Kryptosafe: managing and trading data sets using blockchain and IPFS

Marco Di Francesco
NetService spa
Bologna, Italy
marco.difrancesco@netservice.eu

Lodovica Marchesi
Dept. of Mathematics
and Computer Science
University of Cagliari, Italy
lodovica.marchesi@unica.it

Raffaele Porcu
Dept. of Mathematics
and Computer Science
University of Cagliari, Italy
porcu.raffaele@gmail.com

***Abstract*—Trading data sets is not easy. The owner of valuable data, once they are sold the first time, cannot be sure that they will not be copied and resold. On the other hand, the buyer, cannot be sure that the seller will not sell the same data to a competitor. The advent of blockchain technology, or DLT, can mitigate, or event solve these issues, because it can certify the data ownership, and act as a broker between seller and buyer. In this paper we present Kryptosafe, a system developed following sound software engineering practices, aimed to manage the trade of data sets taking advantage of the unique features of immutability and trustfulness of Ethereum blockchain, and of IPFS distributed DBMS. Kryptosafe allows data sellers to sell a whole encrypted data set or to show potential buyers a subset of it, allowing full access only after the sale is finalized. Using ERC721 and ERC1155 tokens, it also manages one-time sales, when the data set ownership is simply transferred to the buyer, or multiple sales of the same data set to different buyers.**

***Index Terms*—blockchain, IPFS, NFT, dataset trading**

I. Introduction

We live in an information age, and not only data has reached the status of tradable good, but it is commonly believed that data will be the raw material of the 21st century.

Trading data, however, is not easy. The owner of valuable data, once they are sold the first time, cannot be sure that they will not be copied and resold. The buyer, on the other hand, cannot be sure that the seller will not sell the same data to a competitor. Data marketplace and providers are numerous, but they cannot be fully trusted.

The advent of blockchain technology, or DLT, can mitigate or possibly solve these problems, because it can certify data ownership and act as an intermediary between seller and buyer [1]. In this paper we present Kryptosafe, a system developed following solid software engineering practices, aimed at managing the trade of data sets by exploiting the unique characteristics of immutability and reliability of a DLT.

Kryptosafe is based on Blockchain and IPFS technologies, it will allow to provide a secure solution to persons or organizations which have to manage personal or sensitive data of third parties in outsourcing, without causing them to lose efficiency in the management of their core businesses or lose previous investments. The platform is also compliant with the GDPR (General Data Protection Regulation), which came into force in European Union on May 25, 2018.

This research was funded by project KryptoSafe of Puglia Region OH1RWB4, CUP:B86I22000060007

Data sets are sold on the blockchain as NFTs (Non Fungible Tokens), pieces of digital content connected to the blockchain. There is a big difference between NFTs and cryptocurrencies or traditional tokens: While the latter are considered fungible assets, which means that they can be replaced or exchanged for other assets of the same value, the former are unique and non-interchangeable, meaning that no two NFTs are the same.

The remainder of this paper is organized as follows. Section II deals with the background and related work. Section III shows the proposed methodology, highlighting the enabling technologies (III-A); the actors involved (III-B); the functional requirements in the form of User Stories (III-C); the architectural design that includes the client app, the server, the smart contracts and the database (III-D). Finally, Section IV draws some conclusions.

II. Related work

We recall that the blockchain is a shared and immutable data structure. It is defined as a digital register whose entries are grouped into ”blocks”, concatenated in chronological order, and whose integrity is guaranteed by the use of cryptography. Its content, once written through a regulated process, can no longer be modified or eliminated, unless the entire process is invalidated. The characteristics that systems developed using blockchain or distributed ledger technologies (DLT) have in common are data digitization, decentralization, disintermediation, transfer traceability, transparency/verifiability, immutability of the records, and programmability [1]. Thanks to these characteristics, the blockchain is therefore considered an alternative in terms of security, reliability, transparency and costs to databases and registers centrally managed by recognized and regulated authorities (public administrations, banks, insurance companies, payment intermediaries, etc.). However, still few studies focus on the use of blockchain as a means of data trading.

One of the first works was published by Li et al. [2] who developed a blockchain-based model for sharing big data. Their model is based on a specific implementation of a Proof-of-Work blockchain and includes an analysis of its security against attacks.

He et al. [3] proposed an accountable data trading platform based on blockchain. This platform uses blockchain technology to build a distributed secure and trusted environment for big data trading. A peculiar feature of their approach is to perform a secure data set similarity comparison method to detect illegal resale before listing data sets on the data market.

