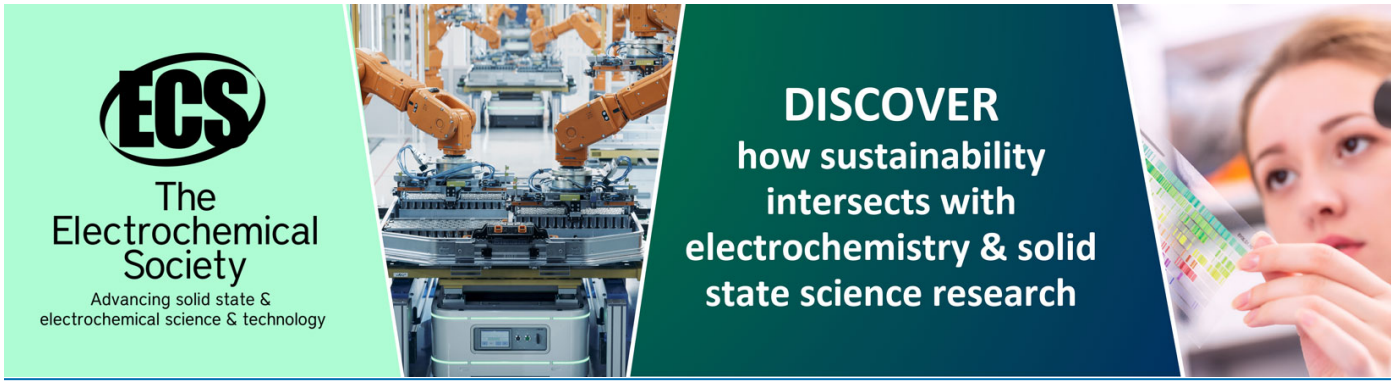
Database Construction in Computer Audit under Big Data Environment



Database Construction in Computer Audit under Big Data Environment

Yanhong Wu*

Guangdong Ocean University Cunjin College, Zhanjiang 524094, Guangdong, China

*Corresponding author e-mail: 08123323@cumt.edu.cn

Abstract. In the information age, the explosive growth of data forces users to deliver the data to the database service provider to manage. This paper mainly studies the database construction of computer audit under the environment of big data. This paper proposes a queryable database encryption storage scheme and a searchable multi-keyword encryption scheme under distributed environment. In this paper, the non-relational database is used in distributed storage in the form of key-value to improve the query performance. For distributed database, this paper proposes a queryable database encryption scheme. In this scheme, there are two encryption query modes: exact matching mode and range matching mode. Users encrypt and store data to the corresponding server nodes in these two ways, which can support rich queries. In the process of query, the user generates the corresponding query password for each server node, and then the parallel query can be realized. While protecting data privacy, the performance of the scheme is also well guaranteed.

Keywords: Big Data Environment, Computer Audit, Database Construction, Database Encryption

1. Introduction

Since the 21st century, the world has stepped into the era of big data, and information technology has been rapidly promoted. The McKinsey Global Institute has released a research report to the world, saying that the era of big data has arrived. The view of this research quickly gained the hot discussion from all walks of life around the world. Big data has a particularly significant impact on the development of the world. In addition to the Internet and social networks, both production and operation and people's lives will also be directly affected by e-commerce and the Internet of Things. Through enterprise operation and management, it can be found that both Internet technology and software or terminals have been used in a large range. From the perspective of enterprise management, it can be seen that the biggest impact of the increasingly popular development of computer technology is paperless office. Another name of paperless office is digitalization, which can be understood as the enterprise's business and a large number of financial data can be transmitted to the database through the form of data. Undeniably, the actual development of enterprise information data has brought great convenience. However, due to the long accumulation time and large overlapping scope, it is easy to significantly expand the scale of enterprise information data and increase the types of data, which will

further increase the difficulty of enterprise internal audit. In the traditional internal audit, the compliance inspection of accounting matters is the focus of work, and most of the research focuses on financial audit. However, after entering the "big data era", the growth rate of data is alarming and the overall growth pattern is explosive. Whether in the data collection or analysis stage, the internal audit department cannot ensure comprehensiveness and accuracy, nor can it meet the timeliness, and it is difficult to monitor the current problems of the audited organization in real time [1]. Therefore, in the enterprise information data auditing standards become more and more significant, at the same time, we must optimize the previous internal audit methods, in the face of large-scale data, the traditional audit methods have no time to deal with. In other words, the development mode of traditional internal audit, which always focuses on financial supervision and inspection, has been unable to meet the audit needs in the era of big data.

