Blockchain-based Personal Health Data Sharing System Using Cloud Storage

Xiaochen Zheng, Raghava Rao Mukkamala, Ravi Vatrapu, and Joaquin Ordieres-Mere

Journal article (Accepted manuscript*)

Please cite this article as:

* This version of the article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the publisher’s final version AKA Version of Record.

Uploaded to CBS Research Portal: August 2019

© 2019 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.
Blockchain-based Personal Health Data Sharing System Using Cloud Storage

Xiaochen Zheng\textsuperscript{1,2}, Raghava Rao Mukkamala\textsuperscript{1,3}, Ravi Vatrapu\textsuperscript{1,3}, Joaquín Ordieres-Meré\textsuperscript{1}
\textsuperscript{1}Centre for Business Data Analytics, Dept. of Digitalization, Copenhagen Business School, Denmark
\textsuperscript{2}E.T.S Industrial Engineering, Universidad Politécnica de Madrid, Spain
\textsuperscript{3}Department of Technology, Kristiania University College, Norway
xiaochenzheng@alumnos.upm.es, {rrm.digi.vatrapu}@cbs.dk, j.ordieres@upm.es

Abstract—With the advent of rapid development of wearable technology and mobile computing, huge amount of personal health-related data is being generated and accumulated on continuous basis at every moment. These personal datasets contain valuable information and they belong to and asset of the individual users, hence should be owned and controlled by themselves. Currently most of such datasets are stored and controlled by different service providers and this centralised data storage brings challenges of data security and hinders the data sharing. These personal health data are valuable resources for healthcare research and commercial projects. In this research work, we propose a conceptual design for sharing personal continuous-dynamic health data using blockchain technology supplemented by cloud storage to share the health-related information in a secure and transparent manner. Besides, we also introduce a data quality inspection module based on machine learning techniques to have control over data quality. The primary goal of the proposed system is to enable users to own, control and share their personal health data securely, in a General Data Protection Regulation (GDPR) compliant way to get benefit from their personal datasets. It also provides an efficient way for researchers and commercial data consumers to collect high quality personal health data for research and commercial purposes.

Index Terms—healthcare; blockchain; cloud storage; data sharing; mobile computing; machine learning

I. INTRODUCTION

With the rapid development of mobile computing, wearable technology and wireless sensing, people have been using different types of mobile and wearable devices, such as smartphone, smart watch, smart band and smart glasses etc., to realise various health-related applications, such as remote diagnosis \cite{1}, disease monitoring \cite{2} and elderly people caring \cite{3}. Large amount of personal health data are produced by these devices and these data are valuable resources for healthcare research and commercial applications. Properly sharing personal health data will benefit all related stakeholders including the device users, patients, researchers, companies and even the whole public healthcare system. As personal asset, the health data should be owned and controlled by the respective users themselves, while in reality they are usually controlled by different service providers, device manufactures or scattered in different healthcare systems \cite{4}, \cite{5}. In general, it brings barriers for the data sharing and puts data security and privacy at risk as these centralised data stores and authority providers are attractive targets for cyber-attacks \cite{6}.

The blockchain technology has gained substantial popularity in recent years, primarily in financial field, due to the cryptocurrencies. For example, Bitcoin was first introduced in 2008 \cite{7} and ever since has attracted the attention of the research community from diverse academic fields \cite{8}, \cite{9}, \cite{10} and gained mainstream popularity due to its unique characteristics, such as the absence of centralised control, an assumed high degree of anonymity and distributed consensus over decentralised networks. Blockchain solutions could reduce data breach risks by utilising threshold encryption of data together using public key infrastructure, where cooperation of multiple parties is required to decrypt data and asymmetric cryptography is used to authenticate communication with system participants \cite{11}. The blockchain based data sharing system could dramatically simplify data acquisition process for research and commercial projects and provide an opportunity for users to gain the ownership and the privileges of their own data and get benefits from them. It could also leads to better control over their data and guarantees fine-grained tracking of all their data usage activities \cite{11}. The aim of this paper is to propose a personal health data sharing system based on blockchain and cloud storage technologies, to enable users easily and securely sharing their personal health data and help researchers and commercial data consumers to obtain necessary required data in an efficient, transparent manner and in compliance with data regulations such as GDPR \cite{12}.