Chen et al. [4] proposed a big data personnel management system to protect sensitive data while providing information management operations such as querying, adding, and updating. They used Hyperledger Fabric DLT due to its performance and availability. However, this work did not provide big data trading among different parties, and the security of the system was not analyzed in the article.

Dai et al. [5] proposed a blockchain and SGX-based data trading ecosystem based on Ethereum blockchain. In the ecosystem, both data brokers and buyers cannot obtain access to the raw data of the seller, as they only obtain access to the analysis findings they require.

Zheng et al. [6] proposed a blockchain-based decentralized data trading platform, on which data providers can better control data trading. This platform includes smart contracts for distributed data trading and for assigning data rewards in trading. They used Ethereum as a proof-of-concept blockchain for performance analysis.

Finally, Hu et al. [7] proposed a blockchain-based trading system for big data, focusing on security, usability, and efficiency. Their system also provides for an evaluation phase of the data received by the user, and distributes the revenue of the data according to the evaluated quality. The structure of their system is a combination of a double chain and a side chain to record the data summary and evaluation information for efficient retrieval and validation.

III. Methodology

A. Enabling Technologies

We used the following enabling technologies for Kryptosafe: Blockchain, NFTs and IPFS. The blockchain is not used as a database for saving large amounts of data due to the huge cost involved, so a secondary platform for saving such data is necessary. Furthermore, the data entered on the blockchain must be reduced to small values, such as strings, numbers, or booleans. In fact, if you want to insert a document, an image or any other large data into the system, only the hash of its content is inserted into the blockchain, saving the original document on a different storage.

To this purpose, we used IPFS (InterPlanetary File System), a technology running on a P2P network that allows its members to store and distribute information in a completely decentralized way throughout the nodes or ”planets”. The system works on the basis of a known distributed hash table technology, or DHT, the same one used in the BitTorrent protocol, from which IPFS takes some functions for its P2P network.

Although it is a development version, IPFS currently allows for the implementation of many of its final functions in a stable manner, so it is a system that we can already use today. IPFS uses content addressing, which means that it identifies content on the network by what it is rather than where it is located.

IPFS leverages a data structure called Directed Acyclic Graph (DAG). Specifically, we talk about Merkle DAGs, where each node has a unique identifier that is the hash of the contents inside itself. To represent content in a Merkle DAG, IPFS splits it into multiple blocks, each of which could reside on a different node. This means that different parts of a file can come from different sources, just like with BitTorrent, where if you download a file you can see that different fetch requests are being made to different peers. This feature also results in faster downloads, since there is no download from only one server at a time, but from several servers in parallel.

In IPFS, everything has a CID (content identifier) that uniquely identifies it. A very useful property of Merkle DAGs, and a consequence of chunking, is that if you have two similar files, they can share part of the same DAG, i.e. different Merkle DAGs can refer to the same subset of data.

Regarding NFT technology, the best known NFTs are Ethereum NFTs, which follow standards known as ERC-721 and ERC-1155 [8]. The system guarantees that an NFT does not change (the certificate is unique and cannot become something else over time), and on the other hand it certifies the ”transfer of ownership” of the hashes managed by the NFT (registered on its unalterable blockchain). There are two main ways that smart contracts and NFTs can interact with each other:

- NFTs can be incorporated into smart contracts. A smart contract can have an NFT within itself, which is then passed on to a user or another contract based on the rules and events defined in the smart contract itself.
- Smart contracts can be incorporated into an NFT to call and access resources within the NFT. For example, users can access a song embedded in an NFT through a smart contract. They would agree terms using the smart contract, pay the agreed amount, and then get access to that song. This is a process that will most likely run in the background when users hit the ”play” button on their applications.

Combining NFTs with smart contracts will give users the flexibility to unlock a wide range of use cases. Contract structures and complex agreements can be created. The underlying blockchain mechanisms will make contracts transparent, tamper-proof, and verifiable in real time.

B. Actors

The actors managed by the system are:

- **Owner**: who owns the data and decides to sell it. It can always get the content in clear text. This is possible because s/he is the owner of the relative NFC token of the data in question and because s/he has the password to be able to generate the decryption key.
- **Buyer**: who is connected to the platform, can view the directly accessible contents of the data sets according to the rules of the data type and of the system, can make queries and can purchase the data.

- **Data Protection Officer**: who has the responsibility to protect the data of the data owners. So his responsibility is to have the data sets protected from unwanted alterations and to guarantee access.
- **Data Manager**: who manages the manipulation of the data according to the connected actor. So s/he is responsible for the type of content displayed depending on the type of user connected, allows a search on data sets and the execution of queries. The manager of the data can in some cases also be the owner himself.

