A CP-ABE Algorithm Based HDFS Distributed Storage Data Security Solution

Nengneng Li¹²,*, Yingcun Wang¹², Xiwei Xu¹², Xiaoyang Liu¹²
¹Weifang Vocational College, Weifang City, Shandong Province, China.
²No.06588, Hai’an Road, Science and Education Innovation Park, Binhai Economic and Technological Development Zone, Weifang City, Shandong Province,262737,China.

*Corresponding author e-mail: linengneng1992@163.com

Abstract. With the continuous development of technology, traditional data encryption technology has been cracked, existing data security measures can no longer meet the user's data security needs, and the problem of data security remains the primary issue facing cloud storage. Based on the analysis of cloud storage data security, this paper proposes a CP-ABE-based HDFS distributed storage data security solution to the problems of traditional attribute-based data encryption technology (CP-ABE). After analysis of the algorithm, the algorithm can ensure the security of the data stored on the cloud platform users through the attribute-based encryption algorithm.

1. Introduction

With the proposal of “Smart Things Moving in the Cloud”, more and more cloud computing, Internet of Things, mobile internet, big data and artificial intelligence projects have taken root and promoted the transformation of traditional industrial production, management and marketing models with new technologies, new formats and new models. The popularity of cloud computing, big data and other applications, as well as the explosive growth of data, has made more and more enterprises and individual users choose to use cloud storage to process data. However, problems such as data security have arisen. Many large-scale data leakage incidents have made older and more people wait and see the use of cloud servers to store data. Many users have refused to use cloud storage. The security problem of cloud data has become the primary issue hindering the development of cloud computing. Therefore, solving the security problem of cloud storage is a prerequisite for the development of cloud computing.

This article first introduces the storage principle of Cloud Computing Distributed Data Storage System (HDFS), and proposes an application of attribute-based encryption algorithm (CP-ABE) to the data distributed file storage system, which not only solves the problem of data security, It also solves the problems of the distributed file system itself.

According to statistics, on average, more than 2 million users worldwide search information on Google every second, Twitter handles more than 340 million tweets every day [1], and Facebook users share more than 4 billion daily content. In the context of the big data era, the growth rate of data volume has far exceeded the capacity growth rate of data storage devices [2]. Faced with the explosive growth of data volume, traditional data storage management systems have been greatly threatened [3]. Seeking more efficient, stable, and secure data storage methods has become a research in many fields such as data storage and data management Focus.
In the face of massive data storage and processing requirements, Hadoop has designed a new open-source distributed file storage system HDFS (Hadoop Distributed File System) [4] with reference to Google Distributed Data Storage System GFS [4]. Availability, high scalability, and low server configuration requirements are widely used in the storage and management of large data sets. However, users store data in a distributed file system, not only facing the risk of data loss caused by a single point of failure, but also the security issue of data being stolen by illegal users. The symmetric encryption technology and the asymmetric encryption technology in the traditional data storage system have found a cracking method. In this paper, the more secure attribute-based encryption technology (CP-ABE) is applied to the Hadoop distributed file storage system. The storage of a large amount of data effectively guarantees the security of the data.

2. Related theories

2.1. HDFS architecture and working principle
HDFS, as one of the core modules of the Hadoop cluster, is mainly composed of a master node Namenode and several data nodes Datanode. Among them, Namenode, as the manager of the file storage system, mainly manages and stores in the form of metadata, which is used to maintain the file system name and manage client access to files. The Datanode is responsible for storing multiple data blocks into which the file is divided. The storage architecture of HDFS is shown in Figure 1.

![HDFS architecture](image)

**Figure 1.** HDFS storage architecture

2.2. Attribute-Based Data Encryption Algorithm (CP-ABE)
ABE belongs to the public key encryption mechanism. The traditional public key encryption mechanism must obtain the user's public key certificate before it can be encrypted, and the overhead of sending the key is large and it takes up a lot of bandwidth. To this end, ABE introduces the concept of attributes, which makes ABE no longer try a single user, but an entire group. ABE's algorithm has been more widely used in cloud computing. The biggest advantage of attribute-based encryption algorithms in cloud computing is that the CP-ABE-based method gives the generation access control right to the data owner, so that the data owner can choose the user to access the file autonomously and conveniently. The encryption process of CP-ABE can be summarized as:

① Setting: This is a random algorithm, input hidden security parameters, output public parameters PK and a master key MK.

② Encryption: This is a random algorithm. Input a message m, an access structure A, public parameters PK, and output ciphertext E.

③ Key generation: This is a random algorithm. Input a set of attributes Y, master key MK, public parameters PK, and output a decryption key D.
Decryption algorithm input: the ciphertext $E$ encrypted based on the access structure $A$, the decryption key $D$ corresponding to the attribute group $Y$, and the public parameter $PK$.

