Data Storage Management in Cloud Computing Using Deduplication Technique

Marcel Chibuzor Amaechi[#1], Matthias Daniel[*2], Bennett E.O[#3]

[#1]*Msc Student, Computer Science department, Rivers State University, Nigeria.*
[*2#3]*Leturers, Computer Science department, Rivers State University, Nigeria.*

Abstract

Duplication of data stored in the cloud occupies more space. However, during data update, duplicate data must be changed in more than one place, which is more complex to rectify and would increase operational cost in cloud. This research aim at developing data storage management in cloud computing using deduplication technique. Object-oriented methodology was used. Data deduplication has been achieved via block level deduplication and key generation (symmetric algorithm). The data file was divided into number of blocks and of fixed length. Each block was divided into segments and the files were saved only once. However, each file was converted into cipher text (key form) using symmetric algorithm, the system checked for existence of key and excluded redundant key maintaining only one copy of the key in the cloud storage, the stored key was shrinked to reduce the storage space using ShrinKey algorithm and rejection algorithm was used to remove replicated key. System was implemented in Java programming language. File in cloud appeared in encrypted key with size of 16bytes thereby storage space was minimized. The system supported data privacy since data stored in cloud was encrypted and user privacy was supported, as data was uploaded by different users.

Keywords — *Deduplication, Data storage management, Cloud computing*

I. INTRODUCTION

The US International Data Corporation (IDC) estimates that, by 2020, the digital world would produce up to 40 trillion giga-bytes of replicated data [1]. This upswing in the digital world continues to increase interest in IT/Network infrastructure, increasing demand on board for a great deal of financially knowledgeable use of data limits and network communication capacity for information exchange. Remote storage systems in this field are used to study Cloud storage-based services which, in particular, provides efficient network storage platforms, is of widespread importance. The project requires the transfer, collection and allocation of the re-appropriated data in the pay-per-use program. The need for better and simpler options on the part of companies and government agencies is practically expressed in this general enthusiasm for shared storage benefits. In other words, cloud appropriation has a very specific effect in the modern age of response, adequacy and competence in the provision of Information Technology resources. Innovative advances later diminish the production of unreliable computerized material. The project requires the transfer, collection and allocation of the re-appropriated data in the pay-per-use program. The need for better and simpler options on the part of companies and government agencies is practically expressed in this general enthusiasm for shared storage benefits. In other words, cloud appropriation has a very specific effect in the modern age of response, adequacy and competence in the provision of Information Technology resources. Now you no longer have to spend much money to buy expensive software or advanced equipment, which you may never need again. The main inspiration behind these cautious benefits is the decrease of capital expenditure (CapEx), a steady charge to maintain operational expenditure (OpEx) and a persistent fee to manage OpEx, which is an ongoing cost to run a business [2].

Cloud computing is just another way of describing information technology "outsourcing"; it also refers to any computer service render over a network [3]. Cloud computing means that it is delivered as a service by another organization and generally accessible over the internet in a completely smooth form, instead of using all of its hardware and software that is stored on your machine or on the network of your business. It is in the nebulous "cloud" that the Internet represents precisely where hardware and software are located and how they all work, regardless of the user. Data storage is one of the most widely used cloud services. The benefit to cloud usage is that they can store and view large volumes of data at anytime and anywhere without updating their devices. Data stored in the cloud that require different forms of protection due to the different nature of the data. Data stored in the cloud includes private personal data, shared information, shared data within a group.

Data deduplication or single instancing applies specifically to the deletion of redundant data. Owing to the exponential increase in the amount of digital information, storage systems must be put in place to store and preserve this information effectively. Deduplication is essentially the removal of redundant data [4]. Deduplication techniques are

widely used in cloud services to reduce server space. Cloud Service Providers (CSPs) also execute deduplications to ensure optimal use of disk capacity or use server redundancy to avoid multiple client redundant file storage. Duplication saves about 90-95% and 68% on storage space for backup systems and standard file systems, while many corporate cloud storage services such as Bitcasa, Dropbox and Google Drive to-deductions make use of cost savings [1]. [5] suggested a PRE-based deduplicatibility program but relied entirely on the Approved Party to control the deduplicatibility of the results. Data access regulated by data holders cannot be tailored flexibly to different scenarios. [6] Proposed access control for Cloud Attributes-Based Encryption (ABE) encrypted data. Set a set of attributes for the user to identify and encrypt data on the basis of a given access mechanism for the attribute. [7] has introduced modern server-side deductibility for authenticated data. The Cloud service will control access to outsourced data even when controlled dynamically by random converging encryption and protected community ownership allocation.

