ISSN: 0369-8963

Performance Analysis of Data Security Mechanisms for Data Deduplication in Hybrid Cloud Computing Environment

MS.L.SUDHA[A], DR.C.ARUNACHALA PERUMAL[B], MS.K.B.ARUNA[C], MS.V.SUREKA[D]

[b] Professor, Department of Electronics & Communication Engineering, S.A.Engineering College,
[a,c,d]Assistant Professor, Department of Computer Science & Engineering, S.A.Engineering College-Chennai,

Corresponding author: Dr.C.Arunachala Perumal

Abstract- *One of the major concerns of users and organizations outsourcing their data to cloud providers is the issue of data integrity and privacy. Data deduplication is a technique to reduce storage space and eliminate duplicate copies of data by identifying redundant data using hash values to compare data chunks, storing only one copy, and creating logical pointers to other copies instead of storing other actual copies of the redundant data. Deduplication reduces data volume so disk space and network bandwidth can be reduced which reduces costs and energy consumption for running storage systems. Data deduplication can be applied at nearly every point in which data is stored or transmitted in cloud storage. Many cloud providers offer disaster recovery and deduplication can be used to make disaster recovery more effective by replicating data after deduplication for speeding up replication time and bandwidth cost savings. Backup and archival storage in clouds can also apply data deduplication in order to reduce physical capacity and network traffic. Moreover, in the live migration process, we need to transfer a large volume of duplicated memory image data. Duplication of data stored in the cloud occupies more space. However, during data updates, duplicate data must be changed in more than one place, which is more complex to rectify and would increase operational costs in the cloud.*

Keywords: *Cloud Storage, Data Deduplication, Security, Encryption, Key Generation*

Introduction

Cloud refers to the network that provides services to the network through the internet. It is a model that enables the characteristics like on-demand self-service, and pay-as-you-use-service. National Institute of Standards and Technology (NIST) defines cloud computing [4] as a convenient, on-demand computing resource for storage services. Cloud computing refers to manipulating, configuring, and accessing applications [1] online. It deals with many services like Infrastructure, Data storage [2], and application. Cloud computing offers many data deployment models like public cloud storage, private cloud storage, hybrid cloud storage, and community cloud storage. Service models are classified into three models Infrastructure-as-a-Service (IaaS), Platform-as-a-Service (PaaS), and Software-as-a-Service (SaaS). Cloud is a massive shared computing resource that includes Data Storage. It is managed by a cloud service provider on cloud data servers built on virtualization techniques known as utility storage. Most of the storage clouds [3] run on the public internet cloud by well-known companies like Amazon, Dropbox, and Google. A few bigger associations have discovered esteem in running private clouds inside their own data centers. A wide range of

 https://doi.org/10.37896/pd91.4/91469

ISSN: 0369-8963

data storage models is used in the concerns of users and organizations outsourcing their data to cloud providers. One of the challenging issues is to accomplish both data privacy and integrity for data outsourcing services.

In cloud storage [5] where data is remotely maintained, managed, and backed up in a cloud environment, and then the data is accessible to end users/organizations over the internet. Itpermits the client to collect the files online so that the client access these files from anywhere, anytime via the internet. Security and Privacy are the distinguished methods used to secure information from attackers. The cloud storage data model specifies how digital data gets stored and retrieved across multiple servers in possibly geographically different locations and managed by a hosting provider. Cloud storage uses a logical memory model that allows providers to store your data on multiple servers in different locations in a way transparent to you. At a minimum, cloud storage includes space for some amount of data and a simple interface to manage files in the storage.

Data Deduplication

Data deduplication [6] in the cloud is a new technology that caters to the rapidly increasing amount of digital data in data storage. Data deduplication is the process of identifying the redundancy in data and then removing it. The resulting unique single copy is stored and will then serve all of the authorized users. A minimum number of data replicas called replication factors are maintained in a large distributed storage system to achieve high data availability. Any duplicate data above the replication factor is removed to reduce storage requirement, storagecost, computation, and energy. Deduplication is fundamentally a storage optimization strategyfor reducing redundant information. Figure 1 illustrates the functioning of knowledge deduplication over data before it will be available in storage. The Deduplicationmechanism can be categorized into two types file-level deduplication and block-level deduplication in view of granularity. File-level deduplication considers the entire record, in this manner even little change or alteration makes the document unique as compared to before processed one then reducing storage optimization. Within the case of block-level deduplication, it'll divide the whole file into a number of chunks and people are considered deduplication. Deduplication is often performed by both client and server side. Client-side deduplication can save bandwidth by sending a hash value if the copy is existing. Deduplication usually is employed in different storage improvements.

