Research on Key Technologies for Big Data Classification Transactions and Privacy Protection

Shaoheng Wang, Wenxia Ge

Abstract—**As the data trading market continues to expand, the issues of privacy information leakage, data leakage, and data quality in data trading are becoming increasingly prominent, and to a certain extent, constraining the development of the data trading market. Blockchain, as a distributed ledger technology, with its characteristics of decentralization, traceability, and tamper resistance, can effectively avoid problems such as single point of failure and data tampering in data trading. This paper addresses the issues of privacy information leakage, data quality, and transaction fairness in data trading. Firstly, this paper introduces a decision tree classification model to classify the source data, and removes data that does not meet the transaction requirements before data trading. During the secure comparison phase, a lightweight ciphertext comparison algorithm is designed so that data owners can obtain classification results without revealing plaintext data to the decision tree classification model.**

During the data trading phase, this paper conducts transactions on a per-data fragment basis, granting data buyers trial rights. Subsequently, upon data buyers obtaining transaction data, consistency verification is conducted on the transaction data to prevent data owners from adding garbage data to the transaction data before data trading, thereby avoiding dishonest trading behavior by data owners. Finally, this paper designs a decentralized data trading model, DCTM (Data Classification Transaction Model), based on smart contract technology. This model addresses issues such as data quality, single point of failure, privacy leakage, and data resale in data trading, ensuring transaction fairness and security.

Keywords: big data, data trading, blockchain, privacy protection, data quality.

I. INTRODUCTION

In today's highly digitized era, the rapid evolution of information technology continuously influences our lives. Each individual actively or passively contributes to the generation of big data through daily interactions with smart devices, leading to an explosive growth in various types of data. Recognizing the value of data, both individuals and nations consider data resources as strategic assets in this information age. To address the issues of data quality and privacy protection in data transactions, numerous scholars have proposed data trading schemes, and several data trading platforms have been established. While these schemes and platforms facilitate data transactions, they still face challenges such as ensuring data quality, high transaction costs, privacy breaches of data trading parties, resale of source data, and difficulty in promoting the development of data trading markets.

Manuscript received May 08, 2024.

Shaoheng Wang, School of computer science and technology, Tiangong university, Tianjin ,China.

Wenxia Ge, School of Software, Tiangong university, Tianjin ,China.

Currently, most data trading follows traditional approaches, primarily categorized into custody and aggregation models. In the custody model, data owners encapsulate source data and entrust it to third-party data trading platforms for sale. The trading platform matches data buyers' demands, facilitating interactions between data owners and buyers to complete transactions. The aggregation model involves integrating source data based on different data types and processing techniques to serve the needs of data buyers. However, both models struggle to protect traders' privacy and ensure data quality. Dishonest third-party platforms may resell source data without the data owners' knowledge, leading to data breaches[1]. Moreover, excessive junk data in transactions diminishes the value of data trading and adversely affects participants' experiences, indirectly impacting the entire data trading market.Furthermore, unified standards for measuring data quality are elusive due to its relative nature. While data owners may perceive their source data as high quality, it may not meet the standards of data buyers.

To address the challenges in the current data trading market, this paper focuses on data quality and privacy protection. Building upon existing research, it explores decision tree classification models and blockchain technology to develop a data quality assessment model and a decentralized data trading model. These initiatives not only ensure data quality but also safeguard the privacy of all parties involved in data transactions, mitigating issues like single points of failure and data breaches inherent in traditional trading models facilitated by third-party platforms.

II. RELATED WORK

The increasingly mature blockchain and smart contract technologies have opened up new directions for data trading. However, current blockchain-based big data trading platforms still face two main challenges: Data Quality Issues: Methods that introduce third-party inspection agencies to conduct data quality checks on source data not only increase the risk of data leakage but also make it difficult to ensure the honesty of data owners in collaboration with these agencies. Regarding proposals from scholars to combine evaluation mechanisms to regulate data owners and improve the quality of source data, it's challenging to guarantee that evaluations of source data are always genuine and effective. Issues such as fake positive reviews exist, and evaluation mechanisms do not fundamentally solve the problem of source data quality.

Privacy Information Leakage: In data trading, personal information of both data trading parties may unintentionally be obtained by third-party attacks, posing significant risks to personal information security. Once leaked, this private

information could lead to various consequences, from spam emails and harassing phone calls to identity theft for fraudulent activities, dragging individuals into innocent cases. Additionally, in an untrustworthy transmission environment, there may be malicious attackers compromising the source data owned by data owners during the transaction process, leading to source data leakage or tampering, greatly compromising the security of data trading.

