Advances in Computer, Signals and Systems (2019) 3: 1-7
Clausius Scientific Press, Canada

DOI: 10.23977/acss.2019.31001
EISSN 2371-8838

A Data Integrity Verification Scheme with Secure Deduplication in Smart Grid Cloud Storage

Shaomin Zhang[a,*], Kang Huang[b], and Baoyi Wang[c]

Department of Control and Computer, North China Electric Power University, Huadian Road, Baoding, China

[a]zhangshaomin@126.com,[b]huangkang9527@163.com,[c]wangbaoyi@126.com

Keywords: Smart Grid, Cloud Storage, Secure Deduplication, Integrity Verification.

Abstract: In order to solve the problem of data deduplication and data integrity in smart grid cloud storage, a data integrity verification scheme which can support security deduplication is proposed in the context of smart grid cloud storage. This scheme refers to an efficient and safe deduplication method based on the bloom filter, realizes the quick verification of the user's hash value and initialization value. In addition, the secure deduplication method is combined with S—PDP integrity verification mechanism, data segmentation and random sampling strategy, to realize the security deduplication and integrity verification of smart grid clients. By using erasure correction code, users can repair the damaged data. The analysis results show that this scheme can effectively reduce the computing cost and communication cost while ensuring that the cloud storage data of smart grid can be secure deduplicated and integrity verified.

1. Introduction

The smart grid is a comprehensive communication infrastructure for transmission of electrical energy and data simultaneously in a real-time, two-way manner [1]. All kinds of information collection devices and intelligent systems have been widely used in power systems leading to a sharp increase in the data of power enterprise [2]. In the power system, most of the power data is duplicate data, which not only reduces network bandwidth and storage space, increases data management costs, but also reduces system performance. Insulator leakage current monitoring, for example, assume that every 10 milliseconds to collect a data, a month will reach 250 million data records, in which a large amount of data may be repeated, largely reduces the storage efficiency of spatial [3].For example, electric energy data acquisition system is an intelligent system with functions of power information acquisition, processing and real-time monitoring. It’s main technical features are: a large number of users, up to millions, even tens of millions, leads to the large amount of data collected; High real-time requirement, the number of messages per second can reach hundreds of thousands. Most importantly: in these huge volumes of data. A significant portion of the data is repetitive, leading to a waste of storage space that affects the performance of the entire system. The researchers developed a technique called deduplication. The idea is that a cloud server stores only one copy of the same file, regardless of how many users want to store it. According to statistics, the weight loss ratio of cloud storage data can reach 1:10 ~ 1:500, which can save more

than 90% of storage space and bandwidth [4].At present, secure deduplication has become a hot research topic. On the other hand, users store power data on a cloud server and do not store any copies locally. Cloud storage service, however, is not completely trusted and reliable. Cloud storage data security problems caused by natural or man-made factors have occurred frequently in recent years. Due to the frequent occurrence of such events, people begin to pay attention to the storage security of cloud data, and the integrity verification of cloud data is one of the important research directions [5].

In cloud storage data secure deduplication and integrity verification study, this paper [6] proposes a data integrity verification method in the cloud storage and a across user data to deduplication method based on similarity. Data integrity verification method can effectively solve the problem of verifying data leakage, which supports an unlimited number of data integrity verification, supports validation of dynamically changing data, and supports third-party authentication, The cross-user data deduplication method based on similarity solves the efficiency problem of data deduplication in cloud storage and saves a large amount of storage space to some degree. Literature [7] proposes a scheme to support security deduplication and integrity verification, but this scheme does not support user privacy protection, and the computational cost is high. Literature [8] design and implementation of a cloud data security deduplication and data integrity auditing prototype system based on Ali cloud platform. However, none of the above is a good combination of data security deduplication and integrity verification. Literature [9] proposes an integrity verification method that supports secure deduplication. It supports the proof of ownership, unlimited verification, and public verification while guaranteeing data integrity and recoverability. It does not disclose user data information during the process of integrity verification and protects user privacy to ensure cloud storage data security and integrity. However, in the security deduplication stage, this scheme adopts a file-level deduplication strategy instead of a more granular one.

