposed and intrigues more attention.

Since the amount of data storage is larger, there may be large amount of duplicate copies. In order to avoid those unwanted data and to save the storage space [9], a peculiar data compression technique has been enabled to remove the redundant data .This helps to reduce the byte storage in cloud. Only one copy of the tautological data is kept ant the remaining data are excluded. Redundant data are replaced with pointers, so that only eminent data can be retrieved.

Pointers are provided to users with same file so that there is no obligation to upload the file [1]. Though there are many privileges, certain security crisis may occur internally and externally. Hence certain encryption techniques are handled and are accompanied by cipher texts. Deduplication can be made possible by formatting contrast cipher texts for divergent users.

To ensure that a particular user is the owner of a specific file, proof of proprietary rights is provided. These are made possible using convergent encryption tactics [10]. The users download the file that is encrypted and then convergent keys are used to decrypt the file. Authorization is provided to guide the user while uploading the file in the cloud. Users without proper authentication are not allowed to perform dualistic checks. These checks are compassed in public cloud. After transmitting the file, checking is done for any existing privileges that correspond to match the privilege of newly uploaded data. Hence for competent storage of uploaded data, *Storage Service Provider* is imported.

II. OBJECTIVE OF DATA COMPRESSION

A. *Review Stage:*

To solve the problem of replication in cloud computing, hybrid cloud is contemplated and it is encompasses public cloud and private cloud. The function of public cloud is to manage the data storage. Private cloud monitors the attribute size. S-CSP is responsible for data storage in public cloud. Attributes are thoroughly checked in the private cloud. Only files with eminent privileges are allowed for duplicate check [1]. There are certain instances where we need to enhance our security in cloud computing. Cloud service providers store more data in the same server and hence it may lead to repetition of data and security complications. Cloud security controls are enabled to reduce the attack from insiders. These are dynamically aligned to the users to reduce complexity and to increase the performance and utilization.

Hence cloud security platform [8] exaggerates way for virtualization and load balance. The aspect of cloud computing is mostly concerned with data portability and information leakage and legal risks concerning compliance. Cloud computing architecture must involve virtualized infrastructure, scalable and dynamic application for users. It must be structured in a methodological way, so that it helps to streamline the process and other requirements. Analysis of existing and proposed system is represented in the following sections.

Acronym	Description
MD5	Message Digest Algorithm
PoW	Power of Ownership
S-CSP	Storage-Cloud Service Provider
M	Message
Y	Key
ABE	Attribue Based Encryption
R	Binary Relation
CP-ABE	Ciphertext-Policy Attribute Based Encryption
HTTP	Hypher Text Transfer Protocol

Table I: Notations Used

B. Design and Implementation:

Symmetric encryption keys are provided to encrypt and decrypt the data respectively. For example if we consider a message M, then the key used is named as Y and the output obtained is highlighted as cipher text [5]. Unlike traditional encryption, convergent encryption is more efficient to practice and implement. A new concept called tag is introduced. Each data is provided with a tag so that replications can be avoided. Privilege keys are stored in the private cloud and hence the user tries to get access through the keys followed by the corresponding data. After the keys are obtained they are applied to the encrypted data [10] stored in the public cloud.

Attributes are found in the private cloud and hence the user should get authentication for accessing the data. Once the user gets the access rights then automatically control passes to the user's data present in the public cloud. These are the main concepts of the cloud architecture. Attribute size should be brief and explicit so that more space is not allocated for storage and attributes [4]. Here Proof of ownership acts as a formal security.

