Security vulnerability and encryption technology of computer information technology data under big data environment

Weijie Chen¹*, Guodong Chen², Yanheng Zhao¹ and Jinghua Zhang¹

¹ State Grid of China Technology College, Jinan, Shandong, 250002, China
² Shandong Electric Power College, Jinan Shandong, 250002, China

*Corresponding author’s e-mail:16865@etransfar.com
(ORCID: 0000-0002-6166-6813)

Abstract. In recent years, the traditional encryption storage technology and management methods have been difficult to satisfy the demand of big data in terms of speed, capacity, storage efficiency and security, along with the rapid increase of data volume. Data security is facing great challenges and challenges. To solve this problem, this paper proposes a de-redundancy method based big data encryption algorithm. Firstly, bloom filter technology is used to decrease the dimension of big data. Based on its remarkable advantages, bloom filter space and query efficiency and variable-length block detection scheme with less additional overhead, duplicate data detection and deletion algorithm suitable for large data is designed and implemented. Secondly, based on eliminating redundancy of big data, combined with the characteristics of the elliptic curve encryption algorithm and AES block cipher mode in terms of computing speed, parallelism and security, an encryption approach based on data de-redundancy technology is presented. According to the evaluation system of the encryption algorithm, the encryption time, keyspace and security of the new algorithm are evaluated. Experimental results reveal that the scheme has better security, and the presented approach can increase the speed of the encryption process, and decrease the encryption time and consumption.

1. Introduction

The network information flow has increased dramatically from GB to TB, Pb, EB and even ZB, since the expansion of modern society. The traditional data encryption storage and management methods have been difficult to satisfy the demand of big data in terms of encryption speed, storage capacity and security. Security and privacy has obtained wide investigation from the scientific and industrial circles, in big data environment. In the big data environment, the data sources are extremely rich, with the characteristics of super strong data set, strong noise, multi redundancy and unstructured. When encrypting big data, the amount of data to be calculated and analyzed is extremely large. Therefore, the encryption processing of big data must balance its characteristics.

Due to the single source of traditional data, the quantity of data is required to be encrypted, and analysed, so it can be processed by using single machine or low complexity encryption mechanism. The meaning of computer software security vulnerability is that the computer system has the defect of malicious attack or intrusion, which leads to the possibility of network information leakage in the transmission. Computer software vulnerability can be divided into security vulnerability and functional defect. The security vulnerability will not affect the software function of computer system,

Content from this work may be used under the terms of the Creative Commons Attribution 3.0 licence. Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.
Published under licence by IOP Publishing Ltd
but it may be exploited by hackers, leading to computer system paralysis Poisoning and so on, even cause the leakage of confidential information in the computer system.

At present, there are two ways to encrypt big data. One is to integrate the encryption technology into the equipment, such as common encryption card, private network, private encryption machine, etc. this method is used in the demand environment of function generalization. The unified security strategy is adopted for different business logic and functions distributed in different regions to realize unified security measures, such as 1pse C VPN, SSL VPN cipher device, the whole disk encryption device, such as the commonly known environment encryption technology; the other way is to provide only password services, allowing specific applications to complete relatively personalized data decryption functions, such as file encryption, source encryption, and so on.

On the basis of the in-depth analysis in the existing encryption technology, this paper explores a Bloom filter technology-based big data de-duplication algorithm, and a big data encryption algorithm based on data de-redundancy technology. The main innovations of this paper include: (1) presenting a big data de-duplication algorithm based on the Bloom filter. Based on the characteristics of fast computation speed of full file detection method, fast matching speed of Bloom filter structure with quick match merits and less extra cost of variable length block detection scheme, a redundancy elimination algorithm suitable for big data is designed and implemented in terms of time cost, space cost and matching accuracy. (2) A big data encryption scheme based on data de-redundancy technology is presented. On the basis of eliminating redundancy of big data, this paper proposes a big data encryption algorithm based on data de-redundancy technology, combining with the characteristics of the elliptic curve encryption algorithm and five working modes of AES block cipher in terms of computing speed, parallelism and security. The effectiveness of the algorithm is analysed and verified by experiments.

2. Data encryption framework and principle for big data

2.1. Data encryption framework under big data environment

![Encryption and decryption framework in the big data environment](image)

**Figure 1.** Encryption and decryption framework in the big data environment
Due to the high-performance requirements for storage resources and computing resources, the traditional encryption algorithm and processing tools are no longer as efficient as before. At present, there are two main ways to deal with big data encryption:

1. Data sampling. In order to make the data encryption processing more targeted, through the method of data sampling, collecting the key information of the data domain, objectively will reduce the data scale, in order to speed up the processing. However, in some areas, the application of sampling technology only encrypts part of the information in big data, so it is difficult to prevent attackers from mining key information and sensitive information from other information.

2. Divide and rule. By using distributed computing technology, big data is encrypted on different computers, which effectively improves the speed of big data encryption and decryption. It is common to use MapReduce mode for the parallel processing of massive data.