Based on the above security issues, this paper proposes an integrity verification method that supports secure deduplication under the smart grid cloud storage environment. It uses file-level deduplication strategy between users and block-level deduplication strategy within users, to eliminate the redundant power data and save storage space and bandwidth to a much larger degree. It supports ownership certification while ensuring data integrity and recoverability. In the process of integrity verification, no user’s data information was disclosed, user’s privacy was protected, and erasure code mechanism was adopted to ensure the security and integrity of smart grid cloud storage data.

2. Related Technology

2.1. Convergence Encryption Algorithm

Convergence encryption algorithm [10] uses the hash value of the data as the key, and encrypts the data with this key. So that the same original text can obtain the same ciphertext under the same encryption key, thereby realizing the deletion of duplicate data. The user calculates the data key through the key generation function. The tag generation function is used to calculate the data tag and upload the data tag to the server for repeatability detection. When the server has the corresponding tag that it can prove the data is repeated. The functions involved in the convergence encryption algorithm are as follows:

GenKey(D) Key. The Key is obtained by convergent key generation algorithm calculates the hash value of data D.

Encrypt (Key, D) C. Using Key and D as input, the ciphertext C corresponding to D is calculated through the symmetric and deterministic encryption algorithm.

Decrypt (Key, C) D. Key and C are used as input, and the symmetric decryption algorithm calculates the original text corresponding to the ciphertext C.

GenTag (D) T. GenTag is a tag generation algorithm. Using data D as input, calculate the corresponding label T; In addition, this article also allows the use of the function GenTagC(C) T, which generates the mark of data D from the ciphertext C.

2.2. Bloom Filter

Bloom filter is a kind of probabilistic data structure with high space efficiency. It is actually a very long binary vector and a series of random mapping functions. Bloom filters can be used to retrieve whether an element is in a collection. Its advantage is that space efficiency and query time are far more than the general algorithm, with less query time and lower space complexity. Its disadvantage is that there is a certain false recognition rate, that is, an element that does not belong to the set is misjudged as belonging to the set, and the more elements in the set are more likely to be wrong, the opposite will not occur. Another disadvantage is that It is not possible to delete existing elements [11].

3. Design of Our Scheme

The research idea of this paper is: The converged encryption is used as the core of the solution. The time cost of power data access authorization is reduced by the ownership proof method based on Bloom filter. And reduce storage space overhead by adopting file-level deduplication strategy between users and block-level deduplication strategy within users. After storing the power data on the cloud server, users can verify the integrity of the power data at any time. If the power data is damaged, the power data can be repaired by erasure correction code.

