Copyright © 2019 Tech Science Press JBD, vol.1, no.1, pp.39-45, 2019

Code-Based Preservation Mechanism of Electronic Record in Electronic Record Center of Cloud Storage

Yan Leng[1, 2], Yongjun Ren[1, 2, *], Chengshan Qian[3] and Jinyue Xia[4]

Abstract: With the rapid development of E-commerce and E-government, there are so many electronic records have been produced. The increasing number of electronic records brings about storage difficulties, the traditional electronic records center is difficult to cope with the current fast growth requirements of electronic records storage and management. Therefore, it is imperative to use cloud storage technology to build electronic record centers. However, electronic records also have weaknesses in the cloud storage environment, and one of them is that once electronic record owners or managers lose physical control of them, the electronic records are more likely to be tampered with and destroyed. So, the paper builds a reliable electronic records preservation system based on coding theory. It can effectively guarantee the reliability of record storage when the electronic record is damaged, and the original electronic record can be restored by redundant coding, thus ensuring the reliable storage of electronic records.

Keywords: Electronic record, cloud storage, preservation mechanism, coding theory.

1 Introduction

The electronic record refers to the digital file, which is generated in digital equipment and the environment, and is stored in the tape, disk, CD and other carriers, relying on computers and other digital devices to read and process [Liu and Li (2017)]. Long-term preservation is a behavior, in which the reliable, scientific and reasonable ways are utilized to maintain the authenticity, complete and effective of electronic records.

The core function of the electronic record center is to improve efficiency and realize the sharing of resources within the whole society. However, the existing electronic record center mainly relies on the construction and development of the portal. Also, the use of electronic record is limited to the within of the organization. The connectivity among electronic record centers is low, which lead to the lack of interoperability and sharing. Secondly, due to the differences of the construction mode of the electronic record centers,

[1] School of Computer and Software, Nanjing University of Information Science & Technology, Nanjing, China.

[2] Jiangsu Collaborative Innovation Center of Atmospheric Environment and Equipment Technology (CICAEET), Nanjing University of Information Science & Technology, Nanjing, China.

[3] Binjiang College, Nanjing University of Information Science & Technology, Wuxi, China

[4] Internetional Business Machines Corporation (IBM), New York, USA.

* Corresponding Author: Yongjun Ren. Email: renyj100@126.com.

JBD. doi:10.32604/jbd.2019.05870 www.techscience.com/jbd

the way and ability to provide service are different. On the other hand, with the development of Internet terminal equipment, there are more and more ways to access electronic records through the Internet. Thus, the services and the content of electronic records are more prosperous and diversified, and taking a new way to provide users is also one of the development directions of electronic record centers.

Cloud computing is the highly demanded technology nowadays. Due to the service oriented architecture, seamless accessibility and other advantages of this advent technology, many transaction rich applications are making use of it [Gopinath and Bhuvaneswaran (2018)]. Advanced cloud computing technology provides cost saving and flexibility of services for users. With the explosion of multimedia data, more and more data owners would outsource their personal multimedia data on the cloud [Xiong and Shi (2018)]. Cloud storage is a new technology based on cloud computing, which integrates the cluster application, distributed file system, multi-storage server data organization and so on [Xie, Yuan, Zhou et al. (2018)]. And it enables a large number of different types of storage devices in the network to work together and provides users the services of data storage [Zheng (2013); Ren, Shen, Liu et al. (2016)]. Cloud storage system is composed of the data storage layer, primary management layer, application interface layer, and user access layer according to its basic structure. Also, it has both functions of data storage and application software, which can be seen as a collection of storage devices and servers. Using cloud storage to build electronic record centers has inherent advantages. But in the cloud storage environment, the owner and manager of the electronic record lose the physical control of the electronic record, making the electronic record easier to lose. Thus, some techniques are essential to recover the damaged or lost electronic records.

2 Related work

Erasure codes, as an essential redundancy strategy, have been paid more and more attention. The commonly used erasure codes are mostly maximum distance separable codes. When the storage node fails, the original data need to be reconstructed based on erasure codes. However, the amount of data transfer is much more significant than the stored data of nodes. When nodes are dispersed in the network, the recovering of nodes need to consume many network bandwidths. To solve the problem, Dimakis et al. proposed a regenerative code to optimize the repair bandwidth. The regenerating code does not limit the degree of the data block and redundant block, and it constructs the generating matrix by selecting particular coding coefficients. When the nodes are repaired, the data blocks stored in the same node are fused, which reduces the amount of data to be transmitted and achieve the purpose of saving bandwidth cost.

