DataToken

OwnershipLab

February 2021

Contents

1 Introduction

DataToken had been selected as Grants Planned Projects of Platon,which is developed by Ownership Labs.DataToken implements a cross-domain distributed data management and remote computing SDK to realize the accurate tracing and audit of the workflow–data sharing and mining by different participant s based on Platon and Rosetta

As the rapid development of Big Data,data has become important asset and business engine for many companies.However,there are barriers to data circulation among various participants, and the traditional IT technology stack cannot simultaneously meet the requirements of data cross-domain computing, user privacy protection and data supervision&audit.The Blockchain technology has properties of Peer-to-Peer,consensus of transactions and tamper-proof,which makes it has enormous potential in data attestation,data authorization and data traceability etc.In order to realize the secure and trusted flow of data assets within the private domain, this project will implement a cross-domain distributed data management and remote computing SDK(Datatoken) based

on blockchain and cryptography, and ensure that the whole process of data sharing and utilization can be accurately tracked and audited.

Business departments and organizations are isolated from each other nowadays,which makes cross-domain data sharing difficult as once the data assets leave the private domain, they will face the risk of data stealth collection or even resale by the third party, this results in the challenge to guarantee the core interests of the data owner.At the same time,data set usually contains a lot of user data,and the requirement from regulators to regulate user data sharing is increasing.The laws and regulations about user privacy protection and data proprietary rights are improving gradually.

The current privacy computing scheme realizes the multi-party data collaboration without the original data out of the private domain,but it does not guarantee strict authority management during the fusion calculation of the private domain data, which doesn't meet the regulatory compliance requirements for the data application subject and the data hosting subject.Meanwhile, users still do not have the right to know the whole process of their data use.

Our project will provide a simple and easy to use cross-domain data fusion computing SDK, implementing Compute-to-Data Distributed Service Specification without the data out of domain while algorithms applying to the data.Our system allows the data requester to send trusted remote algorithms to the data side for local computation, with the enterprises/users in full control of their data assets, while ensuring that the entire process of data manipulation is traceable and auditable.

Data owners will know who will use their data and how their data assets will be processed and distributed in advance, so as to independently authorize the local computing right of the data. Data application subjects can also make more compliant use of cross-domain data assets for data collaborative analysis.

Different from Ocean Protocol's remote computing scheme for single domain data, the DataToken SDK proposed in this project defines a trusted distributed computing workflow on multiple data sources (and their user sub-data) and multiple computing power, and how to track the whole process of data cross-domain fusion calculation.

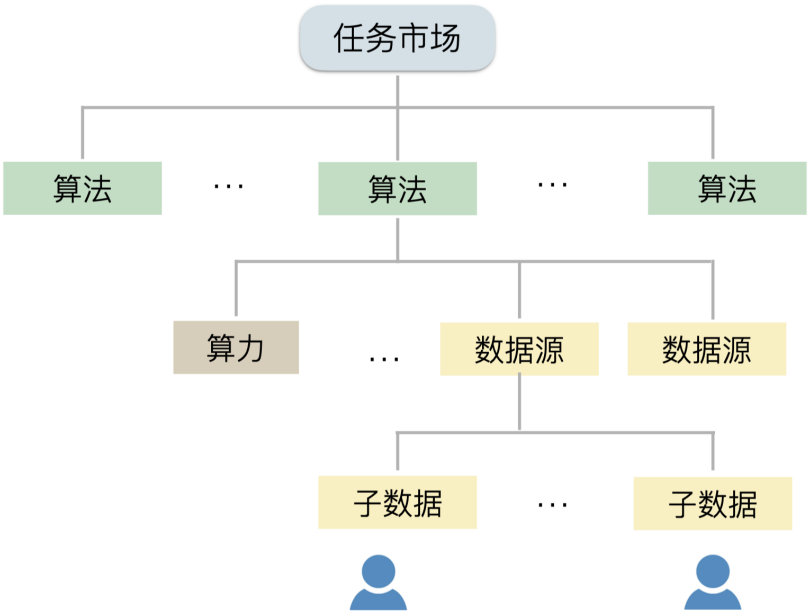
We believe that only through credible and comprehensive data sharing and use of information can the price discovery mechanism of the market price the data reasonably, in turn constructing the trillion-dollar data market.