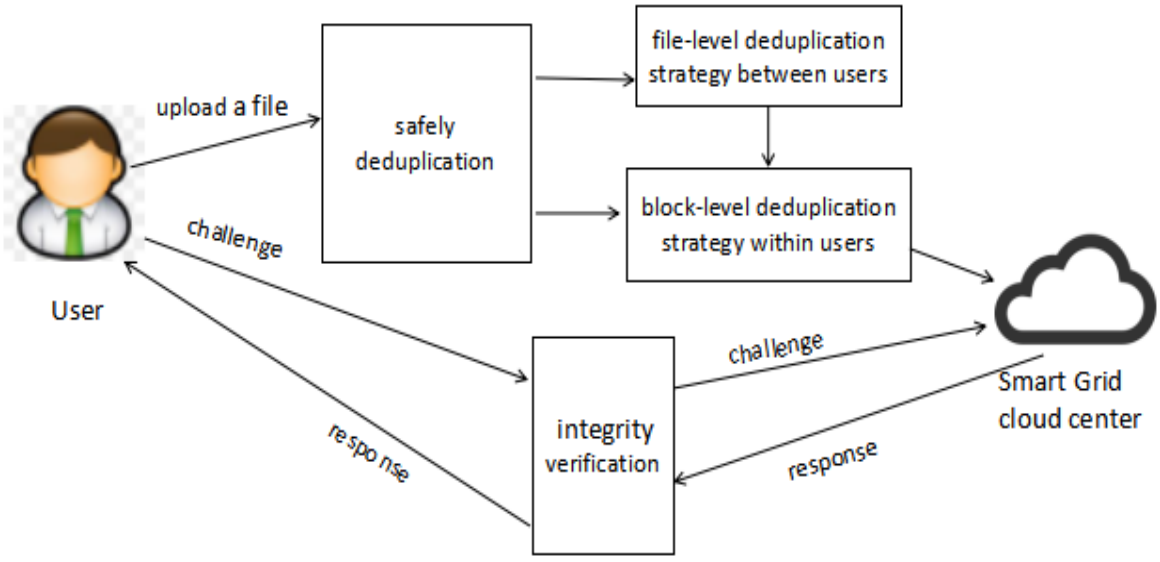

Figure 1 System model.

3.1. The Stage of Secure Deduplication

When a user wants to upload the electric power data file F to the cloud server. The user should encode the file F into F' by using erasure correction code at first, then needs to calculate the tag T of

F' and upload it to the server for repeatability detection. The server performs repeated detections based on T and returns results 0 (representing no duplicate files) or 1 (representing duplicate files) to the user. If there is no duplication of the file F, the file F is divided into blocks at first, and then calculates T of each data block, finally, it is uploaded to the server for repetitive detection of all data blocks that the user has stored on the cloud server.

If the user wants to upload the file F which already other users has stored on the cloud server. The user needs to prove to the cloud server that he really owns the file F. First, the cloud service provider randomly selects c block data blocks from the verified file F or all data blocks to challenge the user chal1; after the user receives the challenge chal1, the algorithm is run to generate the ownership evidence P1, and the evidence is sent to the cloud service provider; After that, the cloud service provider runs an algorithm to verify the ownership proof P1. If the verification is successful, the file block pointer P is sent to the user. The user stores the pointer P and the cloud server marks the user as a legitimate user.

3.2. The Stage of Secure Deduplication

When the user wants to check the integrity of the file F, at first, the user randomly selects the c block data block to challenge the cloud service provider, and sends the chal2 to the cloud service provider; After that, the cloud service provider runs the algorithm GenProof (pk, F'=(B_1, …,B_n), chal2, $\sum = (T_1, \ldots, T_n)$) to get γ. Then, the user verifies the correctness of the integrity evidence through the CheckProof (pk, sk, chal, γ) algorithm. If the verification is successful, the cloud server outputs "success"; otherwise it outputs "failed". The process is as follows:

Algorithm: integrity verification

GenProof (pk,F'=(B_1, ……, B_n), chal2, $\sum = (T_1, \ldots\ldots, T_n)$):

1) Assume chal2= (c， k1， k2， g_s). When 1≤j≤c:

Calculate all the data block numbers used to generate the validation data： $i_j = \pi_{k1}(j)$;

Calculate coefficients： $\alpha_j = f_{k2}(j)$;

2) Calculate: $T_{CBi} = T_{CBc}^{a_c} \ldots T_{CBc}^{a_c} = (h(W_{CB1})^{a_1} \ldots (h(W_{CBc})^{a_c} \cdot g^{a_1 CB1 + \ldots + a_c CB_c})^d \bmod N$;

3) Calculate: $\rho = H(g_s^{a_1 CB1 + \ldots\ldots + a_c CB_c} \bmod N)$;Input $\gamma = (T_{CBi}, \rho)$.

CheckProof(pk, sk, chal, γ):

1) Assume $\tau = T_{CBi}^e$. When 1≤j≤c: Calculate: $i_j = \pi_{k1}(j)$, $W_{CBij} = \nu \| CBi$, $\alpha_j = f_{k2}(j)$,

$$\tau = \frac{\tau}{h(W_{CBij})^{a_j}} \bmod N .$$

2) If $H(\tau^s) \bmod N = \rho$,the cloud server outputs "success". Otherwise the output "failed".

