Multi-source Data Collection Data Security Analysis

Lei Ma[1(✉)] and Yunwei Li[2]

[1] Beijing Polytechnic, Beijing 100016, China
malei235@tom.com
[2] Beijing Youth Politics College, Beijing 100102, China

Abstract. In order to improve the data effect of multi-source data collection and shorten the time of data collection, this paper proposes a data security analysis method of multi-source data collection. The white noise on the blank data field and knowledge background is removed through data processing, the multi-source data acquisition and access control function is optimized, the encrypted symmetric key is obtained, the filtered data is forwarded to the corresponding trusted exchange agent, the data security exchange characteristics are extracted, and the data security analysis mode is set. Experimental results: the average time consumed by the security analysis method of multi-source data collection data in this paper and the other two security analysis methods of multi-source data collection data are 94.283 s, 129.940 s and 130.121 s respectively, which proves that the performance of the security analysis method of multi-source data collection data in this paper is more perfect.

Keywords: Multi-source data · Data collection · Data security · Data sharing · Data fusion · Semi-structured data

1 Introduction

With the deepening of multi-source data security research, people are no longer satisfied with simply integrating and encapsulating interrelated distributed and heterogeneous data sources. Conventional data sharing and integration can no longer satisfy users' needs in data semantics and knowledge need. Data security analysis is the collaborative processing of heterogeneous data from multiple data sources to achieve the purpose of streamlining data, reducing redundancy, synthesizing complementarity and capturing collaborative information. Data aggregation can combine different statements about the same object scattered in different places to get more complete information about the object. The objective existence of multi-source data and the difficulty of seamless integration of data lead to many problems in the effective management and sharing of information data and files in business processes [1, 2]. Finding an effective data fusion method can handle the intricate relationships between different data sources in a large amount of multi-source data, facilitate the analysis of business processes between network security devices, and make the security analysis operations of related devices simpler and

W. Fu and L. Yun (Eds.): ADHIP 2022, LNICST 469, pp. 458–472, 2023.
https://doi.org/10.1007/978-3-031-28867-8_34

more convenient. At the same time, data aggregation generally involves security issues, because after merging the data and passing some reasoning, some conclusions may be drawn that the data publisher does not expect (may be required by other publishers). Multi-source datasets are large datasets. Compared with traditional datasets, big data is characterized by containing a large amount of unstructured data and semi-structured data. The purpose of multi-source data processing analysis is to discover new and hidden value in data sets, and to efficiently organize and manage large data sets. Compared with traditional data integration, in some cases, people are more concerned about the new semantic meaning exhibited by the aggregated data. Data lays a good foundation for knowledge representation because of its good conceptual hierarchy and support for logical reasoning. Data can be reused, thus avoiding repetitive domain knowledge analysis. We should make good use of existing data, improve the quality of data sources as effectively as possible, and reduce the loss of human, material and related resources in the process of information mining. It has become a basic problem faced by today's computer science and technology to easily and quickly screen out useful data feature information from massive multi-source data or to understand the correlation between data.

To this end, a data security analysis method for multi-source data acquisition is designed to effectively reduce the time required for label generation.

2 Data Security Analysis of Multi-source Data Collection

2.1 Improve Data Processing Flow

The selection and quality of data sets are a crucial condition. A good data set should be evenly distributed, cover a wide range, and have real and effective data, so that the model can correctly learn the parameters required by the trainer. However, the aggregation of data is huge, and the quality of all the data in the original dataset cannot be guaranteed, so many data cleaning algorithms for the original dataset are constantly developing. The principle of data cleaning is to use the existing technical means and methods to clean the "dirty data" by analyzing the causes and existing forms of "dirty data", and convert the "dirty data" into data that meets the data quality or application requirements, thereby Improve the data quality of the dataset. Data cleaning is the detection and screening of outliers in data sets. The data sets used in machine learning training all have certain distribution laws, revealing certain data laws. Therefore, the so-called outliers usually refer to outliers. Due to this feature, before analyzing the security of the data, a model needs to be established in advance. Objects that cannot be fitted with high quality by the model can be regarded as abnormal points. The same is true for other models, but this requires a pre-estimation of the data [3–5]. The means of cleaning are: removing noise data and irrelevant data in the source data set, processing missing data and cleaning "dirty data", removing white noise in blank data fields and knowledge background, considering time sequence and data changes, etc., to complete repeated data processing and the default data processing to complete the conversion of data types. Based on distance cleaning, the distance between data is specified, and the distance from other data objects beyond this distance is regarded as an abnormal point. Data cleaning can be divided into supervised and unsupervised categories. A supervised process, under the

guidance of domain experts, analyzes the collected data, removes clearly erroneous noise data and duplicate records, and fills in missing data. The main flow of data processing is shown in Fig. 1:

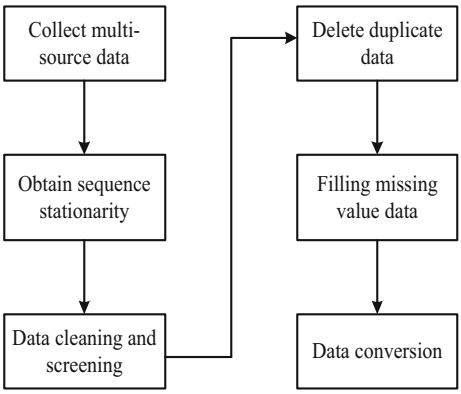

Fig. 1. The main flow of data processing

As can be seen from Fig. 1, the main processes of data processing are: collecting multi-source data, obtaining sequence stationarity, data cleaning and screening, deleting duplicate data, filling missing data, and data transformation. The unsupervised process is to use the established rule base for data cleaning. Generally speaking, data outliers are points with rare attributes in the data set, whose attribute values are very irregular, or relatively low in frequency in the data set, then by calculating the number of occurrences of the values of the corresponding attributes in the data set, we can Effectively assess whether this data is an outlier. Another important aspect of data cleaning is the transformation of data types, usually referring to the discretization of continuous attributes. Generally speaking, the discretization methods that are independent of the class include the equidistant interval method, the equal frequency interval method and the maximum entropy method. According to the above, data set cleaning is a necessary step in machine learning model training, which can help filter out data that is not helpful or even harmful to training. These data do not conform to the overall distribution of the data set, and the model cannot be trained in the data set. It helps, but it will increase the complexity of model training. The methods related to categories include division method and merge method. Through discretization, the size of the data table can be effectively reduced and the classification accuracy can be improved. To use the original data, it must be cleaned, not only to check the storage format of each attribute, but also to check whether its actual content conforms to the specification, such as handling vacancies, identifying and deleting outliers, deleting some duplicate records, and correcting The validity of the attribute value is checked, etc. During the training process, model parameters need to pay more attention to the fitting loss caused by abnormal points, and at the same time, the fitting of normal data will inevitably be affected, which will increase the time and computing resources required for model training, and affect the efficiency of training. In severe cases, it will affect the results of training, resulting in larger training errors

and even more serious business losses. This problem is more noticeable in the scenario where joint data sources participate in data cleaning. Fill gaps with the most probable values: The gaps can be identified using regression, Bayesian formal methods tools, or decision tree induction, etc. For example, using the attributes of other customers in the data set, a decision tree can be constructed to predict the vacancy value of income. In the joint data source scenario, because the data of each participant is mixed, in this solution setting, although all parties hold the same type of data. The problem of detecting and eliminating duplicate records is one of the main issues of research in the field of data cleaning and data quality. In the process of merging multi-source heterogeneous data, it is necessary to import a large amount of data from various data sources. Ideally, for an entity in the real world, there should only be one corresponding record in the data source. However, due to the differences in collection methods and storage methods, the data of each participant may have great discrepancies, such as the format of data storage, the dimensions of the data, and the distribution of characteristics and attributes of the data., the above differences between the data may cause a loss of model accuracy. However, when integrating multiple data sources represented by heterogeneous information, due to various problems such as data input errors, differences in format and spelling, etc. in the actual data, it is impossible to correctly identify multiple records that identify the same entity., so that logically pointing to the same real-world entity, there may be multiple different representations in the merged data, that is, the same entity object may correspond to multiple records. Although each participant can perform data cleaning locally, due to the differences in the data itself, the cleaning algorithm may also have different choices, and the cleaning results of each participant may not meet the requirements after data fusion, so the data of all parties are collected together. It is more feasible and can guarantee the cleaning results to deal with them uniformly.