 https://doi.org/10.37896/pd91.4/91469

ISSN: 0369-8963

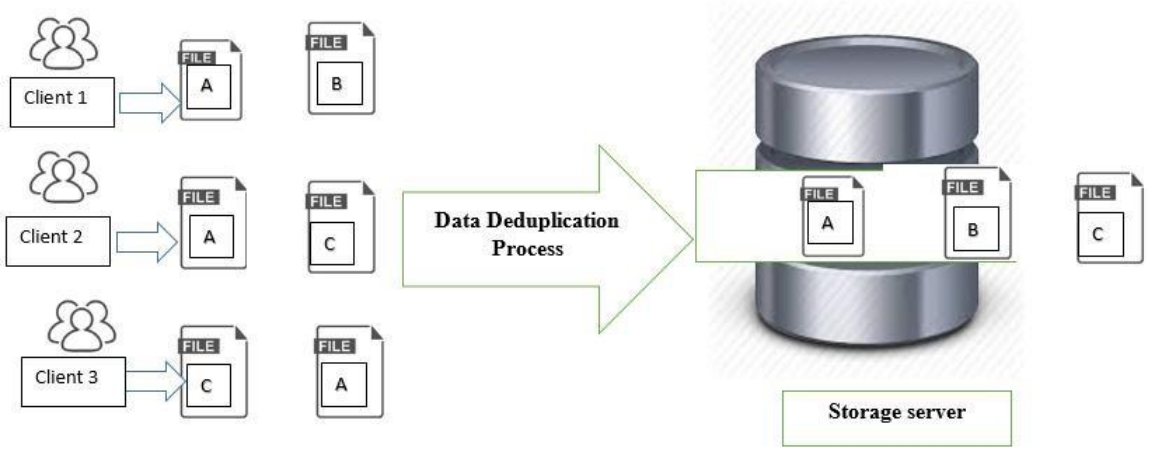

Fig 1: Data Deduplication Process

A. Data Deduplication Classifications

Duplication in data can be identified by many methods one of the method is deduplication process can be used to avoid the storage of redundant data in cloud storage. Deduplication method [7] can be applied to data in two aspects based on location and based on level.
Location can be used either in client side or server side. Levels are categorized into file level and block level.

Merits and demerits of deduplication:

The deduplication has significant advantage in storage system. These techniques require resources to employ and draw benefits. The paper highlighted the important merits and demerits of deduplication techniques.

Merits of deduplication

The following merits of deduplication are identified and presented below:
(i) Reduce storage space Deduplication assists in reducing storage space required for backups, file or other data applications. As only a unique copy of data is stored and duplicate copies are removed. So, it creates more free space to store more data.
(ii) Improves network bandwidth As the unique copies are stored on disk and logical pointers are created for duplicate data, there is no need to transmit duplicate copies over the network. The deduplication helps in reducing network bandwidth requirements.
(iii) Reduce energy consumptions Deduplication is a storage optimization technique that reduces storage and energy requirements. The reduced storage space requires less electricity and coolants. Thus, it saves energy requirements and reduces load on system resources.
(iv) Reduce overall storage cost Deduplication helps in significant savings in terms of time, space, network bandwidth, human resources and budget. It leads to better efficiency and efficacy of storage system.

 https://doi.org/10.37896/pd91.4/91469

ISSN: 0369-8963

Demerits of deduplication

(i) Impact on storage performance in primary storage system, fixed-size approach leads to multiple chunks stored at different memory locations. It leads to fragmentation issue, which adversely impacts the performance. The deduplication technique requires additional resources like CPU, memory and bandwidth for its execution. Any inefficient deduplication technique impacts the performance of a large storage system.

(ii) Loss of data integrity the data blocks are indexed through hash values for better lookup. The identical hashes can be generated for different data blocks due to hash collision that can cause loss of data integrity. So, hash collisions must be carefully addressed to avoid any loss of data and its integrity.