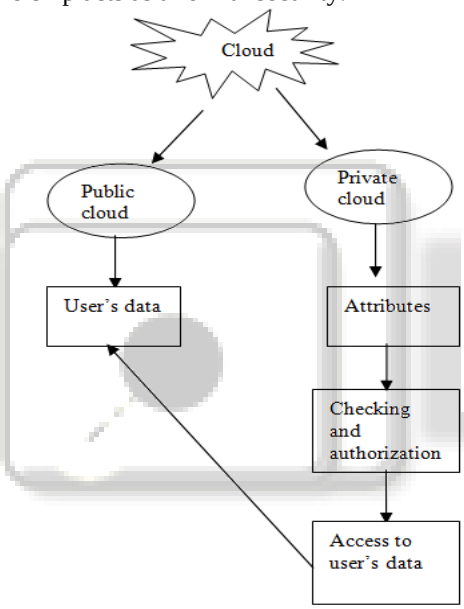

Fig. 1: System Model

C. Data Deduplication:

Data compression also known to be Data deduplication helps to reduce the size of the bits and the amount of data from its original appearance [1]. Compression is appropriate because it supports accurate data transmission. Data compression is carried out mail for backup applications and recovery procedures. Hence public and private cloud acts as two main entities in this architecture. For better understanding let us consider a university that includes principal, Hod and teachers. If the college has set the unique approach to Principal, then he alone will be able to access the data because access rights are restricted to others. Time-based and even data-based privileges can also be dispensed.

File chunking helps to allocate the data into blocks. File-level duplication is more effective than block-level duplication [10]. Because even if a single file is found to be a duplicate in a pool of files then whole file can be deleted instead of block-level elimination. Primary data deduplication is most sought due to its accessibility to data storage and reduction in workloads. Hence CSP copulates with this to reduce the storage cost [1]. When compared with other deduplication architectures, public cloud is unreliable. Therefore hybrid cloud approach has fascinated most of the customers to gain profit as well as satisfaction.

D. Secure Data Compression:

MD5 algorithm is Message Digest Algorithm used by cryptographers in cloud computing. Authority Verifier acts as a controller in this section. Data is first sent to the service provider where the data gets documented. This is consorted by MD5 checksum. User generates MD5 keys for the data received [1]. Consequently service provider also generates keys for the same data. As a result, both the keys are forwarded to authority verifier. Authority verifiers compare both the keys to find out whether they are analogous to each other. If both the keys concede with each other, then the data gets uploaded in the cloud. If the user wants to download a particular section, then he or she has to send the authentication code to the cloud storage section [8]. Request sent by the user is checked by the service provider. If it is a valid request, then the user receives the data.

Therefore we make use of hash functions for integrity check purposefully. Our intention to use hash function is to resolve collision between two identical hash values. Data security and trust worthy complications have always been a denouncing issues in cloud computing.

1) Cloud Architecture:

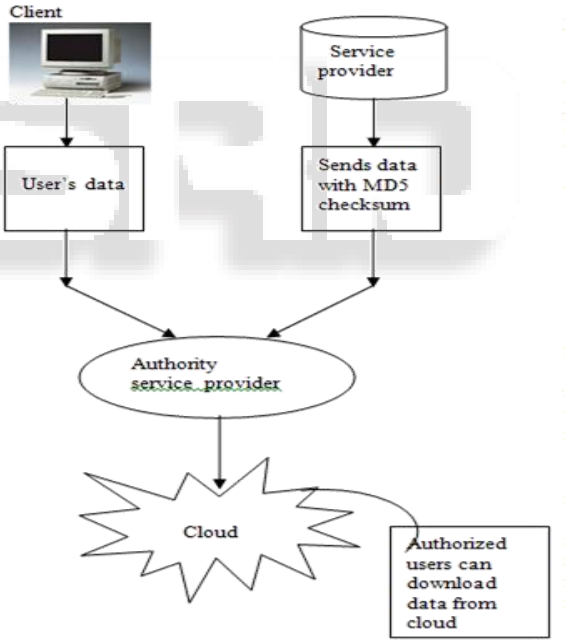

Fig. 2: Cloud Architecture

Even tokens are used as cryptographic keys in cloud computing [1]. Tokens determine the functional output and specify the particular transformation of plain text into cipher text. Tokens vary with each other in functions. Token based approaches are standardized for specific applications. These verification tokens are introduced for the convenience of the users. They also provide access to the network service and provide assurance to the owner regarding personal identification number [8]. When large amount of data are dumped into the servers, security challenges arise. Tokens are assigned to perform some additional tasks such as updating, deleting and appending. These methods are most effective and flexible when compared to previous traditional methods.

