Deduplication Storage optimization for data management

Benita Musabiyinema, Wang jia yang, Gilbert Langat

Abstract- With increased amount of data, nowadays storage has become a challenge. New technologies of storage like cloud computing whereby data is stored in a cloud is now in place. But still not accurate and does not fully solve the issue of storage. There other challenges like duplication, security of the data and the rest.in our paper therefore; we have proposed a noble deduplication system, which will help in dealing with storage space and enhancing the security of the data especially in the cloud storage.

Keywords: Cloud computing, AES encryption, RSS key, data deduplication and cloud security**.**

1 Introduction

Cloud computing is based on the concept of internet in which large groups of remote servers are networked to allow the centralized data storage, and online access to computer services or resources. Cloud service offers typically three categories: Infrastructure as a Service (IaaS), Platform as a Service (PaaS), and Software as a Service (SaaS). Cloud computing poses great possibilities and challenges. The first and very important challenge to be addressed is the data security. Cloud computing is using the public networks to store and transmit data which makes the data in the cloud exposed to public and vulnerable. This nature of the cloud computing, leads to security and privacy concerns. Session hijacking and data segregation are the potential and unavoidable security breaches for the cloud users. Cloud is associated with two kinds of resources namely physical and virtual resources. Both virtual and physical resources pose multiple security issues at different levels.

2 RELATED WORKS & CONTRIBUTION

A new cryptographic method is formalized in which the key that is used for both encryption and decryption is derived from the data itself. This cryptographic method is called Message-Locked Encryption (MLE) [1]. Another new deduplication system called RevDedup, is a deduplication method which follows the concept of reverse deduplication. This method eliminates the duplicate data from the new data rather than removing data from the existing data as in the conventional deduplication methods [2]. The cloud security issues such as virtualization security and data security are the key issues restricting the application of cloud computing. These security issues have become an obstacle restricting cloud computing industrial applications [3]. There are few practical methods for securing cloud computing. These practical methods are as follows: private cloud that are configured with enterprise perimeters, public cloud that possess service gateways, provision for content encryption, cloud access broker, session containers, and security virtualization in runtime [4]. AES is the popular and standard data encryption method. In 2001, American National Institute of Standards and Technology published a novel encryption standard called AES (Advanced Encryption Standard). AES is based on a symmetric- key algorithm. In symmetric key algorithms same key is used for both encryption and decryption of the data. The block size is often 128 bits and the key size can be of different sizes such as 128, 192, or 256 bits. AES operates on a state, which is 4×4 matrix of bytes [5].

- *Benita martine Musabiyinema is currently pursuing master's degree program in Computer Science , Central South University China*
-
- *Wang jia yang is a Professor of computer science Central South university China*
- *Gilbert Langat is currently pursuing PhD degree program in Computer Science, Central South University China*

There is another cloud security system, which is based on the two-way password concept. In this system once the user gets authentication two passwords will be generated. One password is given to the user and the other is given to the cloud provider. The user can only access the cloud if and only both the passwords were given as input. The main disadvantage of this method is that the user has to trust the cloud provider for no reason and this system is complex [6].

In a recently published paper, a novel cloud security in which, the data storage and retrieval is performed using the privilege key. The data is first send to the private cloud for storage through privilege key. The data is next send to the hybrid cloud for storage after encryption. The data is encrypted using convergent encryption technique. One important issue with the service platforms for cloud storage is managing the ever-increasing huge volume of data. Deduplication is the promising technique in order to make the data management in cloud computing into a scalable task. This method has captivated more attention recently. Data deduplication is a concept of data compression, which is made possible by finding and eliminating the redundant copies of the data from the cloud storage. This technique is used to enhance the efficiency of storage utilization and can also applied to data transfers through network to reduce the number of bytes that are transferred. Instead of keeping duplicate data copies, deduplication removes redundant data and store only one physical copy and referring other redundant data by an alert message. Cloud backup service is an affordable choice for data protection for personal computing devices. The cloud computing offers centralized cloud management, which created efficiency and cost effectiveness. This offers simple backup storage for disaster recovery, which is always a major consideration for data backup. The IT resources in the cloud have so much redundancy in backup dataset. Hence there is lot of scope for improving the efficiency of its cloud storage [7].

In the existing systems data deduplication is performed through hybrid cloud architecture. The hybrid cloud architecture does not eliminate the redundant data completely. The convergent encryption technique used for encrypting the data is inefficient. The user uses same privilege key for the storage and retrieval of data. The privilege key is easily predictable by the hackers or intruders.

