International Journal of Engineering Technology Science and Research
IJETSR
www.ijetsr.com
ISSN 2394 – 3386
Volume 4, Issue 5
May 2017

Efficient Utilization of Cloud Bandwidth and Disk Usage

Bhushan Choudhary
Department of M.E.Computer Engineering
G. H. Raisoni Institute of Engineering and Technology,
Pune, INDIA

Prof. DeekshaBharatwaj
Department of M.E. Computer Engineering
G. H. Raisoni Institute of Engineering and Technology,
Pune,INDIA

ABSTRACT

Data deduplication is the technique which compresses the data by removing the duplicate copies of identical data and it is extensively used in cloud storage to save bandwidth and minimize the storage space. To secure the confidentiality of sensitive data during deduplication, the convergent encryption technique is used to encrypt the data before outsourcing. This paper makes attempt to address the problem of achieving reliable and efficient key management in secured deduplication environment. Introducing a baseline approach in which every user keeps an independent master key for encryption of the convergent keys and outsource them to the cloud. The baseline key management scheme generates a huge number of keys in which the increasing number of users that requires users to protect the master keys. The DuplicateKey approach, in which users doesn't need to manage or hold any keys on their own side instead of that it securely distributes the convergent key shares across to multiple servers. Security analysis shows that DuplicateKey is secure in terms of the proposed security model. As a proof of approach, we do the implementation of DuplicateKey using the Load balancing and demonstrate that DuplicateKey carries a limited overhead.

Keywords

Legitimate duplication removal, Proof of Ownership, Confidentiality, Convergent Encryption, Key Management, Cloud Computing.

INTRODUCTION

We Cloud computing is the new emerging trends in the new generation technology. Every user has huge amount of data to share to store in a quickly available secured place. The concept of deduplication 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 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. 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. 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. Yan Kit Li et al.[1] shown To formally solve the problem of authorized data deduplication. Data deduplication is a data compression techniques for removing duplicate copies of identical data, and it is used in cloud storage to save bandwidth and to reduce the amount storage space. 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. De-duplication occurs either at block level or at file level. In file level de-duplication, it removed duplicate copies of the identical file. Deduplication

International Journal of Engineering Technology Science and Research
IJETSR
www.ijetsr.com
ISSN 2394 – 3386
Volume 4, Issue 5
May 2017

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 strategy has been proposed to encode the information before outsourcing.

This paper will work to dissolve the security issue and to evaluate the efficient utilization of cloud band width and disk usage.

Efficiency Issues

Following Issues Addressed in De-Duplication Strategy:

Billing nature of cloud services:

Pay As You Go: User needs to pay charges as per disk space utilized by him. So, because of duplicate copies of file user need to pay more amounts.

Duplicate file upload also increase bandwidth utilization, so it degrades network performance.

User need to afford higher cost for large bandwidth uses.

Access to Authorized Users:

In cloud computing environment, same file could be shared to many users. So, there is need of implementation of access control system.

Authorized users should get download access to shared files in his access domain.

Confidentiality:

Cloud service providers are the third party service providers. So its not secure to store confidential contents as it is on cloud.

To maintain confidentiality we need to implement encryption/ decryption scheme.

But if stored encrypted files on cloud then, we can't that, the new file going to be uploaded on cloud is already present or not. So, In this paper convergence key is generated based on signature/ hash function on original data. So that we can achieve confidentiality as well as de-duplication.

Indexing & Retrieval:

As we are avoiding duplicate data storage, document retrieval will be more efficient as index takes smaller space than files itself.

Motivation

- The Convergent encryption is interested in a confirmation of a data attack in which the attacker may effectively confirm whether a target has certain file by encrypting a plain-text, form and then comparing the output with files of target.
- So to overcome these we can use another encryption technique instead of this like AES/MAES etc.
- We are proposing client side deduplication scheme as future scope.