(iii) Backup appliance issues Data deduplication may require a separate hardware device to transfer and process data. Such backup appliance may lead to additional cost and impact storage performance.

(iv) Privacy and security the deduplication techniques have full access to complete storage. It can be exploited to get complete access of storage. The security of deduplication techniques should be carefully designed to guard system from such security breaches and loss of private data.

Secure Data Deduplication

Deduplication technique in a cloud storage creates various issues since the data going to storein a remote system. It creates various issues [7] like, Dictionary Attack, Poison Attack, and Ownership Forgery Attack. In public and hybrid cloud models, your data resides at a third- party data center. Any attack on co-tenants can result in your data getting exposed too. Attacks can be mitigated by applying efficient cryptographic methods encrypting your data in the cloud prevents hackers from being able to read it correctly.

Related Work

Nishant N. Pachpor, proposed a new method called POD (Performance-Oriented data Deduplication scheme) to enhance the storage performance in cloud. Instead of a capacity oriented one (e.g., iDedup), the I/O performance of primary storage systems within the cloud are often improved by taking under consideration the characteristics of the workload. [8] POD takes two sorts of approach, i.e., selective dedup and iCache. For reducing data deduplication, selective dedup checks the request-based techniques. In iCache by using memory management scheme fragmentation of data which helps to enhance the performance of primary storage systems. iCache technique is suitable for bursty read traffic and therefore the bursty write traffic.

Guohua Tian, a randomized client-side deduplication scheme, which uses a randomized deduplication protocol to stop the collusive authentication attack and offline brute-force attack launched by the surface adversaries, and stores each data consistent with two file tags to resist duplicate-faking attack. Additionally, we realize a more available ownership management and data sharing with the help of dynamic Key-Encrypting Key tree [10]. A dynamic KEK tree to realize a more available ownership management withinthe deduplication system.

 https://doi.org/10.37896/pd91.4/91469

ISSN: 0369-8963

Marcel Chibuzor Amaechi, research aim at developing data storage management in cloud computing using deduplication technique. Object oriented methodology was used. Data deduplication has been achieved via block level deduplication and key generation (symmetric algorithm). The info file was divided into number of blocks and of fixed length. Each block was divided into segments and therefore the files were saved just one occasion. However, each file was converted into cipher text (key form) using symmetric algorithm, the system checked for existence of key and excluded redundant key [11], maintaining just one copy of the key within the cloud storage, the stored key was shrinked to scale back the space for storing using ShrinKey algorithm and rejection algorithm was wont to remove replicated key. The system supported data privacy since data stored in cloud was encrypted and user privacy was supported, as data was uploaded by different users. This approach results in an equivalent content being encrypted several times and to the cloud provider's storage capabilities being reduced.

Secure pseudorandom key-based encryption mechanism to realize semantic security alongside deduplication. During this paper, Data confidentiality are often achieved via user-side encryption. However, conventional encryption mechanism is at odds with deduplication. Developing a user-side encryption mechanism with deduplication [12], may be a vital research topic. SPARK achieves semantic security alongside deduplication. Security analysis proves that SPARK is secure against dictionary attacks and tag inconsistency anomaly.

Secure cloud data de duplication scheme with efficient re- encryption scheme. It had been convergent all-or-nothing transform (CAONT) and randomly sampled bits from the Bloom filter [13]. Due to the intrinsic property of one- way hash function, this scheme can resist the stub-reserved attack and guarantee the info privacy of knowledge owners' sensitive data. Moreover, instead of re-encrypting the whole package, data owners are only required to re-encrypt a little a part of it through the CAONT, thereby effectively reducing the computation overhead of the system. This paper focuses on file-level deduplication, which divides file data into the fixed-size chunk.

Secure and efficient client-side encrypted data deduplication scheme (CSED). In CSED, a dedicated key server is introduced in generating MLE keys to resist brute-force attacks. We propose a Bloom filter-based proofs of ownership (PoW) mechanism and integrate it into CSED [14], to resist illegal content distribution at- tacks. Moreover, a hierarchical storage architecture is used to enhance the I/O efficiency on the cloud server. Security analysis and performance evaluation demonstrate that CSED is secure and efficient.

