Datashop

Open collaboration and data exchange platform based on blockchain

Kai Ansaari, Matt Nguyen, Jazear Brooks

Pbilip Stanislaus, Bin Yu

Chris Yu, Stefan Beyer, John–Paul,Thorbjornsen,

Datashop Team August 25, 2020

1 Abstract

The vigorous development of Bitcoin and Ethereum has made people an important role in the blockchain. The reasonable application of blockchain technology can ensure the credibility and transparency of historical data. It is possible to retrieve credible and transparent historical data for specific chain identities. In the current Internet environment, people are suffering from the lack of distributed metadata, which leads to problems such as the lack of unified platform standards, copyright infringement and data islands. Datashop is committed to building an open collaboration platform through the tracking, retrieval and judgment of specific user data on the platform. Solve problems such as the above by creating distributed and decentralized metadata storage.

Datashop will use the Substrate-based proof-of-stake blockchain/smart contract layer for development, while adopting the IPFS protocol and saving the data to the Filecoin storage network. In the future, Datashop will make full use of Polkadot's cross-chain protocol, will support connection to other different distributed ledgers, and use different databases and storage layers.

2 Introduction

There is no doubt that the World Wide Web is constantly evolving. The current interactive Web 2.0 faces multiple issues in terms of trust, review and distribution. The distributed Web, sometimes called Web 3.0 or OpenWeb, is based on verifiability and provides potential

solutions to these problems. At the same time, these platforms also have some drawbacks of centralized platforms: -Mandatory ownership: In most distributed ledger systems, a key requirement is to own the network tokens to participate in. -Incentive structure: Generally, network providers are either not motivated at all or motivated for providing unnecessary services. -Scalability: Fully distributed applications cannot compete with the scalability of centralized applications. In addition, the distributed Web should try to overcome the existing problems of the current Web, such as -Data islands: Distributed systems usually lead to distributed and disconnected data islands, or rely on centralized permissions to achieve searchability and accessibility. -Copyright: Distributed networks are currently mainly used for copyright infringement and do not provide a clear ownership structure. Therefore, users will lose complete control over their content. This article attempts to solve these problems to a certain extent by combining multiple existing technologies. The key technologies behind it are IPFS, Filecoin and Substrate.

3 Datashop overview

3.1 Participants

The first thing Datashop has to do is to establish a blockchain-based data transaction platform, on which any person or organization can publish data, and any person or organization can retrieve and trade data. The platform mainly includes the following participants: -Datashop: Interoperable blockchain/smart contract layer (reference

implementation based on Substrate). -Dataspace: Distributed storage layer (reference implementation is based on IPFS and Filecoin network). -Council: The management committee, whose main task is to review the compliance of Data and resolve disputes between users uploading data. At the same time, the Council will keep the user's service fee and decide how to use it to promote the development of the platform.

3.2 Assumption

This article is based on three assumptions and the resulting design principles. These assumptions are not facts, so they can be discussed. Distributed Web-The World Wide Web should be diverse and decentralized. Therefore, all systems should try to document interoperability and distribution. In addition, if it is possible instead of finding a single truth for everyone, then users should decide for themselves who to follow, which cluster to participate in, etc. This assumption is also based on the macroeconomic assumption that the ability of monopoly or oligopoly will reduce efficiency. Perfect competition. The Internet in its current state has multiple monopolies or oligopolies in different regions. For example, Google has a market share of more than 90% and quarterly revenue of more than $30 billion. This is not only economically inefficient, but also increases opportunities for censorship and search bias. Therefore, the first goal is to create a distributed diversified network.

1. Openness Datashop will develop a public chain based on Substrate to carry its main business. As long as it generates a public-private key pair and holds platform tokens, it can

participate in the system, without the need for a lot of identity information and credit card information required by traditional platforms. Opening does not mean that it will be used illegally. The release of user data must first be reviewed by the Council. After the review is passed, platform participants can see it. For some Bounty with malicious and sensitive information, the Council will not allow it to pass.

2. Data can be retrieved and traded Users should be able to own, retrieve, and should be able to price and trade the data they own. At the same time, the data owner should have anonymity. Everyone can create their data completely anonymously.

3. Collaborative behavior will be more efficient Council plays the role of data review and management. First, the review work will ensure that the data has a sufficiently clear description. Third, the Council will also clean up the data that cannot be resolved for a long time or that has not been applied for. This ensures the validity of the information on the platform and avoids the interference of too much useless information. At the same time, too many bad behaviors of participants will have a negative effect on their credit, which will inhibit the bad behaviors of participants.

