Semiconductor Optoelectronics ISSN: 1001-5868

THREE FACTOR DEDUPLICATION IN CLOUD USING ELLIPTICAL CURVE CRYPTOGRAPHY

K. Syed Mohamed Bukari

Research Scholar, PG & Research Department of Computer Science, Quaid-E-Milleth Govt. College for Women, Chennai

Dr. K. Nirmala

Associate Professor, P.G Department of Computer Science, Quaid-E-Milleth Govt. College for Women, Chennai, nimimca@gmail.com

Abstract

Cloud storage has many benefits to provide, but it also makes it more difficult to preserve data, wastes resources that are associated to networking, and uses an excessive amount of energy. On the other hand, traditional de-duplication technologies are completely ineffective when presented with encrypted data since they are unable to recognise repetitive pattern repeats. To provide a solution to the problems that were mentioned earlier, we present the three factor-ECC method as a method for dependable de-duplication detection at both the file-level and the content-level of encrypted data stored in a cloud environment.

Keywords: Deduplication, Cloud, Elliptical Curve Cryptography, Three Factor, Authentication

1. Introduction

Data de-duplication is one of the techniques that can be utilised to compress data; its primary objective is to eradicate instances of data that are identical to one another. When data is moved within the cloud, it is possible to enhance the utilisation of the network as well as the storage space by lowering the total number of bytes that are transmitted [1]. During the de-duplication process, the file that is now being transmitted is broken up into a huge number of chunking files [2].

During the process of conducting a duplicate analysis, any chunking of data that cannot be matched is discovered. This data is then saved to the cloud after the procedure is complete. A comparison is made between the chunking data currently being saved and the chunking data that was previously saved during the process of carrying out a duplicate analysis [3]. The chunking that was being matched is replaced with a reference point to the chunking file that has already been stored whenever there is a match. This happens whenever there is a match. There is a possibility that the same section of data will appear more than once within the same file. It is feasible to drastically minimise the amount of time as well as the space that is required to store chunks. The size of the chunking can be used as a basis for determining how frequently matching occurs [4].

Because there are so many cloud services that offer enormous quantities of file storage, it might be difficult to manage and check the integrity of data files such as texts, movies, photographs, and other sorts of sensitive personal records. The typical encryption approach is provided here for both the encryption and decryption methods [5]. This method makes use of the personal key of the user. for the user to have access to the data in this manner, he or she will be asked to enter a secret that has been chosen by them. If the system finds out that one of its secret key pathways has been compromised, it will immediately shift to a technique of encryption that is more ordinary. A new personal key will be generated, but this will come at the expense of communication in a considerable amount [6].

The data leakage resilient (LR) encoding method can be used in the process of making the information accessible and generalizable across all forms, it is not possible to decode it in the opposite direction. It offers fundamental approaches, the most important of which are proofs of ownership (PoW), with the objective of securing cloud servers and, more especially, client-side data duplication [7]. However, LR did not take into account the cost of de-duplicating the information on both ends, and did they take into account the cost of having the conversation [8].

It will only be essential to have one form of authorised access to have access to all the general facts that have been envisioned because of implementing this method because it takes a unified approach to the problem. Despite this, the system is unable to make guarantees of dependability and consistency [9].

A secure de-duplication strategy (SDS), which is explained in this article, it is possible to safely store data in the cloud as well as share that data with other individuals. The primary goal of the method was to prevent unauthorised parties from accessing the data, and it allowed each data owner to generate a one-of-a-kind key that could be used to secure their data before it was stored in the cloud. This was accomplished by preventing unauthorised parties from accessing the data. In comparison, getting information through SDS takes significantly more time and results in increased costs associated with communication [10].

Cloud computing encourages the outsourcing of data storage to the cloud, which is beneficial to both the company and the customer. Cloud storage may be accessed from anywhere with an internet connection. Using cloud storage is beneficial for both parties. It is possible for unauthorised individuals to obtain access to sensitive data due to the decentralised nature of the cloud computing infrastructure. Because so many different people can contribute data, there is a possibility that certain information will be duplicated more than once. Utilizing de-duplication creations that permitted authorised duplicate detection helped make it possible to validate the hybrid cloud architecture [11].