System which overcome the issue:

Taking an example which helps to evaluate the work of system. Considering file 1 and file 2 which has similar contents and have little difference in it. Such as file 1 contains 'I LOVE MY INDIA' and File 2 Contains 'I LOVE MY COUNTRY'. The system will divide these contents in different blocks. Following fig 1. Shows the division and the generated identified codes:

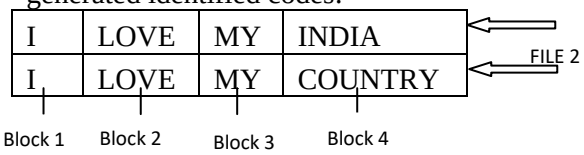

Fig. 1. Division of Blocks of each file and the identified code generated by the system for each content

TABLE I. **The identified code generated by the system for each content**

FILE 1		**FILE 2**	
Content	Identified Code	Content	Identified Code
I	X001	I	X001
LOVE	X002	LOVE	X002
MY	X003	MY	X003
INDIA	X004	COUNTRY	X005

International Journal of Engineering Technology Science and Research
IJETSR
www.ijetsr.com
ISSN 2394 – 3386
Volume 4, Issue 5
May 2017

Considering the File 1 as preloaded file in the storage system. The contents in file1 and file 2 are divided in different blocks and the identified codes are generated by the system. As shown in the table every content has its identified code, we can see the code generated is unique for each. In both the files some contents are same such as 'I','LOVE','MY'. So the identified codes generated are same in both, whereas the identified code generation of 'INDIA' and 'COUNTRY' will be obviously different as they are not same.

So this system works for the removal of the duplicate content which has the same identified code in the preloaded file and addresses the content to respective identified code. So the content of new file which is being loaded are eliminated to store by checking their identified code and the dissimilar contents are stored. Keeping the record of the similar and dissimilar content we can obtain the original file. Therefore the size of storage will be automatically reduced by eliminating the duplicate content.

Preliminary

Here the definition of cryptographic primitives used in the system are given:

Symmetric encryption: Symmetric encryption utilizes a regular secret key κ to encode the decoded data. A symmetric encryption plan comprises of three basic functions:

• KeyGen(1□) →κ -key generation algorithm generates κ utilizing security parameter 1□.

• EnC(κ,M) →C -symmetric encryption algorithm that receives secret keyκ and message M and gives cipher text C.

• DeC(κ,C) →M -symmetric decryption algorithm that receives the secret key κ and cipher text C and gives the original message M.

Convergent Encryption gives information secrecy in deduplication. Customers get a convergent key from each and every unique data copy and encrypt the unique data copy with the convergent key. And also, the customer determines a tag for the unique data copy, which will utilize the tag to recognize duplicate copies. The consideration of the tag accuracy holds[5] that means if both the data copies are the same, then the tags of the data copies are same. To discover the duplicate copies, the customer first sends the tag to the server to verify if the duplicate copy has been already available. The convergent key and tags are individually evaluated, and tags cannot understand the convergent key to distract the data security. The encrypted data copy and the respective tag will store on the server. The convergent encryption system can be defined by four basic functions:

• KeyGen (M) →K -key generation algorithm which maps an information data copy M to convergent key K.

• EnC(K,M) →C -symmetric encryption algorithm that receives the input of both data copy M and convergent key K, then gives output cipher text C.

• DeC(K,C) →M –decrypting algorithm which receives the input of the convergent key K and cipher text C, then gives the output of the original data copy M.

• TagGen(M) → T(M) –tags generating algorithm which maps original data copy M and gives output tag T(M).

Proof of Ownership: The idea of proof of ownership (PoW) [9] allows customers to verify the ownership of the information data copies to storage server. Particularly, PoW is developed as an communicative algorithm (indicated by PoW) run by a verifier (i.e. customer) and a prover (i.e. storage server). The storage server derives a short term ϕ(M) from an information data copy M. To demonstrate the ownership of information data copy M, the customer needs to send ϕ′ to the storage sever such that ϕ′ = ϕ(M). The security definition for PoW follows threat system in content distributed network, where the attacker doesn't knows the whole document, yet has accessories who have the record. The accessories follows "bound retrieval system", that it can help the attacker to get the document, subject to restrict or give limitation that they must send some few bits than the starting min-entropy of the document to the attacker [9].

