IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 10, Issue 6, June 2022

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.165

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 6, June 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1006013 |

Secure Data Deduplication on Cloud Storage

Anjali Gedam, Akshata Nagrale, Rani Tiple, Sakshi Deshmukh, Manish Nimbalkar

UG Student, Dept. of IT, DMIETR Wardha, RTMNU University, India

ABSTRACT: In today's digitally evolving world,the very thing that is of utmost importance is the security of data. Cloud Computing has emerged as a popular and effective tool to manage data for administrations. Each day, around 2.5 quintillion bytes of data are generated on the internet, and to store this large amount of data, we need servers that can deduplicate data efficiently to avoid wastage of storage thus minimizing expenses. The results depict that redundant data is always mapped onto the same hash code and thus it does not get uploaded to the cloud servers thus ensuring successful deduplication. It saves storage as well as saves bandwidth by eliminating duplicate data. In this project, data gets stored in the cloud server named drive and numerous efforts have been taken to ensure complete data access. With effective deduplication, ensuring data confidentiality is also very important thus data is always stored in the cloud in an encrypted format. It is achieved with the help of the Advanced Encryption Standard algorithm.

KEYWORDS: Cloud Computing, deduplication, data confidentiality, data access, Advanced Encryption Standard.

I.INTRODUCTION

As we move towards a more technological-driven era, saving data in cloud servers is the need of the hour.A huge amount of data gets uploaded onto the internet every day, and the preservation and security of this data are becoming more and more challenging with every second. Preservation of data is very important and in this business era, it is even mandated by the law.To secure such a huge amount of data, Cloud Computing is the most effective tool in our hands. Cloud Computing is a practice of using a network of remote servers which are hosted on the internet to store, manage and process data, rather than store on a local server.

Every cloud has limited storage and if we start uploading redundant files to the cloud, the storage is at a loss and data redundancy will be a big problem to tackle. To counter this, researchers have been exploring various methods and the best solution is the deduplication of data. Data deduplication is a technique evolved to optimize storage. This technique today is used by various cloud service providers such as Dropbox, Amazon S3, and Google Drive.It ensures that duplicate data is never uploaded to the cloud more than once.

Big administrations and organizations usually buy a third-party cloud for the storage of client data. But giving valuable information in the hands of a third party is like an invitation to risk. Researchers have been exploring this issue and the best solution is to guard the outsourced data with the cipher text. So, once the data is uploaded to a cloud server, itis in an encrypted format. When the data is downloaded, it is decrypted and is then visible to the client. In encryption strategies, data is converted into another form called cipher text but if encryption is done with different keys, it may result in different cipher text making deduplication less feasible. Thus, encryption is necessary to secure data. So, deduplication and encryption must work in coordination to ensure data security.

II. BACKGROUND

2.1 Deduplication

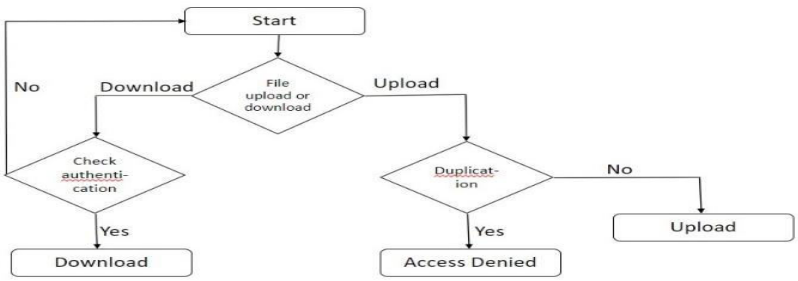

Figure. Deduplication Flowchart

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 6, June 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1006013 |

Data Deduplication is a technique for eliminating duplicate copies of repeating data. It is also called Single Instance Storage. Deduplication can be categorized into two types: file-level deduplication and block-level deduplication. File-level deduplication takes into account full file while block-level deduplication applies deduplication on blocks of data with the help of hashing algorithms.

2.2 Advanced Encryption Standard

This is an encryption algorithm that works by taking plain text and converting it into cipher text which is made up of random characters. AES uses symmetric encryption which uses only one key to cipher and deciphers data. It is based on a 'substitution–permutation network'.