III. PROPOSED SYSTEM DESCRIPTION

A. Attribute Based Encryption:

Advancement to these techniques is Attribute Based Encryption (ABE). A binary relation R is introduced. Let us consider two privileges k and k'. Both the attributes are assigned to one. If the attributes belong to higher level privilege then, value one is assigned to the attributes [4]. As considered in the previous case of university level, Principal, HOD and Staff are the three privilege levels. Principal is in the top level, HOD and staffs are in the low level. Thus the privilege of Principal matches the privilege of others in the lower level. Further the goal of attribute based decryption is to provide security access to the users [7]. One or many encryptions and decryptions are allowed based on the size of the attributes. For example, we can consider a simple file. Each file is provided with a unique name, file id and details about the person we are going to share the file. These are stored under private cloud and are listed under attributes.

B. Abstraction of Attributes:

Attribute size is larger in the existing system. Hence our main aim in the proposed system is to reduce the size of the attribute [4]. Ciphertexts are characterized by the set of interpretive attributes. A set of attributes are designated to the users along with the secret attribute keys. For this methodology, Ciphertext-Policy Attribute Based Encryption (CP-ABE) is imported [6]. When a sender encrypts the data, attributes are scheduled over the ciphertexts in order to restrict a specific set of receivers from downloading the file. The size of the ciphertext is independent of the size of number of attributes. Therefore the performance computation reveals that the proposed system is competent to securely manipulate the data gathered in the server and this eloquently reduces the computational time.

C. Cipher Signature:

Another important concept is the cipher signature. Generally cipher signatures were created to authenticate the legal documents and signatures of a common person. Whenever a user uploads file in the cloud, cipher signatures are created for the uploaded file significantly. Let us consider a user1, who uploads a file f1 in the cloud [3]. Immediately cipher signatures are created for the file f1. Certain parameters and privilege keys are involved in the creation of cipher signature. When the receiver tries to download a particular file, he or she has to examine the data to be downloaded with the cipher signature and the privilege keys provided [5]. Then the user 2 tries to upload the file f2 into the cloud. During this approach, cipher signatures of file f1 are verified with the cipher signatures of file f2. If both the cipher signatures are similar, then file f2 cannot be uploaded in order to avoid duplication. If the file uploaded by the user 2 is different then there exists a variation in the cipher signature. In that case, the newer file uploaded by the user 2 is accepted.

IV. DATA CONFIDENTIALITY

There is a fear among the cloud users regarding the data stored in the remote machines. They are on the verge of leakage because they are operated by various service providers and users are restricted to accessibility in these areas [7]. Data confidentiality deals with third party attackers from outside the system. Sometimes service instance may deviate from the original service description. In this case if a hidden function is inserted into the software that transmits data to the user, then the data gets exposed to the unauthorized users. Therefore the users should test the service platform, whether they are exposed to unauthorized users. An acknowledgement is provided to the service platform after testing so that the users can make use of it flexibly. There is a slight difference between the software service provider and the infrastructure service provider. Software service provider knows the performance of software service instance where as the later knows the location of confidential data but the content of the data is unknown.

V. IMPLEMENTATION

A paradigm is implemented for the authorized data compression technique, in which we figure out three entities in separate java programs. File uploading is illustrated in the client program separately. Private keys and tokens are handled in the private cloud which is illustrated in the private server program. A unique storage server program is designed to represent S-CSP, since all the storage and deduplication techniques are proposed here. OpenSSL tool [2] is implemented for cryptographic and encryption techniques. Users can post HTTP requests to the server. Even images and objects are also encapsulated in HTML forms [1]. Testbed experiments are conducted which includes software and hardware testing components. They support system application and research aspects. All the modules are tested in isolation. Hash functions such as MD5 Algorithm and tokens are implemented on the client side execution. Deployment is intelligible and changes in code can be distributed competently.