Identification Protocol: This protocol can be depicted with two stages: Proof and Verify. In the phase of Proof, a prover/client U (User) can explain his identity to a verifier by demonstration or presenting some recognizable proof of identity. The information of the prover/client is his private key

International Journal of Engineering Technology Science and Research
IJETSR
www.ijetsr.com
ISSN 2394 – 3386
Volume 4, Issue 5
May 2017

sku that is delicate data for example private key of a public key in its debit card number or certificate etc. that the client doesn't wants to share others. The verifier performs the confirmation process with input of public data pku correlated with sku. At the final inference of the protocol, the verifier give output of accepts or rejects to specify that the proof is correct or not. There are numerous effective identification proof protocol, with identity based and certificate based identification.

Literature survey

The new start of cloud computing, securing the information deduplication has pulled in much consideration and attention from research group. For the integrity check a deduplication system in the cloud storage Yuan et al [11] has proposed to reduce the storage size of the tags. Their design allows deduplication of both files and their respective authentication tags. They have proven the security of their proposed scheme based on the Static Diffie-Hellman problem, the Computational Diffie-Hellman problem and t-Strong Diffie-Hellman problem.

Stanek et al. [12] The innovative encryption scheme which provides many different security of known and unknown data. For known information that are not especially delicate or sensitive, the traditional or classic ordinary encryption is performed. An alternate two-layered encryption plan with higher security while giving support to deduplication is proposed for unknown information. Along these lines, they accomplished better tradeoff between the proficiency and security of the outsourced information. Li et al. [13] tended to the key management problem in block level deduplication by circulating these keys crosswise over numerous servers after scrambling the records.

Convergent Encryption: [9] This guarantees information protection in deduplication. Bellare et al. [5] formalized a primary message-locked encryption, and analyzed its application in efficient space secure outsourced capacity storage. Xu et al. [14] additionally tended to the problem and demonstrated a protected convergent encryption for effective encryption, without considering problems of the block level deduplication and key-management. There are likewise different implementations of convergent encryption for secure deduplication. It is realized that some business cloude storage suppliers, for example, Bitcasa, likewise send Convergent encryption.

Proof of Ownership: The thought of "Proof of ownership"(PoW) Halevi et al. [9] for deduplication frameworks, such that a customer can effectively prove to cloud storage server that he owns a record without transferring the record itself. A few PoW developments established on the[9] Merkle-Hash Tree is proposed to allow customer side deduplication, which include the delimited leakage setting. Pietro and Sorniotti [10] proposed an alternate PoW plan by selecting the projection of a record onto some randomly chosen bit-positions as the record verification. Note that all the above plans don't consider information security. Newly, Ng et al. [15] enhanced PoW for encryption documents, yet they don't show how to reduce the key management overhead.

Twin Clouds Architecture: Bugiel et al. [8] given a framework comprising of twin cloud for protected outsourcing of information and subjective processing to an untrusted service cloud. Zhang et al. [16] also introduced the hybrid cloud methods to support security conscious data intensive computing. The work considers pointing the authorized deduplication issue over information in public cloud. The security model of the frameworks is same as related work, in which the private cloud is expect to be completely trustworthy and remarkable.

[6]Security proofs for signature schemes and identity-based identification 2009[G. Neven et. al]This paper gives either security proofs or attacks for a large number of signature schemes and identity-based identification defined either implicitly or explicitly in presently existing research. Underlying these is a system that from one viewpoint helps which explain how these schemes are derived and on the other hand enables the modular security analyses, thereby serving to understand, simplify, and bring together the past work. Additionally analyze a generic folklore development that in specific yields signature schemes and identity-based identification without random oracles.

International Journal of Engineering Technology Science and Research
IJETSR
www.ijetsr.com
ISSN 2394 – 3386
Volume 4, Issue 5
May 2017

[8]Twin clouds: An architecture for secure cloud computing 2002[S. Bugiel et. al] Cloud computing guarantees a more cost effective enabling technology to compute and outsource storage. Existing methodologies for secure outsourcing of information and arbitrary computations are based on a single tamper-proof hardware, or either based on newly proposed fully homomorphic encryption. The hardware based solutions are not scalable, and completely homomorphic encryption is presently only of theoretical interest and very inefficient. In this paper it is proposed an architecture for secure outsourcing of information and subjective computations to an untrusted commodity cloud. In this methodology, the client corresponds with a trusted cloud.