3 PROPOSED SYSTEM

The proposed system has two tasks namely storage of data in cloud and data retrieval from cloud. For storing the data in the cloud the user first log in to the private cloud using the IP address, username and password. The private cloud validates the identification of the user and allows the user to access the cloud. The data to be uploaded undergoes deduplication for the removal of redundant data files. The latency and effectiveness of deduplication can be balanced by the combination of both global and local deduplication. This is followed by the key generation where, the user request for the key from the private cloud and the private cloud generate a RSS key for the files and send it to the user. Encryption of the data is very important to avoid data piracy. This system uses convergent encryption method for encrypting the data before uploading it to the cloud. This results in the deduplicated and encrypted data, which will be uploaded to the public cloud. For the data retrieval the user needs to gain access to the private cloud by validating the user's IP address, username and password, which is same as that of the storing the data as mentioned above. Once the user is validated by the private cloud, the user can access the private cloud and get the RSS keys for the data that are uploaded to the cloud. By using this RSS key the user can access the public cloud, where the deduplicated and encrypted files are stored the other users. By accessing the public cloud the user can retrieve the data they wanted in the encrypted form. The user can retrieve the encrypted file from the server. The user who owns the file can only decrypt the retrieved file using the convergent encryption protocol. Thus, convergent encryption provides two features as follows: 1. It performs de-duplication on the cipher texts in the cloud and 2. The ownership prevents the hackers or unauthorized user to access the file. Each file uploaded to the cloud is bounded by a set of privileges such as they specify the kind of users who are allowed to and access the files and check for the data duplication.

4 SYSTEM ARCHITECTURE

System architecture provides description of the proposed system, which is organized in a way to support logical reasoning of proposed system's the structural properties. System architecture enlist the building blocks of the proposed system that will work in unison for the successful implementation of the overall proposed system.

This study depicts a system architecture shown in figure1. The components of the proposed architecture are:

- Private cloud
- Data load
- Key generation
- Encrypted data
- Public cloud
- Data retrieval
- Decrypted data

Private cloud:

The cloud-computing environment that is operated with private firewall is known as private cloud. In our system the private cloud is used for the authorization of the user. As the authorization of the user is performed in the private cloud, this ensures the maximum privacy and prevention from hacking. Once the user type in his/her credentials the private cloud verifies these user credentials and allow the user to access the database provided the user is a registered user. After passing the authorization process the user can upload or retrieve the required data.

Data load:

If the user wants to upload the data to the database the files are first subjected to the deduplication procedure. The deduplication algorithm checks for the redundancy of the files in the database. If there is redundancy the system alerts the user so that the uploading of the redundant file is avoided. This process is very important as it ensures the efficient use of the cloud space.

Key generation:

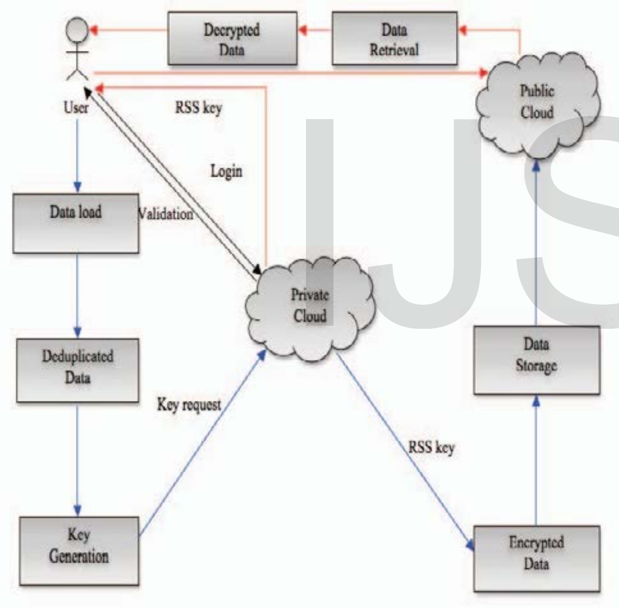
Key generation is the important aspect of our system. We employs Haval algorithm for the generation of keys. During deduplication if the system finds that there is no redundancy the file will be successfully uploaded and generates a RSS key that is attached with the file. This RSS key is random and unique to the uploaded file. This RSS key serves as the necessary input for the decryption of the file that is attached with the RSS key. Hence this key algorithm ensures security and prevents data theft.

Data encryption:

Encryption is the process of converting the normal text into the cipher text. Our system employs AES algorithm as the encryption protocol. The uploaded files are first undergoes the process of encryption before stored in the database. The process of encryption serves as the last line of defense against the data piracy. The encrypted files cannot be decrypted without the RSS key.

***Public cloud*:**

The cloud-computing platform that is available for the general public without any firewall is called the public cloud. It is cheaper than the private cloud. Our system utilizes the public cloud for the data storage, as it is cheaper. The uploaded files are stored in the database located in the public cloud after the encryption.



Data retrieval:

If a registered user wants to retrieve a file from the database in the public cloud, he needs get authorized by the private cloud and should know the RSS key for the files he wants to retrieve. Once the user gets the required authorization he/she can able to access the database in the public cloud and retrieve the target file by downloading the file.