But their encryption framework is the same, as shown in Figure 1:

Big data encryption system can be expressed as a five tuple: $\langle P, C, K, D, E \rangle$, where

1. $P$ is the interactive data set;
2. $C$ represents encrypted database;
3. $K$ is the key container;
4. $E$ is the encryption equation: the designed $E$ and the key $k \in K$ are used to encrypt the interactive data $p \in P$, and the encrypted data $c \in C$ is obtained, i.e., $E: P \times K \rightarrow C$, abbreviated as $E_k(P) = C$;
5. $D$ is the decryption equation: the $D$ and the key $k \in K$ are used to decrypt the encrypted data $C$, and the decrypted data $p$ is obtained. That is, $D: C \times K \rightarrow P$, abbreviated as $D_k(C) = P$.

2.2. Research on encryption algorithm based on Elliptic Curve Cryptosystem

With the development of large integer decomposition technology and parallel processing technology, the current public-key encryption system must further increase the key length to ensure the security of the algorithm, which will make the encryption speed slower and the calculation more complex. Elliptic curve cryptography (ECC) can achieve high security with low computation cost, storage cost and delay (high encryption and signature speed). It is especially suitable for the situation of limited computing power and IC access, limited bandwidth and high-speed implementation. ECC encryption algorithm can use a shorter key than RSA encryption algorithm to obtain the same level of security, which greatly reduces the amount of calculation.

When the elliptic curve encryption algorithm is implemented, it is regarded as a four-layer structure. The lowest layer performs the modular operation, that is, the arithmetic operation. The second layer implements two sets of correlation operations: point addition operation and multiple spot operation. These two operations use the arithmetic operations provided by the bottom layer. The third layer uses the group operation provided by the upper layer to implement point multiplication. The top layer implements the actual protocol. Its operation speed determines the performance of the entire encryption system. Therefore, improving the speed of the point multiplication operation is equivalent to improving the speed of the elliptic curve encryption algorithm.

3. Big data encryption algorithm based on data de-redundancy technology

3.1. Big data De-redundancy algorithm based on Bloom filter

Bloom filter is a kind of random architecture that compresses parameter space based on multi hash function mapping. It uses vector $V$ to express a set concisely and can judge whether an element belongs to this set. For showing $\mathcal{S} = \{x_1, x_2, \cdots, x_n\}$, bloom filter uses $k$ hash maps to match each
individuality in the series range of 1,..., M. For any individuality x, the condition $h_i(x)$ mapped by the i-th hash map to vector V is defined as $i(1 \leq i \leq k)$.

Bloom filter can use Hamming distance to measure similarity, and can also use vector space Jaccard and other methods to measure similarity ratio. The solution of Hamming distance is to compute the quantity of different corresponding bits from two binary sequences.

$$\text{Jaccard}(x, y) = \frac{|X \cap Y|}{|X \cup Y|} = \frac{\sum_{i=1}^{k} X_i Y_i}{\sum_{i=1}^{k} X_i + \sum_{i=1}^{k} Y_i - \sum_{i=1}^{k} X_i Y_i}$$

(1) Bloom filter saves a lot of storage space by allowing a small number of errors. Due to the limited false positive rate, the accuracy of its similarity algorithm depends on the size of Bloom filter itself. The larger the bloom filter is, the higher the accuracy is, but the consumption of storage space is also greater.

Bloom filter uses multiple hash map to match a bit group, and there is a certain error rate, so it is required to select the best quantity of hash combination to minimize the error rate in element query. If more hash combinations are used, the trend of getting 0 is higher when querying an element that does not belong to the set; on the other hand, if hash combinations are less, more zeros will appear in the bit array. The best quantity of hash combination can be calculated according to $f = (1-e^{-nk})$.

3.2. Design of big data encryption algorithm based on data De-redundancy technology

Combined with the characteristics of ECC methods and block cipher technology, and combined with the data de-duplication scheme, an encryption scheme for the big data scene is proposed. Among them, ECC is used to encrypt the content of a hash table with the elimination of processed redundancy scheme.

![Figure 2. Big data encryption algorithm model based on data De-redundancy technology](image)

The encryption algorithm is implemented by MapReduce distributed parallel programming architecture of Hadoop big data platform (Figure 2). Map function is used to eliminate mixed data redundancy and AES hybrid encryption operation, and reduce function completes the merging of encrypted data.

The process of data encryption is shown:

1. Data pre-processing can delete redundant data and improve encryption efficiency
2. The ciphertext is calculated as $C(S1)$. The information interaction under the network platform involves the receiver (Bob) and the sender (Alice),
3. The big data plaintext processed by data de-duplication technology is encrypted by Advanced Encryption Standard AES, in which the working mode of block cipher algorithm is CTR mode with
faster encryption and decryption speed; the block length supports 128/192 bits, and the password length is 128/192. The corresponding wheel number R is 10/12/14. When the key length is 128 bits, the number of rounds is 10.

(4) The data and key are decrypted.