In the regeneration code scheme proposed by Dimakis et al., the data file encoding is stored to n nodes, and each node stores the encoded data of α size. When a node fails, the substitute node can connect to any d nodes from the remaining n-1 nodes $(k \leq d \leq n-1)$ and download the data of β size from each node $(\beta \leq \alpha)$. The nodes that help repair are called help nodes. The entire download data $d\beta$, used to repair node, is called repair bandwidth. The parameter set of the regenerative code scheme can be expressed as $\{(n,k,d),(\alpha,\beta,B)\ \}$; the average repair bandwidth $d\beta$ is less than the size of the file B. The total bandwidth consumption for repairing a node is $\gamma = d\beta$. k is the number of nodes needed for data reconstruction.

The number of help nodes involved in the repair processing is d ($d \geq k$), according to the regeneration code theory. Each help node first linearly combines the data within the node, and the amount of uploaded data is $\beta \leq \alpha$. The repair bandwidth $\gamma = d\beta$ decreases with the increase of d. With the increase of the number of nodes participating in the repair, the data size transferred by each node is smaller. Moreover, the speed of β decrease is faster than that of the d increasing, so that the total bandwidth is reduced. In addition, the total repair bandwidth is the smallest when $d = n - 1$. The minimum bandwidth is achieved by increasing the storage capacity of each node. The minimum bandwidth regeneration (MBR) point can be obtained from the following expression [Hao, Lu and Liu (2013); Wang, Zhao and Hou (2012)].

$$\alpha = \frac{2Bd}{k(2d-k+1)}, \beta = \frac{2Bd}{k(2d-k+1)} \tag{1}$$

The encoding under this extreme condition is called the minimum bandwidth regenerating code (MBRC). Two matrixes represent the MBRC; one is a data matrix and another repair matrix. The data matrix is a symmetric matrix that contains the data itself. The repair matrix is pre-defined for distributing data from the data matrix to each storage node, helping to implement data reconstruction and failure node repairing from a subset of nodes.

File B is segmented and striped with vector B= $[b_1, b_2, \dots, b_B]$. The regenerating code is represented by matrix C(b) with (n×α) dimensions. The α elements contained in the i-th row of C(b), i.e., C_i, are stored by the node i. And the regenerating code is represented by the following formula: $C(b) = R \cdot D(b)$, where R is a pre-defined repair matrix with (n×m) dimension, and R is independent of the data matrix D(b). The repair matrix is used to distribute the information contained in the data matrix to n nodes. D(b) is compose of $b_1, b_2, \dots, b_B$. The data matrix D is symmetry, i.e., $D^t = D$. The i-th row of the regenerating code C, i.e., C_i is the coding vector of node i, so the constructed regenerating code is: $C = \{C(c) | c \in F_q^B\}$.

3 Problem description

3.1 Existing problems

At present, the construction of electronic record centers faces with the following problems. The first problem is the size of the electronic record center. In the era of information explosion, the number of electronic records overgrew. And the demand for storage space is getting higher and higher. Thus, electronic record center faces the contradiction between a large number of electronic records and insufficient storage space. Although the expansion of existing storage capacity can temporarily meet the surge of electronic records, the traditional expansion way of electronic record center will bring new problems to reliability, backup and migration of electronic records [Tu and Zhang (2016)].

Second, the electronic records center based on the government affairs network is relatively independent. The formation and circulation of electronic records enable the electronic records lack of identity in form. For example, carrier type, file format, data storage format and so on, will cause a low utilization rate of electronic records. Moreover, it makes the preservation and development of electronic records lose, which is not conducive to the formation of a unified and efficient management mode.

3.2 Electronic record center based on cloud storage

The traditional preservation way is to store a large number of electronic records in the servers, which requires high-performance storage devices. At the same time, some individual administrators are required to maintain the operation of the servers. And in this way, the storage and maintenance efficiency of electronic records is relatively low.

In contrast, cloud storage system not only has the function of network storage devices but also can serve as the role of a network server, application software, and access interface role. Using cloud storage technology to store electronic records, on the one hand, it can significantly reduce the use of central storage servers. Moreover, the electronic records holders save the workforce and material resources for storage and maintenance. On the other hand, the electronic record center is constructed based on cloud storage service, in which professional service providers manage electronic records. Cloud storage service providers not only support professional and efficient services of electronic record but also reduce the requirement of storage resources and optimize the allocation of network information devices.