Proposed a key-sharing method based on proof of ownership for secure deduplication. In the new scheme, only the initial uploader of the data owner encrypts the data with a randomly- chosen CK and then distributes the CK in the cloud, and only the users possessing the claimed data can retrieve the CK. The CK only needs to store once for a single duplicate data. It adopts a deduplication check on the plaintexts and the consistency policy, and only a few owners need to encrypt the duplicate data.

 https://doi.org/10.37896/pd91.4/91469

ISSN: 0369-8963

Convergent Key (CK) [15], management problem is rectified in this method.In this paper discussed that data duplication is achieved data privacy and security. Data Owners are safeguarded with any data theft that may take place during transmission of data from Data owner machine [16], to CSP's servers. CSP has less liability as they are not the one who is encrypting data. This will help them in distracting attackers. The datastored in the cloud will be accessible only if they are registered as owners of the data.

Proposed a secure and scalable data deduplication scheme with dynamic user management, which updates dynamic group users in a secure way and restricts the unauthorized cloud users from the sensitive data owned by valid users. To further mitigate the communication overhead, the pre-verified accessing control technology is adopted, which prevents the unauthorized cloud users from downloading data. In other words, our present scheme also ensures that only the valid cloud users are able to download and decrypt the cipher text [17], from the cloud server. All this reduces the communication overhead in our scheme implementation. Compared with the existing schemes this scheme does not require a fully trusted third party and the cloud user for excessive computational overhead. It uses the access control technique to verify the validity of the cloud users before they download data. Only when the cloud users are in group, will the cloud server send cipher text to the cloud user. Therefore, the abundant communication cost will be reduced.

Security Technology for Secure Data Deduplication System Model

Symmetric encryption: Symmetric encryption uses a single key to encrypt as well as decryptdata. The key needs to be shared with all authorized people. Asymmetric encryption: Also called public key cryptography, asymmetric encryption uses two separate keys—one public (shared with everyone) and one private (known only to the key's generator). The public key is used to encrypt the data and the private key helps to decrypt it. There are different encryption methods based on the type of keys used, key length, and size of data blocks encrypted. Some of the common encryption methods.

1. Advanced Encryption Standard (AES)

Advanced Encryption Standard is a symmetric encryption algorithm that encrypts fixed blocks of data (of 128 bits) at a time. The keys used to decipher the text can be 128-, 192-, or 256-bit long. The 256-bit key encrypts the data in 14 rounds, the 192-bit key in 12 rounds, and the 128-bit key in 10 rounds. Each round consists of several steps of substitution, transposition, mixing of plaintext, and more. AES encryption standards are the most commonly used encryption methods today, both for data at rest and data in transit.

2. Rivest-Shamir-Adleman (RSA)

Rivest-Shamir-Adleman is an asymmetric encryption algorithm that is based on the factorization of the product of two large prime numbers. Only someone with the knowledge of these numbers will be able to decode the message successfully. RSA is often used in digital signatures but works slower when large volumes of data need to be encrypted.

 https://doi.org/10.37896/pd91.4/91469

ISSN: 0369-8963

3. Triple Data Encryption Standard (TripleDES)

Triple Data Encryption Standard is a symmetric encryption and an advanced form of the DES method that encrypts blocks of data using a 56-bit key. TripleDES applies the DES cipher algorithm three times to each data block. TripleDES is commonly used to encrypt ATM PINsand UNIX passwords.

4. Twofish

Twofish is a license-free encryption method that ciphers data blocks of 128 bits. It's considered the successor to the Blowfish encryption method that ciphered message blocks of 64 bits. Twofish always encrypts data in 16 rounds regardless of the key size. Though it works slower than AES, the Twofish encryption method continues to be used by many file and folder encryption software solutions.

5.ElGammal

ElGamal algorithm is used in encryption and decryption, which is mainly considered for its capability to make the key predictions extremely tough. The asymmetric algorithm uses the mechanism of private and the public key, making the key predictions even tougher.