[14]Private data deduplication Protocols in cloud storage. [W. K. Ng et. al] Nowadays, the utilization of storage capacity becomes an important issue in cloud storage. In this paper, we introduce two categories of data deduplication strategy, and extend the fault-tolerant digital signature scheme proposed by Zhang on examining redundancy of blocks to achieve the data deduplication. The proposed scheme in this paper not only reduces the cloud storage capacity, but also improves the speed of data deduplication. Furthermore, the signature is computed for every uploaded file for verifying the integrity of files.

[12]A secure data deduplication scheme for cloud storage[J. Stanek et. al] The rapidly increasing amounts of data produced worldwide networked and multi-user storage systems are becoming very popular. It contain systematic encryption technique for the data duplication It provide the encryption system to maintain the security for cloud storage by using different techniques .It protects the unpopular contents These approach prevent the storage provider.

Related Works

Following are the objectives of proposed system:

1. The system model for de-duplication system. It contain the Hybrid Architecture for Secure De-duplication. This model is suitable for client file backup and synchronization applications than costly storage abstractions.
2. Advanced de-duplication system supporting authorized duplicate data check. Hybrid cloud architecture is introduced to solve the problem which occurs in the existing system. The private keys for privileges will not be issued to users directly, which will be kept and managed by the private cloud server.
3. The authorization of duplicate data check and the confidentiality of data. Some basic tools have been used to construct to secure de-duplication like convergent encryption scheme, and symmetric encryption scheme.
4. The implement the prototype of proposed authorized de-duplication system.

Architecture for Proposed System

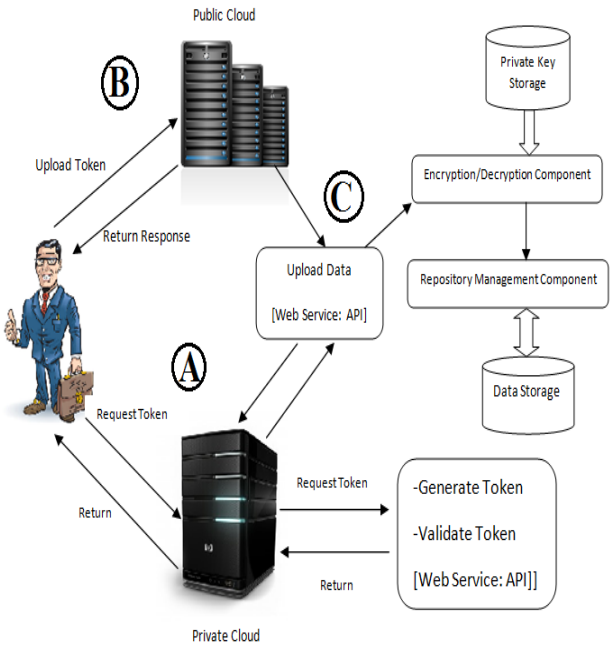

Fig. 2. Architecture of Proposed System

The proposed system contain implementation of authorized de-duplication system which contain following three model :

1. A User system program is used to model the information users to carry out the data file upload process.
2. A Private Server system program is used to model the private cloud which deals with the private keys and handles the data file token processing.

International Journal of Engineering Technology Science and Research
IJETSR
www.ijetsr.com
ISSN 2394 – 3386
Volume 4, Issue 5
May 2017

3. A Storage Server system program is used to model the S-CSP which stores and de-duplicates files.

This paper, addresses the problem of privacy preserving de-duplication in cloud computing and propose a new de-duplication system supporting for Differential Authorization, Authorized Duplicate data check, Unforgeability of file token/duplicate-check token, Indistinguishability of file token/duplicate-check token, Data Confidentiality.