2.2 Optimize Multi-source Data Collection Access Control Function

This model uses smart contracts to implement attribute-based fine-grained access control of multi-source data, and adds access control to the ciphertext acquisition process for dynamic permission judgment, so that the resource owner acts as the only promoter and message in the data sharing process changer. With the exponential growth of multi-source data, coupled with the limited resources of local storage, if a large amount of data is stored locally, it will inevitably bring serious challenges to local storage capacity. Therefore, multi-source data owners upload multi-source data to the cloud to save local storage space. At the same time, the blockchain engine shared general ledger technology effectively ensures the reliability of meta-information storage and the auditability of judgment execution. The main body of the multi-source data sharing process includes the data owner and the multi-source data requester. The direction of multi-source data transmission is from the data owner to the data requester. In order to realize the secure storage of data and the association between on-chain data and off-chain data, we adopt technologies such as blockchain, smart contracts, and IPFS. However, in order to ensure the security of multi-source data, it is necessary to conduct security analysis on multi-source data. One method is to download multi-source data directly from the cloud for security analysis. This method is undoubtedly the best in terms of correctness, but it consumes a lot of resources and time and reduces the efficiency of auditing. Assume that

there are several data points in the original space, including clean data points and noisy data points. The filtering operation of noisy data causes the manifold learning algorithm to map the original high-dimensional data points to the low-dimensional space, so that the topological structure of the data is not disturbed by the noise points, or the influence of the noise points is minimized. Then the expression formula of manifold structure data is:

$$D = \sum \frac{E(L-\alpha)}{2} \tag{1}$$

In formula (1), E represents the number of adjacent points, L represents the mapping result, and α represents the translation vector. The two important processes of the multi-source data security sharing process are the data owner uploading resources and the data requester requesting resources. The data owner generates a symmetric key to encrypt the data to be uploaded. The data owner stores the encrypted data in IPFS and obtains the storage address in IPFS. In order to reduce the client auditing overhead, the main method at present is that the multi-source data owner entrusts the auditing task to a third-party auditing agency for auditing. The third-party audit agency adopts the method of random sampling, that is, extracts a part of all multi-source data uploaded by users to the cloud for security analysis [6]. According to the audit results of this part, the integrity of the overall multi-source data is estimated to determine whether the multi-source data is safe. The data owner obtains the encrypted symmetric key through the access control module. The multi-source data owner calls the data management contract to save the data name, storage address, encryption key, dynamic access policy, and author information to the chain state database. The data requester calls the data list method to locate the required data. According to the eigendecomposition, the minimum weighted mean square value of multi-source data is obtained:

$$G_\beta = \frac{H^2 + \varepsilon}{\|1-\beta\|} \tag{2}$$

In formula (2), H represents the inverse transformation coefficient, ε represents the translation vector, and β represents the weight of the sample point. This method takes both correctness and auditing efficiency into consideration. Compared with the first method, the correctness decreases, but the auditing efficiency is doubled. The multi-source data integrity audit model generally includes three entities: multi-source data owners, cloud service providers, and third-party auditors. The multi-source data requester obtains the storage address information and decryption key through access permission determination. The data requester downloads the encrypted data content and decrypts it with the symmetric key to obtain the metadata content.

2.3 Extract Data Security Exchange Characteristics

From an application-oriented perspective, multi-source data security exchange can be divided into two modes: custom data security exchange and stream security exchange. Data encryption is an effective means of protecting data leakage during transmission or cloud storage. At present, according to whether the keys for data encryption and

decryption are the same, it can be divided into symmetric encryption algorithm and asymmetric encryption algorithm. Customized data security exchange is a process of uniformly adapting, converting, filtering, transmitting and loading static heterogeneous data in a specific format based on exchange policies [7]. The basic characteristics of the symmetric encryption algorithm are that it is easy to implement, suitable for encrypting a large amount of data, and the length of the plaintext and the length of the ciphertext are equal, which are the outstanding advantages of the symmetric encryption algorithm. Of course, it will also bring some disadvantages accordingly. It is necessary to build a data transmission channel that both parties can trust in the real environment, which is basically difficult to achieve in an open Internet environment. The characteristics of this data security exchange mode are that it is generally oriented to specific exchange objects, and has strong control over the data exchange process. Exchange, database synchronization, etc. The multi-source data security exchange mode is shown in Fig. 2, and the specific workflow is as follows.