6.ECC

Elliptic curve cryptography (ECC) is a kind of public key cryptosystem like RSA. But it differs from RSA in its quicker evolving capacity and by providing attractive and alternative way to researchers of cryptographic algorithm. The security level which is given by RSA can be provided even by smaller keys of ECC (for example, a 160 bit ECC has roughly the same security strength as 1024 bit RSA).

 https://doi.org/10.37896/pd91.4/91469

ISSN: 0369-8963

Proposed Work

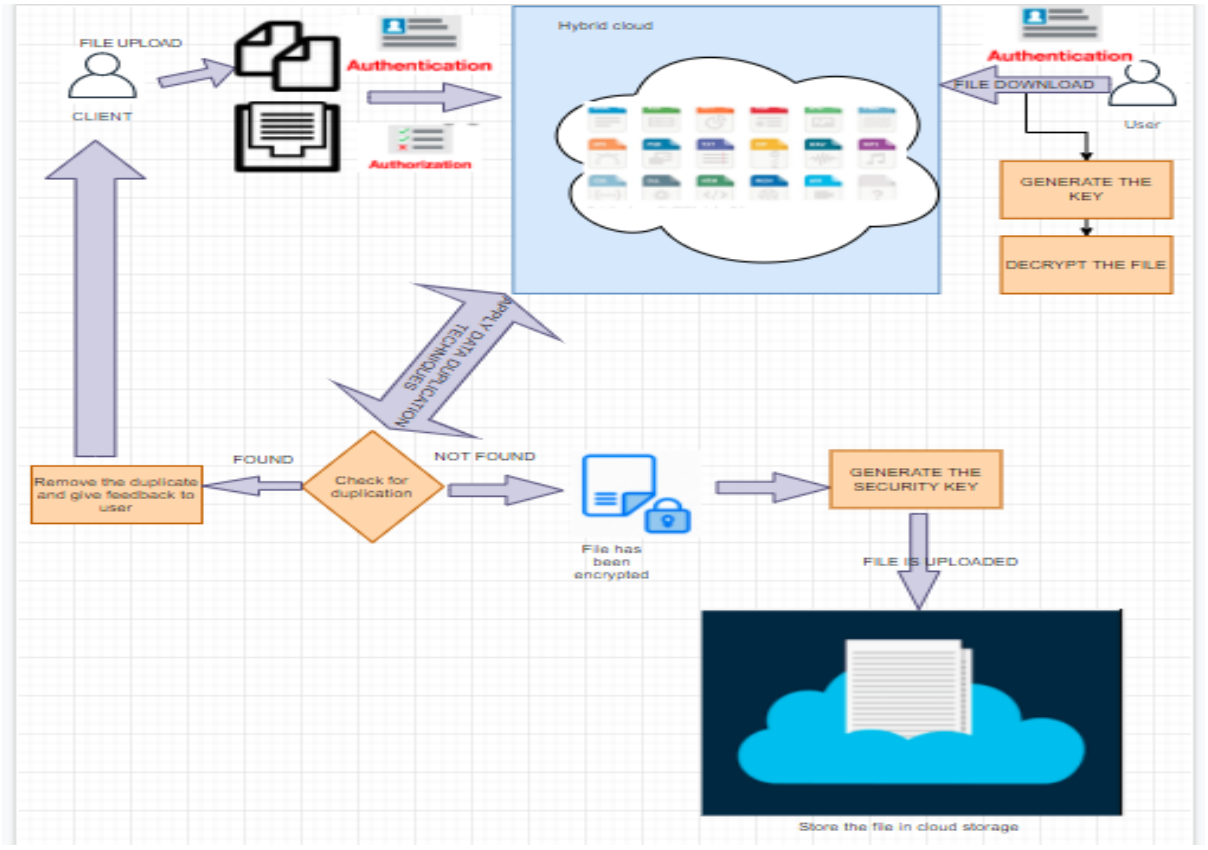

Fig. 2 System Architecture

The proposed deduplication system with efficient user revocation consists of three participants: data owner, user and cloud server. The overview of the proposed architecture is shown in Fig. 2. **Data owner**: When the data owner wants to change the access rights of a document to a new group of users, s/he will send a token to the cloud server. The cipher text can be decrypted by the data owner and the authorized users. **User**: It is possible for a user to download a document and decrypt its cipher text, if the document is authorized to user. If the user's access to the document is revoked, user will not be able to decrypt the document. **Cloud Server**: If a server receives a new update token, it will change the corresponding cipher text under a new group key. The server first performs deduplication on the encrypted document and then stores the non-deduplication part. Thus, proposed DAE can be easily applied to any cloud storage providers to use their cloud storage services. DAE has six main functional modules: Data Deduplication, Data Distribution, Encryption, Key Exchange, Performance Evaluation and Cost Evaluation. To eliminate the redundant data blocks, the Data Deduplication module splits the incoming data into multiple data blocks and calculates their hashes (SHA1 or MD5). The reference values of the data blocks are alsoupdated. With regard to redundancy, the Data Distribution module determines