Data decryption:

The process of converting the cipher text into the readable plain text is called data decryption. In our system we employs AES algorithm for the data decryption. Data decryption is the last step for data retrieval. RSS key is required for the data decryption. This ensures utmost security to the valuable data that is stored in the database.

Figure 1- Architecture diagram for the proposed system. Red arrow represents the data retrieval pathway, blue arrow represents the data storage pathway and the black arrows represent the login pathway, which is common for both data retrieval and data storage.

5 ADVANTAGES OF THE PROPOSED SYSTEM

- The data deduplication is performed using ALG deduplication.
- The technique used for deduplication is more effective compared with the existing technique.
- Data is first stored in the system then it is stored in the cloud, so that if data loss occurs the data is recovered.
- The data is encrypted using AES algorithm rather than convergent encryption technique.
- The data storage and retrieval is performed Ramp Secure Secret key rather than using privilege key.
- The Ramp Secure Secret key is generated dynamically at the time of data storage and retrieval to the user.

6 ADVANCED ENCRYPTION STANDARD (AES) ALGORITHM:

Advanced Encryption System (AES) algorithm is characterized as symmetric block cipher that encrypts 128 bits of the text data blocks into cipher keys of different bit lengths namely 128,192 or 256 bits. This encryption method utilizes the Rijndael algorithm, which was developed by Belgium scientists namely Vincent Rijimen and Joan Daemen. The main advantages of AES algorithm are it is very simple design, high speed in processing without compromising the memory. Another very important advantage is that the both encryption and decryption and performed using the same RSS key.

AES encryption uses the round function. The round function includes of four transformations as follows:

- Byte-sub operation is the non-linear substitution of each byte using the defined substitution table.
- Shift-row operation shifts the rows of the state in a circular fashion.

Figure 2-Advanced Encryption Standard Algorithm – Encryption.

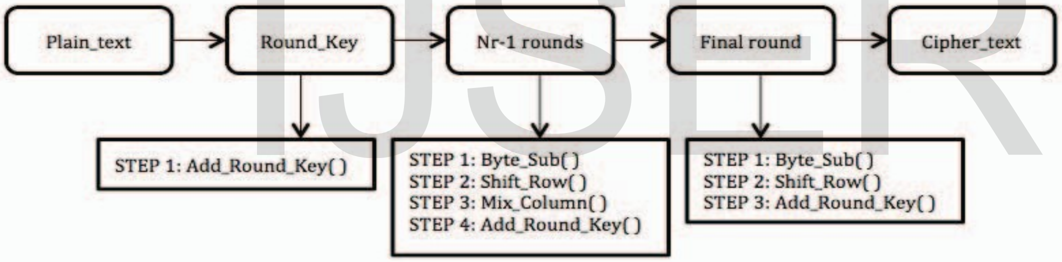
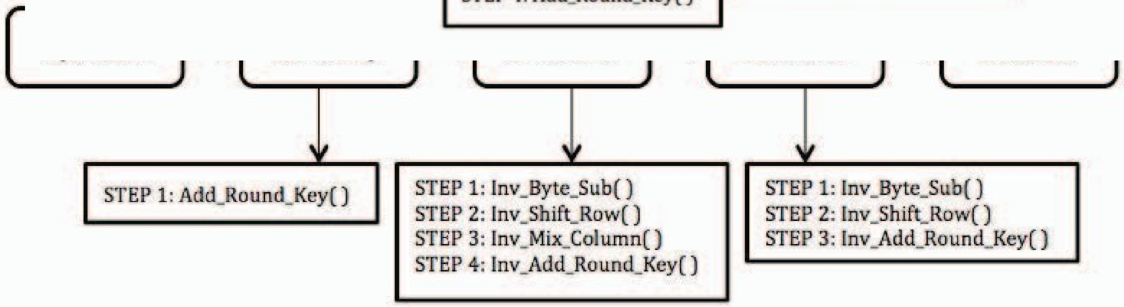


Figure 3-Advanced Encryption Standard Algorithm – Decryption.



- Mix-column operation uses the multiplication with fixed polynomial modulo in order to mix the bytes.
- Add round key operation adds round key to the state [5].

HAVAL algorithm:

Haval algorithm is a one way-hashing algorithm meaning any arbitrary long text into compressed value of specific length. Haval algorithm can compress the given information and synthesis a value of 128, 160, 192, 224 or 256 bits. This algorithm is deterministic, very efficient and faster than other hashing algorithms [8].

7 COMPARISON GRAPH

This comparison graph depicts the encryption speed of the AES encryption and convergent encryption. With the increase in the encryption rate, time time taken for encryption is quiker in AES encryption procedure when compared to the convergent encryption procedure. Its is evident from the graph that the the AES encryption method is much faster than the convergent encryption. Hence it is obvious that AES algorithm saves time and increases efficinency.