However, there are the following problems in the cloud storage of electronic records. The owner or manager of the electronic record has lost the physical control of the electronic records, which makes the electronic record easier to be copied and contents to tamper. Also, if the device failure or mismanagement, which will cause the cloud storage servers losing electronic records. Therefore, it is necessary to take measures to recover the damaged electronic records.

4 Preservation of electronic records based on MBRC

In the paper, we use redundant coding instead of copying to recover lost or corrupted electronic records. If the use of replication to restore the electronic record, it is difficult to judge the authenticity of electronic records. Thus, the evidence of electronic record is lost. While the method based on redundant coding does not produce new data. Also, it does not affect the authenticity of electronic records. Thus, the restored electronic records are still trustworthy.

4.1 Distribution of electronic records based on MBRC

At first, the electronic records are encoded to regenerating codes, and then stored in multiple cloud storage servers. It assumes that B = $[b_1, b_2, \ldots, b_B]$ is an electronic record, and it reconstructed into a data matrix $D(b)$ with $(m \times \alpha)$ dimensions. R is a pre-defined repair matrix. Thus, the generating code is $C(b) = R \cdot D(b)$ of electronic record B based on MBRC. Then, $C(b)$ is allocated to n cloud storage servers. The record block stored in the cloud storage server i is $c_i = r_i D(b)$. The implementation process is shown in the Fig. 1.

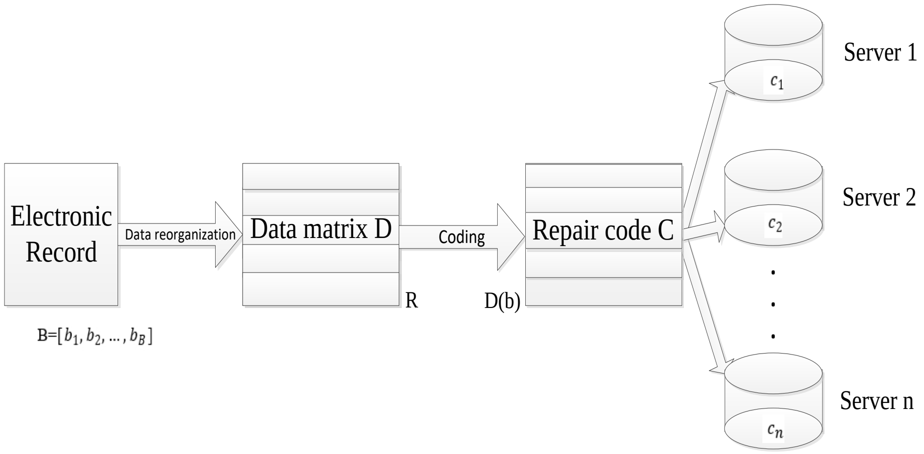

Figure 1: Storage of electronic records based on MBRC

4.2 Repairing of failure cloud server

Each cloud server in electronic record centers is associated with a different coding vector r_i in the repair matrix. The reconstruction process uses the vector r_f(f means the failure cloud server). The replacement server is connected to any of the d storage servers $\{l_1, l_2, \ldots, l_d\}$, called the help servers. The j-th helper server l_j passes the data vector $r_{l_j}^t D(b) r_f$ to the replacement server. The replacement server aggregates data to generate matrix $R_{replace} D(b) r_f$, where $R_{replace} D(b) r_f$ is the sub-matrix of R, and R includes the d line $\{l_1, l_2, \ldots, l_d\}$. Using the matrix R to restore the vector $D(b) r_f$, and achieve partial decoding. The output data of the partial decoding is re-encoded (the process utilizes the symmetry of the data matrix $D(b)$). And the data stored by the replacement server is finally recovered. In the process of failing server recovery, each server l_j participating in server recovery only needs to know the repair coding vector corresponding to the failed server f. And there is no need to know the other nodes involved in the process. This can greatly simplify the system operation.