2 Architecture

2.1 Problem statement

In the following description,the participants of a multi-party data collaboration task include data source party, computing power provider, algorithm provider and data application party.The joint computation problem can usually be represented by the nesting structure shown below:A data source can be composed of multiple sub-data sources, a remote algorithm can be applied to multiple data

sources and multiple computing power, and a task can be composed of multiple algorithm stages.



Data, computing power, algorithms, etc., can be understood as assets, which are uniquely identified by Datatoken (DT). Different assets have different metadata, which are represented by distributed document object DDO.

A DT registry is maintained on the blockchain to quickly locate the IPFS storage location of DDOs down the blockchain:

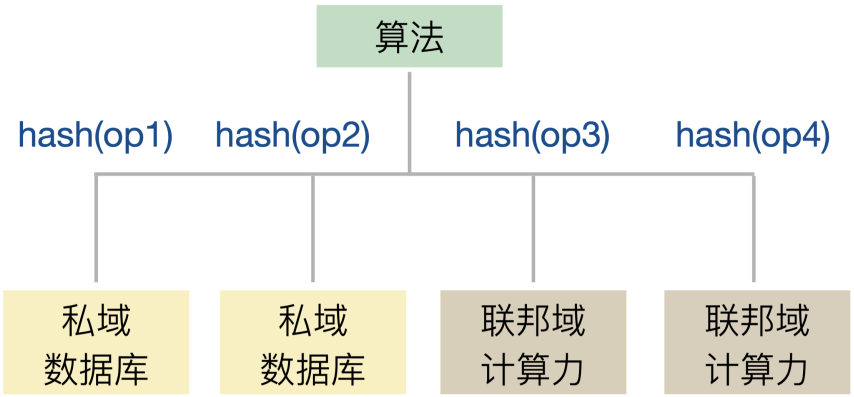
The DT identifier is registered on the blockchain as: $\{DT, owner\_address, storage\_path, proof\}$

DDOs are stored down the blockchain as: $\{DT, proof, services : [type, endpoint, child\_dts, supported\_ops, workflows, extra\_params]\}$

Proof =hash(metadata), Type distinguishes data, computing power, and algorithm resources, endpoint is the endpoint of the service, and Extra_Params contains information such as name, description, provider, and price.Child_dts = [0: dt_0, 1: dt_1,..., n: dt_n], if it is empty then it indicates underlying resource.An enterprise data source DDO can be represented with a structure that contains a DT list of multiple user data.

2.2 Remote computing specification

To enable distributed computing across domains, the data structure of an asset DDO should contain not only multiple child DT identifiers, but also a hash of action code for these resources.Taking the joint modeling of two banks as an example, the third party technology company provides the algorithm:



Among them, OP_1 and OP_2 are secret shared operations, and OP_3 and OP_4 are joint AI operations on SS fragments.The algorithm DDO contains the working certificates of four resources, which means that the original data in the private domain database is secretly shared and then sent to the federated domain computing power for joint modeling.

For example, fintech company C provides customer profiling and joint risk control services for two banks A and B.Assuming that the original customer data of the bank is located in the database of the private domain network, the customer data can be encrypted and transmitted to the federated domain network between banks through secret sharing SS, and MPC joint modeling can be carried out on the basis of the ciphertext.

