A dataset trading system based on DLT and IPFS

Marco Di Francesco
NetService spa
Bologna, Italy

Lodovica Marchesi, Michele Marchesi, Raffaele Porcu
Dept. of Mathematics and Computer Science
University of Cagliari, Italy

I. Introduction

Data set trading can be critical. After valuable data are sold for the first time, the owner cannot be certain that they won't be copied and sold again. On the other side, the buyer cannot be certain that the seller won't sell the same data to a competitor. On the other side, the buyer cannot be certain that the seller won't sell the same data to a competitor. Nowadays, trading datasets is possible through different marketplaces and providers, but they are not completely trustable.

Due to its ability to guarantee data ownership and serve as a mediator between sellers and buyers, the adoption of blockchain technology, or DLT, can reduce or even solve these issues.

We present Kryptosafe, a system designed to manage the exchange of data sets by taking advantage of a DLT's unique characteristics: immutability and trustworthiness. The development of the system followed a sound software engineering methods.

Based on Blockchain and IPFS technologies, Kryptosafe will enable the provision of a secure solution to individuals or organizations that must manage sensitive or personal data of third parties when outsourcing, without causing them to lose productivity in the management of their core businesses or lose prior investments. Also, the platform complies with GDPR (General Data Protection Regulation). On the blockchain, data sets are traded in exchange for NFTs (Non Fungible Tokens), which are tied to digital assets linked to the blockchain.

The system is presently being developed. According to the Agile development principles used, it is being tested for accuracy, effectiveness, and security using a set of unit, functional, and stress test suites. Before its marketing, it will be further validated by making it available to selected beta-testers.

II. Methodology

A. Enabling Technologies

Kryptosafe uses the following enabling technologies: Blockchain, NFTs and IPFS. Only the content's hash is recorded on the blockchain, the original document is kept on another storage.

We employed IPFS (InterPlanetary File System), a P2P network technology that enables its users to store and distribute data totally decentralized across the nodes or "planets" for this purpose. IPFS makes use of the Merkle Directed Acyclic Graph data structure (DAG). IPFS divides content into numerous blocks, each of which might be located on a different node, to represent it in a Merkle DAG. This implies that different pieces of a file may originate from multiple sources, just like with BitTorrent, where if you download a file you can see that different fetch requests are being made to different peers.

Everything in IPFS is uniquely identified by a CID (content identifier). The ability of two similar files to share a portion of a Merkle DAG—i.e., for distinct Merkle DAGs to refer to the same subset of data—is a particularly valuable feature of Merkle DAGs and a result of chunking.

Regarding NFT technology, the most well-known NFTs are Ethereum NFTs, which adhere to the ERC-721 and ERC-1155 standards. The system ensures that an NFT does not change (the certificate is unique and cannot become something else over time), but it also validates the "transfer of ownership" of the hashes that the NFT manages (registered on its unalterable blockchain).

Combining NFTs and smart contracts together will allow users the flexibility to unlock a wide range of use cases. Complex agreements and contract arrangements can be created. Contracts will become transparent, tamper-proof, and instantly verifiable thanks to the blockchain's underlying mechanics.

B. Actors

The system's actors include:

- **Owner**: who owns the data and decides to sell it. It can always get the content in clear text. This is possible because he/she is the owner of the associated NFT token for the data in question and possessing the password needed to create the decryption key.
- **Buyer**: who is connected to the platform and is able to view the contents of the data sets that are directly accessible in accordance with the data type and system rules, make queries, and purchase the data.
- **Data Protection Officer**: who has the responsibility to protect the data of the data owners. Therefore, it is his obligation to ensure access and to have the data sets safeguarded from unauthorized changes.
- **Data Manager**: who manages the manipulation of the data according to the connected actor. As a result, depending on the type of user connected, s/he controls what content is displayed and permits searches on data sets

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

and the execution of queries. In some circumstances, the owner of the data may also serve as the manager.

The possession of the assigned NFT token establishes ownership of the asset. The server and the app control data management, managing the required information and the kind and methods of display based on the connected user, the type of data, and the information to be supplied.