The current approach utilises randomised convergent encryption (RCE) and a central key management machine to solve a revolutionary server-facet de-duplication plot method for integrated statistics. This was accomplished by combining the two. Because of the way in which the merging method operated, we were able to ensure that only a single group of individuals who were appropriately authorised could access the entire data collection. One of the ways that is addressed in the manner that is being used to address the duplication issue is a mechanism known as convergent encryption [12].

It featured multiple layers of security and a system that did away with the need to store duplicate information. Both features were included in the solution. When utilising this method, there are no guarantees made concerning the reliability, consistency, or privacy of the results. It possible for people who are completely different from one another or who are extremely like one another to save several copies of the same content in the cloud [13].

Cloud storage has many benefits to provide, but it also makes it more difficult to preserve data, wastes resources that are associated to networking, and uses an excessive amount of energy. On the other hand, traditional de-duplication technologies are completely ineffective when presented with encrypted data since they are unable to recognise repetitive pattern repeats.

To provide a solution to the problems that were mentioned earlier, we present the three factor-ECC method as a method for dependable de-duplication detection at both the file-level and the content-level of encrypted data stored in a cloud environment.

2. Related works

Encrypting the data hash and generating an appropriate tag from the ciphertext developed by Douceur et al. [14] are two methods that can be used to protect the confidentiality of data. Bellare et al. [15] give an implementation of an encryption model in addition to providing an overview of a comment authentication framework (CAF) and its security concept. Because this approach is susceptible to being broken through brute force, this step is essential.

Convergent Encryption (CE) is a method that has been proposed by Keelveedhi et al. [16] to encrypt data that is coming from a key-server by utilising a forgotten PRF protocol and message-based keys. This method is known as the Convergent Encryption (CE) method. In addition, a mechanism known as DUPLESS has been suggested to better guarantee the security of CE. However, to perform deductions of a more granular nature, DupLESS needs to engage in several lengthy computations. This is required since utilising the PRF protocol causes things to move at an imperceptibly slower pace. One of the things that Duan has to offer is a decentralised version of the EwS Duplex protocol. Due to this, DupLESS necessitates a considerable amount of overhead to do fine grain deduplication, and the utilisation of the PRF protocol is not immediately apparent.

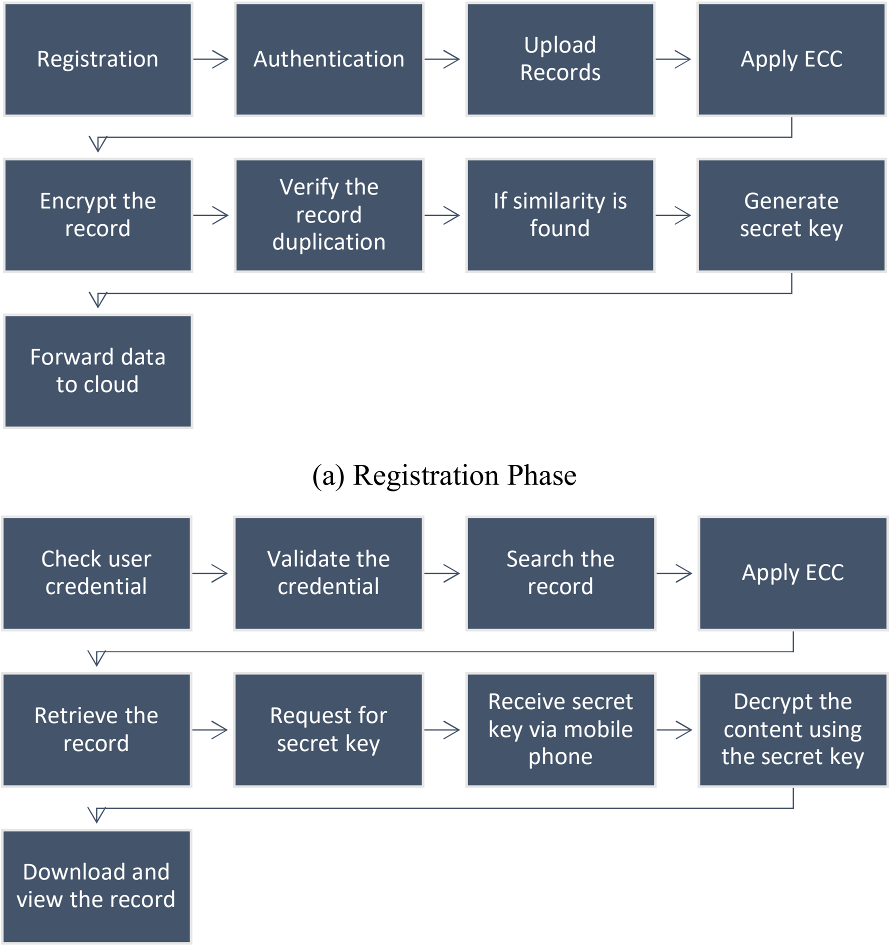
Harnik et al. [17] provide an illustration of a hostile insider attack that contributes to leaks in cloud storage as an example of how data deduplication could be employed in such an attack. The leaks in cloud storage are a result of the attack.