The asset owner should be able to quickly set up remote computing specifications in the DDO, such as supported operations, distributed workflows, and so on, by first normalizing the operation code to form a set of trusted remote code templates.Action code can also be uniquely identified and stored on chain using its hash, and code scripts are stored in IPFS, this step can be performed by a system administrator.In this way, supported_ops and workflows in DDO can be configured using the unified code identifier:

Supported_ops: $[0 : op\_0...m : op\_m]$

Workflows: $[set\_ops : [0 : dt\_0\_op, ..., n : dt\_n\_op], configs]$

When the child_dts list is empty, the workflows should also be empty. In this case supported_ops represents the set of on-chain identifiers of local code operations supported by the underlying asset, such as mobile user data that supports federated learning.When the child_dts list is not empty, supported_ops represents the operations that are consistently supported by all child dts of the asset.At the same time, you can define more complex distributed workflows in Workflow, set_ops contain specific operations for each sub-DT, configs contain parameters for execution or specify the workflow calculation order.

2.3 Task market

High-level assets need to be authorized to use lower-level assets before they can perform actual calculations.Lower-level assets typically verify whether the operational code in the higher-level asset DDO complies with their own support terms, if true,enable on-chain authorization. In complex practical problems, a hierarchical agent structure can be used. For example, DT_1 grants authority to DT_2, and DT_2 grants authority to DT_3, and the DT_3 owner can be considered to obtain local operation authority on the DT_1 asset.In this way, the algorithm provider can perform calculations locally on the user data under the enterprise data source. It is worth noting that in a multi-party data collaboration task, obtaining authorization does not mean that a remote computation can be initiated immediately, otherwise the lower-level asset cannot track its full process usage and the actual computation cannot be performed until all participants authorize.We design an on-chain task marketplace as the termination state of the algorithm DT. When a trusted institution has the DT in the task, it means that someone has vouched for the remote computation and all authorization has been obtained.Lower-level assets can also learn how and by whom they will be used (by querying on the chain whether they have authorized the algorithm DT or its sub-DT, in turn, sub-DT of the sub-DT).

3 MVP interaction flow

The Datatoken SDK allows for trusted traceability calculations on distributed resources and will provide three scenario MVPs: 1) data collaboration among enterprises; 2) User-level edge computing; 3) Traceable privacy AI computing.

3.1 Data collaboration among enterprises

Consider simple two-way vertical federal learning, such as fintech company C providing joint risk control services to two banks A and B.

Assuming that the original customer data of bank is in the database of its private domain with high security level, we transfer the encrypted customer data to the federal domain network with slightly lower security level by cryptography scheme (such as secret sharing SS), in which we assure the data is secure and the operation is auditable, and then we can further realize the MPC joint modeling on the basis of the cipher. Here, Bank A and B are both data sources and computing power providers, while Fintech Company C is algorithm provider and data application provider.The MVP process is as follows: 1.The contract deployer is the system administrator, who adds the organization name and account relationship of A, B and C, registers the SS and MPC code hash on the chain, and stores the script in IPFS.

2.Bank A and B use DT/DDO to describe the terms of service for private domain data and federated domain computing. The former should support SS operations, and the latter should support MPC operations, generating four

DT identifiers and four DDOs, and then store DDOs in IPFS and register dt, storage_path, and proof on the chain.

3.Company C combines the data/computing power DT, fills in CHILD_DTS and hash code of each resource, generate the algorithm DT, and then store them on IPFS and chain.

4.Bank A and Bank B verify whether the algorithm DT meets the terms of resource usage respectively. After the verification is passed, the data/computing power DT is authorized to the algorithm DT on the chain.

5.Company C creates a new task in the on-chain task market and submits the algorithm DT to the work of this task;

6.Company C operates the data/computing power of Bank A and Bank B remotely.Banks need to verify whether the algorithm DT is authorized,get owner signature, and its task working status, etc. If so, download the code script from IPFS and perform the local computation.