C. Use Stories

According to Agile principles, the Kryptosafe system's features are documented using an Epic method, which refers to high-level features that can be decomposed into User Stories and Sub Tasks. A user story is a collection of discrete, incrementally implementable smaller functional units.

The Epics of Kryptosafe are:

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

D. Architectural design

Once the actors and features have been identified, it is important to comprehend how the "unencrypted" data on the Seller's device must be manipulated in order to be traded online. Three alternatives exist: (i) Only the owner of the associated key can see encrypted data; (ii) data are tokenized and can be seen even without being the owner; and (iii) tokenized data are only partially visible, except for their description wich is always available, for instance, using a schema in json format.

As seen in Fig. 1, the Kryptosafe project's overall architecture is divided into four modules.

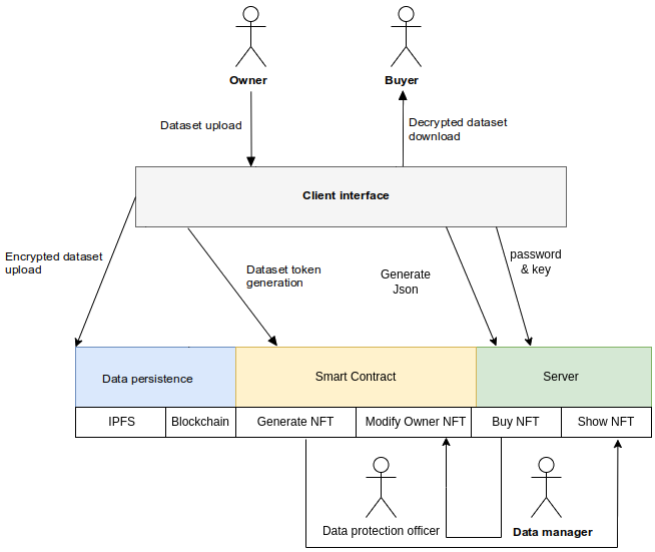

Fig. 1. The overall architecture of Kryptosafe system.

1) Client Interface: The user-side functionalities are covered by the Client Interface, which also enables cross-platform use. For security reasons, this application offers offline encryption and decryption capabilities since we don't want unencrypted content to be immediately accessible from the server. Data can be uploaded on the net after encryption.

Users can access services like password generation, encryption keys, loading and encrypting data sets, downloading and decrypting encrypted data sets from IPFS, and more using a desktop application provided by the Client Interface module.

With a secure offline system, the master key (MK), a symmetric key, is used to control encryption and decryption.

2) Server: The Server's goal is to show the data sets in accordance with the connected user. The Owner of the data set can view his data unencrypted thanks to the key generated by his password. Symmetric encryption is used and with the aid of the Client module, the encryption key can both encrypt and decrypt the data sets. Users who do not own the requested data set can view the description file and run queries on it, but they are unable to access all the data; in order to do so, they must pay for the requested data set.

To enable the display of the description of the data and its typology, the data sets have a descriptive file in json format. The software includes a mechanism for the safe publication of tokenized private data that is still retrievable using information retrieval (IR) tools.

3) Smart Contracts: The development and management of the NFT, or the alteration of the data set's ownership upon purchase, is made possible by SCs. An NFT token connected to the data set serves as the certification of possession. Purchasing the NFT is the initial stage in the process, after which deterministic methods like PBKDF are used to obtain the password or key needed to decode the NFT's information.

There are two main standards that can be used to create NFTs: ERC-721, or ERC-1155. The fundamental distinction between these two standards is that the ERC-1155 token can be duplicated, whereas the ERC-721 token cannot. Hence, for instance, an image tokenized with the ERC-721 standard can only be sold once; an image tokenized with the ERC-1155 standard, however, allows for several identical copies of the picture, allowing for multiple sales.

4) Data Persistence: Both IPFS and the blockchain are used to manage this module. Every content, including text documents, photos, and sensitive data, is kept in encrypted form within the IPFS, and the hash signatures of each piece of content, together with the instructions for retrieving it, are recorded in the blockchain.

This enables both the decentralization and encryption of the data. Everyone with access to the IPFS content link can download the document, but only those with the decryption key or password will be able to view it.