Reference [2] proposes a Secure Data Trading System (SDTE) based on SGX technology, which addresses issues such as data buyers reselling source data, third-party data trading platforms reselling source data owners' data, and single points of failure in data trading platforms. SDTE also utilizes SGX to protect the execution environment of smart contracts, addressing privacy data leakage issues caused by the transparent nature of blockchain.

Reference [3] proposes a fair data trading process conducted solely by data owners and data buyers without the involvement of third parties to address privacy leakage issues caused by third-party intervention in transactions. Smart contracts are used to ensure the fairness and autonomy of transactions, while inadvertent transmission protocols are combined to protect privacy data from being leaked during the transaction process. Furthermore, the proposal introduces Ether checks to enhance the fairness and flexibility of data trading payments.

Reference [4] introduces an accountable and auditable transaction protocol that holds accountable traders who fail to fulfill their responsibilities during the transaction process, eliminating dishonest behavior in big data trading. This protocol also designs data set similarity comparisons to prevent dishonest traders from reselling source data, ensuring the fairness of data trading.

Reference [5] proposes a data trading scheme based on smart contracts, integrating machine learning techniques. This approach combines identity authentication and on-chain/off-chain mechanisms to prevent data retention. Additionally, it introduces an arbitration mechanism to address data quality concerns. In case of disputes regarding purchased data, buyers can seek arbitration from the designated institution. This incorporation of an arbitration mechanism facilitates the resolution of disputes, thereby fostering a healthy development of data trading.

III. Data Trading Model

The data trading model DCTM consists of six roles, including three entity roles and three non-entity roles: Data Owner,Data Buyer, Decision Tree Classification Service Provider, Smart Contract, InterPlanetary File System (IPFS), and Key Distribution Center (KGC). Below are detailed descriptions of these six roles:

Data Owner: The entity role that owns the source data. Data owners classify the data and resell the classified data to earn profits.

Data Buyer: An entity role with a demand for the source data owned by data owners. Data buyers seek high-quality source data and are willing to pay corresponding rewards.

Decision Tree Classification Service Provider: An entity role that earns profits by providing classification services to users. During the data classification process, this service provider does not have access to the original data; rather, it classifies encrypted source data ciphertext.

Key Distribution Center (KGC): Responsible for distributing keys to data owners and decision tree model service providers.

Smart Contract [6]: The core of the entire data trading model, ensuring that data transactions are executed smoothly according to established rules. Smart contracts ensure the fairness of the data classification process, record transaction information, manage public-private key pairs, manage funds, and complete fund settlements.

InterPlanetary File System (IPFS): Provides data storage services.

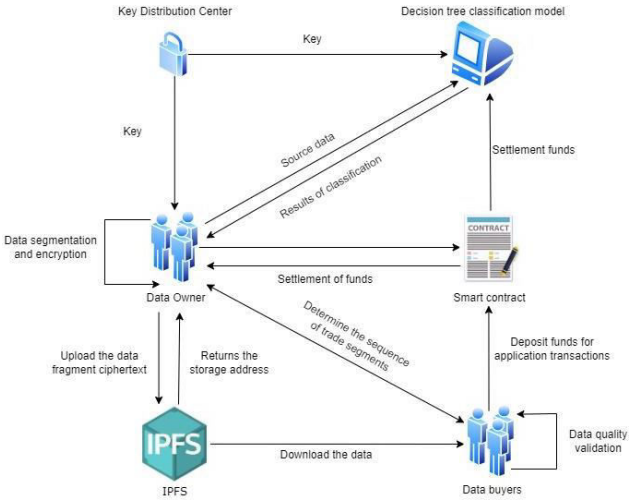

Figure 3.1 Data trading model

A. *Data classification*

Before the data transaction between the two parties, the data owner first categorizes the source data and preliminarily selects the dataset that meets the transaction requirements. This process begins with the KGC generating public parameters and private keys, which are then sent to the data owner. The data owner generates corresponding public keys and uploads the public-private key pairs to the smart contract, which is responsible for storing and distributing the key pairs. The decision tree classification model service provider encrypts decision node thresholds using the public key. Subsequently, the data owner deposits funds into the smart contract to pay for the classification fees of the decision tree classification model service provider. After the secure comparison stage between the feature vectors of the data owner and the decision node thresholds of the decision tree model service provider, the data that meets the requirements falls into the leaf node. The following figure represents a decision tree with 5 leaf nodes. There are a total of 5 paths from the root node to the leaf nodes.