3.2 User-level edge computing

Consider simple horizontal federated learning on mobile devices, such as Smart Health B's desire to train heart disease prediction models on all user data from wearable device provider A.User data is local to the mobile device, denoted as U_1, U_2...U_n. In this case, Provider A is the data source that contains a lot of user data, and Company B is the algorithm provider and data application party. The calculation process is completed in the data private domain. There is no explicit computing power provider, i.e., equipment vendor A run gradient aggregation (such as FedAverage), and user equipment run horizontal model calculation (Edgecomp). The MVP process is as follows:

1.The system administrator deploy the contract,add the organization name and account relationship of A and B, at the same time, the Edgecomp and FedAverage code hashes are registered on the chain, the scripts are stored in IPFS

2.Device vendor A adds identity of its users and sets up an on-chain user registry to control the release of assets;

3.N users and device vendor B all register data assets. Edgecomp operation should be supported by user DDO, and n user DTs should be combined in device vendor DDO, and workflows should be defined. Workflows =[set_ops:[0:ec...n:ec,n+1(self):fa]], that is, to run horizontal models on user devices and converge gradients in device vendor's private domain;

4.Medical Company B registers algorithm assets containing device vendor DT and defines workflows, Workflows =[set_ops:[0:self]], indicating that the algorithm directly uses data source B's workflow;

5.The user verifies the data source DT of the device vendor, while the device vendor A verifies the algorithm DT. If both meet the terms of use of resources,then the authorization relationship between DT is completed on the chain.

6.Medical Company B creates a new task in the on-chain task market and submits the algorithm DT to the task's work

7.Medical Company B makes a remote operation request to device vendor A. Device vendor A verifies the algorithm. If it is passed, it will first inform all users and then run FEDAVERAGE.

8.The user receives a message from Device vendor A, and has the right to know about the data utilization (which is traceable from the task market and DT authorization relationship information on the chain), and then runs Edgecomp.

In the above process, neither the device vendor A nor the medical company B have access to the original user data, only the intermediate data or the result data. Each party knows how its resources are being used, and the user does not perform local computation until authorization is confirmed.

3.3 Traceable privacy AI computing

This project implements traceable privacy AI computing based on Rosetta, RTT-Tracer. The MVP will be developed in combination with Datatoken component to realize the federation model under secure multi-party computation. In addition to the DT full functionality, simple asset service deployment and federated computing capabilities are involved:

1.The joint model based on RTT provides examples of user credit default and registers it as trusted OP template on the chain.

2.Deploy asset services based on flask, including local computing in private domain data and proxy computing in federated domain.

During this process, the data application party sends actual authorization requests and remote computation requests to the flask services of multiple asset parties.After the asset party validates the request, the trusted RTT code is downloaded for local or federated computing. Regulators can also inquire about asset usage.

4 Conclusions,Milestones and Prospects

The complete project will be based on the Platon blockchain and the Rosetta privacy computing framework, involving five code repositories such as dt-contracts, dt-web3, dt-asset, dt-SDK, RTT-tracer, etc. Key milestones can be divided into M1-M4 stages:

M1:On-chain contracts(dt-contracts) and off-chain interactions (dt-web3): the former mainly includes administrator and organization registration, record dt on chain and authorization,trusted OP registration, task marketplace;the latter mainly includes migrating the Keeper-py-lib of Ocean, and the off-chain instance class corresponding to the on-chain contract.

M2:Off-chain metadata management and computing protocol toolset(dt-asset): mainly includes DT identifier generation, DDO metadata management, creation of computing service workflow, and IPFS storage connector.

M3:Business layer development tool for multi-party data collaboration(dt-sdk), mainly includes functional encapsulation of several business participants,

namely system administrator, data provider, computing power provider and algorithm provider, and completes MVP1 and MVP2 at the same time

M4:Traceable privacy AI computing(rtt-tracer):Complete MVP3

In summary, this project is more about capitalization of the whole process of multi-party data collaboration and record it on the blockchain. In the future, we will develop based on Rosetta to form a series of trusted distributed computing OP templates and provide a one-stop multi-party data collaboration portal platform with a visual interactive interface for data collaboration and data traceability.