Bootstrapping a Blockchain Based Ecosystem for Big Data Exchange

Jinchuan Chen
School of Information
Renmin University of China
Beijing, China
jcchen@ruc.edu.cn

Yunzhi Xue
Institute of Software
Chinese Academy of Sciences
Beijing, China
yunzhi@iscas.ac.cn

***Abstract*—In recent years, data is becoming the most valuable asset. There are more and more data exchange markets on Internet. These markets help data owners publish their datasets and data consumers find appropriate services. However, different from traditional goods like clothes and food, data is a special commodity. For current data exchange markets, it is very hard to protect copyright and privacy. Moreover, maintaining data services requires special IT techniques, which is a difficult job for many organizations who own big datasets, such as hospitals, government departments, planetariums and banks. In this paper, we propose a decentralized solution for big data exchange. This solution aims at cultivating an ecosystem, inside which all participators can cooperate to exchange data in a peer-to-peer way. The core part of this solution is to utilize blockchain technology to record transaction logs and other important documents. Unlike existing data exchange markets, our solution does not need any third-parties. It also provides an convenient way for data owners to audit the use of data, in order to protect data copyright and privacy. We will explain the ecosystem, and discuss the technical challenges and corresponding solutions.**

***Keywords*-data exchange, blockchain, ecosystem**

I. Introduction

In the era of Big Data, the data becomes the most valuable asset in many companies. Many researchers and engineers from both academy or industry need real datasets in order to evaluate their systems, train learning models, or analyze market trends. Therefore, data exchange is becoming a huge industry. In China, there are more than a dozen of data exchange centers or markets runned by the governments or companies, such as [1], [2]. Besides those markets, there are also many websites which provide free datasets for people to download, like [3], [4], [5].

Fig. 1 illustrates the simplified process of most data exchange markets. A data owner submits a specification of a dataset into the market. A customer also submits a specification of his/her demands. Then the market may help the customer find the appropriate datasets. Next the customer makes a contract with a data owner and settles the payment. Then he/she can get the dataset. Notice that the transaction may be made inside or outside the market.

As the center role in the above image, the market needs to specify the formats of the specifications of datasets and demands, provide the standard contracts, and make arbitrations when there are any disputes. Sometimes, the market also provides the payment channel, like what ALIPAY[1] does.

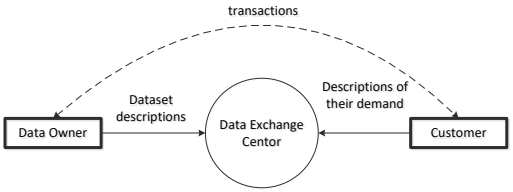

Figure 1. A typical process of data exchange market.

These markets essentially follow the idea of traditional markets, such as clothes, food, and automobiles etc. However, data is a special commodity. It can be illegally reproduced, and it may be sensitive to user privacy. Hence there are several problems in the existing data exchange markets.

First of all, in current dataset transactions, most data owners pass the whole dataset to the customers online or offline, which makes it very difficult to protect the copyright and user privacy. Suppose a data owner sells several copies of a dataset to different customers, and later he/she finds that the content of the dataset has been put on Internet. How can he/she know which customer violated the license specified in the transactions?

Secondly, the data owners need to provide data service. They need to store data and pass data to customers. They also need to clean data. These efforts will be a huge burden for the organizations in traditional industry, like banks, airline companies, government departments, and hospitals etc. These organizations have rich data resource and also would like to share or sell their data. However, they usually do not have enough IT engineers and it is very hard for them to provide these services.

Finally, the data exchange market acts as the authoritative third-party in the data transactions. The data owners and customers have to trust the market. Also, they need to pay some management fees to the market. The market may cause a single-point-failure. It may be hacked and/or may crash. The whole exchange system will halt in this case.

[1]https://www.alipay.com/

978-1-5386-1996-4/17 $31.00 © 2017 IEEE
DOI 10.1109/BigDataCongress.2017.67

In this paper, we propose a blockchain-based ecosystem for data exchange. Blockchain is a distributed ledger maintained by all users in a peer-to-peer way. Thus we do not need any third-parties. There will be different roles in the ecosystem, e.g. data owner, data publisher, and customers etc. Hence the data owners do not need to provide data service. They just pass their data to the data publishers, which are the professional data service providers. Furthermore, the logs of each transaction will be recorded in the blockchain, which is open to every user and cannot be modified. Therefore the data owners and data publishers can audit the use of their data. We will also discuss the techniques to further protect data security, i.e. copyright and privacy.