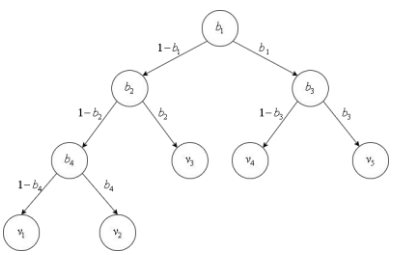

Figure 3.2 Decision Tree Classification Example Diagram

In the stage of obtaining the classification results, we transform the decision tree model into a linear function containing the classification results [7],as follows:

$$\begin{cases} R_{v1} = (1-b_1)+(1-b_2)+(1-b_3) \\ R_{v2} = (1-b_1)+(1-b_2)+b_4 \\ R_{v3} = (1-b_1)+b_2 \\ R_{v4} = b_1+(1-b_3) \\ R_{v5} = b_1+b_3 \end{cases} \quad (1)$$

B. Data trading

After the classification process with the aforementioned decision tree classification model, we filter out the data that meets the requirements for data trading. The classified data is then encrypted and uploaded to IPFS. The next step involves calling the smart contract to publish the data commodity. The first step of data commodity publishing by the contract is to determine the identity of the caller of the publishing contract and check if the address is blacklisted. If the address is blacklisted, it outputs "You are not authorized to sell data" to prevent data resale by dishonest traders. After identity verification, the contract verifies whether the funds deposited by the data owner are greater than the total data amount. Only when the funds deposited by the data owner are greater than or equal to the total data amount can the publishing be successful. If the deposited funds are insufficient, an exception will be triggered, and an error message will be displayed.

When the data buyer creates an order, the contract first verifies whether the user address calling the order creation contract is blacklisted. If it is blacklisted, it outputs "You are not authorized to purchase data." After verification, the contract checks whether the funds deposited by the data buyer are sufficient. If the deposited funds are insufficient, an exception will be triggered, and an error message will be displayed. Then, the contract checks if the data commodity exists, and finally, it verifies whether the data owner and the data buyer are the same person. It is specified in this contract that the data buyer cannot purchase data commodities they own.

After the data buyer obtains the data, they conduct data quality testing on the data fragments. Upon passing the data quality testing, they can choose to continue the transaction or terminate it. At this point, the data owner can call the order settlement contract to apply for order settlement.

C. Data prevents resale

To prevent data buyers from purchasing data and then privately reselling it for profit, a resale prevention contract has been designed. In the resale prevention contract resalePrevention, we restrict data buyers' resale behavior by setting up a transaction blacklist. If it is detected that a data buyer is reselling the traded data, their address will be added to the data trading blacklist (changing the blacklist status to true), making them ineligible to participate in any future data trading activities, including data selling and buying. Before data publishing and creating transaction orders, the identity of the caller is validated. Their transaction address is checked against the blacklist using require(!blacklist[address]). If the address is on the blacklist, it indicates that the caller has previously engaged in data resale activities and has been blacklisted, thus denying them the privilege to publish data or create orders. If the address is not on the blacklist, the subsequent data transactions proceed.

D. Data quality validation

Due to the unique nature of data resources, assessing data quality across different dimensions may yield varied results. Data quality evaluation is influenced by factors such as data type, usage scenarios, and application purposes. Since this paper primarily focuses on testing medical data, the data quality assessment model herein divides data quality evaluation metrics into six dimensions: data integrity, data accuracy, data validity, data uniqueness, data timeliness, and data consistency.

When conducting data quality assessment, we typically evaluate source data across multiple dimensions using various metrics. We assign different weights to each dimension based on its importance in the overall data quality assessment, employing a method commonly known as weighted averaging. Each dimension's weight must range between 0 and 1, and the sum of weights for all dimensions must equal 1. That is

$$\lambda_1+\lambda_2+\lambda_3+\cdots+\lambda_n=1 \quad (2)$$

$$X=\lambda_1 X_n+\lambda_2 X_n+\lambda_3 X_n+\cdots+\lambda_n X_n \quad (3)$$

X_i stands for data quality evaluation, λ_i stands for weight of data quality indicators, We use X for the total score of the quality evaluation of the source data, and Y for the pass rate per rule performed (pass rate = the number of data conforming to the rule as a proportion of the number of data from all sources) , then the data quality evaluation index calculation formula is as follows:

$$\sum_{i=1}^{n} \lambda_i = 1 \quad (4)$$

$$S = \sum_{i=1}^{n} \lambda_i \cdot R_i \quad (5)$$

IV. Experiment

A. Experimental environment

The experimental environment of this study uses an Intel i7 CPU. Smart contracts are written in Solidity and tested in the Remix compiler. Remix is an online editor provided by Ethereum for quickly writing, debugging, and deploying contract code, and deploying contracts to Ganache. Ganache is a local simulator used for developing and testing blockchain applications, allowing for running tests, executing commands, and more. MetaMask is an online wallet management client that does not require downloading; it only needs to be added as an extension to the Google Chrome browser, making it very lightweight.

B. Decision Tree Model Classification experiment

We classified the test data, obtained the classification accuracy of the decision tree classification model for different data sets, and calculated the computational overhead for each data set, as shown in table 4.1 below:

Table 4.1 classification accuracy and computational overhead

Dataset	Classification accuracy	Computing overhead/s
Breast-cancer	0.9135	2.385
Heart-Disease	0.8917	1.871

Based on the experiments of Breast-cancer data set and Heart-Disease data set, the classification accuracy of the two

data sets is about 90% , the classification effect of decision tree classification model is good. The computational overhead of the two data sets is about 2 seconds, which verifies the efficiency of the decision tree classification model.

C. *Data transaction process validation*

After the data transaction starts, the data owner first invokes the data release function to release the data commodity. At this time, if the data owner does not deposit the money in advance or the data owner does not deposit the money enough, Ganache outputs The log information shown in figure 4.1 below. The log output reads"The deposit is insufficient, please top up!"

```
1:04:30]  Runtime error: revert
1:04:30]  Revert reason: The deposit is insufficient, please top up!
```

Figure 4.1 Publish failure log information

When the data buyer resells the data and is blacklisted from the data transaction, Ganache outputs the log information as shown in the figure below, which reads"You are not authorized to purchase data

```
3:48:03]  Runtime error: revert
3:48:03]  Revert reason: You are not authorized to purchase data
```

Figure 4.2 no right to buy tips

V. CONCLUSION

This paper proposes a Data Classification-based Model (DCTM) for addressing data quality and privacy protection issues in data transactions. In the data classification phase, a decision tree classification model is introduced to classify source data, ensuring classification of the source data without the service provider knowing the plaintext data.Subsequently, IPFS is employed to store classified data, enhancing data access efficiency. In the data transaction phase, decentralized data transaction solution is built based on blockchain, with detailed smart contracts designed for each stage of the data transaction process. Different roles engag e in data transactions by invoking different contracts,ensuring that the data transaction process adheres to established rules and enhancing fairness and security. Finally, a local Ethereum private network is constructed based on MetaMask+Ganache, and contracts are deployed using the Remix online smart contract editor to conduct functiona l and performance tests on the data transaction model, demonstrating its feasibility and effectiveness.

REFERENCES

[1] Adlakha R, Sharma S, Rawat A, et al. Cyber security goal's, issue's, categorization & data breaches; proceedings of the 2019 international conference on machine learning, big data, cloud and parallel computing (COMITCon), F, 2019 [C]. IEEE.

[2] Dai W, Dai C, Choo K-K R, et al. SDTE: A secure blockchain-based data trading ecosystem [J]. IEEE Transactions on Information Forensics and Security, 2019, 15: 725-37.

[3] Li T, Ren W, Xiang Y, et al. FAPS: A fair, autonomous and privacy-preserving scheme for big data exchange based on oblivious transfer, Ether cheque and smart contracts [J]. Information Sciences, 2021, 544: 469-84.

[4] Jung T, Li X-Y, Huang W, et al. Accounttrade: Accountable protocols for big data trading against dishonest consumers; proceedings of the IEEE INFOCOM 2017-IEEE Conference on Computer Communications, F, 2017 [C]. IEEE.

[5] Xiong W, Xiong L. Smart contract based data trading mode using blockchain and machine learning [J]. IEEE Access, 2019, 7: 102331-44.

[6] Szabo N. Smart contracts: building blocks for digital markets [J]. EXTROPY: The Journal of Transhumanist Thought,(16), 1996, 18(2): 28.

[7] Bost R, Popa R A, Tu S, et al. Machine learning classification over encrypted data [J]. Cryptology ePrint Archive, 2014.