If the integrity verification result is a failure, it indicates that the user's stored data has been damaged to some extent. Because this scheme uses the erasure correction code mechanism to encode the file, the user can use the erasure correction code to repair the damaged data to reduce the user's loss.

4. Scheme Analysis

4.1 Analysis of Correctness

In the data integrity verification stage, the user needs to verify the evidence sent by the cloud service provider and needs to verify that the following equation is valid:

$$\rho' = H(\tau^s) \bmod N$$

$$= H\left(T^{e \cdot s}\right) \bmod N$$

$$= H\left(\left(\prod_{j=i1}^{ic} \sigma_j^{a_j}\right)^{e \cdot s}\right) \bmod N$$

$$= H\left(g_s^{a_1 C_{B1} + \cdots + a_s C_{Bc}}\right) = \rho$$

When the data is complete, both sides of the equation will be equal. If the two sides of the equation are not equal, it means that the user's power data which stored on cloud server has been broken. At this point, the user can consider repairing the damaged file. Therefore, it can be determined whether the integrity of the stored data is damaged by determining the balance of the equations.

4.2 Analysis of Secure

Theorem: the probability that a malicious user proves ownership without owning data can be ignored.

Proof: when we verify whether an element e belongs to the collection of G through the bloom filter and the element e does not belong to the G set, a misjudgment of e belongs to G may be generated. This is because the position of the hash value which corresponding to all the elements in G is set to 1 in the Bloom filter initialization process, and the hash value of e at the corresponding position bit in the Bloom filter has just been set to 1. At the same time, the probability of misjudgment for a certain Bloom filter is:

$$p = (1-(1-1/m)^{kn})^k \approx (1-e^{\frac{-kn}{m}})^k \quad (1)$$

At this time, suppose k = 1, m = 220, even if the number of blocks for a 1024 MB file is n = 215, substitute the above parameters into the formula p ≈ 0. Therefore, the probability of a malicious user who can pass proof of ownership without owning data is negligible.

4.3 Analysis of Experiment

This section compares the Baseline's[12] scheme time cost with the simulation scheme presented in this paper.Experiment: The purpose of this experiment is to compare the time cost of uploading files under different repetition rate conditions. The experiment selected two groups of 30 identical files, each of which had a size of 16 MB. During the experiment, upload a set of 30 files at first, and then upload another set of 30 files according to the repetition rates of 0, 20%, 40%, 60%, 80%, and 100% to find the trend of time cost.

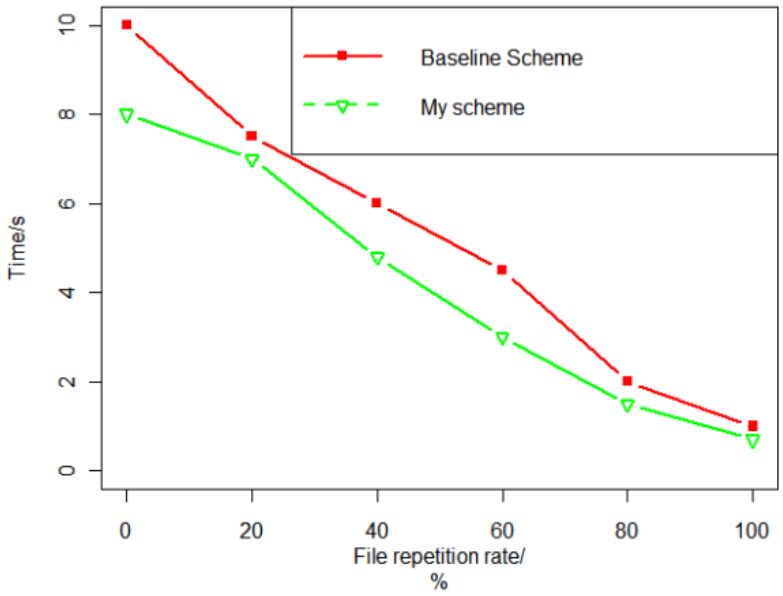

Figure 2: Time overhead for different file repetition rates.