 https://doi.org/10.37896/pd91.4/91469

ISSN: 0369-8963

that the incoming data blocks should be distributed to the cloud storage providers according toreference values in the data blocks. To prevent unauthorized access to data that must be stored in the cloud. Encryption can be applied using ECC, a technique that uses efficient key exchanges.

Performance Evaluation

In this section, the experimental evaluation of the proposed scheme is presented. Fig.3 and Fig.4 gives the encryption time and decryption time respectively with respect to the file size. The different algorithms are compared with respect to the features and performance.

Table 1: Comparison of various cryptographic algorithm parameters

Cryptographic Algorithms	**DES**	**AES-128**	**TDES**	**Blowfish**	**RSA**	**ECC**
Encryption type	Symmetric	Symmetric	Symmetric	Asymmetric	Asymmetric	Asymmetric
Structure	Feistel network	Substitution-Permutation Network	Feistel network	Fiestel	Fiestel	Elliptic Curves
Key Length in bits	56	128	112 or 168	32 to 448	1024	224
Key generation Time	29 ms	75 ms	-	-	287 ms	0.18 s
SecurityLevel	Not Secure enough Brute force attack	Excellent Security	Adequate Security	Vulenerable to birthday attacks	Least Secure	side- channel attacks and twist-security attacks.
Cipher Text	slow	Very Fast	Very slow	Fast	slowest	Fastest
Rounds	16	10	48	-	1	-
Block Size in bits	64	128	64	64	Variable	-

 https://doi.org/10.37896/pd91.4/91469

ISSN: 0369-8963

Table 2: File Size with encryption times

File Size (IN KILO BYTES)	DES Encryption (in ms)	AES Encryption (in ms)	RSA Encryption (in ms)	ECC Encryption (in ms)
32	0.27	0.15	0.13	0.12
126	0.83	0.46	0.52	0.48
200	1.19	0.72	0.74	0.66
246	1.44	0.95	1.11	1.3
280	1.67	1.12	1.39	1.22

Table 3: File Size with decryption times

File Size (IN KILO BYTES)	DES Decryption (in ms)	AES Decryption (in ms)	RSA Decryption (in ms)	ECC Decryption (in ms)
32	0.44	0.15	0.15	0.12
126	0.65	0.44	0.43	0.41
200	0.85	0.63	0.66	0.62
246	1.23	0.83	0.93	0.75
280	1.45	1.10	1.23	1.05