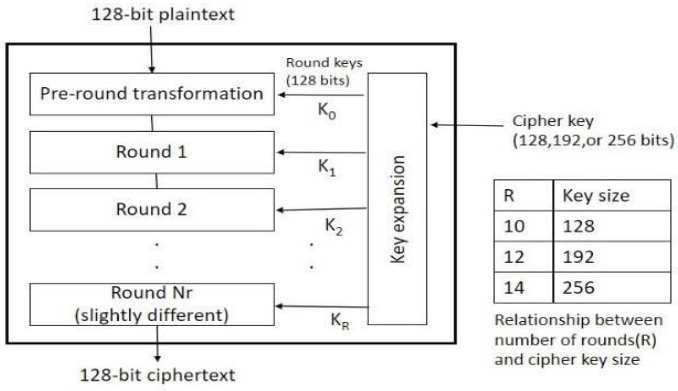

Figure. AES Architecture

It comprises a series of linked operations, some of which involve replacing inputs with specific outputs, and others involve shuffling bits around. Interestingly, AES performs all its computations on bytes rather than bits. Hence, AES treats the 128 bits of a plaintext block as 16 bytes. These 16 bytes are arranged in four columns and four rows for processing as a matrix.

III.LITERATURE SURVEY

1. N. Baracaldo, E. Androulaki, J. Glider, and A. Sorniotti studied thatthere were various instances where data breach was an issue. There were many situations in which the data of the client was breached and exposed by a cloud provider that had access to the storage medium and also where one client had access to the data of another client. To tackle all these issues, end-to-end encryption was proposed.

2. C. Wang, Z. Qin, J. Peng, and J. Wang found out that there were many problems related to the deduplication of encrypted data. So, they proposed a novel encryption scheme. They transformed encryption units into chunks and these chunks are used to generate symmetric keys which will be used to limit mapping between plain text and cipher text. They proposed thatthis scheme is suitable for the application of disk-based which requires confidentiality.

3. J. R. Douceur, A. Adya, W. J. Bolosky, D. Simon, and M. Theimer proposed a mechanism to reclaim space that was lost while replicating files into multiple desktop computers for sake of availability.Their mechanism included convergent encryption which allowed duplicate files to be merged into a single file even if files are encrypted with different user keys and SALAD, a Self-Arranging Lossy Association Database for aggregation of file content and location information in a decentralized, scalable, fault-tolerant manner.

4. D. T. Meyer and W. J. Boloskyanalyzeda large amount of data taken from 857 Microsoft computers to determine what is more efficient between the whole file versus block-level elimination of redundancy. They found out that whole file deduplication is highly efficient in lowering storage consumption, even in a backup scenario. It approaches the effectiveness of conventional deduplication at a much lower cost in performance and complexity.

5. A. Rahumed, H. C. H. Chen, Y. Tang, P. P. C. Lee, and J. C. S. Lui proposed FadeVersion, a secure cloud backup system that can serve as a security layer on top of any cloud storage services of today. It also provides cryptographic protection to date. It also assures deletion of backup so that it can be permanently inaccessible to any client while the shared version will remain unaffected from this deletion.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 6, June 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1006013 |

IV. EXISTING SYSTEMS

In Existing systems, while uploading a file to the cloud storage, it generates one hashtag for a complete file and stores it in the database. It is unable to perform deduplication because if a new file has a few new words then, in that case, it generates a new hashtag for the complete file and it gets uploaded to the cloud wasting storage and filling it with redundant data.

In the former approach, most of the existing schemes have been proposed but they are not effective, and not secure because they are not using any encryption techniques.

V. PROPOSED SYSTEM

We are proposing server-side deduplication of encrypted data. Through this, the cloud server can controlthe access to data when ownership has changed dynamically.

The proposed system breaks file content into very small blocksand then generates a hashtag. If the hashtag does not match with the stored hashtags of previously uploaded files, it will be uploaded to the cloud otherwise it increases the block reference number.

The proposed system provides data security by using the Advanced Encryption Standardalgorithm.

VI. IMPLEMENTATION

6.1 Uploading a file into the server

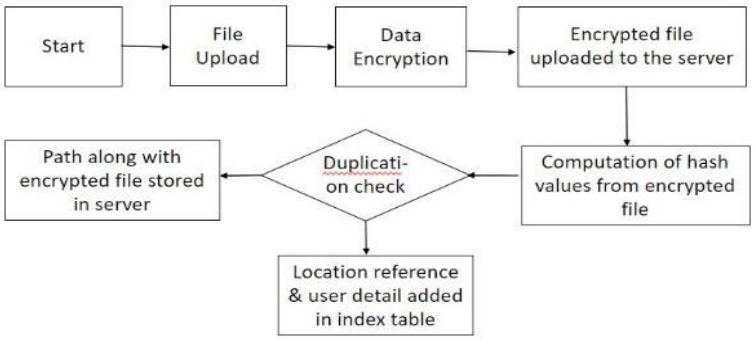

Figure. Flow chart for uploading a file to server

If a user wants to upload a file to the server, it will first check whether the file is already present or not, if it is present it will not be uploaded and the existing file details will be shown. And if not present it will upload the file.