- Differential Authorization: Each authorized client is able to get his individual token of his data file to perform duplicate data check which is based on his privileges. Under this assumption, any Client can't produce a token for duplicate data check out of his privileges or without the support from the private cloud server.
- Authorized Duplicate data check: Authorized client is able to use his individual private keys to produce query for certain data file and the privileges he owned with the assistance of private cloud and the public cloud performs duplicate data check directly and gives the client if there is any duplicate copy. The security necessity and requirements considered in this paper lie in two folds, including the security of data file token and security of information data files. For the security of data file token, two perspective are characterized as unforgeability and indistinguishability of data file token. The details are given below.
- Unforgeability of file token/duplicate-check token. Unauthorized users without appropriate privileges or file should be prevented from getting or producing the file tokens for duplicate data check of any data file stored on the S-CSP. The clients are not permitted to conspire with the public cloud server to break the unforgeability of data file tokens. In our proposed system, the S-CSP is trusted but curious and will sincerely perform the duplicate data check upon receiving the duplicate request from the clients. The duplicate data check token of clients should be issued from the private cloud server in our system.
- Indistinguishability of data file token/duplicate-check token: There is the necessity of any client without querying the private cloud server for some data file token, he can't get any valuable information from the token, which includes the data file information or the benefit privilege information.
- Data Confidentiality: Unauthorized clients without suitable privileges or data files, including the S-CSP and the private cloud server, should be prevented from access to the plaintext stored at S-CSP. In alternate word, the objective of the adversary is to recover and retrieve the data files that doesn't belongs to them. In our system framework, compared to the previous definition of data confidentiality based on convergent encryption, a higher level confidentiality is defined and achieved.

Mathematical Model

System S= {U, H, SE, CE, K, TG, PoW, SS}

Input:

U= Set of users.

F= Input File= {Bi} = {B1, B2…BN}

H=Hash key generation service.

H (F) =K

Where

F= Input file.

H (F) =Hash value of file F.

K= Hash Key.

SE= Symmetric Encryption/Decryption service.

CE=Convergent Encryption/Decryption service.

E (K, F) =C…….. (Encryption)

D (K, C) =F…….. (Decryption)

Where

F- File,

K-Hash Key,

E- Encryption

C-Cipher Text

TG=Tag generation service.

TG (F) =T…… (Tag Generation)

Where

F- File

T-Tag of File F

PoW= Proof of ownership service.

SS=Storage Service (upload, download)

International Journal of Engineering Technology Science and Research
IJETSR
www.ijetsr.com
ISSN 2394 – 3386
Volume 4, Issue 5
May 2017

c. Implementation of System

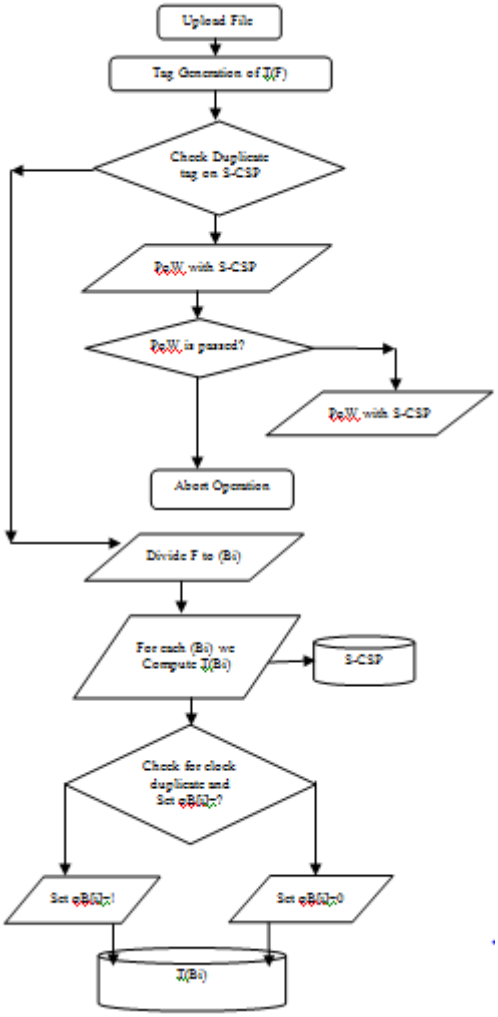

Fig. 3. Implimentation of System

Result analysis

In this section we need focus on the function of the systems with their output that will produce system result. Here in this implementation of the system comprises key generation, encryption, decryption, checking the duplication contents over the S-CSP.

4.1 Input

In this phase user have to upload document file which is to be encrypted and the tokens are checked if the duplicate contents are already present in previously uploaded copy.

