AN APPROACH AND IMPLEMENTATION OF DEDUPLICATION IN ENCRYPTED BIG DATA FOR CLOUD

Mohd Akbar
Research Scholar, Shri
JJT University, Rajasthan,
akb.mtech@gmail.com

Dr. K. E. Balachandrudu,
Associate professor,
Arjun College of
Technology & Sciences,
Hyderabad

Dr. Prasadu Peddi,
Assistant Professor, Shri
JJT University, Rajasthan

ABSTRACT:

Cloud computing enables companies to consume a compute resource, such as a virtual machine (VM), storage or an application, as a utility just like electricity rather than having to build and maintain computing infrastructures in house. In cloud computing, the most important part is data centre, where client's/user's data is stored. In data centres, the data might be uploaded multiple time or data can be hacked so, while using the cloud services the data need to be encrypted and stored. With the continuous and exponential increase of the number of users and the size of their data, data deduplication becomes more and more a necessity for cloud storage providers. By storing a unique copy of duplicate data, cloud providers greatly reduce their storage and data transfer costs. Because of the authorized data holders who obtain the symmetric the encrypted data can also be securely accessed. Keys used for decryption of data. The results show the superior efficiency and effectiveness of the scheme for big data deduplication in cloud storage. Evaluate its performance based on extensive analysis and computer simulations with the help of logs captured at the time of deduplication.

Keywords: *big data, cloud computing, data deduplication, proxy re-encryption.*

INTRODUCTION:

Cloud computing offers a new way of Information Technology services by rearranging various resources (e.g., storage, computing) and providing them to users based on their demands. Cloud users upload personal or confidential content data to a cloud service provider (CSP) data centre and allow to keep this data. With the potentially infinite storage space offered by cloud providers, users tend to use as much space as they can and vendors constantly look for techniques aimed to minimize redundant data and maximize space savings. A technique, which has been widely adopted, is cross-user deduplication. Deduplication has proved to achieve high space and cost savings and many cloud storage providers are currently adopting it. Deduplication can reduce storage needs by up to 90-95 percent for backup applications and up to 68 percent in standard file systems. While the aim of deduplication is to detect identical data segments and store them only once, the result of encryption is to make two identical data segments indistinguishable after being encrypted. A technique, which has been proposed to meet these two conflicting requirements, is convergent encryption whereby the encryption key is usually the result of the hash of the data segment.

OBJECTIVES:

1. To save cloud storage and preserve the privacy of data holders by proposing a scheme to manage encrypted data storage with deduplication.
2. To prove the security and performance of the scheme through analysis and simulation.

3. A Study on Data Auditing and Security in Cloud Computing

LITERATUE REVIEW:

Shobana, R et al (2016) Proposes Cloud Computing Secure Framework (CCSF). Thus CCSF consists of four segments: 1) Identity Management 2) Intrusion detection and prevention system 3) Data deduplication 4) Secure Cloud Storage. Intrusion detection and prevention are performed manually by network operators in the existing system. In our proposed architecture the intrusion detection and prevention is performed automatically by defining rules for the major attacks and alert the system automatically. To ensure data confidentiality the data are stored in an encrypted type using Advanced Encryption Standard (AES) algorithm.

Wu, T.Y (2015) proposed Index Name Servers (INS) to manage not only file storage, data deduplication, optimized node selection, and server load balancing, but also file compression, chunk matching, real-time feedback control, IP information, and busy level index monitoring .The main advantage of this technique is that Index Name Servers algorithm help to reduce workloads of resources and improve the performance of system. INS also handles server load balancing. – The main disadvantage of this technique is that encrypted data cannot be deduplicated.

Shweta D. Pochhi (2014) presented that the data and the Private cloud where the token generation will be performed for each file. Before uploading the data or file to public cloud, the client will send the file to private cloud for token generation, which is unique for each file. Private clouds then generate a hash and a token and send the token to client. A system, which achieves confidentiality and enables block-level de-duplication at the same time. Before uploading the data or file to public cloud, the client will send the file to private cloud for token generation, which is unique for each file.