For use with encrypted cloud data, Li et al. [18] presented a client-side proof-of-work (PoW)-based leak-resistant deduplication technique. This mechanism was designed to be used with cloud storage. The research carried out indicates that convergence keys can be sold and dispersed among the many nodes that make up a network. This helps to guarantee that the deduplication over-block values selected by the nodes keep their authenticity.

3. Proposed Method

In this section, we discuss the proposed methodology that was advised as well as the pre-processing phases of the implementation, including the specifics of the approach that was used. The method that is now being offered makes it simpler to discover and display encrypted data in settings that are hosted in the cloud that has been de-duplicated. It gets rid of stuff that has been copied and pasted several times, which is redundant, and it does this by eliminating duplicate content.

In Figure 1, you can see a depiction of the operational notion of the technique, as well as the flow of the process and the mathematical implementation details. These are all related to one another. The procedures that need to be finished to carry out the pre-processing are outlined in the following, and a summary of those procedures can be seen below.



(a) Registration Phase

(b) Secret Key Generation

Figure 1: Working of Proposed Three factor Deduplication

Data owner

After successfully connecting to their cloud account, the owners of the data can safely upload any encrypted files to the cloud storage associated with their account. A single individual or an entire company could be the owner of a collection of data at any given time. After any duplicates have been eliminated, the owner of the data can check the file that was just uploaded to ensure that it has not been tampered with. After the data has been de-duplicated, the owner of the data has access to the file and can clear it of any unnecessary or redundant information if they so want. The file is then posted to the cloud server when the content replication process has been completed and any duplicate content has been removed. Only the person who owns the data can see the list of requested files, figure out the specifics of the access privileges, and create and distribute the secret key.

Cloud user

Users of various cloud computing services Cloud users have the option of signing up with their genetic data and acquiring an authentication key to gain access to their data that is kept in the cloud if they prefer to make use of this approach. The cloud user will be able to access the cloud and download the necessary data once the key has been verified as being legitimate. If the owner of the data allows another person access to view their data, anyone who has a cloud storage account can view any file that is saved in the cloud if they have the owner permission to do so.

Secure de-duplication system

It is possible for owners of data to save copies of files that they have submitted to a cloud storage service. De-duplication is a subset of data compression that refers to the process of getting rid of duplicate copies of data. This is referred to as de-duplication. Intelligent data compression and single-instance storage are essentially the same thing, even though they are referred to in a variety of different ways. Before uploading a file to the cloud, the owner of the data can make use of the method that was provided to detect and prevent duplication at both the file and the content levels. This is something that can be done before uploading the file. If the content-level de-duplication detection method does not turn up any duplicates, the file is then chopped up into blocks. This system configuration is very similar to that of another system that gets rid of duplicate files at the file level.

Data sharing

Data sharing is a mechanism that is used to disguise the identities of persons while simultaneously breaking up sensitive information and distributing it to those who have a need for it. This can be accomplished by spreading the information to those who have a need for it. It utilised several different data recovery strategies, which enabled it to not only collect but also retrieve sensitive information. The ECC method begins by taking the confidential data and chopping it up into chunks of the same size. Next, the process generates random pieces of the same size. Ad hoc erasure code is utilised so that the pieces of the same volume can maintain dimensions that are uniform across the entirety of those fragments.