4.2 Document File Preprocessing

The deduplication of document file is done. The convergent encryption technique is used to encrypt the data before outsourcing. The duplicate words in a document copy are referenced to the document file copy which is already uploaded to the S-CSP.

Symmetric encryption utilizes a regular secret key κ to encode the decoded data which comprises three steps Key generation, encryption and decryption. Convergent Encryption gives information secrecy in deduplication. Customers get a convergent key from each and every unique data copy and encrypt the unique data copy with the convergent key. The proof of ownership (PoW) [9] is implemented which as allowed customers to verify the ownership of the information data copies to S-CSP. The indentification protocol is depicted with two stages: Proof and Verify.

Acknowledgement

I would like to take opportunity to acknowledge the contribution of certain people without which it would not have been possible to complete this paper work. I would like to express my special thanks to my guide Prof. Deeksha Bhardwaj for her valuable guidance and support.

TABLE II. Input in different Algorithms

Input Size (in byte)	DES	3DES	AES	BF
20527	24	72	39	19
36002	48	123	74	35
45911	57	158	94	46
59852	74	200	120	90
234167	278	755	456	219

References

[1] Li, Jin, et al. "A Hybrid Cloud Approach for Secure Authorized Deduplication."
[2] Li, Jin, et al. "Secure deduplication with efficient and reliable convergent key management." (2013): 1-1.
[3] Bugiel, Sven, et al. "Twin clouds: Secure cloud computing with low latency." Communications and Multimedia Security. Springer Berlin Heidelberg, 2011.
[4] Anderson, Paul, and Le Zhang. "Fast and Secure Laptop Backups with Encrypted De-duplication." LISA. 2010.

[5] Bellare, Mihir, SriramKeelveedhi, and Thomas Ristenpart. "Message-locked encryption and secure deduplication." Advances in Cryptology–EUROCRYPT 2013. Springer Berlin Heidelberg, 2013. 296-312.

[6] Bellare, Mihir, ChanathipNamprempre, and Gregory Neven. "Security proofs for identity-based identification and signature schemes." Journal of Cryptology 22.1 (2009): 1-61.

[7] Li, Jin, et al. "A Hybrid Cloud Approach for Secure Authorized Deduplication."

[8] Bugiel, Sven, et al. "Twin clouds: An architecture for secure cloud computing." Proceedings of the Workshop on Cryptography and Security in Clouds Zurich. 2011.

[9] Halevi, Shai, et al. "Proofs of ownership in remote storage systems." Proceedings of the 18th ACM conference on Computer and communications security. ACM, 2011.

[10] Di Pietro, Roberto, and Alessandro Sorniotti. "Boosting efficiency and security in proof of ownership for deduplication." Proceedings of the 7th ACM Symposium on Information, Computer and Communications Security. ACM, 2012.

[11] Yuan, Jiawei, and Shucheng Yu. "Secure and constant cost public cloud storage auditing with deduplication." Communications and Network Security (CNS), 2013 IEEE Conference on. IEEE, 2013.

[12] Stanek, Jan, et al. A secure data deduplication scheme for cloud storage. Technical Report, 2013.

[13] Li, Jin, et al. "Secure deduplication with efficient and reliable convergent key management." (2013): 1-1.

[14] Douceur, John R., et al. "Reclaiming space from duplicate files in a serverless distributed file system." Distributed Computing Systems, 2002. Proceedings. 22nd International Conference on. IEEE, 2002.

[15] Ng, Wee Keong, Yonggang Wen, and Huafei Zhu. "Private data deduplication protocols in cloud storage." Proceedings of the 27th Annual ACM Symposium on Applied Computing. ACM, 2012.

[16] Zhang, Kehuan, et al. "Sedic: privacy-aware data intensive computing on hybrid clouds." Proceedings of the 18th ACM conference on Computer and communications security. ACM, 2011.

[17] Daemen, Joan, and Vincent Rijmen. The design of Rijndael: AES-the advanced encryption standard. Springer Science & Business Media, 2002.

[18] Standard, Data Encryption. "Data encryption standard." Federal Information Processing Standards Publication (1999).

[19] Wang, Cong, et al. "Privacy-preserving public auditing for data storage security in cloud computing." INFOCOM, 2010 Proceedings IEEE. Ieee, 2010