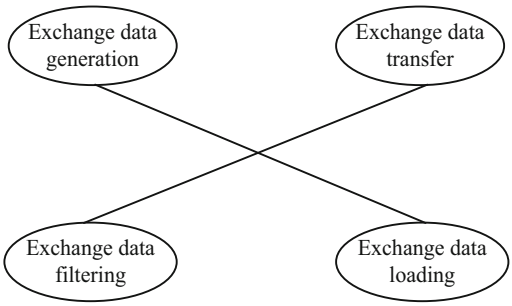

Fig. 2. Multi-source data security exchange mode

As can be seen from Fig. 2, the multi-source data security exchange mode includes: exchange data generation, exchange data transmission, exchange data filtering, and exchange data loading. For exchange data generation, the trusted exchange agent at the sending end is responsible for extracting the data to be exchanged according to the exchange strategy customized by the exchange parties, and then the adapter converts the exchange data into the required format. In addition, once the encrypted data scale increases, the encryption key also increases, and the storage and maintenance of the encryption key becomes a burden for the user. The most fatal disadvantage is that it cannot solve the problems of tampering and denial of messages. Then, the exchanged data is protected and encapsulated by cryptographic technology, and finally the exchanged data is forwarded to the data security exchange server according to the protocol defined by both parties. For exchange data transfer, the data security exchange server, as the controller of the data security exchange, establishes a dedicated secure data channel with both parties of the data exchange, and provides data forwarding for the exchange parties. After the data owner uploads the data to the cloud storage, in order to save the local storage space, the local copy of the data is deleted. When the data owner wants to delete the data copy in the cloud, a deletion order is issued to the cloud service provider. The trusted exchange agent forwards data in a transparent way through “push

mode" or "pull mode", without affecting the data exchange between the exchange parties at the application level. Trusted cloud service providers directly delete data, but due to the untrustworthiness of cloud service providers in the current complex network environment, they may only perform logical deletion, and the real data copies are still stored in the cloud. More serious cases are Directly rent this part of the storage space to other tenants, so that other tenants directly obtain copies of your data, resulting in data leakage. Exchange data filtering, the data forwarded to the data exchange server is filtered according to the customized exchange policy, and the data security exchange server forwards the filtered data to the corresponding trusted exchange agent through a dedicated secure data channel according to the customized exchange task.. After the exchange data is loaded, the trusted exchange agent of the receiver verifies the exchange data after receiving the exchange data. After the verification is passed, the exchange data is adapted, converted and loaded into the target system according to the customized exchange strategy. The main idea of the deterministic deletion scheme based on access control is to assign access rights to shared users. When the data is to be deleted, the user's access rights are revoked, so that the data cannot be accessed and indirectly achieves the goal of deterministic deletion of data. This solution is difficult to avoid illegal access technology by hackers. If the access rights technology is not very strong, the security of data still faces serious challenges. In the customized data security exchange mode, since the source, format and content of the exchanged data are relatively fixed, it is easy to protect it, and the exchange process used to exchange data often becomes the main target of the attack, so the main security threats faced in this mode are: The attack on the exchange process can achieve the purpose of tampering with information and spreading malicious code by attacking the exchange process. Based on this, the nature of data security exchange in this mode needs to focus on the protection and control of the exchange process, which can realize the credibility analysis and verification of the exchange process during the exchange execution process, so as to ensure the security of data exchange.

2.4 Set Data Security Analysis Mode

The security analysis is mainly to prove that the timestamp-based signature mechanism is secure in polynomial time, that is, the cloud service provider must store the data owner's files in order to generate valid evidence to respond to the challenge request of the third-party auditor, if the cloud service provider Arbitrary dishonesty will not yield a valid answer. The problem of detecting and eliminating duplicate records is one of the main issues of research in the field of data cleaning and data quality. In the process of merging multi-source heterogeneous data, it is necessary to import a large amount of data from various data sources. Ideally, for an entity in the real world, there should only be one corresponding record in the data source. The main steps of the data security analysis mode are shown in Fig. 3:

As can be seen from Fig. 3, the main steps of the data security analysis mode include: a data interception module, a data encryption/decryption module and a data key module. The function of the data interception module is to intercept the final data processed by the business logic layer before storing it in the distributed storage system of the cloud