II. RELATED WORK

The research about using personal data generated by mobile and wearable devices to improve the quality of healthcare service has been popular for decades. One of the most challenging tasks during these studies is data acquisition, which is usually costly and time consuming. Most people believe that their medical and other health-related data is private and not willing to share it due to the concerns about data security and privacy \cite{11}. The success of blockchain technology in the financial field demonstrated that, trusted and auditable computing is possible using a decentralized network of peers accompanied by a public ledger \cite{2}. There have been many studies about applying blockchain technology to other fields beside financial recently. In 2015, the study in \cite{13} used blockchain to protect the privacy of personal data. The authors implemented a protocol that turns a blockchain into
an automated access-control manager that does not require trust in a third party which ensures users own and control their data. Since 2016, application of blockchain technology to manage healthcare data has been the primary focus of many research studies. The research presented in [14] introduced an application framework, named Healthcare Data Gateway (HGD), based on blockchain to enable patient to own, control and share their own data securely without violating privacy. It provided a potential way to improve the intelligence of healthcare systems while keeping patient data private. The study in [15] developed a decentralized health record management system, named MedRec to handle Electronic Health Records (EHRs) using blockchain technology. The system provided patients with a comprehensive, immutable log and easy access to their medical information across different providers and various treatment sites.

The above-mentioned studies mainly focused on using blockchain to manage the static health data like EHRs or Electronic Medical Records (EMRs). The EMR contains relatively static data such as almost unchanged during the life of the patient, like gender, blood type, fingerprint etc., or gradually changing, like the age, weight, height, disease history etc. This type of data usually require less storage space which makes it possible to save and share data inside the blockchain. However, in most of the practical healthcare applications, this type of static data only takes a small part of total health data. In contrast, the widely used mobile and wearable devices produce large volumes of dynamic data with high frequency and large data size. For example, the data generated by an accelerometer inside a smart watch are usually with high frequency and millions of records could be produced in one day and the data size may be up to several gigabytes. The high changing frequency and large size makes it difficult to be stored and shared directly inside the blockchain. A recent study presented in [11] proposed a roadmap for a blockchain-enabled decentralised personal health data ecosystem. The authors introduced the concept of a secure and transparent distributed personal data marketplace utilising blockchain and deep learning technologies to help resolve the challenges faced by the regulators and return the control over personal data including medical records back to the individuals. They integrated cloud storage into the ecosystem to provide an off-chain storage solution for large biomedical data files. Different from most previous blockchain applications, a special role named data validator was introduced beside the traditional data contributor/generator and data consumer. The goal of data validators is to validate or to certify the quality of the data contributed/produced by the users. Only the data, validated or certified by data validators is released to the marketplace. This method solves the problem of achieving control over the data quality. Although this method is useful to control the quality of most static and gradually changing health data, it remains challenging to handle the high frequency large size data such as data from accelerometers. To validate such kinds of data manually, the validation may need a lot of efforts and also time consuming. To handle such large amount of data, advanced tools, like data-mining and machine-learning are necessary. The research about analyzing big data produced by wearable devices has been a hot topic for many years [16], [17], [18], [19], [20]. For example, the study of [21] used the acceleration data collected from a smart watch to evaluate the tremor level of patients with Essential Tremor. In [16] the authors recognized the daily activities of elderly people based on the data collected from wearable and mobile devices supported by machine learning techniques. In [22] the researchers adopted deep learning algorithm for Human Activity Recognition (HAR) tasks and obtained satisfying results. The similar machine learning techniques, like deep learning, can be used to evaluate the quality of the data produced by wearable and mobile devices. Inspired by the studies mentioned above, we proposed a new personal health data sharing system supported by blockchain, cloud computing and machine-learning technologies.

### III. Research Scope

Generally, health data can be divided into dynamic and static data as indicated in [11]. The static data refers to the personal data that almost unchanged during the life of the user, like genome and fingerprint and so on. The dynamic data reflects the activity of the user during a period of time, like heart rate Electro-Encephalo-Graphic (EEGs) data; or the state of the organism at the time of sampling, like blood test data. The dynamic data can be further divided into rapidly changing data, like the acceleration data, and gradually changing data, like height and weight etc.

