Simulation of Privacy and Security Features of Big Data Based on Data De-Redundancy Technology

Fang Liu

* Corresponding author: 273237692@qq.com

School of Humanistic Medicine, Anhui Medical University, Hefei, Anhui 230032, China

ABSTRACT: This project takes big data as the research object and data redundancy as the entry point to study the new privacy encryption technology for big data. In this project, Bloom filtering method is used to reduce the dimension of massive data, hash function is used to process the redundancy algorithm, and bit selection is carried out on the bit sequence to optimize the number of extension functions. In addition, based on the existing analysis method based on extensible key, the optimal ciphertext strategy is adopted to realize the security sharing of cloud data more efficiently, reduce the storage scale of data units in attribute space, and reduce the number of shared parameters, so as to ensure the security of cloud data. Experiments show that this method can effectively intercept the attacked information, enhance the security of big data, and ensure the validity of the password, which is a good and feasible method [1].

Key words: privacy big data; Redundant data; Secure shared access; Attribute encryption simulation

1 Introduction

At present, with the rapid development of mass data, people's ability to protect its privacy is becoming stronger and stronger, but the use of attribute cryptography technology for its security protection ability is less and less. Massive data usually consists of various devices [2], and most users cannot achieve the security function embedded in personal information if they adopt the attribute-based password technology in the face of massive data. It is very easy to use the page channel to display the key or other secret remote parameters to attack, resulting in the process of encryption, only the key detection, but not strictly consider the security of the encrypted random number [3].

Due to the continuous rapid growth of massive data, and the Internet client that can collect all the sensitive information such as privacy behavior and state, the development of big data has encountered serious challenges. This is mainly because big data contains not only the protection of users' personal privacy, but also other sensitive information such as users' habits and preferences. Once massive data is destroyed, it will cause great damage to users' privacy. Therefore, it is of great significance to study the data protection of users' private property and personal safety as well as web encryption technology. In this context, the cryptography technology based on cryptography can effectively protect the user's geographical location information. However, in the case of increasing mass data, the conventional cryptographic

ICBBEM 2023, May 19-21, Hangzhou, People's Republic of China

DOI 10.4108/eai.19-5-2023.2334436

system has been unable to meet the requirements of mass data. However, due to a large number of collected information contains a lot of error information and lack of sufficient security, the application of this technology is greatly restricted. This project intends to adopt multi-level cloud cooperation and fusion mode, combined with homomorphic cryptography technology, to achieve multi-level privacy, so that real-time information on the network can be encrypted and obtained [4].

2. Big data redundancy elimination algorithm based on Bloomfilter

At present, with the rapid development of digital information, its development also presents diversification, although the data compression technology can reduce the space occupied by unnecessary data in the file, but because of the rapid development of data, it will occupy more running memory, so this paper adopts a redundant data elimination method to remove the redundant information in the encryption process [5].

2.1 Construction method

In the calculation, Bloomfilter is a collection of multidimensional data containing m, and each group of data in the collection is 0. So, in order to can better show n a collection of data element S = | x, x,,... , x}, using k hash functions to map data elements to 1.... In the range of m [3], it is assumed that an ordered sequence in a certain data segment D is judged as shingle, then the shingle set is defined as S (Dw) in this data segment. Thus, Bloomfilters can be constructed [6].

1) Construct a large capacity of one m bit bf and initialize the data so that the value of the data is 0;

2) Select two hash functions from a set of data that are suitable for mapping functions, initialize them, and represent them as hash 1 hash 2;

3) Randomly extract a string from a set of data, perform hashing 1 and hashing 2 to summarize the value of the big data, and set bf as 1[7];

4) Output the bf as the value of the characteristics of the large-capacity file.

2.2 Solving the misjudgment rate

Bloomfilter calculation can save a lot of storage space, allowing a small number of errors. In addition, because the redundant data is greatly reduced after Bloomfilter calculation, the misjudgment rate of the overall calculation process will also decrease. The more redundant data is eliminated, the higher the accuracy, but the more storage space is used.

