Research of Marketing Big Data Security Storage in Smart Grid Based on Spark

Juan Lei[1, a], Zhu Zhu[1, b], Yun Yang[2, c] and Sen Zhang[1,d]

[1]State Grid Chongqing Electric Power Co. Electric Power Research Institute,Chongqing, 40014, China

[2]State Grid Chongqing Electric Power Company, Chongqing, 40015, China

[a]61902966@qq.com, [b]zz@cq.sgcc.com.cn, [c]yy@cq.sgcc.com.cn, [d]zhangsen@cq.sgcc.com.cn

Key words: smart grid, cloud computing, Spark, data security, security storage.

Abstract. In order to solve the problem of massive data storage and satisfy the requirements of data confidentiality and integrity in smart grid, this paper analyzes the characteristics of existing security storage solutions, then combines with the special occasions in smart grid, designing a strategy of security storage about marketing big data in smart grid based on Spark. This schema adapt to data security and fast storage of real time data by fully using the advantages of high-performance with Spark and modern cryptographic technique. Its main features are high-level security, high efficiency and convenience in key management. In order to satisfy these conditions, a system based on cloud platform for smart grid is developed, analyzing the time-consuming of this program, and associating corresponding experiments prove that the design is effective and feasible.

Introduction

As more and more new techniques and new equipment's are connected to the smart grid, the security threats of smart grid are growing. In smart grid cloud storage system, the existence of information security threats mainly includes the data security and privacy risk. A lot of power data in smart grid need to store, which are subjected to leakage or tamper or other security risks in the process of interaction [1]. A large number of intelligent terminals which connected to the electricity grid assess data from the smart grid cloud storage system, which bring a large amount of data security threats to the smart grid. Therefore, in order to fully protect the power data security in smart grid and user privacy in cloud storage, cloud storage power system security protection construction has become the focus of the smart grid security construction [2].

The most commonly used method of secure store is to save data with encryption, through password technology to realize storage security. For example, literature [3] has improved the file format, the key information or integrity protection information are stored in the file header. For example, literature [4, 5] uses the symmetric and asymmetric encryption technology. Symmetric encryption is used to encrypt the data file and asymmetric encryption technology use two different keys to encrypt and decrypt.

In the smart grid environment, the frequency of data collection and storage is very high, and amount of data is very large. Key management in literature [3] needs to handle the file, efficiency is very low and it is difficult to meet the real-time processing requirements of the smart grid. In literature [4, 5], the private key is composed of random Numbers, which is not convenient to use and can't guarantee the integrity of the data. In this paper, using HBase high performance advantage and password technology, separating the management of the key and the cipher text, we design an efficient safety storage solution based on Spark smart grid marketing data.

Safe storage strategy

Traditional data storage security mainly consider the safety, but few of them involve data real-time processing, such as literature [6,7] mainly considered data encryption to ensure the

security of data storage, literature [8] took the real-time data processing into account to meet instantly. In order to meet the data real-time capability and security, this paper firstly takes the data encryption measures, and then divide the data into real-time and non-real time which stored in the cloud. For the integrity and confidentiality, this paper uses the method of cryptography and selects the appropriate encryption algorithm and message digest algorithm to, generate the corresponding key and cipher. For each encrypted file, each encryption used different key, so it can be said that it basically achieved the effect of a dense, and has high security.

For encrypted information, we divide the data into real-time and non-real time. For real-time data processing, this paper will use the Spark technology of distributed data management based on memory, and convert the data from different data sources into an elastic distributed data set (RDD), which can be stored in memory. The process meets the requirements of real-time data processing, and it can timely dispatch information for users to exchange. In the distributed memory data, this paper will use offline mode to transfer real-time information to HDFS, and its process ensures the security and real-time of the data. In the distributed memory data, this paper will use offline mode to transfer real-time information to HDFS, and its process ensures the security and real-time data [9].

