International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 06 | June 2020 www.irjet.net p-ISSN: 2395-0072

Secure Data Deduplication in Cloud Storage

Shivendra Singh[1], Zulfikar Ali[2], Aishwaray Bhatnagar[3], Ashwani Kumar Shukla[4],

Prateek Srivastava[5], Pradeep Singh Bisht[6]

[1,3,4,5,6]*Student, Computer Science Department, Babu Banarasi Das National Institute of Technology and Management Lucknow-226028, Uttar Pradesh, India*

[2]*Assistant Professor, Computer Science Department, Babu Banarasi Das National Institute of Technology and Management Lucknow-226028, Uttar Pradesh, 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 is generated on internet and to store this large amount of data, we need servers which can deduplicate data efficiently so as to avoid wastage of storage thus minimizing expenses. In this paper, we will be looking onto various techniques and methods to achieve this deduplication. The results depict that redundant data is always mapped onto same hash code and thus it does not get uploaded on 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 drivehq 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 Advanced Encryption Standard(AES) algorithm .*

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

1. INTRODUCTION

As we move towards a more technological driven era, saving data in cloud servers is the need of the hour. Huge amount of data gets uploaded onto the internet every day , preservation and security of this data is 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[1][2].To secure such 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 to store on a local server.

Every cloud has limited storage and if we start uploading redundant files to cloud, the storage is at 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 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[3], Amazon S3[4], Google Drive[5].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 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 cipher text. So, once the data is uploaded to a cloud server, it is 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 a co-ordination to ensure data security. Various techniques for deduplication over encrypted data is studied in this paper.

2. BACKGROUND

2.1 DEDUPLICATION

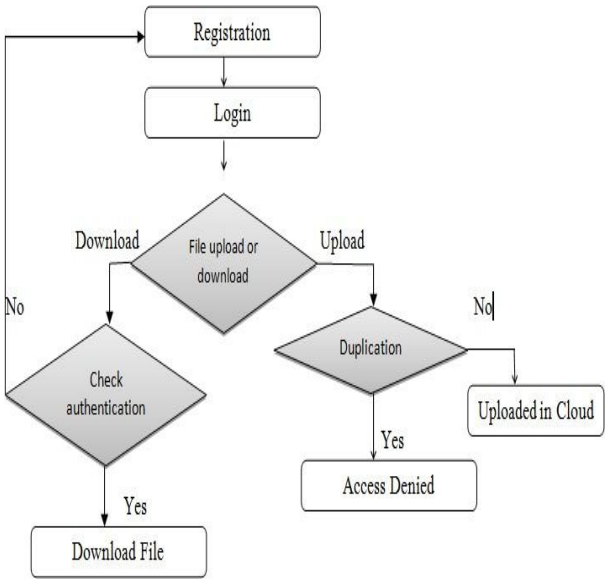

Fig 1.Deduplication Flowchart

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[6][7]. 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(AES)

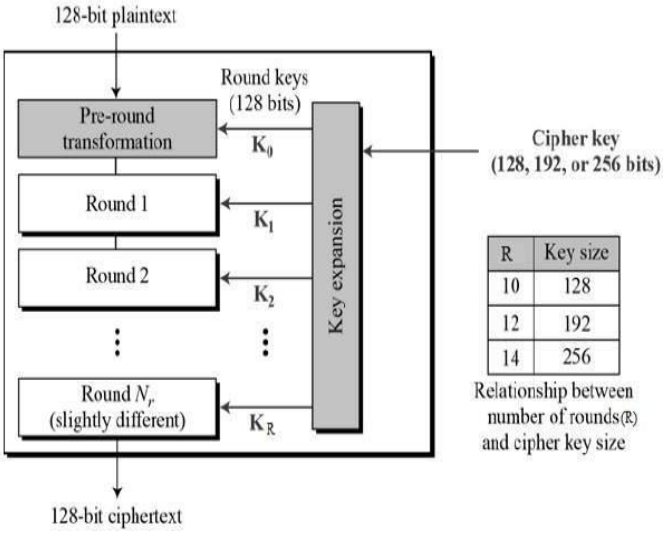

Fig 2. AES Architecture

This is an encryption algorithm which 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 decipher data. It is based on 'substitution–permutation network'. It comprises of a series of linked operations, some of which involve replacing inputs by specific outputs (substitutions) and others involve shuffling bits around (permutations). 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.

3. LITERATURE SURVEY

1. N. Baracaldo, E. Androulaki, J. Glider, A. Sorniotti[8] studied that there were various instances where data breach was an issue. There were many situations in which data of the client was breached and exposed by cloud provider that had access to storage medium and also where one client had an access to 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[9] found out that there were many problems related to deduplication of encrypted data. So, they proposed a novel encryption scheme. They transformed encryption unit into chunks and these chunks are used generate symmetric keys which will be used to limit mapping between plain text and cipher text. They proposed that this scheme is suitable for application of disk based which requires confidentiality.