When S = | x, x,... , x_ | set each data element is mapped to the corresponding data range, can further to get out of a particular still zero probability p, computation formula is as follows

$$p' = \left(1 - \frac{1}{m}\right)^{kn} \approx e^{-n/m} \tag{1}$$

So if I want to map the set S completely into the array. So you're going to have to do kn hashes. This paper uses the common approximation of e, as shown below

$$\lim_{x\to\infty}\left(1-\frac{1}{x}\right)^{-x}=\mathrm{e} \tag{2}$$

According to the calculation result of the above formula, if ß is represented as the proportion of 0 data in the set, then the expected value of β is E(B)=p'.

In order to facilitate subsequent calculation, let p=e-nk/m, so that we can know the magnitude of error rate in the case of known value

$$(1-\beta)^k \approx (1-p')^k \approx (1-p)^k \tag{3}$$

$$\mathrm{f}=\left(1-\mathrm{e}^{-\mathrm{nk/m}}\right)^{\mathrm{k}} \tag{4}$$

According to the calculation results, k is set to 2 in the equation. If the above equation is needed to know whether the data v fully conforms to the requirements of the set, k function calculation can be carried out. The calculation results show that the mean value of h(r) is 1(1≤i≤k), then y belongs to this set; otherwise, it will be misjudged.

2.3 Determining the optimal number of hash functions

When calculating BloomFilters, multiple hash bearing numbers are used to map the collection to the array, and there is a certain error rate. Therefore, we need to select the optimal number of column expansion functions before the calculation, so as to minimize the error rate in the subsequent data retrieval [8]. Similarly, if you use more than one hash function when doing a calculation, you will generate data that is not part of the set, resulting in a state of 0 for error. In addition, if the number of hash functions is too small, more zeros will appear in the bit data table [5].

Let g=kln(1-e-"), in which, when g takes the minimum value, relative f will reach the highest value. In this way [9], p=e-nk" can be converted with g, and then

$$\mathrm{g}=\frac{\mathrm{m}}{\mathrm{n}}\ln(p)\ln(1-\mathrm{p}) \tag{5}$$

According to the above calculation, the similarity between the data objects is calculated according to the Hamming distance and cosine similarity formula. If the two values are the same, the file is replaced by the index of the saved file: If the two values are different, the file is stored and the hash table is updated to add the new hash value to the file [10].

3 Privacy big data attribute encryption

Since the privacy leakage problem mainly comes from the access process, this paper discusses the common cloud access scenario to analyze the data attributes, in order to achieve the security, scalability and precise control of encrypted data access.

3.1 Data attribute analysis under data sharing access

The attribute encryption mechanism is used to control the access to the system and realize an effective symmetric encryption. Building the appropriate system and module models is shown in Figure 1.

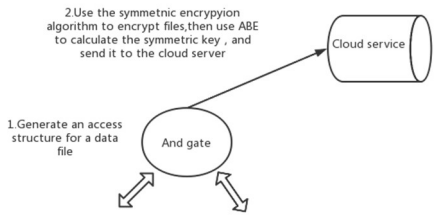

Figure 1: Schematic diagram of the data sharing access scenario

The big data access network consists of data user, data end user and network transmission server. In order to obtain the access data track searched by users in the network, it is necessary to use the server used by network end user and collect the data track and access track. Abe selects a set of properties that are used to describe a container for a unit of information.

The student information network consists of student types, training levels, majors and grades. The researchers in the computer field are taken as a group and taken as an indefinite term. Since the attribute value in big data is taken as the standard according to the classification of computers, the attribute value is the standard of classification and value. The third type of students and attribute value are holidays. It is a property of the user [7]. In Figure 2, the cloud computing data structure access based on the feature password is shown.