According to the data acquisition method, the health data can be divided into continuous data and instant data. The continuous data is collected in a period of time indicating the status or activity of the user during this period. Continuous data are usually rapidly changing dynamic data in time series format. In contrast, the instant data are obtained in one single measurement. The instant data can be unchangeable static data or gradually changing dynamic data. The classification methods are based on the type of the data and the data acquisition methods, not based on the health indicators represented by the data. The data reflecting the same health indicator may belong to different categories. For example, we can count the number of heart beats in one minute and use this single number to represent the heart-beat status and this single number is instant-dynamic data. But when EEG is used to monitor the heartbeat over a period of time, the data collected during this period is the same heart-beat, but it is continuous-dynamic data.

Our study focuses only on the continuous-dynamic data, as shown in figure 1, as this type of data accounts for the most of the data generated by wearable and mobile devices. They are usually with high frequency and large size and cannot be stored and shared using the same methods as other health data and there are few studies focusing on this topic.
A. General Data Protection Regulation:

In May 2018 the European Union’s new General Data Protection Regulation (GDPR) [12] came into effect. The GDPR is one of the largest changes in data privacy regulation in recent history and will replace the current Data Protection Directive, which was established in 1995. The primary goal of the regulation is to harmonise data privacy laws across Europe and particularly to empower and protect EU citizens’ privacy. One of the most central issues being the question of user’s consent. The regulation states that the service provider must be able to clearly show what the user’s consent is for and that it should be as easy to withdraw consent as to give it for the user. Upon a withdrawn user consent or change in purpose of data collection is the service provider required to delete the data related to the specific user. Furthermore is the user’s right to access, meaning that on the user’s request the service provider or company must provide an overview of whether the user’s personal data is being processed and the purpose of processing. The service provider must also provide all data to the user in a machine-readable format. Similar to the right to access is the right to data portability; the user should be able to get all data regarding them from the controller in a machine readable format and have the right to give this to another controller. Being non-compliant with the regulation can result in large fines for companies of up to 20 million Euros or 4% of global turnover, whichever is larger [12].

IV. Conceptual Design

The architecture of the proposed health data storage and sharing system is described in figure 2. Three roles are defined in this system:

- Users: to produce, upload and share (sell) personal health data and get monetary or service benefits.
- Key keepers: to keep the private keys to decrypting the data after they are uploaded by the user and release the keys to customers when a transaction is approved. They will get monetary benefits for every validated transaction.
- Customers: to buy user data and provide monetary or service rewards to users and key keepers.

For each of the roles, a corresponding App was designed to help them to realize their targets.

- User App: running on a mobile computing device, e.g. a smartphone or tablet. It is able to connect with different wearable devices or other sensors, e.g. smart watch, via wireless communication, e.g. Bluetooth, to collect different health-related data. The data collection is supported by the APIs of the corresponding sensors. The collected raw data will pass through a quality validation module, which will be introduced later, to get a quality score. After the validation, the validated data will be integrated with several identification labels, e.g. Title to briefly describe the data, DataType to indicate the type of the data, Size to indicate the size of the data, Quality to indicate the quality of the data etc. Some static personal data, e.g. gender, age and weight, could also be integrated if necessary. The integrated data will then be compressed and encrypted. The encrypted data will then be uploaded to a cloud. The key to decrypting data will be split into various shares and distributed to key keepers. A transaction will then be generated and broadcasted to the blockchain nodes. The transaction contains the public key of the user, the link to the encrypted data (hash pointer), basic information and price of the dataset etc.
- Key keeper App: running on a local device which is connected to the internet, or on a cloud sever. It is able to receive key shares from the system and keep them securely. When a data sharing transaction is validated it will receive a notice to release the corresponding key share to the customer of that transaction.
- Customer App: running on a local device which is connected to the internet, or on a cloud sever. It is able to show the customer all the available various datasets and allow them to search for a certain type of datasets. Once a dataset is chosen, the App will guide the customer to generate a transaction for buying that dataset. The transaction will then be broadcasted to the blockchain nodes. Once the transaction is validated, the App will receive a link to the encrypted data and the decryption keys from the key keepers. Then the customer could download the data and decrypt them with the private keys.

The core modules and functions of the proposed system are introduced in the following sections.

1) Data quality validation: As mentioned above, our study only focused on the continuous-dynamic data. These data are usually generated by standard sensors. The information of the sensor is accessible through the APIs of the devices. Moreover, the pattern of the collected data can be evaluated using advanced machine learning techniques to make sure that the data is valid according to certain validation patterns or checks. It enables us to validate the quality of the data from both hardware and software aspects.