II. LITERATURE REVIEW

Data storage is a way to store information in memory. Data are managed in large storage rather than saving it on the hard disk of a computer. Storage Area Network (SAN) is a distributed storage network that provides centralized block storage. Such SANs are easily saccessible by clients because they are directly connected to the operating system. Furthermore, A SAN has its own network storage system and is typically inaccessible over a standard network for normal users. It should also be remembered that the SAN does not abstract the file itself but blocks operations only. Network Attached Storage (NAS) uses, like SAN, NFS and SMB / CIFS protocols, where remote storage is supported and computers are ordering data rather than disk blocks [8]. End-users can link their data to a local archive with a copy from a remote server by means of cloud storage [9] Document updates are synchronized in the local folder automatically to the cloud service provider. It may be mounted on multiple devices to ensure the data is modified in the same folder, independently of the user. Only with the same approach is data available to the customer [10]

[11] Introduced a very strong cloud storage system. This system uses various cloud storage providers for data storage and streaming of proxies, this eliminates the one-time cost for switching network providers on a separate computer the machine is running.

[12] Published two algorithms that permit users to create Byzantine cloud tolerant storage through the quorum of cloud storage providers, DepSky-A and DepSky-CA. these algorithms are supported by the theory and performance assessment. The system uses existing cloud computing systems and avoids running codes on database server

Meta Storage was launched by [13]. It is highly accessible Hash table that duplicates data in various storage systems. It offers a description of a wide selection of stock approaches such as cloud storage, SSH file servers and local file systems by unifying them in a single Application Program Interface (API).

III. RESEARCH METHODOLOGY

The research methodology used in the study is the CRA and Object-Oriented Design (OOD), Object-Oriented Design. The study is based on objects. CRA is a methodology to solve problems that allows the development of systems, procedures, tools and technology that can be applied far beyond case study. The study used in early-development prediction of several diseases with various premature and ill-tea full-tea, has been born after the Greek goddess of principle [14] The design approach is aimed at providing the best solution for any problem situation in industrial design, electronics or architectural design.

A. SYSTEM DESIGN

The proposed system is based on Deduplication Technique on Block Level. Figure 1 represents proposed system.

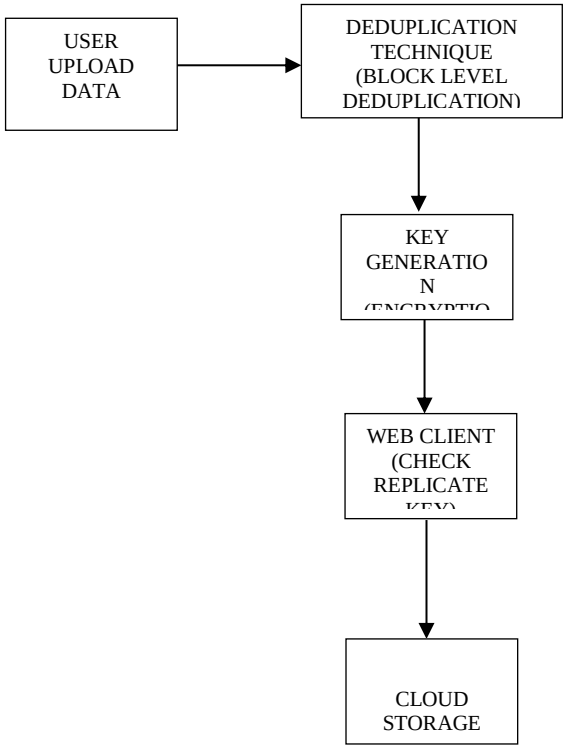

Fig. 1: Architecture of the Proposed System

The file is uploaded to a server by the user (data owner). The user (database data owner) uses the encryption algorithm and web- client searches for key presence within the cloud storage and stores key in the server. Five components of the data owner are included in this scheme.

a) USER

Before accessing cloud storage information, the user must be authenticated. The user is needed to provide device authentication needs including user name and password. The user will start the server when opening the cloud environment. Then the user can upload file to the cloud. System input is shown in Table 1.