3. J. R. Douceur, A. Adya, W. J. Bolosky, D. Simon, and M. Theimer[10] proposed a mechanism to reclaim space which 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. Bolosky [11] analyzed large amount of data taken from 857 Microsoft computers to determine what is more efficient between 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 towards 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, J. C. S. Lui[12] 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.

4. EXISTING SYSTEMS

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

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

4.1 DISADVANTAGES

1. User deduplication on client side cannot be given a new hash tag when updation of file is done. Hence, dynamic ownership fails.

2. Existing dynamic ownership cannot be extended to a multi-user environment.

5. PROPOSED SYSTEM

In this paper, we are proposing a server side deduplication on encrypted data. Through this, cloud server can control the access to data when ownership has changed dynamically.

Proposed system breaks file content in very small blocks(500bytes) and then generates a hash tag. If hash tag does not match with the stored hash tags of previously uploaded files, it will be uploaded to cloud otherwise it increases block reference number.

Proposed system provides data security by using Advanced Encryption Standard(AES) algorithm.

5.1 ADVANTAGES

1. Generation of Hash codes for every block of data is done thus deduplication at block level is performed.

2. Securing the data through the using of encryption algorithm.

3. Maintaining integrity and confidentiality of data.

6. IMPLEMENTATION

6.1 UPLOADING A FILE INTO CLOUD

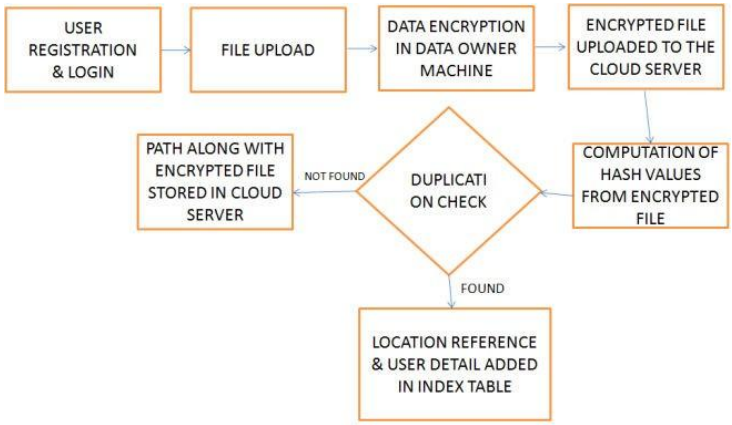

Fig 3. Flow chart for uploading a file to cloud

- Let us suppose a user uploads two files file1.txt and file2.txt into the cloud. File 2 contains duplicate data from File 1 along with some new data.
- First, we upload file 1
- Next, we upload file 2
- The generated hash values of both the files are shown as below:
  File 1 : 46-47-48-49-50-51-52
  File 2 : 46-47-48-53-54-55-56-57-58

- As we can see, hash codes of first three blocks are exactly same, thus block 46,47,48 will be uploaded onto the cloud only once and the rest will be uploaded as it is.
- Before uploading, these blocks are encrypted using MD5 Hashing algorithm.

6.2 DOWNLOADING A FILE FROM CLOUD

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 cloud. The flowchart for the same is shown below:

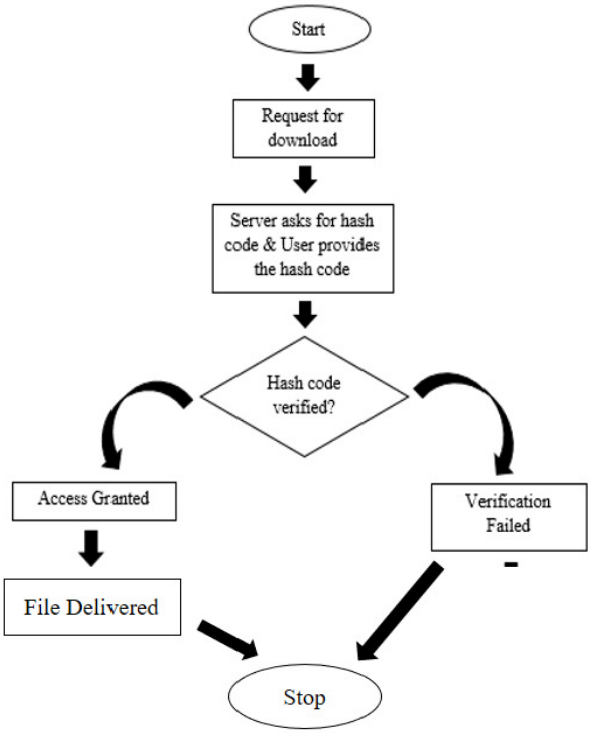

Fig 4. Flowchart for downloading a file from cloud

There are two login sections, first the admin login and second user login. Firstly, it is checked that request for file has been made by a admin or a user. This is determined with the help of a unique value which was assigned to them at the time of registration. So, when the cloud finds out that unique key besides the hash code is same as the present requested user key, it will ask for first 6 digits of hash code and if it is valid, the file will be granted.