4.3 Recovery of electronic record

To reconstruct the data D, the data recovery server will connect to any k servers $\{i_1, i_2, \ldots, i_k\}$ among the n storage servers in electronic record center. The j-th server passes it's $r_{i_j}^t D(b)$ to the data recovery server. And the data recovery server generates the matrix $R_{DC} D(b)$ by aggregating the k vectors (R_{DC} is the sub-matrix of R, the matrix has k rows). Using the repair matrix R and decoding the data matrix D, the original electronic record is reconstructed. The implementation process is shown in Fig. 2.

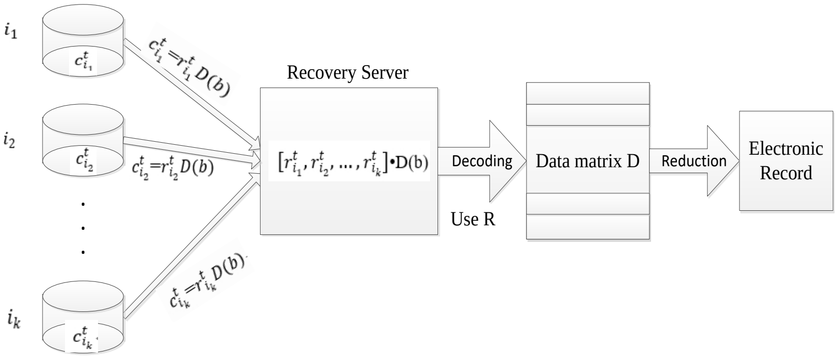
i_1
$c^t_{i_1}$
$c^t_{i_1}=r^t_{i_1}D(b)$
i_2
$c^t_{i_2}$
$c^t_{i_2}=r^t_{i_2}D(b)$
i_k
$c^t_{i_k}$
$c^t_{i_k}=r^t_{i_k}D(b)$
Recovery Server
$[r^t_{i_1}, r^t_{i_2}, \dots, r^t_{i_k}]$•D(b)
Decoding
Use R
Data matrix D
Reduction
Electronic Record

Figure 2: Recovery of electronic record based on MBRC

5 Conclusion

This paper constructs an electronic record preservation mechanism based on the minimum bandwidth regenerating code in the cloud storage environment. The electronic records are encoded firstly, then partitioned to blocks and stored into multiple cloud storage servers in the electronic record center. When a chunk of the electronic record is lost or damaged, the block can be recovered by regenerating code, and the electronic record is restored.

Acknowledgement: This work is supported by the NSFC (61772280), and the PAPD fund from NUIST.

References

Gopinath, V. (2018): Bhuvaneswaran R.S. Design of ECC based Secured Cloud Storage Mechanism for Transaction Rich Applications. *Computers, Materials & Continua*, vol. 57, no. 2, pp. 341-352.

Hao, J.; Lu, Y. B.; Liu, X. J.; Xia, S. T. (2013): Survey for regenerating codes for distributed storage. *Journal of Chongqing University of Posts and Telecommunications (Natural Science Edition)*, vol. 25, no. 1, pp. 30-38.

Liu, Y.; Li, J. (2017): Conceptual comparison and linkage between electronic data in law field and electronic records in archival field. *Archives Science Study*, no. 4, pp. 92-99.

Ren, Y. J.; Shen, J.; Liu, D. Z.; Wang, J.; Kim, J. (2016): Evidential quality preserving of electronic record in cloud storage. *Journal of Internet Technology*, vol. 17, no. 6, pp. 1125-1132.

Tu, Y. M.; Zhang, M. X. (2016): The influences and strategies deal with the long-term preservation of electronic records from the view of life cycle. *Archives and Construction*, vol. 4, no. 11, pp. 8-13.

Wang, Y.; Zhao, Y. L.; Hou, F. (2012): Minimum bandwidth regeneration code of distributed storage system. *Journal of Chinese Computer Systems*, vol. 33, no. 8, pp. 1710-1714.

Xie, X. L.; Yuan, T. W.; Zhou, X.; Cheng, X. C. (2018): Research on trust model in container-based cloud service. *Computers, Materials & Continua*, vol. 56, no. 2, pp. 273-283.

Xiong, L. Z.; Shi, Y. Q. (2018): On the privacy-preserving outsourcing scheme of reversible data hiding over encrypted image data in cloud computing. *Computers, Materials & Continua*, vol. 55, no. 3, pp. 523-539.

Zheng, S. Q. (2013): Research on application of cloud storage in the construction of electronic records center. *Archives and Construction*, no. 3, pp. 16-18.