Table 1: Input Data

S|N	INPUT DATA	FORMAT
1	Document	Docx, pdf, txt
2	Audio	Mp3
3	Video	Mp4
4	Image	Jpeg, png

Cloud storage accepts files in different file format, but the listed format is applied in this study.

b) DEDUPLICATION TECHNIQUE

Deduplication is the process used to delete redundant data copies and to maintain only one data copy on cloud storage. With respect to data storage in cloud, deduplication check undergoes the following stages as shown in Figure 2.

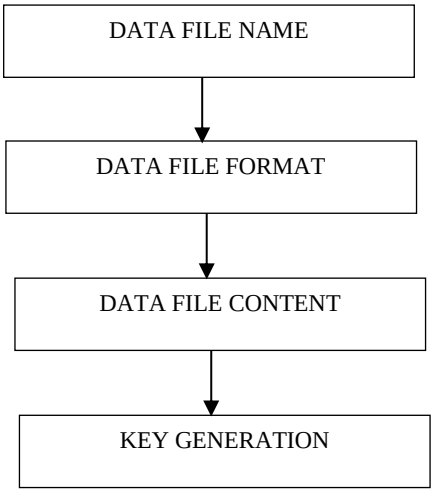

Fig. 2: Deduplication Check

The data file input is contrasted with name, format and content. The data file would be removed if the parameters remain the same.

For example, given a series of data file V_i with name V and content *i* ($i\epsilon$1,..,n). Assuming V_1 is identified as first data file. V_i's ($i\epsilon$ 2,...,n) offset compared to V_1 in the data chain is L = $F(V_i)$.

Where F is the calculated offset function for finding out Vi's offset relative to the first data file in the chain. If the data file's content is large, its offset value relative to the first data file in the chain is the maximum content length Lmax. In the data file, the content category can be simply executed by dividing the segment maximum length.

However, once the data storage threshold is given, the number of content groups G can be obtained by:

G = Lmax/T

Where,
T is the storage threshold.
G is the content groups.
Lmax is the maximum length of the content.
V is the data file
i is the content.

We used a chunk level often referred to as "Deduplicate Block Level." the data file is divided into blocks and fixed lengths. The block must be broken into pieces, and the files will only be saved once. The main benefit of using block level is that any form of file can be carried.

Fig 3: Block Level Deduplication

File 1, File 2 and File 3 consists of 8 blocks, each block represents 1MB. With deduplication, File 1 and File 3 have same content in each block, so the content in File 3 is removed. In File 2, block A, B, C, D and E have same content with File 1 block A, B, C, D and E, thus block I, J and K in File 2 remain with 3MB. In Data de-duplication, the data blocks are checked to find the same data block. Thus, there is more storage space.

In block level segments are assigned using hash code which generates uniqueness to identify chunks. This can be achieved via key generation using encryption algorithm.

c) KEY GENERATION (ENCRYPTION ALGORITHM)

The user encrypts and saves the cloud key, and encrypted message is stored in the data block. First, depositors use k key for encoding data, based on a SymEnc symmetrical encryption method, to encode the data file F. The F data key is extracted by a special hash function using a converging cryptographer solution from the original F data file.

Algorith 1: To Encrypt a Message (kaaniche, 2014)

1. Input: {f, H, SymEncg}, where f is the data file, H is a one-way hash function and SymEnc is a symmetric encryption algorithm.
2. Output: (C_f, k)
3. k = H(f);
4. C_f = SymEnc(f; k);
5. return (C_f, k)

To retrieve outsourced data file, the authorized recipient U_j extracts the enciphered file C_f key, using the public key k by the depositor U_i.