The first two actors are physical actors, possibly acting on behalf of legal persons, who connect to the platform for the exchange of goods. The Protector and Manager of the data are not necessarily human actors, but in most cases they are managed by a component of the system capable of guaranteeing this behaviour. The ownership of the asset is determined by the ownership of the assigned NFT token. From the SC it is possible to verify the owner's identity because the NFT token is uniquely linked to the data set and the NFT Smart Contract tracks the owner of the given NFT token.

Data management is determined by the server and the app which, depending on the connected user, the type of data and the information to be given, manage the information required and the type and methods of display.

C. Use Stories

The Kryptosafe system features are described according to Agile practices, following an Epic approach, meaning high-level feature which can be divided into User Stories and Sub Tasks. An User Story is a smaller pieces of functionality which can be implemented separately, in an incremental way.

The Epics of Kryptosafe are:

- Purchase data set
- Visualization of personal data sets
- Sale of data sets
- Data set tokenizatio
- Encryption of the data set
- Data decryption

For the sake of brevity we report only one epic decomposition.

1) EPIC - Sale of data sets: As an owner, I want to upload my data sets to be able to sell them.

Breakdowns:

- US-C1 - Uploading an encrypted data set
- US-C2 - Creation of a related NFT
- US-C3 - Sale of the data set

D. Architectural design

After identifying the actors and features, it is necessary to understand how the "unencrypted" data present in the Seller's device must be manipulated in order to be traded online. There are three possibilities: (i) encrypted data are visible only by the owner of the associated key; (ii) data are tokenized and visible even without being the owner; (iii) data are tokenized and are not fully visible, except for their description using, for example, a schema in json format.

The types of data can be different: files, images, database records, documents, and each of these will have a different level of visibility depending on the permissions that the data manager gives to that particular content.

The general architecture of the Kryptosafe project is divided into four modules.

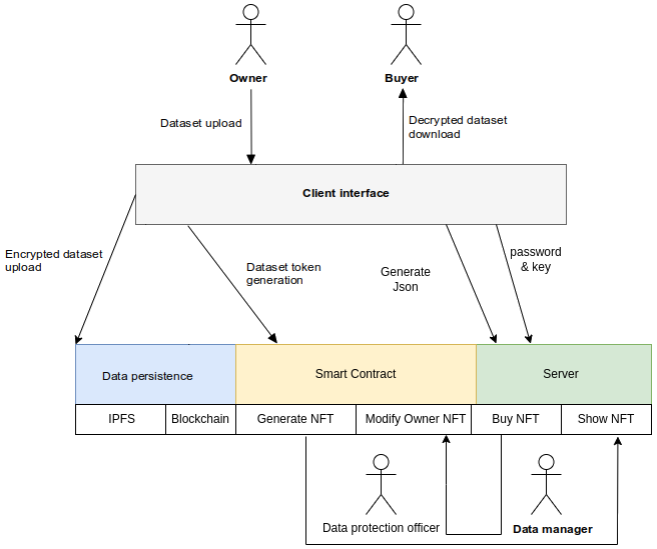

Fig. 1. The overall architecture of Kryptosafe system.

Figure 1 shows how the modules interact with each other. In the following, we describe the four modules in deeper detail.

1) Client Interface: The Client Interface covers user-side functionalities and allows cross-platform operation. This application has offline encryption and decryption functions for security reasons, as we want to prevent unencrypted content from being directly accessible on the server side. Once encrypted, the data can be put on the net.

The Client Interface module provides users with a desktop application and enables the following features:

- Password generation and encryption keys
- Loading data sets
- Data set encryption
- Uploading of encrypted data sets directly into the persistence layer (in this case, we assume the use of IPFS)
- Downloading of encrypted data set from IPFS
- Json generation of NFT token description
- Uploading of descriptive Json to IPFS
- Decryption of encrypted data set

The encryption and decryption are managed with a symmetric key called the master key (MK), through a private mechanism managed offline.

2) Server: The Server has the goal of displaying the data sets according to the connected user. The Owner of the data set can view his data unencrypted thanks to the key generated by his password. The encryption type is symmetric; this encryption key can encrypt and decrypt the data sets with the help of the Client module. If the user is not the owner of the

data set, s/he can view the description file of the data set and execute queries on it but cannot access all the data, in order to obtain them s/he must purchase the requested data set.