7. RESULTS

7.1 DATA ENCRYPTION EFFICIENCY

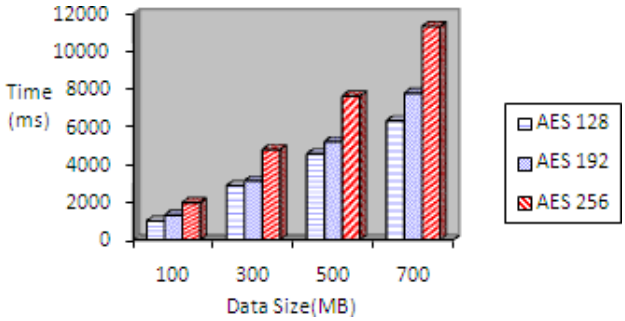

Chart 1. Period analysis of AES Algorithm

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

8. CONCLUSIONS

Large amount of data is gathered from internet on a daily basis and this data needs to be secured from unauthorized users, criminals of cyber world. Thus encryption is necessary. This paper discusses about Advanced Encryption Standard which encrypts data before uploading it in the cloud.

If duplicate data is allowed to be uploaded on the cloud on a regular basis, cloud storage will be filled with unnecessary data which need not be present there thus killing our storage and resulting in less bandwidth and bad client service. To tackle this, this paper talks about deduplication which makes use of hashing algorithm to deduplicate data.

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

9. REFERENCES

[1] Health Insurance Portability and Accountability Act of 1996 (HIPAA)

[2] Abdullah, K.A. & Al-Jafari, Mohamed. (2011). The effect of Sarbanes-Oxley Act (SOX) on corporate value and performance. European Journal of Economics, Finance and Administrative Sciences. 42-55.

[3] Drago, Idilio & Mellia, Marco & Munafo, Maurizio & Sperotto, Anna & Sadre, Ramin & Pras, Aiko. (2012). Inside Dropbox: Understanding personal cloud storage services. Proceedings of the ACM SIGCOMM Internet Measurement Conference, IMC. 481-494. 10.1145/2398776.2398827.

[4] Persico, Valerio & Montieri, Antonio & Pescapè, Antonio. (2016). On the Network Performance of Amazon S3 Cloud-Storage Service. 113-118. 10.1109/CloudNet.2016.16.

[5] N. Jeber, Jalal. (2019). The Future of Cloud Computing Google Drive. 10.13140/RG.2.2.26342.06724.

[6] Burramukku, Tirapathi & Ramya, U. & Sekhar, M.V.P.. (2016). A comparative study on data deduplication techniques in cloud storage. 8. 18521-18530.

[7] Ku, Chan-I & Luo, Guo-Heng & Chang, Che-Pin & Yuan, Shyan-Ming. (2013). File Deduplication with Cloud Storage File System. 280-287. 10.1109/CSE.2013.52.

[8] 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.

[9] C. Wang, Z. Qin, J. Peng, and J. Wang, "A novel encryption scheme for data deduplication system," Proc. International Conference on Communications, Circuits and Ssytems (ICCCAS), pp. 265–269, 2010.

[10] 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.

[11] D. T. Meyer, and W. J. Bolosky, "A study of practical deduplication," Proc. USENIX Conference on File and Storage Technologies, 2011.

[12] 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.

[13] Yang, Xue & Lu, Rongxing & Shao, Jun & Tang, Xiaohu & Ghorbani, Ali. (2018). Achieving Efficient and Privacy-Preserving Multi-Domain Big Data Deduplication in Cloud. IEEE Transactions on Services Computing. PP. 1-1. 10.1109/TSC.2018.2881147.

[14] Khakim, Lukmanul & Mukhlisin, Muhammad & Suharjono, Amin. (2020). Security system design for cloud computing by using the combination of AES256

and MD5 algorithm. IOP Conference Series: Materials Science and Engineering. 732. 012044. 10.1088/1757-899X/732/1/012044.

[15] Burramukku, Tirapathi & Rao, M.V.P.. (2017). Data deduplication in cloud storage using dynamic perfect hash functions. Journal of Advanced Research in Dynamical and Control Systems. 9. 2121-2132. 10.5013/IJSSST.a.19.04.08.

[16] 1B.Sai Nikhila, 2K.Kiranmaiee, 3Divya Vadlamudi, 4Dr.K.Thirupathi Rao, 5Deevi Radha Rani(2017) A Framework to Implement Secure De-duplication Using SHA-512 In Cloud Environment International Journal of Pure and Applied Mathematics Volume 115 No. 8 2017, 37-43 ISSN: 1314-3395 (on-line version)

[17] Siddiqui, Shadab & Darbari, Manuj & Yagyasen, Diwakar. (2019). A Comprehensive Study of Challenges and Issues in Cloud Computing: Methods and Protocols. 10.1007/978-981-13-3600-3_31