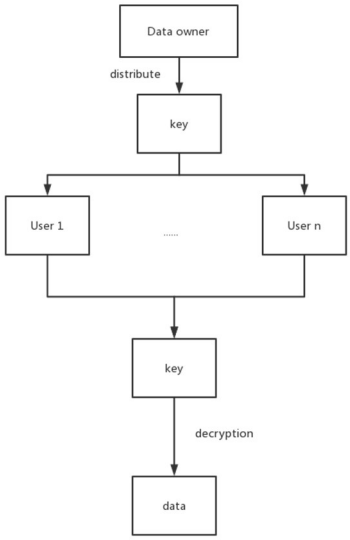

Figure 2: Data share access diagram based on attribute encryption

According to the design goal of the system, Abe cryptosystem is used to achieve accurate access to data, and Abe is the key in the public cryptosystem. The cryptosystem of functional cryptosystem is a huge algebraic system, and the cryptosystem in the cryptosystem is a huge cryptosystem. In this case, we adopt a new cryptographic algorithm, which is a combination of

symmetric cryptographic algorithm and asymmetric cryptographic algorithm, to overcome the shortcoming of high computational complexity. According to the design purpose, the corresponding system structure is obtained, and its schematic diagram is shown in Figure 3.

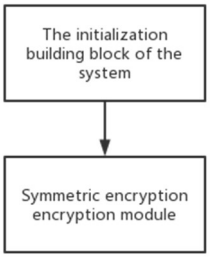

Figure 3: System functional architecture diagram

3.2 Mathematical model of attribute encryption

If G and G represent two completely different multiplicative groups, and g is described as the generator of G, then bilinear mapping can yield eGxG G. Due to the irreversible property of eGxG→G calculation, If so the calculation results meet the characteristic condition, you can call it bilinear | a6} please - Z *, e (g, g ') = e (g, g).

If Setup(1) is represented as the generator of a double-line map, then there is at least one security parameter 1, so that any output parameter set qG,GT,g,e} can be obtained.

On the network, if x is used for data transmission, m can be expressed as the true length of x, and if there is a polynomial p, p (m) will be the end point in the cryptographic operation. In password calculation, the hash function with password features belongs to a special extension function, which can adjust the length of the string arbitrarily by different operation values. Then set it to the desired length. If h represents a hash function and n represents the real string length of data output, h has the following characteristics: Extensibility: In password calculation, the hash function value range is within the interval range of [0.2"] according to different data points, and the binary data and polynomial time are the same; Unidirectional: if the hash function value w is known through calculation and x is further obtained, all necessary conditions of the formula h (x)= w must be satisfied; Inefficient calculation: Under the condition that x≠ gamma, find two kinds of large data for cryptographic operation, and make h (x)= h (y); Actual validity: When the value of encrypted data is known, h (x) can be used to obtain the actual computation time.

In view of the characteristics of data security, scalability and fine-granularity in the cloud environment, this project plans to adopt the attribute password based CP-ABE cryptosystem to meet the security, scalability and fine data access requirements of data in the cloud computing environment. That is, an access structure that contains data controlled by the data owner is added to the cryptographic data. Poni is the formal definition of the Generalcp-abe algorithm, which can be decrypted by users whose private key attributes conform to the encrypted access structure [9].

The password processing process of CP-ABE algorithm is as follows:

Data encryption initialization Setup (1) : During initialization calculation, a security parameter is entered first, and then the calculation parameter MS of the primary public key MP and primary private key of the encryption system is returned. Password generation algorithm KeyGen (MSU) : During calculation, the corresponding master public key MS and a set of attributes U of the big data are input respectively, and the decryption key of each data in the set of attributes is generated accordingly. Big Data attribute encryption algorithm Encrypt (MPSM) : In the process of big data encryption calculation, MP big data message M and network access structure M are first input, and on this basis, a private key SK with only data users is established. After that, the attribute set associated with the private key meets the necessary conditions of UI= S and is ciphertext C [10] that is convenient for decryption. Attribute decryption algorithm of big data: MPSK (MPSK) : After password processing, the big data is decrypted, mainly referring to the main public key MP, decryption ciphertext C, and the user's private key SK.