When computer technology and computer network in the society to achieve widespread use of the ability to talk about internal audit information system technology. In foreign countries, Internet computer technology has been born at a very early time and gained popularity. Therefore, the theory of audit information system has long been studied, compared with the domestic can be said to be a long way behind. Hatsu has come up with the idea of using computer information systems to track audit leads, research them and make bold predictions. With the development of computer, the society has entered the information age. Audit informatization can reduce the workload of auditors and screen out the information needed by auditors from the massive data. This assumption has laid the foundation of audit informatization [2]. Heatley believes that the complexity of computer auditing and the quality of auditors are not independent of each other, and then conducts an in-depth investigation on the relationship between the two. Through a large number of investigations, it is found that compared with the people with low quality, the people with high quality use and contact the computer more frequently. At the same time, he also pointed out that people prefer audit software with simple operation [3].

This paper proposes a searchable encryption scheme for distributed storage based on database encryption, which supports multi-keyword search. Because the scheme relies on distributed and symmetric encryption, it has high performance and can be applied to computer audit database effectively.

2. Construction of Computer Audit Database

2.1 Overview of Internal Audit of Computers

(1) Content of Computer Internal Audit

Internal audit informationization can be divided into two parts from different audit objects. The first part is computer-aided audit, which mainly carries out specific audit work through computer technology and software technology. The second is the information system audit, which is mainly to check and confirm the authenticity and compliance of the data input to the information system and the output data of the information system. Therefore, the main audit object of this part is the information system, so as to ensure the quality of enterprise management level [4].

(2) Characteristics of Computer Internal Audit

Along with the computer technology, software technology, the Internet and the rapid development of big data technology, enterprise management mode and the way in continuous improvement, also affected by the internal audit of enterprise informatization is in continuous development, through the information ability and the timeliness of enhancement, enterprise internal audit has become more efficient and accurate, and effectively reduce the audit turnovers.

First of all, it innovates the traditional audit mode. Through the increase of audit software functions and the synergistic effect of the remote Internet, the problem of strong subjectivity of audit workers in the traditional audit mode has been changed, and the audit mode that the same project can be jointly participated by many people has been realized.

Secondly, effectively improve the level of audit. During the audit of the project operation stage, the enterprise needs to carry out investigation work. Before the audit of the project, it needs to understand the content and plan of the project. During the audit process, it needs to continuously check the data results and revise them, and these investigations all involve information sources [5]. The investigation of the information can be transmitted to the remote client through the Internet, so that the remote investigation becomes a reality. Through this mode, only timely and continuous data collection in the process of the project can enable the investigation to be carried out as soon as possible, and the effective time of the audit is more abundant, so that the audit process is more orderly, and the results are more reliable. In addition, with the help of software and the Internet, the audit operation becomes simple and clear, and the audit quality of the project can be paid attention to throughout the whole process. The management can analyze and save the process data anytime and anywhere, and put forward valuable suggestions and ideas in time.

Thirdly, effectively improve the efficiency of audit. In the process of internal audit, the investigation in advance and the data collection and input of the project process will become fast, simple and systematic through information software and Internet technology. At the same time, the discovery of risks and loopholes in the audit process will be more timely, and even automatic warning, so that people's energy can be released for more valuable and useful work [6].

Finally, reduce audit errors effectively. Computer and software can not completely replace people to complete the audit work, but it has unique advantages in information transmission, calculation and sorting, through the use of these platforms to make it possible to audit from sampling to comprehensive audit, and then reduce errors.

2.2 Database Encryption Storage Scheme

Suppose a company has a large amount of internal data. In order to save the storage cost and computing cost, the company encrypts the data on the common cloud, such as Aliyun, which is relatively popular in the market. When the company needs to query data, the user only needs to generate the password to query the data and send it to the cloud server, so that the distributed server can query in parallel, and finally the server will return the data to the company [7-8]. This system consists of two entities, one is the user, that is, the company that obtains the cloud storage service in the above scenario, and the other is the cloud server.