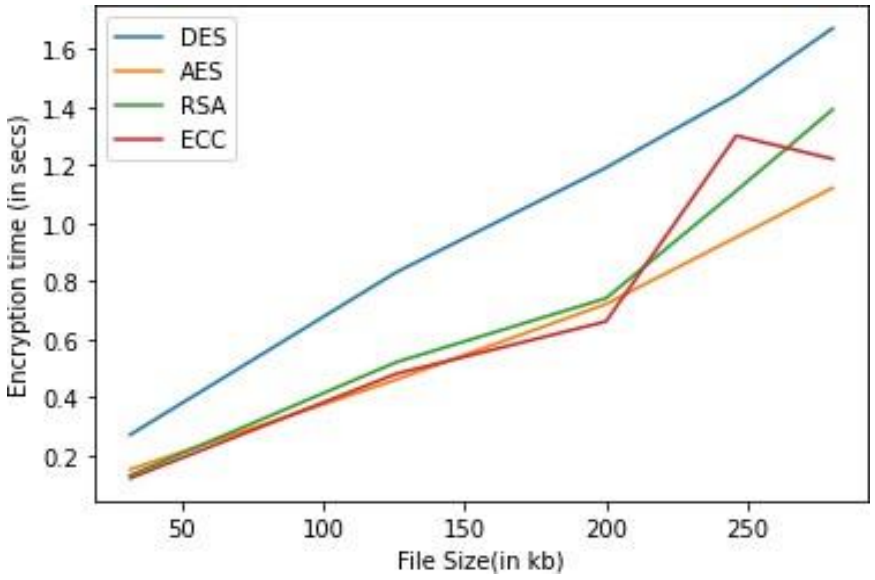

Fig. 3 File Size vs Encryption

 https://doi.org/10.37896/pd91.4/91469

ISSN: 0369-8963

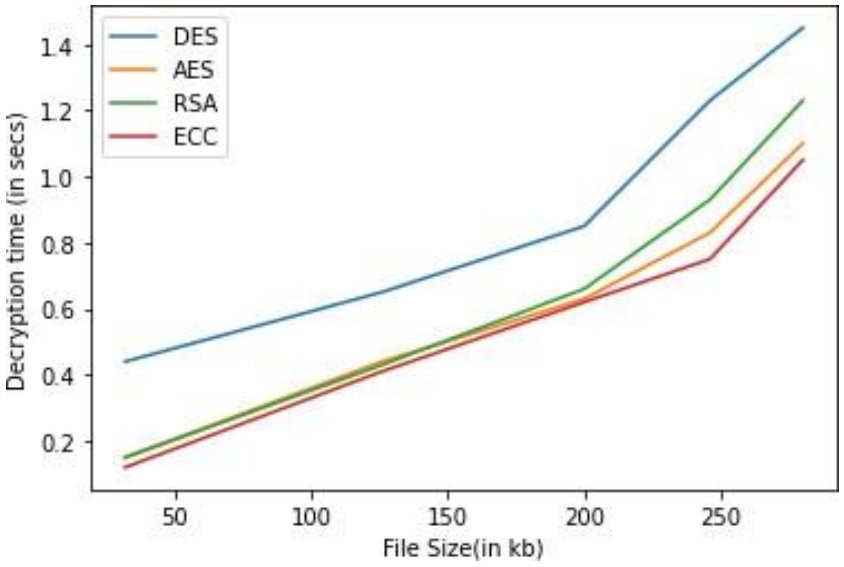

Fig. 4 File Size vs Decryption

From the figure 3 and 4 one can understand that the time required for encryption as well as decryption of DES is higher compared to other algorithms. AES, RSA and ECC are relatively similar inboth encryption and decryption. It is also noted that AES algorithm is almost linear compared to any other algorithms, ECC requires minimal time for decryption when the file size is above 200 kb.

Conclusion

To reduce storehouse costs Data deduplication technologies are designed. In order to reduce the network business in recent times, many plans have been made. Data centers are generally similar to pall storehouses, stationed as remote waiters. Similarly, remote waiters must be designed against data leakage since that data can be blurted or converted by bigwig or external factors at any time due to the semi-trustworthy. In general, there are confidentiality and integrity conditions that can repel data leakage and tampering. By cracking and storing the data confidentiality can be answered. But still, data deduplication technology is a fashion that grasps the source of data and has a point that conflicts with encryption technology that transforms the source of the data. Thus, to cipher data, encryption is suitable for deduplication must be applied which is typical technology for this is CE. Still, fresh exploration has been conducted while introducing new security pitfalls, indeed with the operation of these technologies. Bandied ways for performing data deduplication and the process of performing secure data deduplication is been discussed in this paper. These technologies aren't fully redesigned but some modified security technologies have been applied to the need for data deduplication and also achieved their intended objects, it’s delicate to anticipate high situations of absoluteness when satisfying both calculation and business effectiveness due to the emergence of fresh security pitfalls and the use of fresh security technologies to address them. Still, the recent demand for sequestration and secure technologies is anticipated to increase the demand and force of secure data deduplication technologies.

 https://doi.org/10.37896/pd91.4/91469

ISSN: 0369-8963

References

[1] Nurma Ayu Wigati, Ari Wibisono and Achmad Nizar Hidayanto Faculty of Computer Science, "Challenges of Infrastructure in Cloud Computing for Education Field: A Systemmatic Literature Review "Universitas Indonesia, Depok, Indonesia.