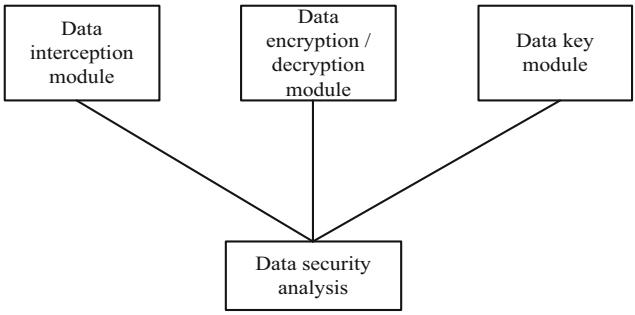

Fig. 3. The main steps of the data security analysis mode

platform, and generate the data ciphertext through the data encryption module. In addition, when the user requests data, the ciphertext data read by the data layer is decrypted and transmitted to the business logic layer for processing. However, when integrating multiple data sources represented by heterogeneous information, due to various problems such as data input errors, differences in format and spelling, etc. in the actual data, it is impossible to correctly identify multiple records that identify the same entity.. The main function of the data encryption/decryption module is to encrypt the stored data and decrypt the read data. The system obtains the secret key generated according to the secret key generation algorithm through the data secret key module, uses the secret key to encrypt the data, and uses the decryption key provided by the user to decrypt the encrypted data when accessing the data. Use the following formula to represent the encrypted data set of the participants:

$$R = \sum_{q=1}^{p} \frac{1}{\eta} \times K_{pq} \tag{3}$$

In formula (3), p, q represents two adjacent data nodes, η represents the frequency of occurrence of attribute values, and K represents the weight of attributes. The entities that logically point to the same real world may have multiple different representations in the merged data, that is, the same entity object may correspond to multiple records. Duplicate data can lead to incorrect merge patterns, so it is necessary to deduplicate data in the dataset to improve the accuracy and speed of subsequent merges. The data key module is mainly responsible for managing master keys, generating and distributing data keys. In order to ensure the security of data storage, the form of secondary encryption is adopted. The RSA algorithm master key encrypts the AES algorithm data key. When decrypting, the ASE data key is obtained by decrypting the RSA algorithm private key. The data key is responsible for encrypting the data. Encrypt and decrypt. Each duplicate record detection method needs to determine whether two or more instances represent the same entity. An effective detection method is to compare each instance with other instances to find duplicate instances. In order to detect and eliminate duplicate records from a dataset, the first problem is how to determine whether two records are duplicates. The uploading process of secure data can be divided into: data legitimacy verification, data business logic processing, encryption and storage in the cloud database [8–10]. Data security verification is to ensure that the data format is correct and the content

meets the requirements of business logic processing. The operation is checked by the interface before execution. After the data verification is completed, the corresponding functional interface completes the data business logic processing and then the file data is processed. After the encryption module is encrypted, it is sent to the corresponding class of the Dao layer to store the data in the corresponding database in the cloud. The encryption module will intercept and encrypt files before uploading files to HDFS for storage or data uploading to distributed databases for storage operations. This requires comparing the corresponding attributes of the records, calculating their similarity, and then performing a weighted average according to the weight of the attributes to obtain the similarity of the records. If the similarity between the two records exceeds a certain threshold, the two records are considered to be matched., otherwise, it is considered a record pointing to a different entity. The data to be encrypted first determines whether parallel encryption is required according to the size of the data. The data that does not need parallel encryption is directly encrypted by the hybrid encryption algorithm, and then the storage module interface is reflectively called. Large files that need to be encrypted in parallel are uploaded first through MapReduce for parallel hybrid encryption. The sort-merge method is a standard method for detecting exact duplicate records in a database. Its basic idea is to first sort the dataset and then compare adjacent records for equality. This method also provides an idea for detecting duplicate records on the entire dataset, and most of the existing methods for detecting duplicate records are also based on this idea. The data that needs to be encrypted in parallel will be divided into data blocks, and then the data blocks will be allocated to different processors by the MapReduce master node according to the allocation rules. The AES algorithm will encrypt each data block, and the encryption of all data blocks is completed. After that, through the reduce function, the encrypted data blocks are combined and processed to obtain the final ciphertext and the decryption key of the AES algorithm is generated. Each independent file has its own key to ensure that all files will not be caused by the cracking of one key. Encrypted data is broken. After obtaining the AES decryption key, the AES key will be encrypted with the RSA public key of the key management module. After completion, the ciphertext data will be stored in the cloud platform data server.The specific flow of data security analysis of multi-source data collection is shown in Fig. 4.