Algorithm 2: To Shrink a file

To reduce data file in a storage space R, ee denote each key element by key.elt as;

$$key.elt_{i\epsilon 1,2,3} = \{C_i, R(C_i)\}_{i\epsilon 1,2,3} \quad (3.1)$$

1. Input: {C_2, skc}
2. Output: C_3
3. $C_3 = (C_3)^{\frac{1}{skc}}$
4. return C_3

In order to get a redirected C3 key element, the Cloud Service provider (CSP) gets the second key element (C3) element, and runs the ShrinKey algorithm as shown in Algorithm 2.

Subsequent ciphertext elements are shrink to reduce the storage space.

Algorithm 3: To Decrypt a file

1. Input: {C_f, k, SymEnc}
2. Output: f
3. f = SymEnc(C_f, k)
4. return f

Algorithm 4: To remove replicated file

1. Let H_{Cf} := hash(C_f) be computed for once and stored for later use.
2. If $Y_1 = H_{Cf}$,
3. set Y_2 = accept,
4. Else Y_2 = reject.

The cloud storage will store file in an encrypted key as shown in Table 2.

Table 2: Expected Output Key Generation

S/N	ENCRYPTION KEY
1.	Key 1
2.	Key 2

Any data uploaded to the cloud is saved and to reduce storage space in an encoded key. The program monitors the presence of a new file in cloud storage before storing the key.

d) WEB CLIENT

The deduplicate system uses a data key-keying technique while the depositor produces the respective enciphering key to encrypt data information before saving on remote servers. This approach leads to multiple encryption and reduction of the storage capacity of the cloud provider.

The KF enciphering key is extracted from one-Way hash-function H() when the data owner wants to store a new data file F in the cloud. Please note that data is stored on a symmetrical algorithm on cloud servers with the derived KF key. The data owner must first encrypt the data file to be outsourced. The key KF for data identification is then generated. The file created (key) must be special in the database of Cloud Provider. Client sends message to test the uniqueness of created KF to start the storage process. The type of file storage is as follows:

a. ClientRequestVerif
b. ResponseVerif
c. ClientRequestStorage
d. ResponseStorage

a. ClientRequestVerif: This message checks whether KF is unique. If the key is sent, the customer must keep the provider secure. Upon completing the inspection, the cloudserver labels the requesting company and only asks the customer to give allowed access to the data owner.

b. ResponseVerif: This message must be sent by the cloud services provider to tell the customer that the re-applied KF is in its databases.

c. ClientRequestStorage: If the file does not exist on cloud servers, the client sends the encrypted and decrypted key KD using the registered users ' public key.

d. ResponseStorage: This message confirms the data storage capacity of the data owner.

IV. RESULTS

Cloud storage allows different files to be stored. We used audio, video, document in the form of pdf and docx in this research. The cloud was uploaded with eighty files as Table 3 reveals. Every user is authenticated prior to storing the file.

Table 3: Deduplication Check

S|N	USER	FILE SIZE BEFORE UPLOAD	FILE FORMAT	FILE ENCRYPTED KEY	DEDUPLICATION CHECK	DUPLICATE KEY
1	Joy.cey	7.53MB	mp3	2xZQCz51dSc3weDq2gaH	100%	NONE
2	Mike_mi	13.63MB	mp4	BxLjDq5Am7BUWGXUkpPT	100%	NONE
3	CJ	2.75MB	Pdf	OjteXOIcRB5UIY2XCmQF	100%	NONE
4	CJ	22.29MB	mp3	OagPxzNbaRx3dNPNXH36	100%	NONE
5	Joy.cey	600KB	Docx	SMREyhnjN8qIy1Ne	100%	NONE
6	Smart_ivy	430KB	Docx	Ckp82oNpIjsseKerWcge	100%	NONE
7	Clement.uju	28.30MB	mp4	JNWZZuxtnuG7zZc7aETU	100%	NONE
8	Smart_ivy	16.11MB	Pdf	IHk0UXi6QTS0wUa9MJMU	100%	NONE

The length of a key in bits is just a size specification. 16 bytes of space is needed with a 128-bit key. Table 4shows file size before and after uploading.