To be able to query the data, the data structure must be tokenized and the various database items must preserve the logical format of the original data, through a key-value mapping, which can be maintained on the server side. In this way, it is possible to make queries and know the result in terms of frequency of the data, but without knowing its real content. This is facilitated by the fact that the data are structured and encoded by the data source provider.

The data sets have a descriptive file in json format to allow the display of the description of the data and their typology. The software has a mechanism for secure publication of tokenized private data that can still be recovered through Information Retrieval (IR) systems. In fact, in order to be able to query the data without having to modify the back-end application software that manages the database, it is necessary to use a public data annotation mechanism, through a standard JSON structure that declares the format of the tokenized data, open to the manager of the data itself.

3) Smart Contracts: SCs allow the generation and management of the NFT, i.e. the change of the data set ownership after its purchase.

Possession is certified through an NFT token linked to the data set. The first step of the process is to acquire the NFT by purchasing it, and then to obtain the password or key to decrypt its content, through deterministic algorithms such as PBKDF.

NFTs can be created using two different standard: ERC-721, or ERC-1155.

The main difference between these two standard is that ERC-721 token is not duplicatable, while ERC-1155 is. So, for instance, an image tokenized with ERC-721 can only be sold once; an image tokenized with ERC-1155 standard allows different identical copies of the image, and therefore it can be sold a given number of times.

4) Data Persistence: This module is managed by blockchain and IPFS technologies. All types of content, whether they are text documents, images, or sensitive data, are stored in encrypted form within the IPFS, and their hash signatures, together with the information to retrieve their content, are written in the blockchain.

This allows us to have the decentralization of the data as well as its encryption. The document can be downloaded by anyone who knows the link to the IPFS content, but can only be viewed by those in possession of the decryption key or password. Basically, the data set contents can be shown in three ways:

- *Visible*: all contents are accessible.
- *Encrypted*: only the information present in the metadata of the data set is shown.
- *Visible in tokenized form*: the data have undergone a map labeling aimed at obscuring the original contents, preserving their characterization.

IV. Conclusions

In this short paper, we proposed a platform for managing and trading data sets using blockchain, IPFS and NFT technologies.

The main features of Kryptosafe are: (i) the data are encrypted off-line, outside the network, and signed with a private symmetric key; (ii) the unencrypted data are visible to third parties, if enabled by the owner, but the data will not be modifiable by third parties, unless the owner decides to do so; (iii) the data are organized according to a structure agreed by the owner and by the data manager and made public on the Internet; (iv) the data are searchable through their meta-information, associated with a data instance, protected by a key, and kept in a safe block.

The system is currently under development, the use of the aforementioned technologies introduced innovation, trust and safety in the market of dataset trading. Kryptosafe's correctness, efficiency and security is being assessed by providing a set of unit, functional and stress test suites, according to the Agile principles followed in its development. After the development it will be further validated by making it available to selected beta-testers, before its marketing.

References

[1] N. Deepa and et al., "A survey on blockchain for big data: Approaches, opportunities, and future directions," *Future Generation Computer Systems*, vol. 131, pp. 209–226, 2022.

[2] Y. Li, J. Huang, S. Qin, and R. Wang, "Big data model of security sharing based on blockchain," in *International Conference on Big Data Computing and Communications*, 2017.

[3] Y. He, H. Zhu, C. Wang, K. Xiao, Y. Zhou, and Y. Xin, "An accountable data trading platform based on blockchain," *IEEE INFOCOM 2019 - IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, pp. 1–6, 2019.

[4] J. Chen, Z. Lv, and H. H. Song, "Design of personnel big data management system based on blockchain," *Future Gener. Comput. Syst.*, vol. 101, pp. 1122–1129, 2019.

[5] W. Dai, C. Dai, K.-K. R. Choo, C. Cui, D. Zou, and H. Jin, "Sdte: A secure blockchain-based data trading ecosystem," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 725–737, 2019.

[6] S. Zheng, L. Pan, D. Hu, M. Li, and Y. Fan, "A blockchain-based trading platform for big data," in *IEEE INFOCOM 2020-IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*. IEEE, 2020, pp. 991–996.

[7] D. Hu, Y. Li, L. Pan, M. Li, and S. Zheng, "A blockchain-based trading system for big data," *Computer Networks*, vol. 191, p. 107994, 2021.

[8] Q. Wang, R. Li, Q. Wang, and S. Chen, "Non-fungible token (nft): Overview, evaluation, opportunities and challenges," 2021. [Online]. Available: https://arxiv.org/abs/2208.00543