3. CP-ABE Based HDFS Security Policy

3.1. HDFS data storage scheme based on CP-ABE encryption algorithm

The user uploads the data to the distributed file storage system HDFS of the cloud computing platform Hadoop for storage, and then uses the distributed computing framework MapReduce to process the data. The CPFS-based BEFS-based HDFS distributed storage data security scheme proposed in this paper aims to shard the data during the distributed storage phase. Each shard is encrypted using the CP-ABE encryption method to obtain the ciphertext. Different ciphertext fragments are stored on different data storage nodes DataNode. The specific data writing process is shown in Figure 2.

**Figure 2. HDFS write data flow based on CP-ABE encryption algorithm**

HDFS write data flow based on CP-ABE encryption algorithm:
1. The client sends a request to upload a file to the master node NameNode (hereinafter referred to as NN);
2. After receiving the client's request, NN detects the user. If the test passes, it sends a response message that can be uploaded to the client.
3. After receiving the response, the client transmits a request to upload each data block (Block n);
4. The NN detects whether the Block n is duplicated. If it is new data, it assigns the corresponding data node DataNode (hereinafter referred to as DN);
5. After receiving the metadata information of the sent DN, the client establishes a connection with DN n and transmits Block n;
6. Block n uses the CP-ABE algorithm for encryption on the DN;
7. After the encryption is completed, store the ciphertext, public key $PK$, and attribute set $Y$;
8. The DN storing the data block sends a connection request to other DNs;
9. Other DNs return response information and agree to establish a connection;
10. Transfer data and backup to ensure that there are at least two copies of data to avoid data loss due to single point of failure.

When a user wants to read data, he first needs to apply for the public key $PK$ and attribute set $Y$ information to the HDFS distributed file system to form a decryption key $D$, and then request to read...
the fragment ciphertext $E$, and use the decryption key to encrypt the ciphertext $D$ to generate plaintext. The specific data reading process is shown in Figure 3.

**Figure 3. HDFS read data flow based on CP-ABE encryption algorithm**

HDFS read data flow based on CP-ABE encryption algorithm:
1. The user sends a request to the NN to read the data;
2. NN authenticates the user requesting access to the data;
3. If verified, the NN sends the metadata of the target file to the user;
4. The user establishes a connection with the corresponding DN according to the returned original data information, and sends a request to transmit the decryption key $D$;
5. DN uses the CP-ABE algorithm to generate the decryption key $D$ from the master key $MK$, the public key $PK$, and the attribute set $Y$, and returns it to the user;
6. The user makes a request to read the data to the DN;
7. DN returns the ciphertext $E$ to the user;
8. The user decrypts the ciphertext $E$ with the decryption key $D$ to obtain the plaintext of the data and completes the data reading.

### 3.2. Solution Analysis

The main idea of the HDFS data storage solution based on the CP-ABE encryption algorithm is to use the CP-ABE attribute-based data encryption algorithm during the HDFS data storage phase, and combine the "Reduce and Conquer" idea of the working principle of the MapReduce distributed computing framework during data processing. In the storage phase, the data is first sharded, and then each shard is encrypted using the attribute set and key. This solution can ensure that the data is not completely encrypted into ciphertext. If the user does not have permission when accessing the data, and the number of ciphertext data blocks obtained does not reach the threshold for forming a complete ciphertext, the complete ciphertext cannot be formed. It is also impossible to decrypt the complete plaintext, which increases the difficulty of decrypting the data and further strengthens the security of the data.

### 4. Conclusion

This article first analyzes the storage architecture and working principle of HDFS, a distributed file storage system for cloud computing, then analyzes the encryption and decryption process of CP-ABE based on the attribute encryption algorithm, and finally proposes a HDFS distributed data storage based on the CP-ABE attribute encryption. The solution applies the CP-ABE algorithm to the process of data transmission from the client to HDFS, which greatly guarantees the security of the data.
5. Acknowledgments

This work was financially supported by ‘Weifang Science and Technology Development Plan Project Design and implementation of Research on Weifang Intelligent Bus Management Platform Based on Big Data Technology (2019GX041)’.

References

[1] Feng Dengguo, ZhangMin, LiHao. Big Data Security and Privacy Protection. Chinese Journal of Computers, 2014,31(1):246-258(in Chinese).

[2] Yang Ting, Wang Meng, Zhang Yajian, Zhao Yingjie, Pen Haibo. HDFS Differential Storage Energy-saving Optimization Algorithm for Cloud Computing Data Center [J]. Chinese Journal of Computers, 2019,42 (04): 721-735.

[3] Panianz. A new data management challenge: How to handle bigdata//Proceedings of the International Conference Humanities, Geography and Economics. Dubai, UAE, 2018:4751.

[4] ShiYing Jie, MengXiao Feng. A survey of querytechniques in cloud data management system. Chinese Journal of Computers, 2013,36(2):209-225.

[5] Elomari A, Maizate A, Hassouni L. Data storage in big data context: A survey//Proceedings of the IEEE International Conference on System of Collaboration(SysCo). Casablanca, Morocco, 2016:1-4.