The system deploys the encryption algorithm and the local index table algorithm on the client side. In addition, it deploys the algorithm to process the user query password and the Redis database for storage on each cloud server node. The cloud server receives the user-generated encrypted data and local index tables and stores them in the Redis database. In the process of query execution, the user generates query password (query trap gate), the server calculates the trap gate value according to the deployed algorithm and carries out query in the Redis database, and finally returns the result to the user.

When users get such a row-based data pattern, they first need to store the data in the cloud server after data separation and encryption. Data separation algorithm is a consistent Hash algorithm, which can store data evenly on multiple server nodes respectively [9]. For example, such algorithm is adopted in the REDI server cluster, so that the data stored on each node is similar and will not lead to a large performance deviation, thus eliminating the performance bottleneck:

$$Route(id) \to i. \tag{1}$$

The user uses the data separation algorithm to calculate the corresponding node where the data is stored. Where id is the increment type data used to uniquely identify a row of data and map the data to the corresponding node. In addition, Route () is the data separation algorithm.

$$P_{kl}(att \,||\, id). \tag{2}$$

The user computes the key value of the encrypted tag pair. Where KL is the encryption key of the label, P is the pseudo-random function, and ATT is the attribute of the data item.

$$E_{kv}(v \| Att \| id). \tag{3}$$

The user computes the value of the encrypted tag pair. Where, KV is the encryption key of data, E() is the symmetric encryption algorithm, ATT is the attribute of the data, ID is the data item ID of the data, and V is the data corresponding to the data item ID and the attribute. Finally in a data item id corresponding cloud server node, encrypted data item to < Pkl (Att | | id), the Ekv (v | | Att | | id) > such labels to encrypted storage form, and these key/value pair is out-of-order storage, so that can protect some additional information, such as the number of lines. In addition, if the cloud server is malicious, the wrong data is returned to the user, and the data is not the result of the user query. If the encryption method of the original scheme is followed, its scheme cannot verify the query results, but this scheme can solve the above problems [10]. Therefore, in the symmetric encryption of the above data, the scheme introduces the attributes of the data and its data item ID, so that the blind ciphertext can be realized, and the user can verify the results after getting them.

3. Functional and Performance Testing

In this section, we analyze the multi-keyword searchable encryption scheme for distributed storage from two aspects of function and performance. The experimental environment of this scheme is a computer with Inter Core i7-6700 CPU, 2.60GHz and 8Gbram. VMware virtual machine is configured on this computer, Ubuntu 14.04 system is installed, and Redis (V3.2.0) cluster is set up as cloud server node. In this scheme, all cryptographic primitives are written in C++, and OpenSSL (v1.01f) library is called. Pseudo-random functions are implemented using 128bit AES ciphertext.

The application scenario of this scheme is that in a distributed environment, the cloud server has multiple nodes, and each node stores different documents and corresponding keyword indexes. When a query is executed, MapReduce is responsible for distributing and reducing the query requests and results. In addition, this scheme provides multi-keyword search, so when the MapReduce server receives the results of all nodes, it needs to extract the ciphertext with S times occurrence, that is, documents satisfying the existence of S keywords, which will be returned to the searcher after summarization.

In this scheme, for the consideration of performance, we mainly adopt the symmetric encryption algorithm (such as AES) with low computing overhead and the secure pseudo random function, in which the pseudo random function also adopts AES ciphertext and their output size is 128bit. Before the query is executed, the data owner encrypts the document to the cloud server node in the form of the aforementioned tag pair. Because of the data separation algorithm, different nodes store different document ciphertext. Similarly, the locally encrypted index table is different for each node. In addition, after the index establishment is completed, the data owner needs to encrypt the intermediate variable counter table and store it in the server. This table can be used for the expansion of the database later. This scheme sacrifices some computing and communication costs in exchange for storage costs at the client end. When the query is executed, the user calculates the trap gate value of the keyword index for n server nodes respectively, that is, it calculates for ns times, where s is the number of satisfying keywords.