Puzio. P et al (2013) have proposed ClouDedup security system to provide secure and efficient storage service, which assures block-level deduplication and data confidentiality at the same time. The security of ClouDedup relies on its new architecture with metadata manager and an additional server. The server adds an additional encryption layer to prevent well-known attacks against convergent encryption and thus protect the confidentiality of the data; on the other hand, the metadata manager is responsible of the key management

SYSTEM ARCHITECTURE:

In recent time, there are many problems of storage places in cloud. If data holder store file in cloud which is already available in cloud. The security of ClouDedup relies on its new architecture whereby in addition to the basic storage provider, a metadata manager and an additional server are defined: the server adds an additional encryption layer to prevent well-known attacks against convergent encryption and thus protect the confidentiality of the data; on the other hand, the metadata manager is responsible of the key management task since block-level deduplication requires the memorization of a huge number of keys. Therefore, the underlying deduplication is performed at block-level and we define an efficient key management mechanism to avoid users to store one key per block.

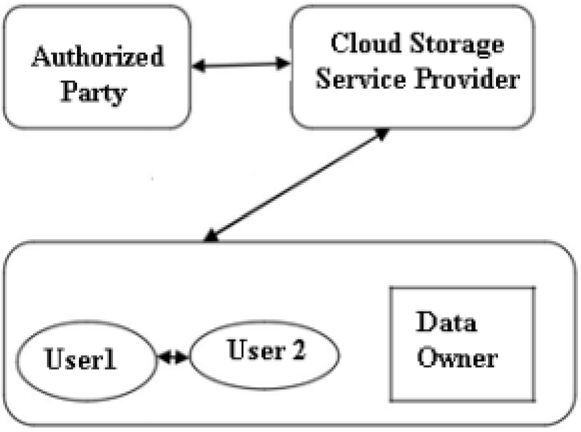


Figure 1 - System Architecture

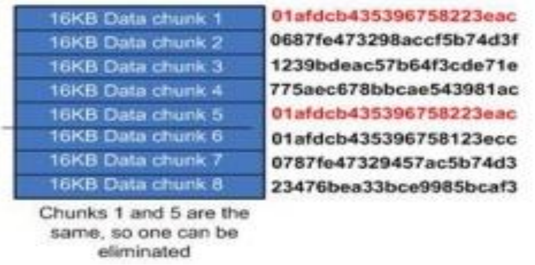
Data Holder: The data holder can uploads and saves their data and files in the CSP. In this system is possible to number of data holders could save their files in encrypted raw data in the CSP. The data holder that produces or creates the File regards the file as data owner. The data holder is in normal form than the higher priority of owner.

Cloud Service Provider: When the data holder deletes data from CSP, CSP firstly manages the records of duplicated data holders by removing the duplication record of this user. If the rest records are not empty, the CSP will not delete the stored encrypted data, but block data access from the holder that requests data deletion. If the rest records are empty, the encrypted data should be removed at CSP.

Encrypted Data Update: In case a data owner with DEK 0 updates that DEK and the new encrypted raw data is provided to CSP to replace old storage for the reason of achieving better security, CSP issues the new re-encrypted DEK 0 to all data holders with the support of AP.

Data Owner Management: In case that a real data owner uploads the data later than the data holder, the CSP can manage to save the data encrypted by the real data owner at the cloud with the owner generated DEK and later on, AP supports re-encryption of DEK at CSP for eligible data holders.

DATA DEDUPLICATION: Data deduplication or Single Instancing essentially refers to the elimination of redundant data. As the amount of digital information is increasing exponentially, there is a need to deploy storage systems that can handle and manage this information efficiently. Data deduplication is one of the emerging techniques that can be used to optimize the use of existing storage space to store a large amount of data. Data deduplication is removal of redundant data. Thus, reducing the amount of data reduces many costs storage requirements costs, infrastructure management cost.