For data reading, this paper will use Spark technology to read the cipher text from HDFS and obtain the key of the related data from HBase, and deposit it into the distributed memory. And then checking the integrity of the data to determine whether the data in the cloud has been tampered with. Because Spark has powerful fast data reading and writing ability about memory, there is not a lot of disk operations, the efficiency of implementation is higher to meet the real time.

The Design of Smart grid marketing efficient and secure large data storage scheme

In smart grid, full of a great deal of data source, bearing the enormous of data volume and time efficiency requirements, which need encryption fast and efficient encryption algorithms. Therefore, symmetric encryption algorithms is suitable for data encryption. Asymmetric encryption algorithm is suitable for the metadata or key encryption.

The scheme of marketing big data storing

When storing data, the data will be divided into real-time and non-real-time. In order to meet the characteristic of real-time processing, Due to Spark ignoring a lot of I/O disk operations, so the operations of efficiency more higher. Spark will need to reuse, in order to improve the computational efficiency. Reading the data from different data source, and convert the data into resilient distributed datasets (RDD). The real-time information which users need to interact is stored in the distributed memory. The data stored in the distributed memory, use offline mode stored into HDFS.

The summary information obtained message from signature data, the cipher text is data that obtained by encrypting, the key information obtained the data that after used for encryption key information hiding, Whether it is real-time data or non-real-time data after Spark processing, The encrypted data is stored in the cloud, the scheme of Specific storage has shown in Fig. 1.

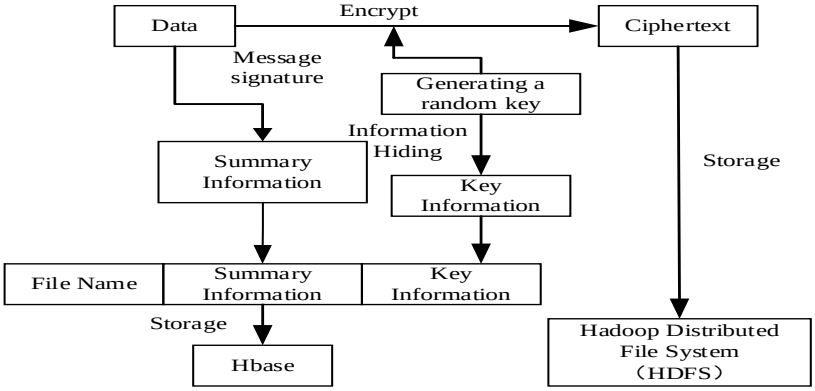

Fig. 1 Data storage program design

The first step: Encrypt data. The key generating function is used to generate the random key, and the data is encrypted with the key to obtain the corresponding encrypted data;

The second step: Random to hide key. For the first step, which get a random key is encrypted to hide;

The third step: Store data. Cipher text is stored in the cloud (HDFS), after the cipher text success to the cloud, the key summary information, digital information, and file name stored in HBase.

The scheme of marketing big data reading

When reading data, the user read the data directly from the cloud, and then decrypt the data into file and then check the file, the scheme of specific storage has shown in Fig. 2.

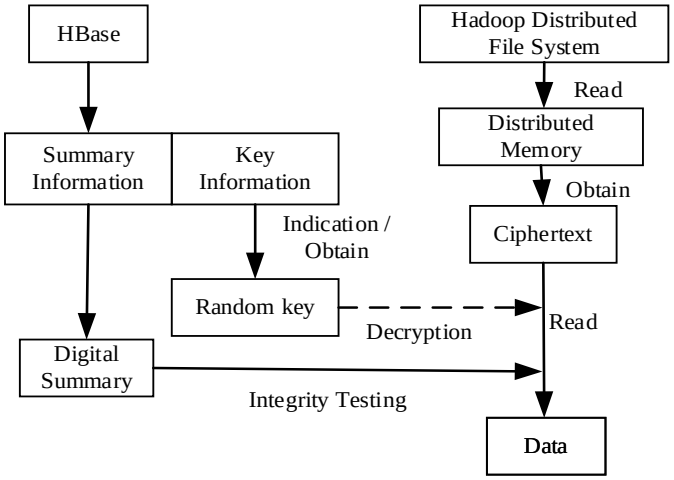

Fig. 2 Data reading program design

First, obtain the appropriate data from HBase and HDFS, and then check the data whether it is cipher text by key information, dealt according to the actual situation, finally, check the integrity of the data, to determine the integrity of the data. Specific steps are as follows:

The first step: Data acquisition: Using Spark to read the cipher text from HDFS and access to relevant data key from the HBase, write into distributed memory;

The second step: Obtain random key information: According to the key information to determine whether the data in HDFS are cipher text. If it is cipher text, the system will use the method of information obtained, access to information random key;

The third step: Decrypting data. Decrypt the cipher text and check the data integrity, make sure the data has not been tampered in the cloud.

Security Storage Policy Analysis

In the Programs, using the method of cryptography to achieve the integrity and confidentiality of data, if the system select the appropriate encryption algorithm and Message Digest Algorithm, it can fully guarantee the integrity and confidentiality of cloud storage in the smart grid data. In addition, the program also has the following advantages:① High Efficiency. For real-time information, we will firstly store it into the Spark-based distributed memory, and then store it to the cloud. At the same time the data stored in the cloud, we also use Spark technology, data parallel operation, it is more easy and fast;② Safety. For each encrypted file, every time the encryption key used is different, therefore it can be said that basically reached the one-time pad effect, which ensures high security. Even if any one of the cipher text is cracked, the other files are still safe;③Convenience for key management. For users, it can be given out only one common password, or no password, it is easy to use.

Experiment and Verification.

In order to verify the validity of the scheme, we use a concentrator connecting PC into a local area network (LAN). This platform use 4 PCS. Because of the extensibility of Hadoop and the

Spark, we can easily add new nodes to the cluster. Each PC machine is equipped with virtual machine, Ubuntu 32 bit operating system and 4 GB of memory. Download Spark1.2.0 version. Select one PC as a master (that is namenode), the other 3 as slave (that is datanode). In order to facilitate the management of the cluster, each PC builds Hadoop users, and gives root privileges to the Hadoop users. The Hadoop and Spark operations are conducted under the Hadoop users. Modify each machine's host name into master, slave1, slave2 and slave3. The system uses Java language and use open source cloud computing platform Hadoop framework and HBase database.

System of data storage and reading

We use a power supply bureau SCADA data to test encryption storage solution. The figure below shows part of the data. We use key generation function to generate 256 random key. Random key: Wq QBnz VM1y LOz Nx WJQU5Vce JJnx OMowr XX4SM2/+Tb Q=. After storing the data to the cloud, the cloud data is shown in Fig. 3.

Fig. 3 Data in cloud storage

The key data in the Base is shown in Fig. 4. In Fig. 4, key information: f4bY7AXJE+AStv zkw4RFOir3dnMFdj8n3AFhG0PI=; summary information: NSitRtIegYv+0LyNds PSnytWiwstnD kcwbK/l/zkme=.

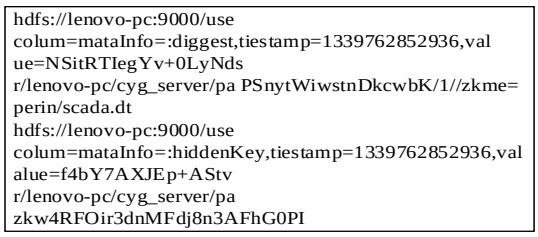

Fig. 4 Data in HBase

System reads the data, obtain from the HDFS cipher text at first, and then obtain the corresponding key information from HBase. As we can see from figure 4, key and summary information stored in the HBase are as follows:

Key information: f4bY7AXJEp+AStv zkw4RFOir3dnMFdj8n3AFhG0PI=.

Summary information: NSitRTIegYv+0LyNds PSnytWiwstnDkcwbK/l/zkme=.

Compared with the previous random key, we can find the key information which stored in the HBase is different with random key, this also shows the effectiveness of information hiding scheme.