Our complete solution is not to construct any markets. It contains three parts. Firstly, we propose an ecosystem and explain how different users can cooperate together to make data exchange, and how they benefit from the ecosystem. Secondly, the solution will include a set of initial specificatoins for bootstrapping the ecosystem, including protocols of running the blockchain, requirements for services, and format of logs etc. Also, the solution will provide some implementations of the protocols and the business processes.

The system most similar to ours is Jingdong Wanxiang [2], which also adopts blockchain to records transaction logs. But Wanxiang is the mandatory third-party in their solution. It defines all the specifications, such as the format of services, API's, blockchain protocols etc. Hence it also suffers from the problems of introducing the third-parties, like single-point failure, cost, trustness. On the contrary, our solution is completely peer-to-peer and decentralization, which has more vitality than Wanxiang.

II. Preliminaries

In this section, we will briefly explain Blockchain, and discuss the related issues.

The term "blockchain" comes from the Bitcoin network, which is first proposed by Satoshi Nakamoto [6]. In Bitcoin, the transactions are recorded in a distributed ledger in the form of a chain of blocks, i.e. blockchain. There are also some other definitions for blockchain. According to Wikipedia, blockchain "is a distributed database that maintains a continuously growing list of ordered records called blocks"[2]. In [7], blockchain is defined as *state replication* in a distributed system, which is achieved if all nodes execute the same set of commands in the same order.

Recent years, blockchain attracts more and more interests from both industry and academy. In Fintech, blockchain acts as a peer-to-peer distributed ledger to record transactions between different banks, in order to save the cost of clearing banks. Blockchain is also adopted in intellectual property protection, shareholding transaction, and information sharing etc.

[2]https://en.wikipedia.org/wiki/Blockchain

There are three major merits of blockchain.

- Decentralization. There is no intermediary or authoritative third-party in the network. Decentralization can save the money paid for the third-party agencies, and also reduce the corresponding time cost.
- Byzantine Tolerance. A Byzantine node means that this node can present arbitrary behaviors, such as crash, cheating, and even collusion etc. In practice, we cannot assume all the participators are honest and behave correctly. Hence Byzantine tolerance is necessary for large-scale distributed system.
- Reliability. The data stored in a blockchain cannot be modified, and is open for all the nodes in the network. It is also a necessary property for recording transactions.

Now, we are ready to explain why we choose blockchain as the foundation to construct the data exchange system. First of all, we can remove the third-party agency from the transactions. Currently, the data exchange markets need lots of computer servers to store huge datasets, develop a software platform, and hire many employees to operate the market. All these costs are finally paid by the users. Now we can save the resources for constructing the data exchange markets. Secondly, we can avoid the risk of leaking out data without the third-party agency, which may cause a single-point failure. Moreover, all transactions are logged into the blockchain, which is available for all users. Hence it is easier and more reliable to audit the use of datasets. Finally, we can use smart contracts, which can automatically guarantee the rights and benefits of both participators of each transaction.

III. Overview of the Ecosystem

In this section, we illustrate an overview of the ecosystem for data exchange and explain each role.

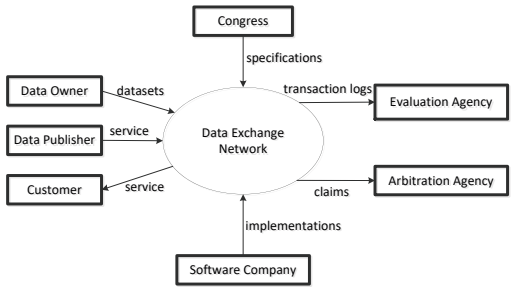

Figure 2. An Ecosystem for Data Exchange

Fig. 2 illustrates the ecosystem for data exchange. In the core part there is the *Data Exchange Network*, which is essentially a blockchain based distributed ledger. All participators can browse and submit documents from/to this network. These documents include services announced by *Data Publisher*, bidding documents submitted by *Data Owner*, and smart contracts signed by several participators etc. Moreover, the network also record the logs of the

corresponding actions, e.g. the transactions for exchanging data. Notice that this network, like other blockchain based networks such as Bitcoin, is operated by all participating nodes in an autonomous way. This is the **most important difference** between our solution and other data exchange systems. We do not aim at constructing a new market for data exchange, but bring out a solution which enable organazations to build up a peer-to-peer market. Next we will explain the different roles inside this ecosystem separately.