3 Experimental Tests

3.1 Experiment Preparation

The experiment uses a local virtual machine to load the open source project OpenStack for performance testing, in which the Hadoop sub-project in OpenStack is mainly used to build the required experimental environment. First, the Hadoop cloud environment is built to verify the DPOML algorithm and RFMML algorithm proposed in the paper. According to the existing equipment of the laboratory and the previous scientific research work, the UCI machine learning security data source is selected for the experiment. The hardware configuration of the experimental computer is Dell OptiPlex 3020 Mini Tower desktop, the processor is Inter Core (TM) i7-4590@3.30 GHz quad-core, the memory is 8 GB (Hynix DDR3 1600 MHz), and the main hard disk is GALAXY CX0128ML106-P (128 GB solid state). Hard disk) and Western Digital WDC WD5000AAKX-75U6AA0

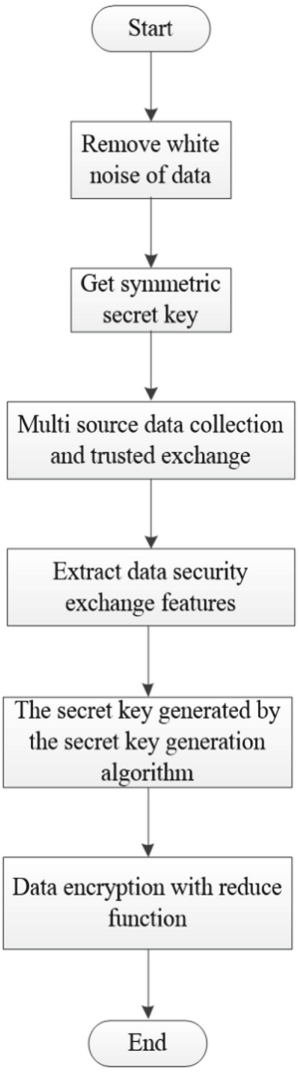

Fig. 4. Data security analysis flow of multi-source data collection

(Blue Disk) (SOOGB mechanical hard disk). The data source is obtained through the Flume component in Hadoop, and then the security data is stored uniformly. This experiment uses four types of security data, namely firewall logs, IDS logs, NetFlow security data, and DNS security data. MultiInputFormat is used to process multi-source data. The deployed Linux system is Centos and Hadoop, and the function is the function library provided by PBC, which is programmed and developed in python language. Finally, the attack test is carried out on the target host. The attack behavior includes application layer, session layer, transport layer, data link layer and network layer, basically covering all layers of the Internet protocol.

3.2 Experimental Results

In order to test the effectiveness of the designed multi-source data acquisition data security analysis method, a comparative experiment is carried out to discuss. The data security analysis method of multi-source data acquisition based on blockchain and the data security analysis method of multi-source data acquisition based on clustering algorithm are selected respectively, and the data security analysis method of multi-source data acquisition in this paper is selected for experimental comparison. The time consumption of security label generation of three multi-source data collection data security analysis methods is tested under different query data volume conditions. The experimental results are shown in Tables 1, 2, 3 and 4.

Table 1. Query data volume 20 GB Security label generation time (s)

Number of experiments	Data security analysis method for multi-source data collection based on blockchain	Data security analysis method for multi-source data collection based on clustering algorithm	Multi-source data collection data security analysis method in this paper
1	72.833	71.009	56.028
2	74.929	74.677	55.993
3	72.091	72.319	53.362
4	71.829	72.062	52.019
5	73.713	71.055	52.044
6	72.640	74.314	53.372
7	72.031	72.298	51.001
8	73.218	72.341	53.083
9	72.090	71.476	51.462
10	74.216	73.318	52.646

According to Table 1, when the amount of query data is 20 GB and the number of experiments is 10, the security label generation time of the blockchain method is 74.216 s, the security label generation time of the clustering algorithm is 73.318 s, and the security label generation time of the method in this paper is 52.646 s; The security analysis method of multi-source data collection data in this paper, compared with the other two security analysis methods of multi-source data collection data, consumes an average of 53.101 s, 72.959 s and 72.487 s for security label generation respectively.

According to Table 2, when the amount of query data is 40 Gb and the number of experiments is 8, the security label generation time of the blockchain method is 98.717 s, the security label generation time of the clustering algorithm is 96.590 s, and the security label generation time of the method in this paper is 62.546 s; The security analysis method of multi-source data acquisition data in this paper, compared with the other two security analysis methods of multi-source data acquisition data, consumes an average of 64.013 s, 95.036 s and 94.254 s for the generation of security labels, respectively.