4 Datashop Architecture

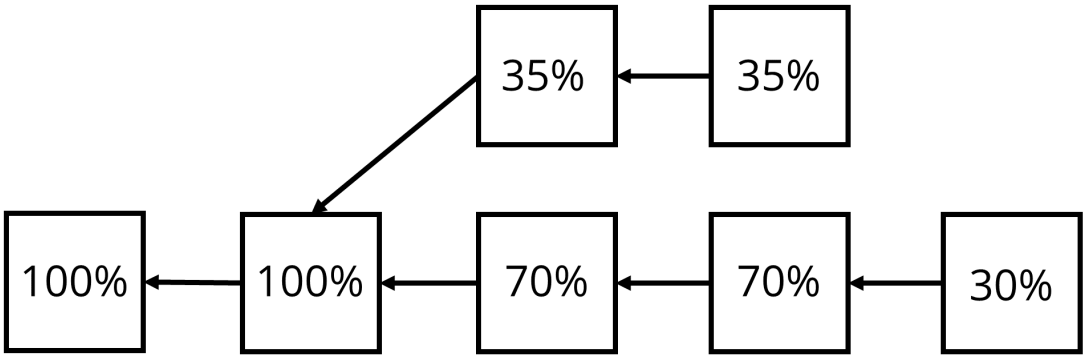
The network itself consists of a blockchain/smart contract layer based on Substrate, and a distributed storage layer composed of the IPFS protocol and Filecoin. By dividing the technology stack into different layers, each layer can be adjusted according to their key

requirements. For example, Datashop will fully consider loose coupling when designing, and the blockchain/smart contract layer only processes top-level user information. The storage layer can process metadata information independently and efficiently.

4.1 Blockchain/Smart Contract Layer: Datashop is based on Substrate

Substrate is an open source framework of Parity Technologies, used in the blockchain developed by the Rust language.

It can be connected through so-called bridges and Parachains to realize future multi-chain networks. The framework supports a diverse and extensible Web, where various distributed hosting applications use smart contracts or simple runtimes. In addition, it includes upgradeability from the beginning, no hard fork is required, and utilizes a hybrid GRANDPA (GHOST-based recursive ancestor derivation protocol) and BABE (Blockchain Extension for Braille Distribution) consensus algorithm. BABE is a block production mechanism that determines the new block producer. GRANDPA combines probabilistic finality similar to proof-of-work with Byzantine provable finality. In GRANDPA, validators vote for blocks of different heights. Once a block gets more than 2/3 of the votes, the block becomes part of the chain. If there are different parts in the chain, people will follow the chain with more than 2/3 votes, so in the image, the bottom of the chain.



In order to ensure that Datashop can make full use of the cross-chain collaboration features of Polkadot in the future, we recommend implementing Datashop as Polkadot's Parachain, so that it can be easily connected to various blockchain projects. In addition, Parachain will benefit from the shared security of the relay chain. Since Datashop permanently stores data items based on network consensus, it only tracks the most important information. This information includes:

Attribute	Type	Description
DID	Vec<u8>	Decentralized Identifier (max. length 100 characters)
unique name	Vec<u8>	An optional unique name (max. length 50 characters)
license code	u16	Numbers referencing the license of the data
timestamp	Time	The timestamp of the entry

All types of attributes only allow a limited size to keep each transaction information that needs to be chained to a minimum to prevent potential denial of service attacks. -DID points to metadata storage. For example, the following string represents a valid Datashop DID:

did:fil:0b51b44e0330995979a5ddaa206260b1c18e2471ad51043c27d68d8a9c40261f The prefix "did" stands for decentralized identifier, and "fil" stands for "Filecoin". The rest of the DID is the FilecoinID of the first metadata upload, which is the SHA3-256 hash value of the transaction. This hash can be used to track all future changes to metadata entries on Filecoin. Datashop allows users to register unique names similar to domain names, which can be combined with metadata. -The concept behind unique name and DID is based on the Ethereum Non-Fungible Token (NFT) standard. Therefore, each Datashop DID and name entry is unique, so it can be collected and traded. The Substrate chain also stores ownership and content licenses in the form of digital license codes, which also allows public recording of delete requests for owned content. -license code Number of license code represents a specific license status. The system enables content creators to provide signed information about the right to use their works, thereby providing possible technical solutions for EU copyright directives and GDPR compliance. -timestamp The main benefit of timestamp is to provide reliable timestamp services for various digital content. For example, this is useful for proving the existence of certain documents.