16KB Data chunk 1	01afdcb435396758223eac
16KB Data chunk 2	0687fe473298accf5b74d3f
16KB Data chunk 3	1239bdeac57b64f3cde71e
16KB Data chunk 4	775aec678bbcae543981ac
16KB Data chunk 5	01afdcb435396758223eac
16KB Data chunk 6	01afdcb435396758123ecc
16KB Data chunk 7	0787fe47329457ac5b74d3
16KB Data chunk 8	23476bea33bce9985bcaf3

Chunks 1 and 5 are the same, so one can be eliminated

Figure 2 - Deduplication in the cloud

IMPLEMENTATION:

For implementation, we preferred ASP.NET C# language, Visual studio framework and Windows O.S. Platform as it provides inbuilt server called IIS. ASP.NET provides inbuilt MSDN managed code to support cryptographic hashing algorithm needed to perform encryption and decryption. Data deduplication is referred to as a strategy offered to cloud storage providers (CSPs) to eliminate the duplicate data and keep only a single unique copy of it for storage space saving purpose. Data deduplication is one of the techniques, which used to solve the repetition of data. The deduplication techniques are generally used in the cloud server for reducing the space of the server. Cloud Storage usually contains business-critical data and processes; hence, high security is the only solution to retain strong trust relationship between the cloud users and cloud service providers.

In this methodology we have to detect the duplicate copy of the file any type of file can be detect file .txt,.doc,.xls, ppt, .pdf. Therefore, we have to start with uploading the file when we upload the file we have to extract first 50 bytes from the file and last 50 bytes from the file

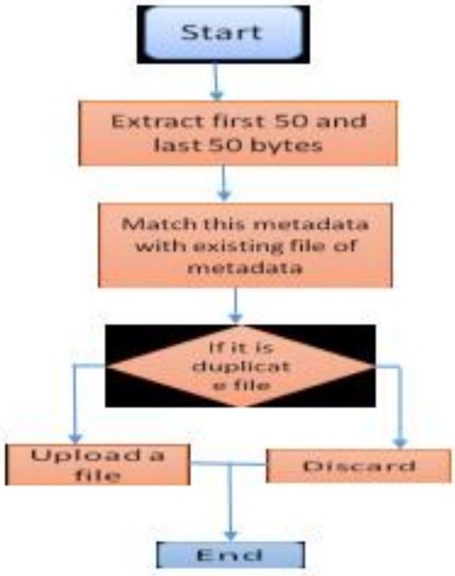

Start
Extract first 50 and last 50 bytes
Match this metadata with existing file of metadata
If it is duplicat e file
Upload a file
Discard
End

Figure 3 - Flow Chart

RESULTS:

Efficiency of data encryption and decryption. In this experiment, we tested the operation time of data encryption and decryption with AES by applying different AES key sizes (128 bits, 196 bits and 256 bits) and different data size (from 10 megabytes to 600 megabytes).we observed that even when the data is as big as 600 MB, the encryption/decryption time is less than 13 seconds if applying 256-bit AES key. Applying symmetric encryption for data protection is a reasonable and practical choice. The time spent on AES encryption and decryption is increased with the size of data. This is inevitable in any encryption schemes. Since AES is very efficient on data encryption and decryption, thus it is practical to be applied for big data.

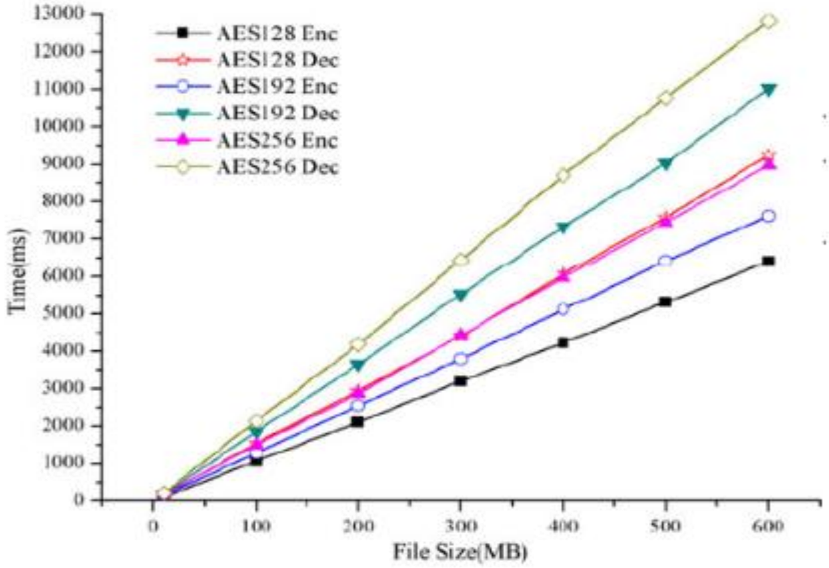


Figure 4 - Operation time of file encryption and decryption with AES

Data Ownership Challenge:

In this experiment, we selected 192-bit field of elliptic curve (160-bit ECC has a security level comparable to 1024-bit RSA), 256-bit AES, 1024-bit PRE and 10M uploaded data. We can observe that data upload is the most time-consuming if the file is big, but it is inevitable in all schemes. Therefore, our scheme can save a lot of computation load and communication cost for cloud users. In addition, the data ownership challenge in the proposed scheme is very lightweight, which does not involve much burden to cloud users.