4 Simulation Research

To further test the encryption method presented in this paper, a PC with AMDAthlon (tm)IX3.3.10GHz2GB memory and C++ programming language was used in the simulation environment. Using 1 TB of data for the test, regardless of the different test environment, the total number of files obtained is 905797, and the average file capacity is 1105 kB. The methods in this paper are compared with those in references [1]2 from the perspectives of data redundancy elimination and key generation time cost in encryption process. It can be found that, among the additional storage costs of detection schemes, the method in this paper is the largest, and its redundancy storage is the highest, reaching 60%, followed by literature [1], which will bring about 50% additional storage costs at most, while the detection scheme of the method in literature [2] brings about the lowest storage costs. It is obvious that our algorithm is better than our two algorithms. Our algorithm is the most effective, and our algorithm is not as effective as the password error.

Based on the efficiency comparison, the computational cost of adding and decrypting is compared. A comparison of the time to generate the primary public key and the primary private key by three different algorithms is given. The test results show that compared with CP-ABE, the method in literature [1] and [2] takes more time to generate master public key and master private key, resulting in low efficiency of encryption and decryption.

5 Conclusion

With the high dependence of human on Internet, the information exchange and storage using Internet technology also produce security problems at different levels. Therefore, this paper proposes a password security scheme for large users. However, there are some drawbacks to this approach. For example, there may be some noise in the data set that will cause the final password to be wrong, but this noise will not affect the password effect much. In addition, this topic will also carry out the research of big data preprocessing methods with the aim of continuously reducing the problems of error and unrestricted decomposition faced by ciphertext data security, ensuring the security of users' personal information, and promoting the in-depth development of China's network technology.

References

[1] Chen S G. Research on smart grid security and privacy protection data aggregation based on fog computing [J1. Journal of Nanjing University of Posts and Telecommunications: Natural Science Edition, 2019.39(6):62-72. (in Chinese with English abstract)

[2] Xiong Jinbo, Zhang Yuanyuan, Tian Youliang, et al. Cloud data security deduplication based on role-symmetric encryption [J1. Journal of Communications, 201839(5):63-77.

[3] Wang Liying, Xia Yuhong, Xuan Ye, et al. Identification and elimination of redundant point cloud data from multi-echo Lidar [J1. Science of Surveying and Mapping, 201843(6):137-141.155.

[4] GU Z Z. Research on the development of data deduplication technology [J]. Information and Communication, 2018,(1):158-159. (in Chinese)

[5] Zhang Guipeng, Chen Pinghua. A data Security Deduplication Scheme Based on Merkle Hash Tree in Hybrid Cloud environment [J]. Computer Science, 2018.45(11):187-192.203.]

[6] Li Chuan, Wang Zhi. Research on multi-level Encryption Simulation of User privacy data in Big Data environment J1. Computer Simulation, 2019.36(11):159-162.

[7] Li Yuancheng, Zhang Pan, Zheng Shiqiang. Data privacy protection based on Empirical Mode decomposition and homomorphic encryption [J]. Power Grid Technology, 2019,43(5):1810-1818. (in Chinese)

[8] Wang Caifen, Cheng Yudan, Liu Chao. Full homomorphic data encryption aggregation scheme based on WSN J]. Computer Engineering, 2018,44(12):190-195.

[9] Li Zongyu, GUI Xiaolin, Gu Yingjie, et al. Homomorphic encryption technology and Its Application in Cloud Computing Privacy Protection [J]. Journal of Software, 201829(7):8-29. (in Chinese)

[10] Lin Yuxiang, Duan Xindong. Digital information encryption technology for network privacy protection [J]. Modern Electronic Technique, 2018,41(9):45-48,53.