In addition, Datashop stores all government activities and token movements. The user signs all uploaded Datashop items and therefore owns the ownership of the data. However, if they do

not share a specific public key, others will not know the owner of the data. In order to connect the above interactions with other layers of the network, on-chain interactions will trigger certain events, which are then processed by the underlying contract. The contract will provide a variety of services around the content of the data. For example, the development of content data for issuing unique identifications, and the provision of storage/custodial or verification data transaction markets.

4.2 DataBridge – WebSocket Client

Datashop talks to other layers via an event-based system, which is triggered based on specific chain events. For example, in case a user logs a delete request on Datashop, it needs to be ensured that the specific Captain's Log and Dataspace also receive this information. If only the owners are responsible to share this request between the different layers, they might in certain cases unintentionally or maliciously not inform all involved parties and later claim that someone ignored their request. Therefore, independent Databridges subscribe to these events via a WebSocket connection and ensure the appropriate action on the storage or metadata layers. Different groups of Databridges are randomly responsible for different block numbers. One group member is responsible for sending the information, and the rest of the group is checking and potentially challenging the execution of this selected member. Once the update is executed, only unsuccessful executions (e.g. due to network issues) are logged on Datashop. If this is the case, the next group of Databridge takes care of the job.

To calculate the probability for firing an event when a certain Databridge is currently unavailable, this paper assumes that the probability that a Databridges operates without failure for a time t is given by the following equation based on the paper by Jaynes, E. T., 1976 `Confidence Intervals vs. Bayesian Intervals'.

$$Pr(\theta \geq t) = e^{-\lambda t};\ \ 0 < t, \lambda < \infty$$

λ is, in this case, the unknown "rate of failure". With this you can derive the following equation for the probability for hitting the system when it's unavailable:

$$Pr(\theta \in (0, S) | \hat{F}, T_U, T_D) = 1 - \frac{\left(\frac{T_U}{T_U+T_D}\right)\left(\frac{T_U}{T_U+S}\right)^2}{\hat{F} - (\hat{F}-1)\left(\frac{T_U}{T_U+S}\right)}$$

Where TU is the "uptime" and TD is the "down time", in seconds. θ is the time of the first downtime (in seconds) observed by the user and F is the expected number of "downtime periods". This equation shows that if Databridge providers ensure high availability, the probability that an available honest provider didn't get an event gets quickly close to zero. Especially if there are multiple rounds by different Databridge providers.

4.3 Data description

The description information includes two categories, metadata (Metadata) and business data (Business Data).

4.4 Metadata

The metadata part is based on the "ERC721 Metadata JSON Schema", which contains the following elements:

Attribute	Type	Description
name	string	A descriptive title of the data
description	string	A detailed description of the data
thumbnail	hash	A thumbnail of the data or file
hashes	array	An array of the hashes, which represent the metadata
time	array	An array of the timestamps representing the creation or update of the metadata
storage location	array	An array of permanent storage locations, which ensure the availibility
filetype	string	The file type
similarity digest	string	A context-sensitive hash
additional meta	object	Additional attributes, like for example categories, which can be defined by marketplaces.

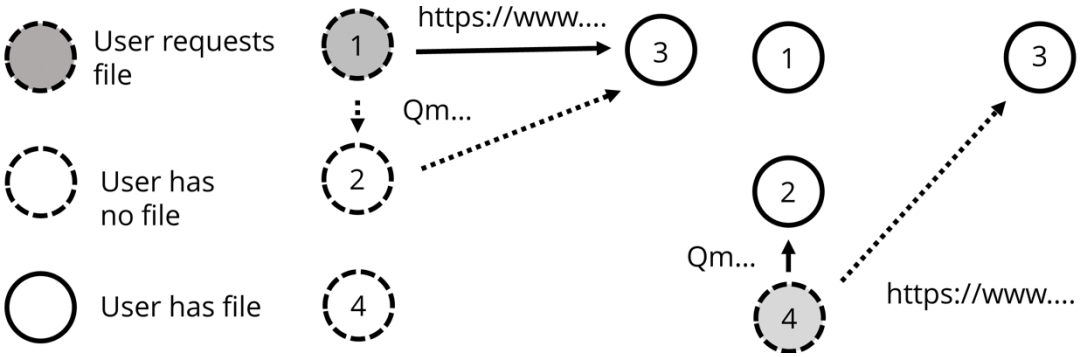
The similarity digest is a context-sensitive hash, which allows comparing two different hash values to obtain an estimate of the similarity between two documents. This is especially useful for finding nearly duplicates of search engine documents. The "Other Metadata" attribute

allows additional data to be attached, which may be useful for different markets, for example. [14] In addition, metadata also stores the following information, which has been stored on the chain:

- unique name
- license code

Complete data can be used, for example, for decentralized searchability, verifiability, near-duplicate detection, and a combination of address-based and name-based storage. Each metadata entry must contain a unique file hash that points to the file stored on Filecoin.