Distributed cloud storage

With the assistance of this strategy, it is possible to take full advantage of the capabilities offered by cloud storage facilities. It is feasible to make use of it to cut down on the amount of space that is necessary for data transfers that take place in the cloud. When performing a deduplication analysis, unique data chunks or byte patterns are sought out and catalogued as they are discovered. The validation of the partitioned blocks is carried out by utilising the cloud computing resources. When a copy is created, the original block is overwritten with a reference that offers a new position for the data that was input. This occurs whenever a copy is generated. This takes place each time a duplicate is manufactured. The amount of data that is being compared as well as the size of the locks that are currently being employed are both factors that can affect the likelihood that a duplicate match will be discovered.

File restored

Any user of the cloud storage service who can access the initial upload can remove it at any moment from the service storage space at their discretion. After the ECC process has successfully validated the one-of-a-kind identification of the file, the secret key is then transmitted to the cloud storage provider together with the identifier of the file. When files are uploaded to the cloud, the process of deleting them afterwards is handled in an automated method.

Revocation of cloud user

When an administrator has access to the revocation list, they could expel a cloud user from their account. File encryption is one approach that owners of data can employ to prevent past users of a cloud service from accessing and viewing their data. Encrypting one files is one way that owners of data can protect their data. The administrator will always can maintain an accurate and up-to-date blacklist of cloud users. A signature that validates the veracity of the cloud user list revocation is included in the document, and the document itself is included in the package as well. An algorithm for the formation of signatures is utilised by the administrator to accomplish the task of creating the signature. A list of cloud users who have had their access to the cloud temporarily revoked is uploaded by the administrator to a cloud server that is visible to the public.

ECC

The study builds an elliptic curve over a finite field Fp by finding the solution to the equation

$$E(Fp):y2=x3+a\cdot x+bmodp,$$

where a, bFp, and =4a3+27b20modp are all constants, and p is a very large prime number. All the points on $E(\underline{F_p})$ and the point O, which is situated at infinity, originate from the same location in an additive Abelian group G of order q, where P is the generator point and $n\cdot P=P+P+\ldots+P$, where $n\in Z*q$ is an integer. This location may be found in an additive Abelian group G.

Threat Model

The Dolev-Yao threat model was utilised to carry out a comprehensive analysis on the authentication and key agreement procedure that was demonstrated. This model is based on the premise that there are two communication concepts that interact with one another through a channel that is both unsecured and open. The following is a summary of some of the characteristics that this model possesses, from most important to least important:

1. It is impossible to crack the one-way hash function that is currently being utilised.
2. A standard protocol ensures that all parties adhere to the same set of standards when it comes to the sharing of information.
3. In the third place, any messages that are sent through an unprotected channel are open to the possibility of an adversary listening in on them, intercepting them, playing them back, and modifying them.

Fuzzy Extractor

With increasing number of systems make use of biometric characteristics as a means of beefing up their levels of security. Because of their unique nature, biometric traits lend themselves particularly well to the process of individual verification. When compared to passwords with a low entropy, biometric characteristics have several advantages, including the fact that it is difficult to fake them and that it is difficult to lose them.

When applied to raw biometric fingerprint data, the fuzzy extractor has the capacity to smooth out any abnormalities between the user repeated biometric feature extractions. This is accomplished by combining the results of many user extractions. This is accomplished by merging the results of the extractions performed by many users. The procedure for the fuzzy extractor can be divided into two stages, which are as follows:

1. A formula that can be used to generate probability The initial biometric fingerprint, denoted by BIOi, is read in by the method Gen, which then generates the data for the biometric identification key and a public parameter by applying the formula.

   $$\mathrm{Gen}(\mathrm{BIOi})\rightarrow(\sigma\mathrm{i},\theta\mathrm{i}).$$

2. The mechanism behind deterministic reproduction in its most physical form Rep: The function is able to successfully duplicate the confidential data since it makes use of both the fingerprint BIOi and the public argument.

   $$\mathrm{Rep}(\mathrm{BIOi},\theta\mathrm{i})\rightarrow\sigma\mathrm{i}.$$

The technique that has been recommended includes a number of different phases, such as initialization, registration, user login, authentication and key agreement, and a phase in which the user can change their password.

3.1. Initialization Phase

During the phase of initialization, SA selects an additive cyclic group of order q, an elliptic curve E over a prime finite field Fp, and a point PE(Fp). This will be done in accordance with the technique that was outlined before. The hash function is selected to be utilised in the scheme, where refers to the length of the output produced by the hash function. This is the last step, although it is certainly not the least.