4. Performance Test Results

4.1 Index Size

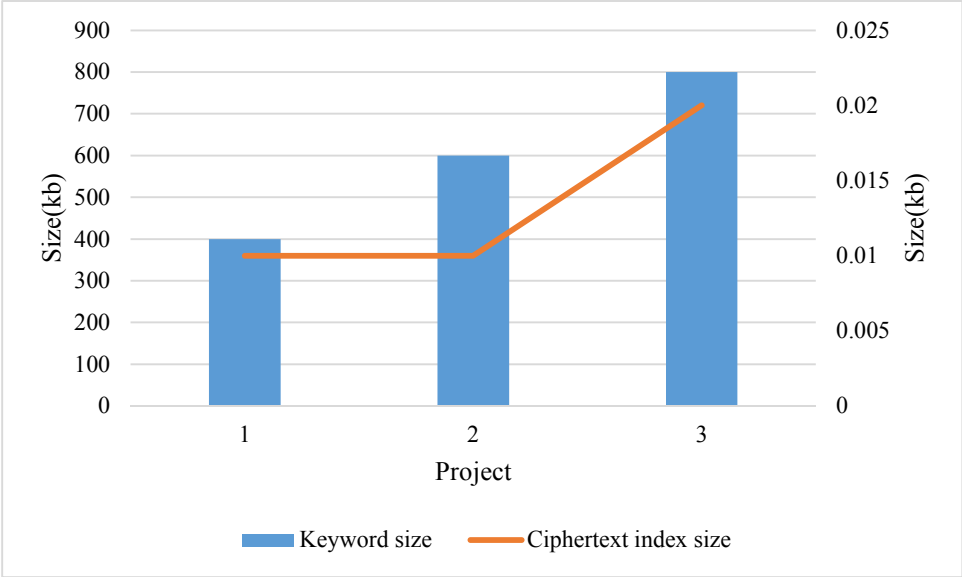


Figure 1. Index size comparison

As shown in Figure 1, the index size increases linearly from 0.010GB to 0.020GB as the number of keywords increases. Since this scheme will only index the documents with the keyword in each document, the local encrypted index table generated by this scheme is smaller than that of other schemes. In summary, it can be seen that this scheme reduces the computing and storage costs of the indexing process to a large extent.

4.2 Index Generation Time

Table 1. Time to generate the index

	2000	4000	6000	8000	12000
The time overhead	0.028	0.031	0.057	0.074	0.085

As shown in Table 1, when the number of data is 2000, the time cost is 0.028 seconds. When the number of data is 4000, the time cost is 0.031 seconds; When the number of data is 6000, the time cost is 0.057 seconds; When the number of data is 8000, the time cost is 0.085 seconds; When the number of data is 12000, the time overhead is 0.085 seconds.