8 CONCLUSION

This proposed system uses ALG data deduplication technique, which avoids data redundancy saving space ultimately increasing the efficiency of the data storage. This system also uses the standard AES algorithm for data encryption thus adding data security. Finally this also uses the RSS key method, which is generated dynamically in a unique way, which is far secure than the privilege keys.

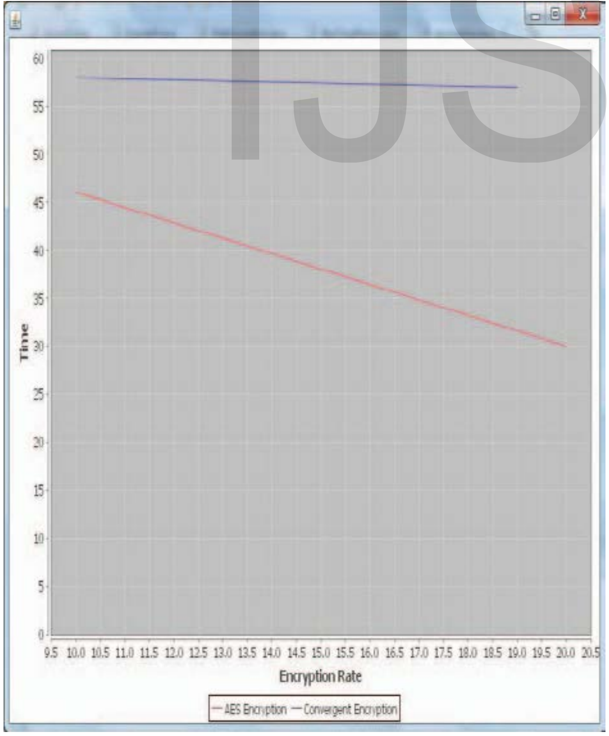


Figure 3- comparison graph between AES algorithm and convergent algorithm.

9 FUTURE ENHANCEMENT

The problem with the secret keys is when the number of files stored in the cloud increases the number of secret key also increases. Thus the user has to remember many secret keys which is cumbersome. The concept of key-aggregate cryptosystem is introduced recently. In this method many secret keys could be combined to form a single secret key. While sharing the data, the user can generate a secret key for the sharing file. So that only the file that need to share can be accessed and the rest of the files remain secured [9]. In our future enhancement we like to incorporate modified version this concept in our system in order make it more user friendly. Our modification will include implementing login based secure aggregate key and individual key generation portal for the creation of the aggregate keys and individual keys from the aggregate key. The current encryption algorithm encrypts 128 bits in a given time; we intend to develop an encryption algorithm in future that could encrypt more bits in a given time. Thus enhancing the overall speed of data encryption.

REFERENCES

[1] Bellare.M, Keelveedhi.S, Ristenpart.T (2013), 'Message-locked encryption and secure deduplication', *In EUROCRYPT*, pp 296– 312 .

[2] Ng, C., & Lee, P. P. C. (2013). RevDedup : A ReverseDeduplication Storage System Optimized for Reads to Latest Backups. *APSys*. doi:10.1145/2500727.2500731

[3] Luo, X., Yang, L., Hao, D., Liu, F., & Wang, D. (2014). On Data and Virtualization Security Risks and Solutions of Cloud Computing. Journal of Networks, 9(3), 571–581. doi:10.4304/jnw.9.3.571- 581

[4] Amoroso, E. G. (2014). Practical methods for securing the cloud. IEEE Cloud Computing, 1(1), 28–38. doi:10.1109/MCC.2014.1

[5] Ankita N., & Sheeja S. (2014a). Design and Implementation of AES Algorithm Using, 2(1), 419– 423.

[6] Brar, Y., Krishan, S., Mehta, A., & Talwar, V. (2014). An Advanced Security - A Two-Way Password Technique for Cloud Services, 3(4), 7–15.

[7] Li, J., Li, Y. K., Chen, X., Lee, P., & Lou, W. (2014). A Hybrid Cloud Approach for Secure Authorized

Deduplication. IEEE Transactions on Parallel and DistributedSystems,1–1.doi:10.1109/ TPDS.2014. 2318320.

[8] Zheng, Y., Pieprzyk, J., & Seberry, J. (1992). HAVAL - A One-Way Hashing Algorithm with Variable Length of Output. Advances in Cryptology - AUSCRYPT '92, 718, 83–104. doi:10.1007/3-540- 57220-1_54

[9] Chu, C., Chow, S. S. M., Tzeng, W., Zhou, J., Deng, R. H., & Member, S. (n.d.). Supplementary Material for Key-Aggregate Cryptosystem for Scalable Data Sharing in Cloud Storage, 1–4.