3.2. Registration Phase

The process of registering is divided into two distinct phases: the first phase is reserved for users, and the second phase is reserved for sensors. At this stage, all communications are sent across a private channel to ensure their safety.

3.3. Login Phase

The user is the one who kicks off the sign-in process by providing the appropriate information while simultaneously putting a smart card into a terminal. If the equation cannot be solved, this indicates that at least one of the input parameters is wrong. In this scenario, the terminal will decline the login request and will not proceed with any further authentication procedures.

3.4. Authentication and Key Agreement Phase

When both the user and the sensor in question are situated within the territory that is subject to the authority of the same CG, the activities described below will be carried out by both the user and the sensor in question:

By utilising the FG and CG, users are able to get access to sensors that are situated in various locations of the world.

3.5. Password Update

In the third step, users will be required to change their passwords. The user must insert their smart card SCi into the terminal while simultaneously entering their IDi, PWi, and BIOi credentials for 3.5 Ui to successfully validate their identification. When a user enters their biometric information into the terminal, it verifies their identify by reading their secret parameter and replicating their biometric key data. This is done to prevent unauthorised use of the user information. This is done to verify that the user is who they say they are and not someone else entirely. This request for an update will not be fulfilled if the equation does not continue to remain true. This request is going to be processed regardless of the circumstances, and after that, we are going to go on to the following step. Ui is going to go into his account and change his password while the software is being upgraded. After you have done that, the terminal will generate a completely new passcode for you to use.

4. Results and Discussions

The computer that was used to develop the code that has been released to the public is a laptop with an Intel 17 processor, 16GB RAM, 250GB of storage space, and Windows 10 Ultimate as its operating system.

Both platform as a service (PaaS) and container as a service (CaaS) are brought together on a single platform by Jelastic Cloud. Over 60 data centres, including public cloud, private cloud (virtual and on-premises), hybrid cloud, and multi-cloud data centres, are dispersed across the globe and make up Jelastic Cloud infrastructure.

Once their identities have been confirmed, owners of cloud data and users of cloud data will have access to the framework. In this case, the owner of the data has the capacity to store any number of files in any format on a server that is in a different geographical location. Once the validity of their private key has been verified, users of cloud storage will be able to access the data they require. The method is effective in avoiding the complexity as well as the longer calculation time that relates to using the cloud.