- Hardware aspect: when a new device is connected to the user App, the hardware information of that device and the sensors embedded in it will be acquired by the user App. If a device or a sensor is from a validated manufacture, it is recognised as a qualified hardware and

Fig. 1. Personal data categories and the study scope of this paper
the data produced by it are reliable. Otherwise, it will be refused to connect with the App. For this purpose, a database of validated manufactures and devices should be predefined and well maintained.

- Software aspect: supported by advanced machine learning techniques, it is possible to classify the patterns of a time series dataset with high accuracy. There have been many studies on this topic. For instance, it is able to recognise a user’s daily activities using the data collected from an accelerometer embedded in wearable devices [21], [22], [23]. Using similar machine learning techniques, we can create quality classifiers for different health data. Only the data with predefined features will be saved and the meaningless data and noises will be eliminated. Here the quality of the data is a relative standard. Take the above-mentioned acceleration data as an example and imagine that a user’s acceleration data are collected by a smart watch during 24 hours. The quality validation algorithms will be able to distinguish sleep from other daily activities. The data corresponding to the sleep period could be classified as high-quality data or noise depending on if the user want to share sleep related data or only other daily activities.

2) Data sharing transaction validation: One of the core components of the system is the blockchain module. It is used to secure the data sharing process. There are several decentralized blockchain application platforms available currently, such as Ethereum$^1$, Hyperledger Fabric$^2$ and others. These platforms allow developers to create blockchain applications conveniently. In this study, we choose the Ethereum as the development framework for our system. Ethereum enables developers to design and issue their own cryptocurrency or a tradable digital token that can be used as a currency, a representation of an asset or a virtual share. These tokens use a standard coin API, so the contract will be automatically compatible with any wallet, other contract or exchange also using this standard. Another important aspect in selling or sharing the personal data to the data consumers or commercial entities is that, the personal data will be anonymised sufficiently so that it will conform to the regulations of GDPR [12]. For example, personal demographic data such as name, address, person identifier etc. will be removed or hashed properly such that the final dataset that will be sent to the consumer will be in compliance with data regulations such as GDPR [12].

3) Cloud storage: The main reason for integrating cloud storage into the data sharing system is to provide an off-chain storage solution for the large size dataset. The continuous-dynamic data are usually collected with high frequency during a long-term process. Take the above-mentioned acceleration data as an example, millions of records could be collected in a single day and the data size may reach several gigabytes. The blockchain is replicated distributed datastore, where the transactional data will be replicated across many nodes such as mining nodes. Therefore, blockchain is not ideal for storing large amount data due to its replication across various nodes. On the other hand, large datasets are stored as off-the-chain such as cloud storage, where the data will be stored in an encrypted format and data pointers such as hash pointers will be used to point to the location of dataset to make sure the integrity and non-repudiation of the datasets. Only the metadata of the original dataset and the bare minimum

\[\text{Fig. 2. The general architecture and workflow of the proposed system}\]
data required for the transactions will be stored and shared in the blockchain. The cloud storage could be existing cloud platforms, such as Amazon Web Services and Google Cloud Platform.

4) Data encryption: To ensure security and privacy, the data will be encrypted before uploading to the cloud by the user App using symmetric-key algorithms like Rijndael AES [24] in combination with a threshold encryption scheme [11], [25], [26], [27], [28]. Then the symmetric key for decrypting the data will be split into multiple shares using the Shamir’s secret sharing technique [29] and the key shares will then be distributed among different key keepers. The minimum number of key keepers for decrypting the data is determined by the total number of key keepers and the blockchain security model [11]. To be able to download the encrypted data, one has to obtain both the link and authentication to the data. Then he/she has to get enough key shares of the encryption key to decrypt the data. Theoretically, they can only get these information through a validate transaction approved by the blockchain nodes.

5) Crypto token: In order to motivate the users to share data, it is necessary to provide them certain benefits when they share their personal data. The benefit could be health-related services, like disease monitoring and diagnosis, but this method may not be applicable to all situations. It is better to provide a kind of monetary benefit in case of certain cases. Exchanging personal data for currency may be problematic for many reasons including the need to perform a massive number of micro-transactions in multiple countries and among a large number of different types of the participants. Following the suggestion of [11], we propose our own crypto token called Personal Health Data coin (PHD coin), which can be generated or mined by putting the data on the blockchain-enabled system to facilitate for transactions. It is expected to support exchange with other crypto currencies or real currencies in future when the network has enough nodes and participants.