Table 2. Query data volume 40GB security label generation time (s)

Number of experiments	Data security analysis method for multi-source data collection based on blockchain	Data security analysis method for multi-source data collection based on clustering algorithm	Multi-source data collection data security analysis method in this paper
1	92.736	89.636	67.973
2	98.090	91.017	65.611
3	89.567	92.367	63.544
4	92.381	96.873	62.628
5	91.884	94.091	61.710
6	98.980	98.563	63.008
7	96.862	93.488	64.767
8	98.717	96.590	62.546
9	95.099	95.607	65.330
10	96.045	94.312	63.012

Table 3. Query data volume 60GB security label generation time (s)

Number of experiments	Data security analysis method for multi-source data collection based on blockchain	Data security analysis method for multi-source data collection based on clustering algorithm	Multi-source data collection data security analysis method in this paper
1	120.678	118.937	98.673
2	122.087	121.663	96.556
3	121.990	119.579	95.474
4	118.664	122.367	97.329
5	122.491	123.291	99.09
6	118.368	120.398	105.089
7	121.276	118.467	103.673
8	119.564	122.094	102.182
9	123.820	122.678	96.321
10	124.093	120.022	98.334

According to Table 3, when the amount of query data is 60 GB and the number of experiments is 9, the security label generation time of the blockchain method is 98.717 s, the security label generation time of the clustering algorithm is 122.678 s, and the security label generation time of the method in this paper is 96.321 s; The security analysis method of multi-source data collection data in this paper, compared with the

other two security analysis methods of multi-source data collection data, consumes an average of 99.272 s, 121.303 s and 120.950 s for generating security labels, respectively.

Table 4. Query data volume 80 GB security label generation time (s)

Number of experiments	Data security analysis method for multi-source data collection based on blockchain	Data security analysis method for multi-source data collection based on clustering algorithm	Multi-source data collection data security analysis method in this paper
1	163.323	162.563	112.676
2	162.093	163.643	104.442
3	158.248	161.334	108.110
4	161.654	158.873	113.533
5	160.765	162.232	107.708
6	163.238	163.112	112.699
7	158.548	158.111	109.457
8	159.235	163.090	108.245
9	160.345	161.765	110.220
10	163.896	160.874	111.433

According to Table 4, when the amount of query data is 80 GB and the number of experiments is 10, the security label generation time of the blockchain method is 163.896 s, the security label generation time of the clustering algorithm is 160.874 s, and the security label generation time of the method in this paper is 111.433 s; The security analysis method of multi-source data collection data in this paper, compared with the other two security analysis methods of multi-source data collection data, consumes an average of 109.852 s, 161.135 s and 161.560 s for security label generation respectively.

According to Table 5, when the amount of query data is 100 GB and the number of experiments is 5, the security label generation time of the blockchain method is 207.637 s, the security label generation time of the clustering algorithm is 207.771 s, and the security label generation time of the method in this paper is 138.789 s; The security analysis method of multi-source data collection data in this paper, compared with the other two security analysis methods of multi-source data collection data, consumes 145.175 s, 199.265 s and 201.353 s respectively for the generation of security labels.

The data in Tables 1, 2, 3, 4 and 5 show that our party has high multi-source data collection efficiency under different data volumes. This is because the method in this paper removes white noise on the blank data domain through data processing, and shortens the time-consuming of generating security labels by using trusted switching technology.

In order to verify the security of multi-source data collection data of different methods, the blockchain method, clustering algorithm and the method in this paper are used to verify the integrity of multi-source data collection data, and the results are shown in Fig. 5.

Table 5. Query data volume 100 GB security label generation time (s)

Number of experiments	Data security analysis method for multi-source data collection based on blockchain	Data security analysis method for multi-source data collection based on clustering algorithm	Multi-source data collection data security analysis method in this paper
1	201.883	207.119	145.121
2	202.729	193.391	153.362
3	189.647	189.093	142.004
4	193.289	202.088	143.737
5	207.637	207.771	138.789
6	201.562	203.627	142.088
7	193.322	188.028	153.489
8	188.976	201.672	139.421
9	212.389	214.091	152.355
10	201.220	206.646	141.387