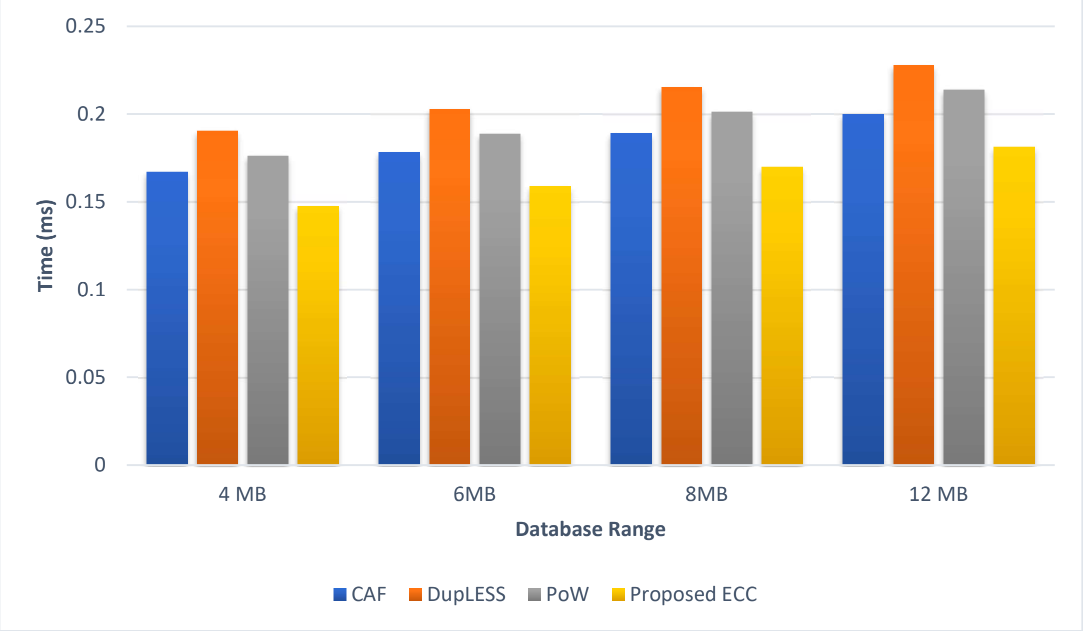


Figure 3: Communication Overhead

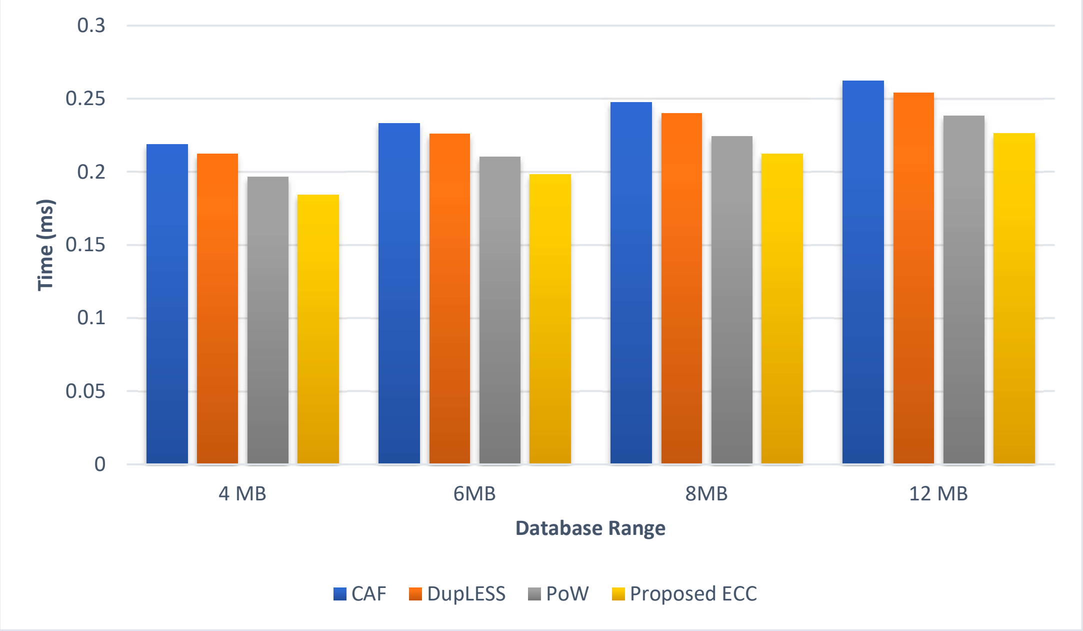


Figure 4: Computational Overhead

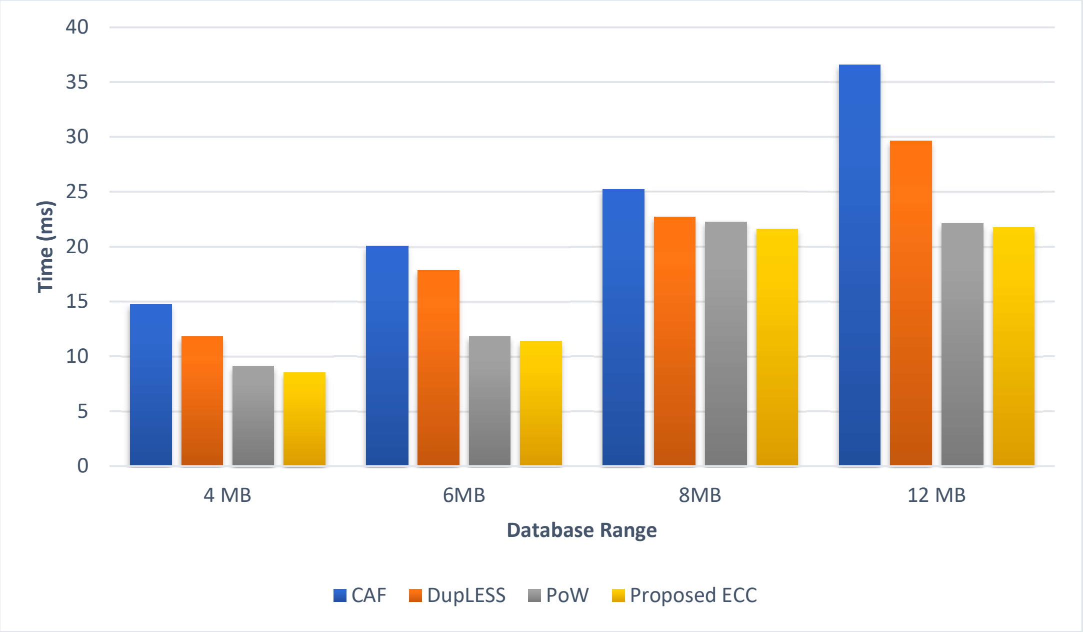


Figure 5: Encryption Time

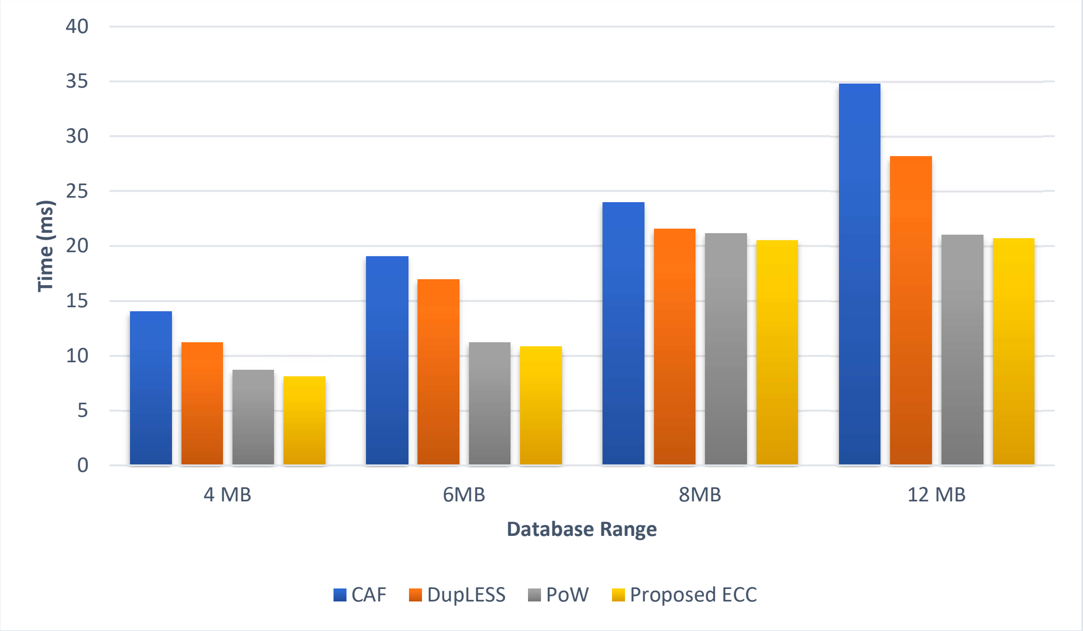


Figure 6: Decryption Time

The databases range in size from 4 - 12MB and the ECC method that was developed helps the to provide effective cloud data portability with high privacy for client applications to use in an environment where the cloud cannot be trusted.

This is achieved by lowering the requirement for storing duplicate content on the cloud server, which is made possible by lowering the requirement for storing duplicate content on the cloud server. This evaluation was carried out to determine which of the solutions would be most effective as in Figure 3 - 6.

5. Conclusions

The ECC technique offers protection against unauthorised access to data files, the ability to identify and prevent data duplication, and protection against data corruption through the validation of encrypted data with a key that is kept secret. Every user of the cloud who implements the ECC method that is advocated for the purpose of encryption has their very own unique master key. This key can only be used by that user. If they have the necessary permissions, the person who owns the data is the only one who can run duplicate file checks. Before the owner of the data can determine which files are comparable to one another, they must first resolve a token for the first copy of the data. This token is used to identify the data. When token integrity checks are present, it is guaranteed that all copies of a data file will contain the same tokens, even if they were copied exactly. According to the findings of the study, the ECC approach achieves the highest possible performance about all of the evaluation matrices and input parameters that were covered in the introduction to this section.