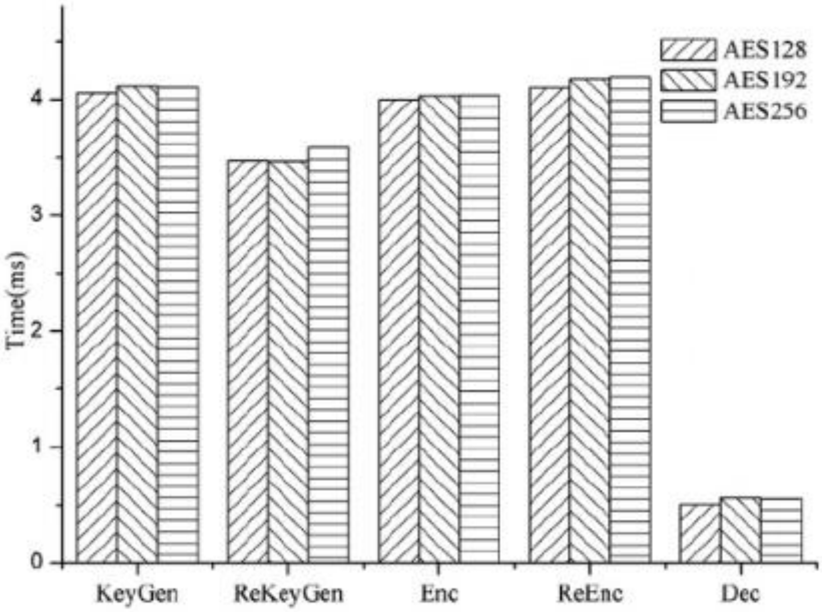


Figure 5 - The execution time of PRE operations.

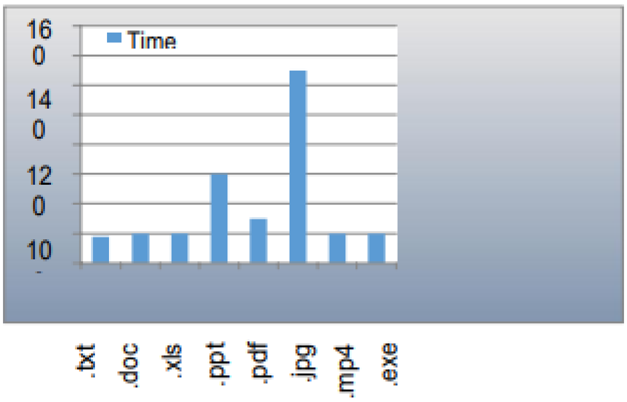


Figure 6 - Deduplication time factor at Byte level

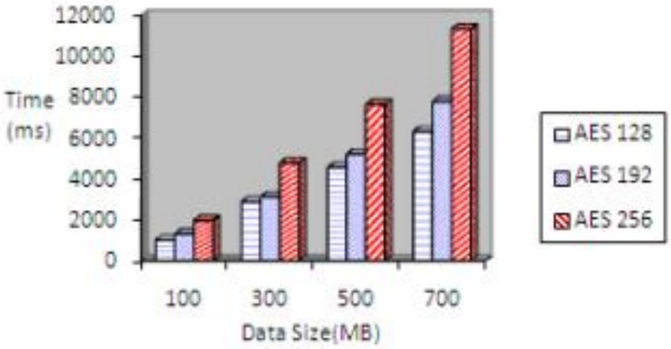


Figure 7 - Data encryption and decryption efficiency (Period Analysis)

In this experiment, the time taken by various encryptions standard of AES is demonstrated. It was found that greater the bit size the encryption takes even more time to encrypt file. Achieving this amount of uniqueness and speed made AES our first choice majorly for getting lower encryption - decryption time as compared to others. This will highly reduce the amount of time spent in the entire process when huge data files are to be considered.