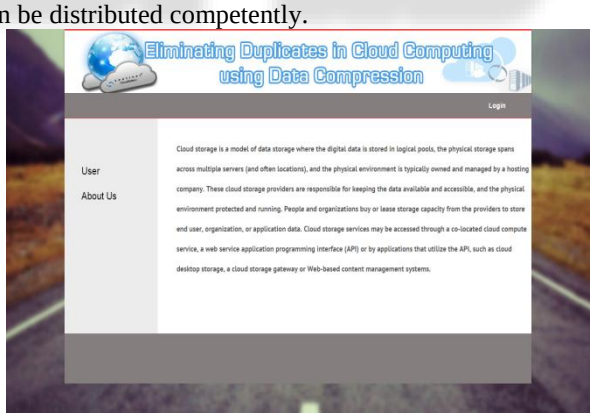

Fig. 3: Home Page

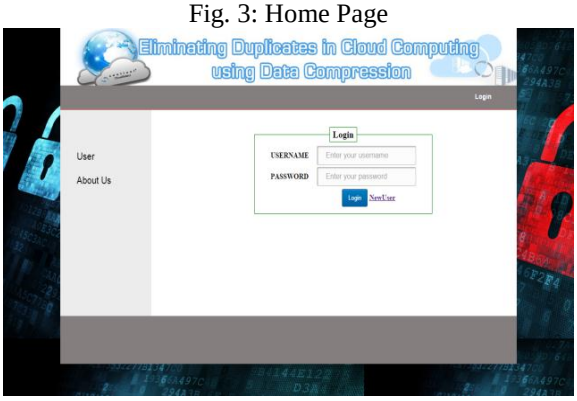

Fig. 4: Login Page

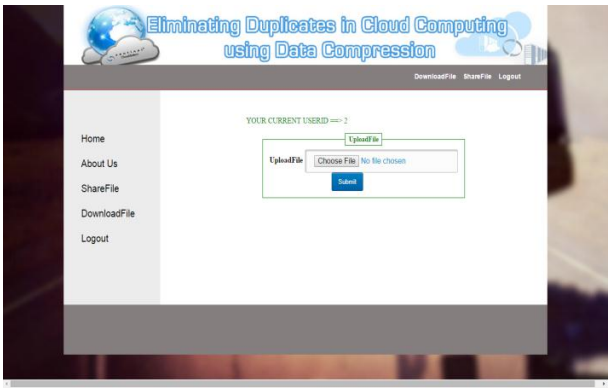

Fig. 5: File Upload

Fig. 6: Shared File List

Fig. 7: Avoiding Duplication

VI. EVALUATION

Interpretation of these procedures includes token generation and convergent encryption. Hence overhead can be reduced by considering certain factors such as size of the attribute, file size, storage capacity, privilege keys and the way, the encryption algorithms are handled. Experiments are manipulated with minimum of two machines harnessed with Intel Core 2-Quad 2.40 GHz Quad Core CPU and 4 GB RAM. Storage space for hard disk is 20 GB. Tomcat 5.0 is used as application server. Database used is Mysql and JDBC is used as database connectivity. These machines are connected with 1 Gbps Ethernet speed. Token generation, duplicate check and encryption are the three main mechanisms executed. Starting and Ending time are considered for calculating the total time spent in the review. Finally the average time exhausted in the project are depicted in the figures.

A. File Size:

To calculate the file size let us consider 50mb files to be uploaded in the cloud. Since there is no deduplication we have to upload all the files in the space allotted. Therefore the time spent on encryption, duplicate check and token generation increases because these operations involve total time spent. In other cases time spent may vary. Time taken for encryption, duplicate check and token generation vary from each other because the operations and functions are unique when compared to each other.

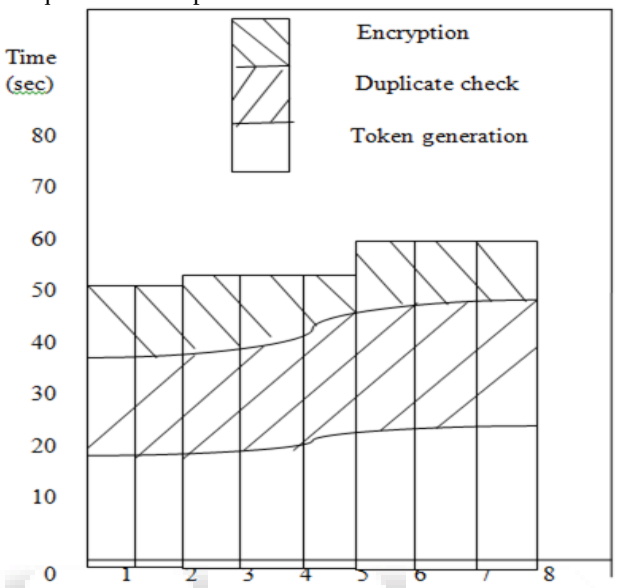

Fig. 3: Time Consumption

In case if the user uploads 100mb then there exists a time variation in following functions carried out above. Since the size of the file increases simultaneously attributes and token generation also increases. Deduplication and encryption ratio are vice versa to each other. If the deduplication ratio is higher, then consequently time taken in data transfer and encryption are lower. During initial stages, time taken for uploading the data is higher, because the data will be unique in approach. Finally the results are always consistent