Congress. The responsibility of the congress is to formulate a set of specifications which must be followed by all the participators. Of all the specificaitions, the most important one is the protocol to achieve consensus. All the nodes in the Data Exchange Network must adopt the same protocol to decide what should be the common state of the network. Furthermore, the congress need to specify the format of the important documents, e.g. bidding documents, books of tendor, data services, transaction logs etc. Also, it should give the standards of the API's to access the datasets. There are still other specifications required by the ecosystem, such as the measurement to evaluate the data quality, the regulations for the transactions, and the constitution for organizing and running the congress etc. There is only one congress inside the ecosystem, which is formed by selected participators according to a predefined constitution.

Data Owner. A *Data Owner* owns several datasets and bring them into the ecosystem. He/She focuses on generating datasets and does not provide data services. Instead, data owners will transfer the datasets and corresponding copyrights to some **Data Publishers**, who are responsible to publish the datasets and serve the **Customers**. Data owners can be the companies and organizations who generate data continuously, like hospitals and banks. They do not need to worry about how to publish datasets and provide data accessing services. The separation of data owners and data publishers can encourage the data producers to trade their data, and promote data value in data circulation.

Data Publisher. A *Data Publisher* stores and publishes datasets and serves the customers. The role of Data Publisher is somehow like a cinema. He/She does not produce datasets, but purchases datasets from data owners and earns money by selling datasets to customers. A data publisher must carefully audit the process that customers accessing datasets, in order to ensure that there are no illegal actions, such as leaking out data or hacking privacy. Typically, a data publisher is a data center because it has necessary storage capacity, network bandwidth, computing power and sustainable service capacity.

Customers. The customers buy licenses from data publishers and access datasets through API's. The Data Exchange Network will record each acess action and store the records on the underlying blockchain. In this way, data owners and data publishers are able to audit the use of their datasets.

Evaluation Agency. An *Evaluation Agency* is responsible to analyze the actions of each participator based on the information stored on the blockchain. He/She may sell evaluation reports which can help other participators to choose partners.

Arbitration Agency. The responsibility of the arbitration agency is to settle the disputes among some participators. The arbitration agency is composed by elected participators. Since lots of pacts are in the form of smart contracts, it is possible to settle many disputes automatically.

We can see that our solution is not just a market or platform for data exchange. Our objectives are two folds. We propose a brand new ecosystem for data exchange, which requires no authoritative or official third-parties and provides great conveniences for data owners, data publishers and customers. Also, we will design a set of initial specifications for bootstrapping the network and construct the congress.

IV. Business Process

In this section, we illustrate the main business processes in the blockchain based data exchange.

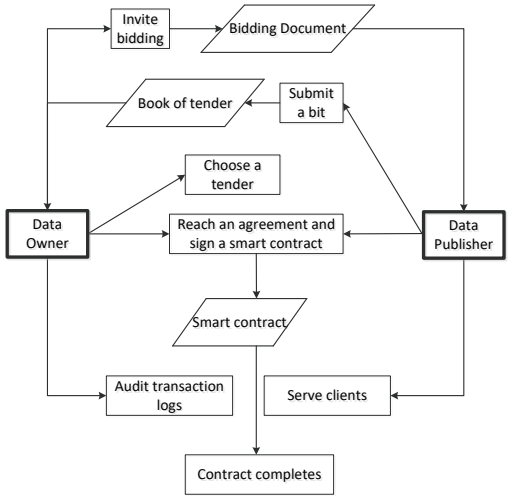

Figure 3. Process of Publishing the Datasets

Fig. 3 illustrates the process of publishing datasets. In the beginning, a data owner would like to publish a dataset. For this purpose, firstly he/she needs to issue a bidding document, which contains the description of the dataset, the requirements for the data publisher (e.g. storage ability), and the lowest price etc. The bidding document will be recorded on the blockchain. Some data publishers may be interested in it and they can submit their books of tender, which describes the services they can provide, the prices they want to offer and their abilities for serving customers. The data owner then chooses one or multiple tenders. Next the data owner and the chosen data publisher reach agreements, and sign smart contracts. The chosen data publishers can then present this dataset to the customers by creating a new data service (which will be illustrated in Fig. 4) . The data owner can

audit the transaction logs to ensure there are no violations of his/her benefits. Finally, the aforementioned smart contract automatically decides when to terminate itself.