The experimental results are shown in Figure 2, because the calculation time of the label does not change but the encryption operation gradually decreases, with the increase of the file repetition rate, the cost of file uploading time gradually decreases in both schemes. As the file repetition rate increases, the block-level deduplication operation will be reduced, and a large number of file duplications will reduce the corresponding data segmentation, block label calculation, and encryption operations. Therefore, the time difference between the two schemes is reduced. When the file repetition rate is 100%, the scheme of this paper does not need to perform file block level deduplication, and the cost time of file upload is basically the same as the time cost of Baseline's scheme. However, in the overall trend, the time cost of this paper scheme is better than the Baseline's scheme.

In the process of uploading the encryption phase, the scheme of this paper uses the key encryption mechanism to encrypt the previous file block key to the next file block key, and then generates the key ciphertext block to upload to the cloud.The user only needs to store the key of the first file block of each local file locally, which is the hash value of the file block. In the Baseline's scheme, only one fixed-bit hash value is required for each file for decrypting the file. In this scheme, the client will eventually need to upload to the server only the ciphertext that the file block key is not repetition. Therefore, the consumed key space decreases linearly with the increase of the file block repetition rate.

5. Conclusions

Based on previous work, this paper proposes a data integrity verification method that supports secure deduplication and privacy protection in smart grid cloud storage, and through the correctness analysis, security analysis and performance analysis to prove the feasibility and effectiveness of the method. By adopting file-level deduplication strategy between users and block-level deduplication strategy within users, it solves the problem of linear growth caused by the increase in the number of keys and the increase in the number of data sharing users.Of course, this method also has some shortcomings. The next step will research how to implement public verification of smart grid data that supports third-party verification, as well as more efficient algorithms.

Acknowledgements

The paper is supported by the Fundamental Research Funds for the Central Universities (2018ZD06).

References

[1] Zheng Ling, Yao Jiangang, The Construction of Power Cloud Integrated with Heterogeneous Application Service Systems [C]//THE 2016 International Conference on Communications, Information Management and Network Security (CIMNS).Shanghai, CHINA,2016:345-348.
[2] HE Yao, Wang Wenqing, Xue Fei, Study of massive data mining based on cloud computing [J]. Automation of Electric Power Systems, vol.34, no.15, pp. 15-18, 2013.
[3] Wang Dewen, Song Yaqi, Zhu Yongli. Intelligent power grid information platform based on cloud computing [J]. Automation of Electric Power Systems, 2010, 34(22):7-12.
[4] Dutch M, Freeman L. Understanding Data Deduplication Ratio[EB/OL].[2018-07-08].http://www.snia.org/.
[5] Zhang Taoyi. Research on data integrity detection in cloud storage [D].ShenZhen University, 2016.
[6] Li Zhike. Research on data integrity verification and Deduplication technology in cloud storage [D]. Guangdong University of Technology, 2015.
[7] Zheng Qingji, Xu Shouhuai.Secure and Efficient Proof of Storage with Deduplication[C]//Proceedings of the 2nd ACM Conference on Data and Application Security and Privacy. San Antonio, USA: ACM Press, 2012: 1-12.
[8] Li Bin. Research on data integrity and deletion technology in cloud storage [D].Beijing University of Posts and Telecommunications, 2015.
[9] Zhang Lihong, Chen Jing, Du Ruiying, et al. Data Integrity Verification Method with Secure Deduplication in Cloud Storage [J]. Computer Engineering, 2017, 43(1):32-36, 42.
[10] Mao Zhen.Design and implementation of document ownership certificate protocol based on convergent encryption [D]. Xidian University, 2016.
[11] Liu Zhusong, Yang Zhangjie. Efficient and secure deduplication cloud storage scheme based on proof of ownership by Bloom filter[J]. Journal of Computer Applications, 2017,37(03):766-770.
[12] GONZALEZ-MAZANO L, ORFILA A. An efficient confidentiality-preserving proof of ownership for deduplication [J]. Journal of Network and Computer Applications, 2015, 50: 49 -59.