4. Experimental Results and Analysis

4.1. Big data experimental environment

The experimental environment is to adopt multiple virtual environment on the high-performance PC server using the virtual machine software VMware Workstation 12, and deploy Hadoop big data platform on the virtual machine for implementation. Each virtual machine is configured with a single-core CPU and 1 GB of memory. Linux system is ubuntu16, Hadoop version is 2.7.3, java version is based on jdk8, and IDE is eclipse 3.8. Two big data text files with the sizes of 1 GB and 2 GB were selected for the experimental data set, and the size of map blocks was set to 128 MB by default.

4.2. Time-consuming analysis of encryption

In the experiment, the encryption time of four kinds of encryption algorithms (ECC, CBC, CFB, OFB) and the algorithm of this scheme are statistically analysed. Figure 3 and Figure 4 show that the encryption time of the new scheme is the sum of the encryption time of the hash table in the pre-processing stage and the encryption time of big data after pre-processing; the decryption time of the new scheme is the sum of the decryption time of hash table and the decryption time of big data after pre-processing. In addition, the key length of AES encryption system in Figure 4(a) and 4(b) are 128 and 192 bits, respectively, and the statistical results of encryption time are as follows:

The experimental results show that, under the same data size, the encryption and decryption time of the four working modes of ECB, CBC, CFB and OFB are very close. In Figure 4, the curves overlap. The new algorithm shows good performance in encryption and consumes less time. With the increase of data volume, its advantages become more and more obvious. At the same time, the data De redundancy scheme and block cipher mode CTR used in the new algorithm have good parallelism, which is suitable for the implementation of distributed and parallel computing platform, and have good scalability. At the same time, when the key length is 128 bits, the encryption operation of 50GB data is carried out, and the encryption time of four working modes (ECB, CBC, CFB and OFB) reaches 3185, 3360, 3204 and 3180 seconds respectively, which takes a lot of time. The new algorithm shows good performance and consumes less time, only 2186 seconds.
Figure 3. Comparison of encryption time

4.3. Statistical properties of ciphertext

The histogram of encrypted ciphertext is analyzed, as expressed in Figure 4. The histogram of original plaintext text shows a certain regularity of distribution statistical characteristics, while the text data histogram after encryption in this paper expresses a random distribution condition of noise like, and the experimental results reveal that the scheme obtains better statistical features of ciphertext, it is distributed in strict boundaries.
Figure 4. Statistical histogram of plaintext and ciphertext data

Key sensitivity is an important index of algorithm security. Experiments show that the initial plaintext data can be restored, since the key is matched exactly. When the error of the key parameter is 10, the
key parameter will have an avalanche effect on the decryption ciphertext, and the original plaintext cannot be decrypted correctly. At the same time, the ciphertext with a great difference from the plaintext is generated. At this time, the histogram is shown in Fig. 4 (b). Experimental results show that the key parameters have an avalanche effect on ciphertext, and the algorithm has good key sensitivity.

5. Conclusion

This paper proposes a de-redundancy method based big data encryption algorithm. Firstly, bloom filter technology is used to decrease the dimension of big data. Based on its remarkable advantages, bloom filter space and query efficiency and variable-length block detection scheme with less additional overhead, duplicate data detection and deletion algorithm suitable for large data is designed and implemented. Second, based on eliminating redundancy of big data, combined with the characteristics of the elliptic curve encryption algorithm and AES block cipher mode in terms of computing speed, parallelism and security, an encryption approach based on data de-redundancy technology is presented. According to the evaluation system of the encryption algorithm, the encryption time, key space and security of the new algorithm are evaluated. The developed encryption method provides a convenient approach to make up for the loopholes in big data. In addition, to eliminate the security loopholes of computer information, the research of data encryption method should adhere to the principle of "keeping pace with the times", so that it can always satisfy the actual demands.

References

[1] Verma, Surabhi, S. S. Bhattacharyya, and S. Kumar.(2018) "An extension of the technology acceptance model in the big data analytics system implementation environment." Information Processing & Management 54(5):791-806.

[2] Chen, Genlang, et al.(2018) "Generic user revocation systems for attribute-based encryption in cloud storage." Frontiers of Information Technology & Electronic Engineering, 19(11):64-86.

[3] Gardiner, Adrian, et al.(2018)"Skill Requirements in Big Data: A Content Analysis of Job Advertisements." Journal of Computer Information Systems 58(4):374-384.

[4] Tai, Lingjuan, L. Li, and J. Du.(2019) "Multimedia based intelligent network big data optimization model." Multimedia Tools and Applications , 78(4):4579-4603.

[5] Lujie, Li, et al.(2018)"Application of Big Data Technology in Evaluation of Operating Status of High-pressure Hydrogenation Heat Exchanger." China Petroleum Processing & Petrochemical Technology, 120(3):20-26.

[6] Hejun, Zhu, et al.(2018)"Online and automatic identification and mining of encryption network behavior in big data environment." Journal of Intelligent and Fuzzy Systems, 34(2):1111-1119.

[7] Wu, Chuanrong, et al.(2018) "Knowledge Structure and Its Impact on Knowledge Transfer in the Big Data Environment." Journal of Internet Technology, 19(2):581-590.