Consequently signature and validation algorithms are implemented in the introduction and termination of cipher signature. Signature algorithm is used for converting plain text into digital signature. Further validation algorithms are used for converting digital signatures into a valid signature. Threats can be expected from technically advanced malicious programs that remain undetected for a particular duration of time. Hence these techniques are followed strictly in order to avoid certain threat related complications.

VII. SUMMARY

To consummate the proceedings, encryption and token generation, only minimal overhead is introduced for minimal overhead in the entire process. Hence this approach is applicable for deduplication.

VIII. CO-RELATED WORKS

Data Compression has attracted much attention from research community due to the enhancement of data

deduplication in cloud computing. Encryption shows how to convert the message into a secret ciphertext confidentially.

Symmetric Encryption provides key in which parties should have the same key before they can achieve secret communication. Convergent Encryption provides identical ciphertext derived from original data. Hence the user derives tag for data copy to be unique to avoid duplication. Proof of retrievability can assure a verifier via a consire proof that a user file is available. Hybrid cloud approach is an additional enhancement to this accession.

IX. CONCLUSION

The developed system must have a modest requirement, as only minimal or null changes are required for implementing this system. The level of acceptance by the users solely depends on the methods that are employed to educate the user about the system and to make the user familiar with it. Application of cloud computing theory is clear and carefully designed. Security model manipulates that our proposed system is secure in terms of outsider attacks.

REFERENCES

[1] Jin Li, Yan Kit Li, Xiaofeng Chen, Patrick P. C. Lee, Wenjing Lou "A Hybrid Based Approach For Secure Authorized Deduplication" in Proceedings of, IEEE Transactions on Parallel and Distributed Systems,2014
[2] OpenSSL Project. http://www.openssl.org/.
[3] M. Bellare, C. Namprempre, and G. Neven. Security
Proofs for Identity-based identification and signature
schemes J.Cryptology,
[4] Rakesh Bobba, Himanshu Khurana and Manoj Prabhakaran, "AttributeSets: A Practically Motivated Enhancement to Attribute-Based Encryption", July 27, 2009
[5] J. Bettencourt, A. Sahai, and B.Waters"Ciphertext-policy attribute based encryption "in Proceedings of IEEE Symposium on Security and Privacy, pp. 321V334, 2007.
[6] K. D. Bowers, A. Juels, and A. Oprea, "Proofs of Retrievability: Theoryand Implementation," Cryptology ePrint Archive, Report 2008/175, 2008, http://eprint.iacr.org/
[7] Shucheng Yu., Cong Wang†, Kui Ren†, Wenjing Lou., "Achieving Secure, Scalable, and Fine-grained Data Access Control in Cloud Computing" ,IEEE Communications Society for publication in the IEEE INFOCOM 2010.
[8] Pardeep Kumar, Vivek Kumar Sehgal, Durg Singh Chauhan, P. K. Gupta, Manoj Diwakar, "Effective Ways of Secure, Private and Trusted Cloud Computing", IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 3, No. 2, 2011.
[9] Cong Wang, Qian Wang, Kui Ren, Wenjing Lou, "Privacy Preserving Public Auditing for Data Storage Security in Cloud Computing", 2010.
[10] M. Bellare, S. Keelveedhi, and T. Ristenpart. Message-locked encryption and secure deduplication. In EUROCRYPT, pages 296–312, 2013.
[11] A. Rahumed, H. C. H. Chen, Y. Tang, P. P. C. Lee, and J.C.S. Lui. A secure cloud backup system with assured deletion Version control. In 3rd International Workshop on security in Cloud computing, 2011

