Addressing Cloud Data Deduplication using Hybrid Approach

R. Sivaramakrishnan[1] M. Abbinayaa[2] R. Abirami[3] M. Jayarathnaravi[4]
[1]Assistant Professor [2,3,4]UG Scholar
[1,2,3,4]Department of Computer Science and Engineering
[1,2,3,4]KPR Institute of Engineering and Technology, Coimbatore, Tamilnadu, India

Abstract— In cloud computing, data deduplication is one of the important data compression techniques for eliminating repeating data which are duplicate, and has been widely used in cloud storage to reduce the amount of storage space. To protect the confidentiality of those sensitive data while supporting deduplication, the convergent encryption technique has been proposed to encrypt the data before deploying. In existing approaches, only the file names are checked so that many data have been lost. In this paper, we use AB (Attribute Based) algorithm to check the files based on their attribute so it requires less time for checking all those files. After finishing this process it will check the content of the particular file, then the files having the same content re identified to be duplicated and will be eliminated, which will be more efficient.

Key words: Attribute Based (AB), Convergent Encryption, Deduplication

I. Introduction

Cloud Computing is one of the type of computing that trust on sharing computing resources rather than having local server or personal devices which are to be handled the over all applications . In other words, Cloud Computing refers to the delivery of computing resources in excess of the Internet. It can be accessed by individuals and business organizations and managed by third parties called Cloud Service Providers (CSP). In Cloud Computing, the word Cloud refers to the Internet. Hence, the phrase cloud computing means “ a type of Internet-based computing “ where different services such as server , storage and applications are to be delivered within a organization's computers and devices through Internet. Cloud computing helps to achieve economies of scale, reduce spending on technology infrastructure, improve flexibility and accessibility, and globalize our workforce on the cheap.

A. Types of Cloud Computing

Cloud computing is classified into two ways:

1) Deployment Model

Deployment models define the type of access to the cloud. Cloud services are typically made available via a private, community, public cloud or hybrid cloud.

2) Public Cloud

The public cloud allows the system and services to be easily accessible, owned and operated by cloud provider.

3) Private Cloud

The private cloud allows the system and services to be accessible within an organization and managed by third party.

4) Community Cloud

The community cloud is accessible by a group of organizations and made available only to those groups.

5) Hybrid Cloud

The hybrid cloud is a combination of public and private cloud, in which assessing literary activities are performed using private cloud while the non-artistic work is performed using public cloud.

B. Service Model

Cloud computing is based on service models. Service model are made available into three basic service models which are:

- Infrastructure-as–a-Service (IaaS)
- Platform-as-a-Service (PaaS)
- Software-as-a-Service (SaaS)

1) Infrastructure-As-A-Service (Iaas)

IaaS allows accessibility to fundamental resources such as physical machines, virtual machines, virtual storage, etc.

2) Platform-As-A-Service (Paas)

PaaS offers the runtime environment for applications, development and deployment tools, etc.

3) Software-As-A-Service (Saas)

SaaS model provides to use software applications as a service to end-users.

II. Related Work

A. Proofs of Ownership In Remote Storage Systems

Shai Halevi et. al. proposed a cloud storage system [1] and technology to keep their cost down in deduplication; that is to removing unnecessary copies of repeating data. Moreover, client-side deduplication attempt to identify deduplication opportunity already at the client and save the bandwidth in uploading the another copy of an existing file to the server. For eg.,an attacker who knows the hash value of a file can convince the storage service that owns file, hence the server later the attacker download the entire file. To overcome these attacks, we introduce the term proofs-of-ownership (PoWs), where a client proves to the server that actually holds the data of the file. We sanctify proof-of-ownership, present solutions based on Merkle trees and specific encodings, and analyze their security. Finally it produces small overhead for storage.

B. Secure and Constant Cost Public Cloud Storage Auditing with Deduplication

Jiawei Yuan et.al. proposed a novel scheme based [2] on techniques including polynomial based authentication tags and homomorphic linear authenticators to achieve data integrity and storage efficiency in cloud storage. These techniques assure data integrity for cloud storage and improve storage efficiency by securely removing unnecessary duplicated data on the cloud storage server. The proposed scheme is characterized by constant real-time communication and computational cost on the user side. Hence, this proposed scheme outperforms existing Proof of retrievability and PDP schemes while providing the additional functionality of deduplication.