6) General workflow: As shown in figure 2, the interactions among the users, key keepers and customers include three pipelines: the encrypted data, blockchain transactions and the decryption key shares. The general workflow is as below:

- **Step 1**: User collects, compress and encrypt data using the User App. Then the data will be uploaded to the cloud storage and the key shares will be distributed to the key keepers.
- **Step 2**: a transaction will be generated for notifying the other participants that the data have been uploaded and ready for sharing. After the confirmation of the user, this transaction will be included into the blockchain via consensus algorithm and the transaction will be visible to customers and data validators.
- **Step 3**: as part of the validation of data, the data validator will verify the data by running the data-mining and machine-learning algorithms to make sure that data is valid according to guidelines and requirements and finally certify that the data is valid according the given specifications.
- **Step 4**: a customer chooses the data he/she wants to buy, make sure that it is certified by the appropriate data validators and then creates a transaction to buy them. After signed by the customer, the transaction will be included into the blockchain.
- **Step 5**: the transaction for buying the data will be validated according to the consensus algorithm. If the customer has enough balance (PHD coins) to buy the data, the transaction will be approved. According to the price of the data, certain amount of PHD coins will be sent to the smart contract and the workflow goes to step 6. Otherwise, the transaction will be rejected and the workflow goes back to step 3.
- **Step 6**: key keepers of the corresponding dataset will receive a notice that a transaction for purchasing those data has been approved. Then key keepers deliver their key shares of the related dataset to the customer via an authenticated communication channel.
- **Step 7**: the PHD coins stored in the smart contract will be distributed among the accounts of the user and the key keepers according to a predefined rule.
- **Step 8**: the customer receives the link to the encrypted data and enough key shares for decrypting them. Then he/she will be able to download the data from the cloud storage and decrypts them. The data are ready to use and the workflow ends.

V. DATA PRIVACY AND SECURITY ANALYSIS

The data security of the proposed system relies on the following three setups.

- For the data stored on the cloud the access is restricted. One has to know the address and get authentication to access to the encrypted data. Moreover, the data is encrypted before uploading to the cloud storage, so that even the compromise of the storage would not lead to the data leakage. The symmetric key for decrypting the data is split and distributed among multiple key keepers. Therefore, compromise of a single key keeper would not lead to the data compromise.

- For the data transaction process, it is secured by the hash function and public-key signature schemes utilised in the blockchain contracts. Blockchain technology allows consumers to use pseudonyms (e.g. public key) to perform transactions in blockchain and therefore no visible personal information is involved in the transaction until unless explicitly included on purpose.

- The symmetric decryption key to data transmission and storage is secured by the blockchain-based Public key infrastructure (PKI), which allow key keepers to establish authenticated communication channels with other participants e.g. encrypting the symmetric key to data storage using the public key of the recipient, so that only the intended recipient can decrypt the data securely with his own private key.

This study mainly focused on the data sharing process. The data security issues after the data has been purchased and
transferred to the customer, e.g. data leakage on purpose or accidentally by the customer, is not the main concern of this paper. Such protection could be achieved with the help of existing security measures for data at rest and in use [11], which is out of the scope of the present paper.

VI. CONCLUSION

In this paper we proposed a personal health data sharing system based on blockchain, cloud storage and machine learning techniques. It enables users to own, control and share their personal health data easily and securely, and get benefits during this process. In this work, first, we classified personal health data into different categories according to data characters (dynamic and static data), and the data acquisition methods (continuous and instant data) in the context of health related data from wearables and mobile devices. We proposed to use different solutions to share the large size continuous-dynamic data using hash pointers to the storage location. Secondly, our proposed system overcame the size limitation of the continuous-dynamic health data by integrating blockchain and cloud storage. We also proposed that larger size of health-related data can be stored in encrypted format on the cloud and only the transactional data and metadata can be saved and shared on the blockchain. Third, a data quality validation module was included to the proposed system to control the data quality from both hardware and software aspects supported by machine learning techniques.

ACKNOWLEDGMENT

The authors were partially supported by the project Big Social Data Analytics for Public Health: Copenhagen Health Innovation: ReVUS funded by Region Hovedstaden and Copenhagen Commune. Any opinions, findings, interpretations, conclusions or recommendations expressed in this paper are those of its authors and do not represent the views of the funding agencies.

REFERENCES