References

[1] Borissov, N., Haas, Q., Minder, B., Kopp-Heim, D., von Gernler, M., Janka, H., ... & Amini, P. (2022). Reducing systematic review burden using Deduklick: a novel, automated, reliable, and explainable deduplication algorithm to foster medical research. *Systematic reviews*, *11*(1), 1-10.

[2] Jiang, T., Yuan, X., Chen, Y., Cheng, K., Wang, L., Chen, X., & Ma, J. (2022). FuzzyDedup: Secure fuzzy deduplication for cloud storage. *IEEE Transactions on Dependable and Secure Computing*.

[3] Al Azad, M. W., & Mastorakis, S. (2022). The promise and challenges of computation deduplication and reuse at the network edge. *IEEE Wireless Communications*.

[4] Lin, L., Deng, Y., Zhou, Y., & Zhu, Y. (2022). InDe: An inline data deduplication approach via adaptive detection of valid container utilization. *ACM Transactions on Storage*.

[5] Azeroual, O., Jha, M., Nikiforova, A., Sha, K., Alsmirat, M., & Jha, S. (2022). A Record Linkage-Based Data Deduplication Framework with DataCleaner Extension. *Multimodal Technologies and Interaction*, *6*(4), 27.

[6] Kim, J., Ryu, S., & Park, N. (2022). Privacy-Enhanced Data Deduplication Computational Intelligence Technique for Secure Healthcare Applications. *CMC-COMPUTERS MATERIALS CONTINUA*, *70*(2), 4169-4184.

[7] Kousik, N. V., Jayasri, S., Daniel, A., & Rajakumar, P. (2019). A survey on various load balancing algorithm to improve the task scheduling in cloud computing environment. J Adv Res Dyn Control Syst, 11(08), 2397-2406.

[8] Zhou, W., Wang, H., Mohiuddin, G., Chen, D., & Ren, Y. (2022). Consensus Mechanism of Blockchain Based on PoR with Data Deduplication. *Intelligent Automation & Soft Computing*, *34*(3).

[9] Sangeetha, S. B., Sabitha, R., Dhiyanesh, B., Kiruthiga, G., & Raja, R. A. (2022). Resource management framework using deep neural networks in multi-cloud environment. In Operationalizing Multi-Cloud Environments (pp. 89-104). Springer, Cham.

[10] Wang, Y., Narasayya, V., He, Y., & Chaudhuri, S. (2022). PACk: an efficient partition-based distributed agglomerative hierarchical clustering algorithm for deduplication. *Proceedings of the VLDB Endowment*, *15*(6), 1132-1145.

[11] Natarajan, Y., Kannan, S., & Dhiman, G. (2022). Task scheduling in cloud using aco. Recent Advances in Computer Science and Communications (Formerly: Recent Patents on Computer Science), 15(3), 348-353.

[12] Sohail, A., & Qounain, W. U. (2022). Locality sensitive blocking (LSB): A robust blocking technique for data deduplication. *Journal of Information Science*, 01655515221121963.

[13] Raja, R. A., Karthikeyan, T., & Kousik, N. V. (2020). Improved privacy preservation framework for cloud-based internet of things. In Internet of Things (pp. 165-174). CRC Press.

[14] Douceur, J. R., Adya, A., Bolosky, W. J., Simon, P., & Theimer, M. (2002, July). Reclaiming space from duplicate files in a serverless distributed file system. In *Proceedings 22nd international conference on distributed computing systems* (pp. 617-624). IEEE.

[15] Bellare, M., & Keelveedhi, S. (2015, March). Interactive message-locked encryption and secure deduplication. In *IACR international workshop on public key cryptography* (pp. 516-538). Springer, Berlin, Heidelberg.

[16] Bellare, M., Keelveedhi, S., & Ristenpart, T. (2013). DupLESS: Server-aided encryption for deduplicated storage. *Cryptology ePrint Archive*.

[17] Harnik, D., Pinkas, B., & Shulman-Peleg, A. (2010). Side channels in cloud services: Deduplication in cloud storage. *IEEE Security & Privacy*, *8*(6), 40-47.

[18] Li, S., Xu, C., & Zhang, Y. (2019). CSED: Client-side encrypted deduplication scheme based on proofs of ownership for cloud storage. *Journal of Information Security and Applications*, *46*, 250-258.