Table 4: Reduction in Storage Space due to Deduplication

S|N	USER	FILE SIZE BEFORE UPLOAD	UPLOADING DURATION (SECOND)	FILE ENCRYPTED KEY	ENCRYPTION BITS	STORAGE SPACE (FILE SIZE AFTER UPLOAD)
1	Joy.cey	7.53MB	9	2xZQCz51dSc3weDq2gaH	128	16bytes
2	Mike_mi	13.63MB	11	BxLjDq5Am7BUWGXUkpPT	128	16bytes
3	CJ	2.75MB	6	OjteXOIcRB5UIY2XCmQF	128	16bytes
4	CJ	22.29MB	17	OagPxzNbaRx3dNPNXH36	128	16bytes
5	Joy.cey	600KB	4	SMREyhnjN8qIy1Ne	128	16bytes
6	Smart_ivy	430KB	3	Ckp82oNpIjsseKerWcge	128	16bytes
7	Clement.uju	28.30MB	21	JNWZZuxtnuG7zZc7aETU	128	16bytes
8	Smart_ivy	16.11MB	14	IHk0UXi6QTS0wUa9MJMU	128	16bytes

A. COMPARISON

As data is encrypted, the proposed and current system supports data privacy. Nevertheless, as shown in Table 5, the proposed system allows different users to access data.

Table 5: Comparison Analysis

System	Deduplication	Data Privacy	Data access
Heterogeneous Data Storage Management with Deduplication in Cloud Computing [3]	Yes	Encryption	Single user
Proposed System	**Yes**	**Encryption**	**Multi-user access**

V. DISCUSSION

Data can be stored in documents images, voice and pictures; this research is done in pdf and docx format using audio, video, image and data. In this study, we use audio, video, document in PDF and docx formats. The deduplication test is provided in Table 3. The cloud was uploaded with 80 files. Before storing files, each user has been authenticated. All stored files have been translated into key and deduplication was performed on each of the keys. When testing deduplication, duplicate key was removed. There was no duplicate key in the encrypted key column of the register, so the deduplication was met.

A decrease in storage area is shown in Table 4. The table indicates the number of files uploaded, the file size before uploading, the file transfer time and the file size after uploading. Each file has a different file size, video and audio files are higher in megabytes of 6 to 11 seconds, while the docx and pdf text files are in bytes of 2 to 3 seconds.

There are 16 bytes of space for 128-bit key files have been accessed from various users, and data are 20 in a cloud. The storage space of files was therefore reduced to a minimum.

After selecting a file by a data owner (user), the storage process starts. A file is stored and processed; an imported file has been encrypted and "ClientRequestVerif" checks that the cloud storage key is available, A message has been returned back to the data owner to suggest that the file "exists already". However, if there exist no duplicate key in the cloud storage, the new file would be successfully stored in cloud as found. It shows the list of files stored in cloud storage. Symmetric key encryption was used and the key size was long enough to satisfy security requirement and has facilitated data protection in cloud.

Data was decrypted during download, data owner or any user could download the file using access control assigned to users.

VI. CONCLUSION

We developed a data storage management system for deduplication of cloud computing in this research Data duplication by block level deduplication was prevented in cloud storage. The cloud is based on the application of a data keying approach, while the remote server stores the associate encoding key for the encryption of data contents. This approach leads to the same content being encrypted several times and to the cloud provider's storage capabilities being reduced. Data stored on cloud servers based on a symmetrical algorithm was shrunk to minimize space in storage using the ShrinKey algorithm, and replicating file removal algorithm is used. The system has successfully removed duplicate data by searching for the presence of encryption key in cloud storage. Thus, file in cloud storage appeared in encrypted key with a storage space of 16bytes thus storage space was minimized. The system supports data privacy of cloud users as data stored inthe cloud was in an encrypted form.

ACKNOWLEDGEMENTS

First and foremost, all praises and thanks be to God Almighty for his unsurpassed favour and grace and for the successful completion of this research work. I would like to express my deep and sincere gratitude to my supervisor, Dr. Daniel Matthias and co-supervisor, Dr. Bennett E. O. for giving me valuable guidance, encouragements, and patience during this research. Also, I will in the same vein genuinely appreciate my entire family my wife Mrs. Juicy Chinaka Marcel and my child Master Daniel Chibuike Marcel, Miss Shalom Oluebebe Marcel and Miss Joy Ozichukwu Marcel as well my siblings. I am grateful for the love, patience, support, understanding and prayers towards the success of this study.