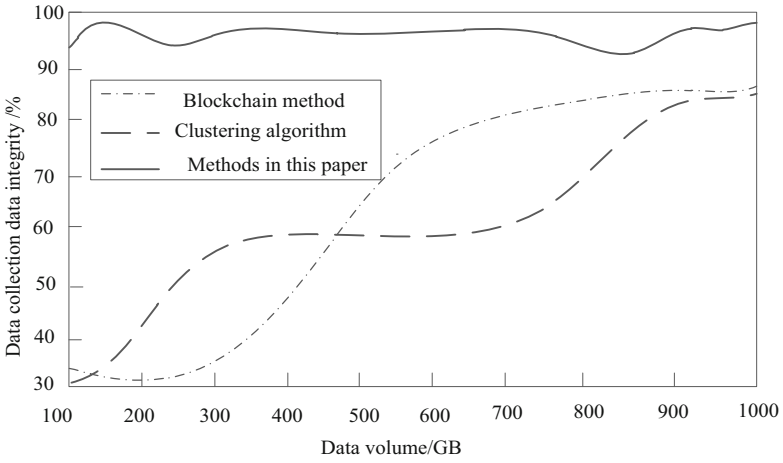

Fig. 5. Data integrity of multi-source data acquisition

It can be seen from Fig. 5 that when the amount of multi-source data collection is 200 GB, the integrity of multi-source data collection data of the blockchain method is 31%, the integrity of multi-source data collection data of the clustering algorithm is 42.5%, and the integrity of multi-source data collection data of the method in this paper is 95.1%; When the amount of multi-source data collection data is 600 gb, the integrity of multi-source data collection data of blockchain method is 78.1%, the integrity of multi-source data collection data of clustering algorithm is 58.9%, and the integrity of multi-source data collection data of this method is 96.8%; This method always has a

high integrity of multi-source data collection, which indicates that the multi-source data collection data security of this method is higher.

4 Concluding Remarks

The data security analysis method of multi-source data collection in this paper, in terms of data security analysis, mainly solves the problem of dynamic operation of multi-copy data, and prevents forgery and forgery between distributed storage nodes of a single cloud service provider during the data integrity audit process. Substitution and collusion attacks while reducing data leakage to third-party auditors. At the same time, in the merging and preprocessing of multi-source data, the methods used in data cleaning are discussed, such as dealing with missing values, removing outliers, removing duplicate records, etc. In addition, the methods of data transformation, such as normalizing the data, have also been improved. The method of data merging is to collect data from multiple data sources and store them in a consistent data store. In terms of data deterministic deletion, it mainly solves the fine-grained operation of data by users and adds a trusted verification mechanism after the deletion operation is completed. The future research direction is mainly to continuously improve the subject at the level of multi-dimensional data query optimization.

References

1. Wan, Q., Ma, Y., Wei, L.: Knowledge acquisition of multi-source data based on multigranularity. J, Shandong Univ, (Natural Science) **55**(1), 41–50 (2020)
2. Yu, L., Li, S., Chen, C., et al.: Analysis of ocean data merge based on multi-source parameters. J. Data Acquisition Process. **35**(5), 824–833 (2020)
3. Luo, J., Liu, X.: Strategies for scientific data security management from the perspective of intellectual property. Library Inf. Serv. **65**(12), 38–46 (2021)
4. Zhou, X., Liu, W., Sui, H., et al.: Five safes framework and its enlightenment to security data access in China's library field. Inf. Stud. Theory Appl. **43**(3), 85–90 (2020)
5. Feng, T., Jiao, Y., Fang, J., et al.: Medical health data security model based on alliance blockchain. Comput. Sci. **47**(4), 305–311 (2020)
6. Tang, X., Zhou, L., Shan, W., et al.: Threshold re-encryption based secure deduplication method for cloud data with resistance against side channel attack. J. Commun. **41**(6), 98–111 (2020)
7. Lv, G., Chen, L., Xiao, R., et al.: Simulation of quantitative assessment method for data security situation of wireless network communication. Comput. Simul. **37**(7), 337–340,372 (2020)
8. Jiang, L., Tang, Z.: Multi source data acquisition system based on Flume, Kafka and HDFS. Inf. Technol. Informatization **06**, 115–117 (2021)
9. Xu, H., Xu, Z., Chen, M.: Multi source and multi dimensional reading data collection and digital portrait based on xAPI. Educ. Commun. Technol. **16**(04), 59–63 (2020)
10. Wang, J., Guo, Y., Wen, X., Wan, F.: Multi source data acquisition and comprehensive evaluation system for smart business district. Comput. Eng. **45**(01), 284–291 (2019)