4.3 Delay in Returning Query Results

Table 2. Returns the delay of the query result

	4	5	6	7	8	9
8K	7	5	4	3	2	1
16K	18	13	10	9	8	7
32K	32	27	23	21	20	19

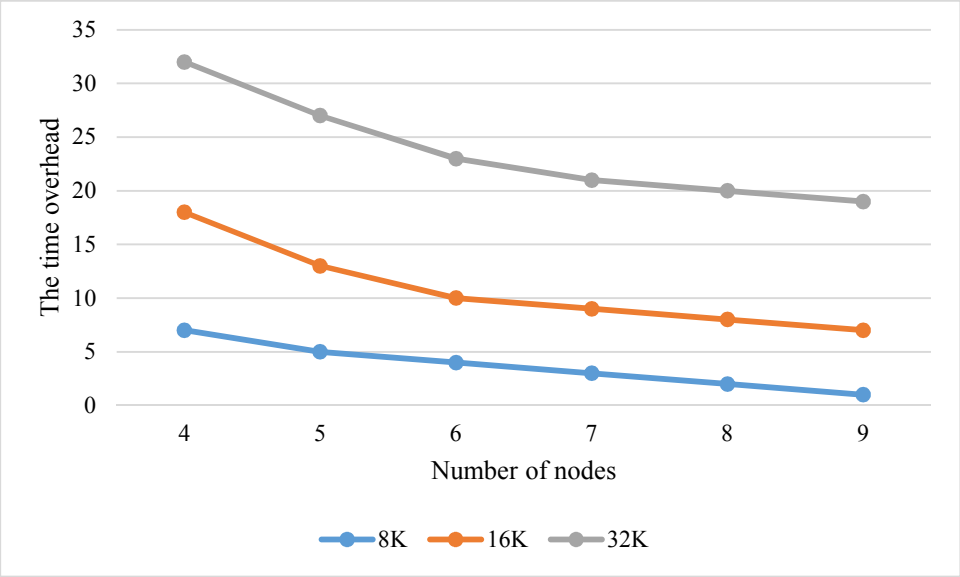


Figure 2. Returns the delay of the query result

As shown in Table 2 and Figure 2, the number of queries processed by the cloud server per second increases with the number of nodes. In particular, we can see that the throughput reached 265KB when the server nodes reached 8. The main overhead comes from the calculation of pseudo-random function and symmetric encryption. This is because the searcher needs to generate the trap gate value for each cloud server node when executing the query, and it needs to perform 2PRF× S × N operations, where S is the number of keywords and N is the number of nodes. The experimental results show that the scheme has good performance on a certain scale.

5. Conclusions

In the distributed database encryption storage scheme proposed in this paper, the encryption algorithm proposed in this paper can support keyword query and range query. The range matching algorithm adopts the improved ORE algorithm, which can prevent the leakage of data size information. The system runs in a distributed environment, and the client stores the corresponding ciphertext data and the corresponding encrypted index table on each node server. During the query execution, the client generates the query password for each node, and the server executes the query in parallel. It can be seen from the experimental results that the more nodes the server has, the higher the throughput of the system is. In addition, through these two encryption query algorithms, the system can support a wealth of queries.

Acknowledgement

This work was supported by the 2019 College-level Quality Engineering Project: Practical Teaching Reform and Research of Management Accounting Course (Project No.ZLGC2019013) ; the Innovation and Strengthening School Project: Application of Block Chain Technology in Supply Chain Finance in Beibu Gulf Region (Project No.:CJ20CXQX006); the 2019 Innovation and Entrepreneurship Project of GuangDong Ocean University Cunjin College: Zhanjiang Internet Supply Chain Finance Research under the Background of Big Data (Project 2020CJXYDCYB67)

References

[1] Wu T H, Huang S M, Huang S Y, et al. The effect of competencies, team problem-solving ability, and computer audit activity on internal audit performance [J]. Information Systems Frontiers, 2017, 19(5):1133-1148.

[2] Hatsu S, Ujapka M B, Mpimwood E D. An examination of the extent of implementation of the information security system and IT audit system in Ghananian Banks [J]. Journal of Mass

Spectrometry, 2015, 11(11):375-382.
[3] Heatley S K, Otto J R. Data mining computer audit logs to detect computer misuse[J]. Intelligent Systems in Accounting Finance & Management, 2015, 7(3):125-134.
[4] Shelmerdine S C, Singh M, Norman W, et al. Automated data extraction and report analysis in computer-aided radiology audit: practice implications from post-mortem paediatric imaging [J]. Clinical Radiology, 2019, 74(9):733.e11-733.e18.
[5] Pantiukhin I S, Zikratov I A, Levina A B. GRAPH-BASED POST INCIDENT INTERNAL AUDIT METHOD OF COMPUTER EQUIPMENT [J]. Naučno-tehničeskij Vestnik Informacionnyh Tehnologij, 2016, 16(3):506-512.
[6] Jammel, Mohammed, Ali, et al. Computer Audit Programs (Software) And A New Variables Sampling Concept A Proposed Approach Empirical Study [J]. International Journal of Applied Engineering Research, 2018, 13(7 Pt.6):5491-5500.
[7] Huertas Celdrán, Alberto, Dólera Tormo, Ginés, Gómez Mármol, Félix, et al. Resolving privacy-preserving relationships over outsourced encrypted data storages [J]. International Journal of Information Security, 2016, 15(2):195-209.
[8] Chen Y C, Wu Y S, Tzeng W G. Phrase Search for Encrypted Cloud Storage [J]. Journal of Information Recording, 2018, 34(2):401-417.
[9] Yeting G, Fang L, Zhiping C, et al. Edge-Based Efficient Search over Encrypted Data Mobile Cloud Storage[J]. Sensors, 2018, 18(4):1189-1189.
[10] Rajput A S, Raman B. Cloud based image color transfer and storage in encrypted domain [J]. Multimedia Tools & Applications, 2018, 77(7):1-29.