Finally, to my course mates, Tuase Emmanuel, Ene Donald and Charles Davidba, I cannot forget the cooperation and team work that existed between us, you guys are the best, I Thank God I was part of you.

REFERENCES

[1] Chen, R. "*Secure Data Storage and Retrieval in Cloud Computing*. *Doctor of Philosophy thesis, School of Computing and Information Technology*", University of Wollongong. http://www.ro.uow.edu.au/theses/4648,(2016).

[2] Kaanichie, N. "*Cloud Data Storage Security based on Cryptographic Mechanisms*". Telecom Sudparis in Partnership with Pierre Et Marie Curie University. (2014).

[3] Sangeetha, A and Geetha, K. "*Heterogeneous Data Storage Management with Deduplication in Cloud Computing*." *Asian Journal of Applied Science and Technology (AJAST)* (Open Access Quarterly International Journal) 2(2), pp (840-845). (2018).

[4] Li, J., Jia, C., Li, J and Liu, Z. "*A Novel Framework for Outsourcing and Sharing Searchable Encrypted Data on Hybrid Cloud," Fourth International Conference on Intelligent Networking and Collaborative Systems*, pp(1–7). (2012).

[5] Yan, Z., Ding, W. X and Zhu, H. Q. "*A scheme to manage encrypted data storage with deduplication in cloud*," in *Proc. of ICA3PP*, pp (547-561), Springer, (2015).

[6] Bethencourt, J., Sahai, A and Waters, B. "*Ciphertext-policy attribute-based encryption*," in Proc. of IEEE Symp. Secure Privacy (07), pp(321-334), (2007).

[7] Hur, J., Koo, D., Shin, Y and Kang, K. "*Secure Data Deduplication with Dynamic Ownership Management in Cloud Storage*," IEEE Trans. Knowl.Data Eng., 28, (11), pp (3113-3125), (2016).

[8] Singh, A and Tech, G. "*Server-Storage Virtualization: Integration and Load Balancing in Data Centers*," In *International Conference for High Performance Computing, Networking, Storage and Analysis*, pp(1-12), (2008).

[9] Drago, I., Mellia, M., Munafo, M., Sperotto, A., Sadre, R and Pras, A. Inside Dropbox. *The ACM Conference*, 481. (2012).

[10] Geel, M. Cloud Storage: File Hosting and Synchronisation 2.0. [Online] Available at: https://www.vis.ethz.ch/de/visionen/pdfs/2012/visionen_2012_3.pdf?end=15&start=11, (2013).

[11] Abu-Libdeh, H., Princehouse L and Weatherspoon, H "*Racs: a case for cloud storage diversity*," in Proceedings of the 1st ACM symposium on Cloud computing, ser. SoCC '10, pp(229–240). [Online]. Available: http://www.cs.cornell.edu/projects/racs/pubs/racs-socc2010.pdf, (2010).

[12] Bessani, A., Correia, M., Quaresma, B., André, F and Sousa, P. *"Depsky: dependable and secure storage in a cloud-of-clouds*," in Proceedings of the sixth conference on Computer systems, ser. EuroSys '11. ACM, pp(31–46). [Online]. Available: http://www.gsd.inesc-id.pt/~mpc/pubs/eurosys219-bessani.pdf, (2011).

[13] Bermbach, D., Klems, M., Tai, S and Menzel, M. *"Metastorage: A federated cloud storage system to manage consistency-latency tradeoffs*," in Cloud Computing (CLOUD), *IEEE* International Conference on, pp (452–459), (2011).

[14] Blount, M., McGregor, C., James, A., Sow, D., Kamaleswaran, R., Tuuha, S., Percival, Jand Percival, N. On the Integration of an Artifact System and a Real-Time Healthcare Analytics System.In *Proceedings of the 1st ACM International Health Informatics Symposium*,Arlington, Virginia, USA.pp (647–655), (2010).