[2] Shalini Bhaskar Bajaj, Aman Jatain, Sarika Chaudhary, and Pooja Nagpal, "Cloud Storage Architecture: Issues, Challenges and Opportunities "

[3] Garcia, Gene Joseph, V. "Past, Present, and Future of Cloud Computing: An InnovationCase Study"

[4] Mrs. Ashwini Sheth, Mr. Sachin Bhosale, Mr. Harshad Kadam "Research Paper on Cloud Computing",

[5] H. Guesmi, C. Ghazel and L. A. Saidane,"Securing Data Storage in Cloud Computing", International Journal of Engineering Research & Technology (IJERT),2016.

[6] K.Vijayalakshmi and Dr.V.Jayalakshmi ,"Analysis on data deduplication techniques of storage of big data in cloud", Proceedings of the Fifth International Conference on Computing Methodologies and Communication (ICCMC 2021).

[7] Won-Bin Kim and Im-Yeong Lee,"Survey on Data Deduplication in Cloud Storage Environments",

[8] Nishant N. Pachpor and Prakash S. Prasad," Securing the Data Deduplication to Improve the Performance of Systems in the Cloud Infrastructure", SPRINGER 2020.

[9] Cheng Guo, Xueru Jiang, Kim-Kwang Raymond Choo, Yingmo Jie,"R-Dedup: Secure client-side deduplication for encrypted data without involving a third-party entity", Journal of Network and Computer Applications 162 (2020) 102664.

[10] Guohua Tian , Hua Ma , Ying Xie , Zhenhua Liu ,"Randomize d de duplication with ownership management and data sharing in cloud storage", Journal of Information Security and Applications 51 (2020) 102432.

[11] Marcel Chibuzor Amaechi, Matthias Daniel, Bennett E.," Data Storage Management in Cloud Computing Using Deduplication Technique" *SSRG International Journal of Computer Science and Engineering (SSRG-IJCSE) – Volume 7 Issue 7 – July 2020* .

[12] Jay Dave, Parvez Faruki , Vijay Laxmi , Akka Zemmari , Manoj Gaur , Mauro Conti," SPARK: Secure Pseudorandom Key-based Encryption for Deduplicated Storage",Computer Communications 154 (2020) 148–159.

[13] Haoran Yuan, Xiaofeng Chen, *Senior Member, IEEE,* Jin Li, Tao Jiang, Jianfeng Wang, and Robert H. Deng, *Fellow, IEEE,"* Secure Cloud Data Deduplication with Efficient Re-encryption *"* IEEE Transactions on Services Computing.

 https://doi.org/10.37896/pd91.4/91469

ISSN: 0369-8963

[14] Shanshan Li, Chunxiang Xu , Yuan Zhang," CSED: Client-Side encrypted deduplication scheme based on proofs of ownership for cloud storage ",Journal of Information Security and Applications 46 (2019) 250–258.

[15] Liang Wang a , Baocang Wang a , ∗, Wei Song a , Zhili Zhang b," A key-sharing based secure deduplication scheme in cloud storage",Information Sciences 504 (2019) 48–60

[16] Vaishnavi Moorthy, Arpit Parwal and Udit Rout ,"DEDUPLICATION IN CLOUD STORAGE USING HASHING TECHNIQUE FOR ENCRYPTED DATA", ARPN Journal of Engineering and Applied Sciences,VOL. 13, NO. 5, MARCH 2018 ISSN 1819-6608

[17] Haoran Yuan a , Xiaofeng Chen a , ∗, Tao Jiang a , Xiaoyu Zhang a , Zheng Yan a , Yang Xiang a , b "DedupDUM: Secure and scalable data deduplication with dynamic user management", Information Sciences 456 (2018) 159–173.

[18] https://symbiosisonlinepublishing.com/computer-science-technology/computerscience-information-technology32.php

[19] http://norma.ncirl.ie/3890/1/vickybidhuri.pdf

[20] Suzhen Wua, Kuan-Ching Li c, Bo Maob,∗, Minghong Liao b ,DAC: Improving storage availability with Deduplication-Assisted Cloud-of-Clouds, Future Generation Computer Systems, Volume 74, September 2017, Pages 190-198.

[21] Yunling Wang a,*, Meixia Miao a, Jianfeng Wang b, Xuefeng Zhang a , Secure deduplication with efficient user revocation in cloud storage, Computer Standards & Interfaces Volume 78, October 2021, 103523

 https://doi.org/10.37896/pd91.4/91469