C. Reverse Deduplication Storage System Optimized for Reads to Latest Backups

Chun-Ho Ng et. al. proposed RevDedup [3] method to achieve deduplication for VM (Virtual machine) image storage. RevDedup, a deduplication system that optimizes reads to the latest VM image backups using an method called reverse deduplication. In contrast with the conventional deduplication that removes duplicates from new data, whereas RevDedup removes duplicates from old data, thereby shifting fragmentation to old data even as keeping the layout of new data as sequential as possible. Reverse Deduplication achieves high deduplication efficiency with around 98% of saving and also high backup and read throughput on the order of 1GB/S.

D. Private Data Deduplication Protocols in Cloud Storage

Wee Keong Ng et. al. proposed private data deduplication protocol [4] in cloud storage for secure the data. A private data depulication allows a client who holds a private data proves to server and owner of the data without revealing further information to the server.It is compliment of the public data deduplication protocol and it will be secure. The security of private data deduplication protocols is dignified in the simulation-based framework in the context of two-party computations. A construction of private deduplication protocol based on the standard cryptographic assumptions.

E. Twin Clouds

An Architecture for Secure Cloud Computing Sven Bugiel et. al. proposed an architecture and cryptographic protocols [5] that accumulate slow secure computations over time and offer the possibility to query them in parallel on stipulate by leveraging the benefits of cloud computing. In Existing approaches for secure outsourcing of data and arbitrary computations are also based on single tamper-proof hardware or fully homomorphic encryption. Since homomorphic encryption is inefficient, In our approach, the user communicates with a resource constrained Trusted Cloud(either a private cloud or built from various secure hardware modules) which encrypts algorithms and data to be stored and later on queried in the powerful but it is untrusted Commodity Cloud. We split our protocols such that the Trusted Cloud performs security-critical pre-computations in the setup phase, whereas Commodity Cloud computes the time-critical query in parallel under encryption in the query phase.

F. Fast and Secure Laptop Backups with Encrypted Deduplication

Paul Anderson et. al. proposed prototype backup system and NOVEL algorithm [6] will reduces the storage requirements and increase the speed of backups. Because many people now store the large quantities of personal and then corporate data on laptops or personal computer which leads to poor connectivity and failure of hardware situation. By this algorithm it will be secure and it support unique feature which is necessary for confidential personal data.

G. Dupless: Server-Aided Encryption for Deduplicated Storage

Mihir Bellare et. al. proposed Message-locked encryption [7]. Cloud storage service providers such as Dropbox, Mozy,and others perform deduplication to save space by only storing the one copy of each file uploaded. If any encryption process is performed some savings are lost. Message-locked encryption resolves this problem. It is inherently subject to brute-force attacks that can recover files falling into a known set. The architecture is to secure deduplicated storage resisting brute-force attacks, and realize it in a system called DupLESS. In this clients encrypt under message-based keys obtained from a key-server. It patrons to store encrypted data with an existing service and have the service perform deduplication and achieves strong confidentiality.

III. Existing System

Data deduplication is one of important method of data compression techniques for eliminating duplicate copies of repeating data and also has been widely used in cloud storage to diminish the amount of storage space. To better protect data security, this paper which make the first attempt to formally address the problem of authorized data deduplication. By checking file to eliminate the duplicate data using convergence encryption.

A. Disadvantage

The existing methods only check the file name not the content so there is a chance of duplicate data available in content so it is not efficient.

IV. Proposed System

In our paper we eliminate the duplicate data by using convergence encryption and AB (Attribute Based) algorithm. In convergence encryption, convergent key and tag will be generated. In AB algorithm, check both file name and content. If any duplicate data is there it will eliminate the whole file.

A. Advantage

There is no way to availability of duplicate data.

B. Secure Authorized Deduplication

1) Efficiency in Storage

Storage plays a main role in cloud computing. In our paper we proposed algorithm called convergent encryption [4] help for storage efficiency. It providing the confidentiality in data deduplication .when the user derives a key (ie. convergent key) from each original copy and after encrypts the data copy with the convergent key .And also user derives a tag for the data copy, the copy will be used to detect duplicates. The correctness property [4] hold that if the two copies are same .If we detect the duplicate data, first send the tag to the user side and to check if the identical copies are already present in it. Convergent encryption scheme can be defined by four step function.