CONCLUSION:

We have proposed a new system called deduplication, which saves storage space of cloud enabling to reduce the data duplication by saving only single copy of data for multiple users and provides security mechanism**.** Managing encrypted data with deduplication is important and significant in practice for achieving a successful cloud storage service, especially for big data storage. Our scheme can flexibly support data update and sharing with deduplication even when the data holders are offline. Encrypted data can be securely accessed because only authorized data holders can obtain the symmetric keys used for data decryption. To secure the confidentiality of sensitive data during deduplication, the convergent encryption technique is used to encrypt the data before outsourcing. The results of our computer simulations further showed the practicability of our scheme. Future work includes optimizing our design and implementation for practical deployment and studying veritable computation to ensure that CSP behaves as expected in deduplication management.

REFERENCES:

1. Shobana, R., K. ShanthaShalini, S. Leelavathy V. Sridevi (2016), “De-Duplication Of Data In Cloud”, Int. J. Chem. Sci., ISSN: 0972-768X, Volume: 14, Issue: 4, PP: 2933-2938
2. Wu, T. Y (2015) J. S. Pan, and C. F. Lin, Improving accessing efficiency of cloud storage using deduplication and feedback schemes, IEEE Syst. J., Volume: 8, Issue: 3, PP: 1-10
3. Shweta D. Pochhi, Prof.Pradnya V. Kasture (2014), “Encrypted Data Storage with De-duplication Approach on Twin Cloud”, International Journal of Innovative Research in Computer and Communication Engineering, Volume: 3, Issue: 2, PP: 22-31
4. Puzio.P., R. Molva, M. Onen, and S. Loureiro (2013), “ClouDedup: Secure deduplication with encrypted data for cloud storage,” in Proc. IEEE Int. Conf. Cloud Comput. Technol. Sci., Volume: 5, Issue: 3, PP: 363–370.
5. Deepak Mishra, Dr. Sanjeev Sharma, “Comprehensive study of data de-duplication”, International Conference on Cloud, Big Data and Trust, Nov 2013.
6. Zhang, D., Liao, C., Yan, W., Tao, R., & Zheng, W. (2017, August). Data Deduplication Based on Hadoop. In Advanced Cloud and Big Data (CBD), 2017 Fifth International Conference, PP: 147-152.
7. Ajay Jangra, Vandna Bhatia, UpasanaLakhinazand, NiharikaSinghx (2015), “An Efficient Storage Framework design for Cloud Computing: Deploying Compression on De-duplicated No-SQL DB using HDFS” 2015 1st International Conference on Next Generation Computing Technologies, Dehradun, India, 4-5 September 2015
8. Wang, C., Z. Qin, J. Peng, and J. Wang (2010), “A novel encryption scheme for data deduplication system,” Proc. International Conference on Communications, Circuits and Systems (ICCCAS), PP: 265–269.
9. YAN, Zheng; DING, Wenxiu; YU, Xixun; ZHU, Haiqi; and DENG, Robert H (2016), “Deduplication on encrypted big data in cloud”, IEEE Transactions on Big Data, Volume: 2, Issue: 2, PP: 138-150.
10. Liu, C. Y., X. J. Liu, L. Wan (2013), “Policy-based deduplication insecure cloud storage,” in Proc. Trustworthy Comput. Serv., Volume: 8, Issue: 3, PP: 250–262

AUTHOR DETAILS:

My name is Mohd. Akbar. I am a Research Scholar from Shri Jagadishprasad Jhabharmal Tibrewala University, Rajasthan. I have completed my M. Tech in Computer Science and Engineering from Jawaharlal Nehru Technological University, Hyderabad. I am having more than 17 years of experience (including overseas) in Teaching field. My research area of interest is Big Data, Machine Learning, and Cloud Computing. Other areas of interest are Computer Networks, Artificial Intelligence. I taught several subjects such as Databases, Programming Languages, Operating Systems, Computer Networks, etc.

Dr. K. E. Balachandrudu, Associate Professor, Arjun College of Technology and Sciences, Hyderabad.

Dr. Prasadu Peddi, Assistant Professor, Shri Jagadishprasad Jhabharmal Tibrewala University, Rajasthan.