System of performance of test and analysis

We use the different size of the file to test time needed for data storage. When a file is 10M、160M、580M、930M、1200M respectively, the experimental results are shown in Fig. 5.

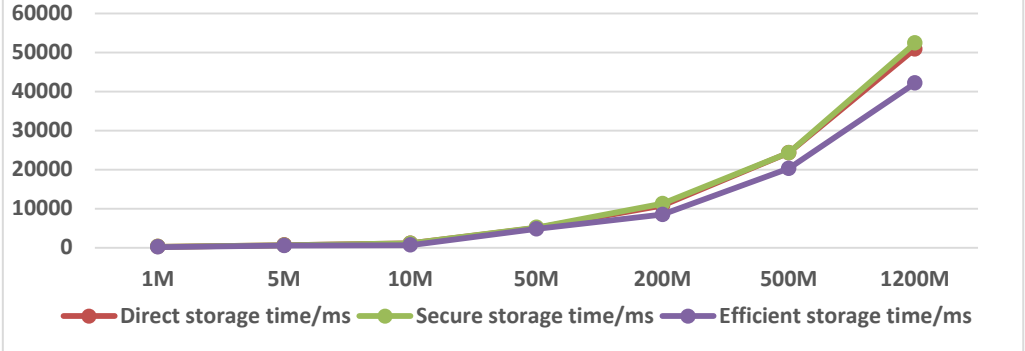

Fig. 5 Comparison of time-consuming in different size of data files storage

By experimental comparison, we can see that even in the case of small files, data encryption scheme based on the Spark is more efficiency than which based on source data encryption scheme. This is mainly because the Spark is adopting the distributed memory computing. Storing data in a distributed memory greatly reduces the disk I/O operation and make operating higher efficiently.

Summary

This paper learn the research of marketing big data security storage in smart grid based on Spark. It analyzes the security problems of smart grid information systems, designing a smart grid marketing big data storage security strategy based on the Spark. First, it introduces the basic features and current research condition, and show the marketing a big data storage and read design idea, and analyzed their advantages and performance, Ensuring data security, but also to solve the efficient handling of real-time data. Related experiments demonstrate the effectiveness and feasibility of the scheme, which is suitable for protecting smart online marketing data security.

Acknowledgement

The authors acknowledge support from the Science Technology Project of State Grid Chongqing Electric Power Company (No. 2016Yudiankeji18).

References

[1] X. Guan, G. He, C. Zhou, et al, Research of boundary security monitoring model based on Big Data for smart grid[C]// International Conference on Advances in Energy and Environmental Science, 2015.
[2] Y.C. Xu, Research on Data Integrity Protection in Cloud Storage for Smart Grid [D], North China Electric Power University, 2015.
[3] W.W. Zhang, Research of user data transmission and storage security in cloud computing solution [D], Beijing: Beijing University of Posts and Telecommunications, 2011.
[4] Information on http://hadoop.apache.org/docs/hdfs/current/hdfs_design.html
[5] AIAA. Stability of Symmetric and Asymmetric Vortex Pairs over SlenderConical Wing-Body Combinations - 33rd AIAA Fluid Dynamics Conference and Exhibit (AIAA)[J], Journal of Nanophotonics, 2015, 9(1).
[6] A. Mirhoseini, A.R. Sadeghi, F. Koushanfar: Secure outsourcing of big data machine learning applications[C]// IEEE International Symposium on Hardware Oriented Security and Trust, IEEE Computer Society, 2016:149-154.
[7] Y. Mu, Recent Advances in Security and Privacy in Big Data [J], Journal of Universal Computerence, 2015, 21(3):365-368.
[8] W.I. Chung, H.K. Kim, An Efficient Storing Scheme of Real-time Large Data to improve Semiconductor Process Productivities [J], 2009.
[9] Y. Wang, P. Hu, Research of Key Separate-Management Method Based on n Dimension Space [J], Microcomputer Applications, 2006.l