6.2 Downloading a file from the server

If the user wants to download the file, firstly he has to prove his authentication by providing the hash details of the file requested from the server. The flowchart for the same is shown below:

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 6, June 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1006013 |

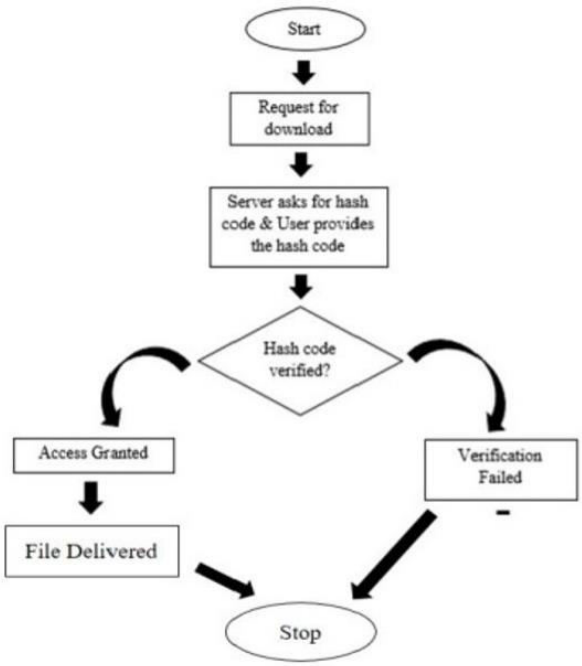

Figure. Flowchart for downloading a file from server

VII. RESULTS

7.1 Data Encryption Efficiency

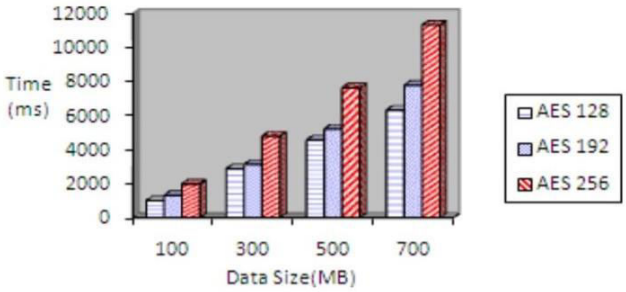

Figure. Period analysis of AES Algorithm

In this experiment, the time taken by various encryption standards of AES is shown. It is visible that the greater the data size, it takes time to encrypt the data.

7.2 Security Analysis

This project successfully passes all the security parameters to ensure data confidentiality. Because if any unauthorized user however gets to the file from the cloud, he won't be able to see the file and will only see the encrypted data. He neither has the decryption key to decipher it. Thus, all security issues are solved.

VIII. CONCLUSIONS

A large amount of data is gathered from the internet daily and this data needs to be secured from unauthorized users, and criminals of the cyber world. Thus, storing and encryption are necessary.

There are many options but they follow file-level storage which eventually increases the storage. So, to overcome that we have used block level to reduce the storage issue and AES and MD5 algorithms for encryption and hashing respectively.

Thus, this project is successful in performing Secure data deduplication in cloud storage at the block level which optimizes storage space and security of data. Future enhancements include the production of a system that can handle large amounts of data generated every day on the cloud.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 6, June 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1006013 |

REFERENCES

[1] N. Baracaldo, E. Androulaki, J. Glider, A. Sorniotti, "Reconciling end-to-end confidentiality and data reduction in cloud storage," Proc. ACM Workshop on Cloud Computing Security, pp. 21–32, 2014.
[2] C. Wang, Z. Qin, J. Peng, and J. Wang,"A novel encryption scheme for data deduplication system," Proc. International Conference on Communications, Circuits, and Systems (ICCCAS), pp. 265–269, 2010.
[3] J. R. Douceur, A. Adya, W. J. Bolosky, D. Simon, and M. Theimer, "Reclaiming space from duplicate files in a serverless distributed file system," Proc. International Conference on distributed Computing Systems (ICDCS), pp. 617–624, 2002.
[4] D. T. Meyer, and W. J. Bolosky, "A study of practical deduplication," Proc. USENIX Conference on File and Storage Technologies, 2011.
[5] A. Rahumed,H. C.H. Chen, Y. Tang, P. P. C. Lee, J. C. S. Lui, "A secure cloud backup system with assured deletion and version control," Proc. International Workshop on Security in Cloud Computing, 2011.

BIOGRAPHY

Prof. Sunny G. Gandhi Is A Research Assistant In The Information Technology Department, DMIETR Wardha RTMNU University. He received a Master Of Computer Application (Mtech) Degree And He Also Published Lots Of Research Paper On Different Computer Science Topics Etc.

INNO SPACE
SJIF Scientific Journal Impact Factor

Impact Factor: 8.165

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details