1) Key generation: set of files as {d1,d2,d3,d4,….dn}
   Key Gen ce=k, Data copy=M
2) Encryption E(K,M)=>C = cipertext
3) Decryption D(K,C)=>M
   M=original data copy
4) Tag generation M->T(M)

2) Authority Check

Proof of ownership [1] is used to prove their ownership of data copies to the storage server. Basically POW which is implemented as an interactive algorithm and it is run by user and verifier. Identification protocol has been introduced in two phase which are proof and verify. The input of user (prover) his private key kpr that is most sensitive information such as private key of a public key in his certificate or credit card number etc. and it is not shared to other users. The user performs the verification with the input of public based information Kpu relate to Kpr. Finally the conclusion of the protocol, that the verifier output is either accept or reject to denote whether the proof is passed or not.

C. Hybrid Cloud Approach

1) Architecture Model

In this technique, deduplication can be frequently used which is for data backup and disaster recovery application for reducing the data storage space. Three entities are used in our system such as private, public, user and S-CSP in public cloud. Privilege key is a set of privilege to specify which kind of user is to allow for performing the duplicate check operation and access the file. S-CSP which is storage cloud service provider is a new deduplication system for checking differential duplicate check in public cloud. Pu which is referred as privilege key in private and user send the token request to the private cloud after sending the token it will generate the file token for the particular user.

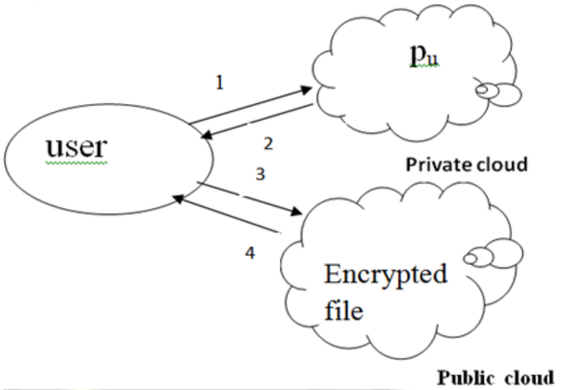

Fig. 1:

1) Request the token
2) To send a file token
3) Upload/download the file
4) Successfully performed

In this paper we mainly focus on content level encryption. Because the file level and block level deduplication for checking whole file and eliminate the duplicate file. In block level, it identifies the unique block to be uploaded.

- S-CSP: It provides the data outsourcing service and stores the data narrate on user. It is used for reducing the storage cost.S-CSP is always a online resource and has huge storage capacity and computation power.
- USER: It is one of the entity that want to outsource the storage of data to the S-CSP and to accessing the data. User performs their task as only uploading the unique data to save bandwidth.
- PRIVATE CLOUD: Private cloud is specially introduced for user security and the set of privilege key also present in that cloud,that is the action are performed for particular user and key will be provided for that user only.
- PUBLIC CLOUD: Public cloud will allow the system and service to be easily accessable to the different organization. In these part we encrypt the file for uploading/downloading the information.

D. Deduplication Security

1) Tag Generation

In our data deduplication system , we present a straightforward attempt with the technique of token generation TagGen(F, kp).The main idea of this basic construction is to perform the duplicate check based on the privilege keys and files. Suppose that there are N users in the system and the privileges in the universe is defined as P =(p1,….ps). For each privilege ps in P, the private key kp will be selected. For a user U with a set of privileges Pu, It will be assigned the set of keys {kpi} pi∈p.

2) File Uploading

Suppose the data owner U wants to upload a file with privilege set PU and share a file F with users and identified which user have the privilege set PF = { pi }. The user computes and the files to sends S-CSP token Ø^' f,p=TagGen(F,kp).

3) File Retrieving

Suppose the user wants to download a file F. It initially sends a request and the file name to the S-CSP. When receiving the request and file name, the S-CSP will check whether user is eligible to download F or not. If the user is not suitable then the S-CSP send back a signal to the user to indicate the download failure. Else the S-CSP returns the corresponding ciphertext CF. When receiving the encrypted data from the S-CSP, the user uses the key kF which has been stored locally to recover the original.In our paper we eliminate the duplicate data by using convergence encryption and AB (attribute based) algorithm.