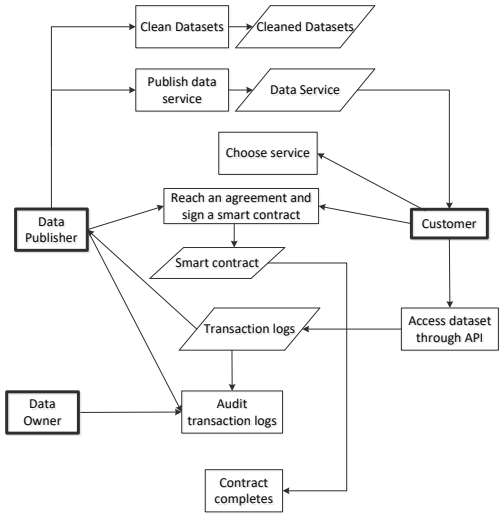

Figure 4. Process of Providing Data Service

Fig. 4 illustrates the process of providing data service. Notice that the original datasets obtained from the data owners may need to be cleaned. For example, some datasets contain privacy information and/or err data. Hence the first step for the data publishers is to conduct data cleaning. After that, a data publisher can then publish a new data service. This service item may contain the descriptions of the dataset, the API's for accessing this dataset, and price etc. Suppose a customer chooses this service item and signs a smart contract with this data publisher. Next this customer can access the dataset through the provided API's and the accessing logs will be recorded. Finally, the smart contract automatically decides the termination time point.

V. Technical Challenges

In this section, we will highlight some major technical challenges and try to present reasonable solutions.

Protecting Copyright. One major issue for data exchange is how to protect copyright. In traditional data exchange market, data owners usually send the whole dataset to the clients who have purchased it. In this case, it is very hard to find out which customer violated the license when the dataset leaks out.

One solution is to encapsulate the dataset with API's. We can carefully design the API's so that the clients cannot obtain the whole dataset while the data service can still meet their needs. Another solution is to utilize the honey pots. For each customer, we may generate some artificial tuples. Each customer is bound with some identical artificial tuples. We can then track the flow of each copy.

The Consensus Protocol. The core part of blockchain is the consensus protocol. Currently there are two kinds of approaches to achieve consensus. One is to compete for the token of appending new blocks, such as the mining process in Bitcoin. The other one relies on the Byzantine Fault Tolerant protocols, which is a voting-based solution. In our case, each participator needs an identity, and we do not want to issue new currency like Bitcoins. Thus it is reasonable to choose the BFT approach.

Currency. Since the transactions are running online in the data exchange network, we need a currency or payment channel. One choice is to adopt Bitcoin or any other cryptocurrency. It is easy to construct a payment channel based on cryptocurrency. However, the price of cryptocurrency may fluctuate acutely. Another choice is to introduce some third-party payment channels, like ALIPAY. Then the currency used in transactions are legal money like US dollars. However, this choice will increase the cost of transactions, and make the ecosystem more complicated.

VI. Conclusion

In this paper, we propose a blockchain based solution for data exchange. Unlike current data exchange markets, this solution requires no authoritative third-parties. Data producers and customers can cooperate together to build up a network, or market. The core part of the network is a set of protocols which are followed by all the participators. Also, the network will automatically record transaction logs which help data owners to audit the use of their data. The solution can help to promote data circulation, and to promote data-intensive applications.

Acknowledgements. This work is funded by the National Key Research & Develop Plan (No.2016YFB1000702).

References

[1] "Guiyang big data exchange," http://www.gbdex.com/website/.

[2] "Jingdong wanxiang," http://wx.jcloud.com/.

[3] "National data from the national bureau of statistics of china," http://data.stats.gov.cn/.

[4] DATA.GOV, "Usa goverment open data," https://catalog.data.gov/dataset.

[5] Amazon, "Amazon aws public datasets," https://aws.amazon.com/cn/public-datasets/.

[6] S. Nakamoto, "Bitcoin: A peer-to-peer electronic cash system," https://bitcoin.org/bitcoin.pdf, 2008.

[7] R. Wattenhofer, "The science of the blockchain," *Inverted Forest Publishing*, 2016.

[8] M. Castro and B. Liskov, "Practical byzantine fault tolerance," in *Proceedings of the Third Symposium on Operating Systems Design and Implementation*, ser. OSDI '99, 1999, pp. 173–186.