Since one of the main assumptions of this article is distribution and diversity, the captain's log provides location-based links as well as name-based addresses. This means that the content can be found by the hashed name and the current URL. See below.



User number 1 requests the file of user number 3 from the network through location-based requests (http://www...) and name-based (Qm...). Since the file has not been distributed yet, the traditional method will load the file faster.User number 2 recognizes that he is interested in the file and decides to load the file as well.

Now, when user No. 4 requests this file in the same way as user No. 2, he or she receives the file from his neighbor No. 2 through a naming-based system. In addition to the above information, the metadata layer also stores availability and verification data. This is actually data provided by someone other than the publisher, which describes the uploaded content in more detail.

Availability data helps provide information about the availability of content in distributed networks. This is essential when no one provides permanent storage for uploaded content. Data verification is essentially a layer of trust in which different people can evaluate the quality of the data provided. Instead of trying to find a single truth directly on the blockchain (for example, the Token Management Registry (TCR)), the metadata system uses a subscription-based system in the Filecoin retrieval market. The reason is the high frequency of expected metadata changes and updates, as well as different interests when it comes to search topics.

4.5 Business Data

Business data serves the upper-layer business of the platform, such as information retrieval and Bounty collaboration. As the business expands, this part of the data may expand. The preliminary business data mainly includes: -Category: Bounty category, such as design, copywriting, development, etc. -Report Schedule: Indicates the schedule for completing the status report after being authorized by Bounty, such as daily, weekly, etc.

4.6 Dataspace

Compared with the previous layer, Dataspace stores various types and sizes of data. Therefore, Dataspace uses the IPFS protocol and Filecoin as the initial basic technology layer. However, the metadata system can easily support other protocols and their respective interoperability.IPFS is a protocol used to create a content-addressable, peer-to-peer data storage method.Dataspace only stores information based on a single request. Therefore, if the license code on Datashop is changed and the event-based system is triggered, the information can also be deleted. The distribution of content on the storage network also depends on this license code. For example, you can specify that data should only be stored in one location, and not distributed across the network. In this way, the settings comply with GDPR requirements.

5 Platform Behavior and Credit

5.1 Collaboration Unit

The collaboration unit can be simply understood as what the participants want to accomplish together. For example, "solving a Bounty" can be regarded as a collaboration unit that needs to be completed by Fund, Hunter, and Council. A large collaboration unit may include multiple small collaboration units. For example, "Solving a Bounty" may include a "progress report" sub-collaboration unit. Therefore, the cooperative unit may exist alone or in the form of a tree of cooperative units.

Each collaboration unit has a corresponding result set, which represents the final collaboration result of the unit. For example, the result set of the collaboration unit "Progress Report" is `(Yes, No)`; Suppose there is a collaborative unit of "Bounty Completion Evaluation", which means that after Bounty is completed, the mutual evaluation of Funder and Hunter, then the result set of this unit is `(1,2,3,4,5)`. We define platform behavior as the specific result produced by the collaboration unit, and this behavior can be directly used in the calculation of user credit.

5.2 Selection criteria for collaboration units and behaviors

We emphasize once again that behavior is the basis of credit investigation. Reasonably choosing the behavior to be tracked and correctly judging the user's behavior is the guarantee of building user credit. Therefore, the choice of collaboration unit is very important, The collaboration unit selected for tracking on the platform must comply with several basic principles.

1. Can reflect the credit of the actor The purpose of tracking behavior is to measure the credit of the actor. If this behavior does not reflect the credit of the actor in a certain aspect, it is meaningless to define and track this behavior.
2. Can be clearly verified The occurrence of behavior requires clear trigger criteria. If Funder stipulates that Hunter must report the progress weekly after accepting the task, then the similar definition of "reporting lost" is: Since Hunter accepts the task, every 67200 blocks (assuming the block generation time is 6s) is a reporting interval, and any reporting interval must have Hunter's progress report on Bounty. If there is no Hunter's progress report in a certain reporting interval, the platform believes that the behavior has occurred.
3. The impact is quantifiable Quantifiable impact means that the credit impact corresponding to the behavior can be quantified, such as "reporting loss" must correspond to a clear credit score.