4) Convergent Encryption

Convergent encryption is defined as content hash keying, is a cryptosystem that produce identical cyphertext fromidentical plaintext files. This having applications in cloud computing to remove duplicate files from storage without encryption keys.

5) Working

1) The system computes the cryptographic hash of the plain text in question.
2) The system encrypts the plaintext by using its hash as a key.
3) Finally, the hash itself is stored, encrypted with a key chosen by the user.

In convergence encryption, convergent key and tag will be generated. In AB algorithm, check both file name and content .If any duplicate data is there it will eliminate the whole file.

6) Advantage

- There is no way to availability of duplicate data.
- The user is only allowed to perform the duplicate check for files marked with the corresponding privileges

V. EVALUATION AND RESULT

This section which discusses the storage efficiency [4] and improve the bandwidth. The operations which include authorization check file token generation, share token generation, convergent encryption and files upload steps. In our base work some duplicate files are present because of checking the files only. In our project we eliminate all availability of duplicate data and reducing the storage. If we upload 100 unique files of particular file size and record the time break down suppose the duplicate files are present in that unique file name because of same content. Twin clouds architecture [5] provided an architecture consisting of twin clouds [5] for secure outsourcing of data and arbitrary computations to an un-trusted commodity cloud.

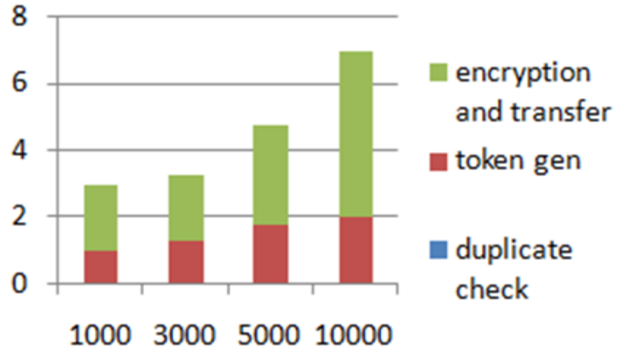

Fig. 2:

VI. Conclusion

In this paper we check authorized data deduplication is one of important data compression techniques for eliminating duplicate copies of repeating data, to reduce the amount of storage space. We also protect the confidentiality of sensitive data supporting deduplication, the convergent encryption technique has been proposed to encrypted the data. In our paper we are using AB (attribute based) algorithm to check the files based on their attribute so its requires less time for checking all those files. After finishing this process, it will check the content of the particular files if any duplicated files is present and also which files are having same contents eliminated it will be more efficient.

References

[1] S. Halevi, D. Harkin, B. Pinkas, and A. Shulman-Peleg, “Proofs of ownership in remote storage systems,” in Proc. ACM Conf. Comput.Commun. Security, 2011, pp. 491–500.

[2] J. Yuan and S. Yu, “Secure and constant cost public cloud storage auditing with deduplication,” IACR Cryptology reprint Archive, 2013:149, 2013.

[3] C. Ng and P. Lee, “Revdedup: A reverse deduplication storage system optimized for reads to latest backups,” in Proc. 4th Asia-Pacific Workshop Syst., http://doi.acm.org/10.1145/2500727.2500731, Apr. 2013.

[4] W. K. Ng, Y. Wen, and H. Zhu“Private data deduplication protocols in cloud storage,” in Proc. 27th Annu. ACM Symp. Appl. Comput., 2012, pp. 441–446

[5] S. Bugiel, S. Nurnberger, A. Sadeghi, and T. Schneider, “Twin clouds: An architecture for secure cloud computing,” in Proc.Workshop Cryptography Security Clouds, 2011, pp. 32–44.

[6] P. Anderson and L. Zhang, “Fast and secure laptop backups with encrypted de-duplication,” in Proc. 24th Int. Conf. Large Installation Syst. Admin., 2010, pp. 29–40.

[7] M. Bellare, S. Keelveedhi, and T. Ristenpart, “Dupless: Serveraided encryption for deduplicated storage,” in Proc. 22nd USENIX Conf. Sec. Symp., 2013, pp. 179–194.