5.3 Credit Score

5.3.1 Bounty creation and review

Because OpenSquare is an open Bounty collaboration platform, anyone can create Bounty as long as they hold the platform's native assets as a fee, so the quality of Bounty may vary. And some Bounty may not comply with laws and regulations and seriously violate social morality. The quality of Bounty cannot be verified by mathematical means, so it is necessary to introduce the Council for review. The main audit standards are: -Comply with laws, regulations and general social moral standards. -Clearly describe the content of Bounty, and accurately define Bounty acceptance criteria. -The remuneration given by Bounty cannot be severely marketed.

For Bounty that fails the review, the platform will not be shown to platform users. If the creation fails due to negligence and omission of information, etc., Funder can modify and recreate a new Bounty. Once Bounty is created, the corresponding digital assets will be locked; after being resolved, the corresponding part of the assets will be transferred to Hunter, and the Council will charge a certain service fee.

5.3.2 Bounty application and authorization

The choice of partners is two-way. Both Funder and Hunter are worried about encountering untrustworthy partners. The historical user behaviors stored in OpenSquare historical

transactions and the resulting user behavior scores precisely solve this problem. When Hunter browses the Bounty list that can be applied for, he can see Bounty’s Funder and Funder’s historical behavior at the same time.Hunter can measure the credit of the Fund according to the behavioral credit indicators given by the platform to decide whether to apply for this Bounty.Similarly, when a Bounty has multiple Hunter applications, Bounty Funder can view the historical behavior and credit and ability indicators of each Hunter to select the most suitable candidate for Bounty.

5.3.3 Submission, review and arbitration

The final solution of Bounty usually requires multiple rounds of submission and review by Hunter and Fund. Although the Council has an initial review of Bounty's delivery standards, it does not mean that the delivery standards are mathematically verifiable.The arbitration of the dispute between Funder and Hunter will be the biggest challenge faced by the Council, which may involve a lot of tedious and complicated off-chain behaviors.When necessary, the Council will introduce third-party authorities and even make decisions through community voting. But what can be guaranteed is that all Council's actions related to arbitration judgments will be recorded on the chain and will be publicly monitored by the community.

6 Economic Model

6.1 DS coin

The token name with voting rights in the Datashop network is "DS". All holders of DS have the right to participate in the governance process (see Governance-Federation) and to verify data. All currency transactions in the network, such as paying for unique names, fees or participating in the smart contract market, require the use of DS as a payment token. However, the goal is to simplify the process of making payments in other currencies as much as possible.Therefore, systems such as Chainlink will be implemented at a later stage to make it even possible to use statutes to purchase digital goods without trusting a central authority.All transactions on Datashop require a basic fee. The network will charge related hand fees based on the size of the transaction. In addition, because Unique name is a scarce resource that belongs to the network, users need to pay when registering. The longer the name, the lower the fee.The cost of the unique name can be calculated according to the maximum allowable characters of the unique name UNmax, the length of the unique name UNlen and the fixed fee according to the following methods:

$$F = (UNmax + 1 - UNlen)^2 * Fcon$$

Since the number of available names is almost unlimited, this algorithmic pricing is more practical than auction-based systems. The fixed fixed fee will be relatively high at first, but will decrease over time, depending on the on-chain voting system.

6.2 Coins at Stake

Datashop Network adopts a federal governance model, which is based on a proof-of-stake system. Datashop will also implement a representative system or liquid democracy. This means that everyone can vote or delegate voting rights to others. Each user can only occupy one level at a time.Inflation funds are used to incentivize voting, which means that those who participate in the voting process will receive an appropriate proportion of DS. In the case of voting on malicious parties, the tokens participating in the voting will be lost. Before voting starts, holders participating in voting need to stake DS tokens and lock them for a certain period of time to participate in the voting process. The total voting rights VP owned by a user in a specific time period T can be calculated by the following recommended Vitalik.

$$T^2 * S = VP$$

S stands for the number of locked up tokens and T is the lock up time. This leads to a high incentive to lock up your tokens as long as possible and therefore means more voting power requires living with your decisions for longer. Furthermore, the user can decide how to allocate his voting power, which means he could theoretically use the complete voting power on a single vote.

6.3 Token Distribution

- Total Supply：100,000,000
- pre-sele 10%：10,000,000;
- Liquidity mining 70%：70,000,000;
- Community governance 10%：10,000,000;

- DEV-team 10%:10,000,000;

7 Conclusion

This article introduces the first version of the shared storage data market based on the blockchain system, which attempts to achieve a high degree of distribution, free access and collectability of digital content. The current design is still very imperfect, welcome everyone to participate in improving this article.