TALLINN UNIVERSITY OF TECHNOLOGY
School of Information Technologies

Ryo Shiraishi 195557IVSB

Improving Data Collection from Healthcare Applications

Bachelor's thesis

Supervisor: Kaido Kikkas
PhD

Tallinn 2022

TALLINNA TEHNIKAÜLIKOOL
Infotehnoloogia teaduskond

Ryo Shiraishi 195557IVSB

Andmete kogumise parandamine tervishoiurakendustest

bakalaureusetöö

Juhendaja: Kaido Kikkas
PhD

Tallinn 2022

Author’s declaration of originality

I hereby certify that I am the sole author of this thesis. All the used materials, references to the literature and the work of others have been referred to. This thesis has not been presented for examination anywhere else.

Author: Ryo Shiraishi

12.05.2022

Abstract

In recent years, we are seeing increase in use of wearable devices, such as fitness tracker and smartwatches, and smartphones to collect healthcare and activity data. These data are collected by various entities, doctors and health institutions make use of the data for monitoring patients' health, and device manufacturers and software developers are also interested in collecting for research purpose for improving its products. However, many of users are not feeling confident with sharing their data even though they are willing to participate to improve the hardware and software for better healthcare application capability. The reason behind this is due to the concerns over privacy and data management done by entities who are collecting the data, which applies to confidentiality and integrity of data.

The objective of this thesis is to create a design model of infrastructure as a reference, that can be used to support for future implementation of data collection from healthcare applications. The healthcare applications aim for a data sharing where the users who are willing to participate for improving healthcare hardware and software can feel confident sharing, from the perspectives of data privacy and trust how data are handled.

This thesis is written in English and is 51 pages long, including 7 chapters, 4 figures and 0 tables.

Annotatsioon

Andmete kogumise parandamine tervishoiurakendustest

Viimastel aastatel on tervishoiu- ja aktiivsusandmete kogumiseks üha enam hakatud kasutama kantavaid seadmeid, näiteks terviseseisundi jälgimise seadmeid, nutikellasid ja ning nutitelefone. Neid andmeid koguvad erinevad üksused, arstid ja tervishoiuasutused kasutavad neid andmeid patsientide tervise jälgimiseks ning ka seadmete tootjad ja tarkvaraarendajad on huvitatud andmete kogumisest teadusuuringute eesmärgil oma toodete täiustamiseks.

Paljud kasutajad ei tunne end siiski kindlalt oma andmete jagamisel, kuigi nad on valmis osalema riist- ja tarkvara täiustamises, et parandada tervishoiurakenduste võimekust. Selle põhjuseks on mure eraelu puutumatuse ja andmeid koguvate üksuste andmehalduse pärast eriti kõiges, mis puudutab andmete konfidentsiaalsust ja terviklikkust.

Käesoleva lõputöö eesmärk on luua infrastruktuuri disainimudel, mida saab kasutada tervishoiurakenduste andmekogumisel. Tervishoiurakenduste eesmärk on andmete jagamine, kus kasutajad, kes soovivad osaleda tervishoiu riist- ja tarkvara täiustamisel, võivad tunda end kindlalt andmete jagamise seisukohast ja usaldada andmete käitlemise viise.

Lõputöö on kirjutatud inglise keeles ning sisaldab teksti 51 leheküljel, 7 peatükki, 4 joonist ja 0 tabelit.

List of abbreviations and terms

API	Application Programming Interface
CA	Certificate Authority
CIA	Confidentiality, Integrity and Availability
COVID-19	Coronavirus Disease 2019
DLT	Distributed Ledger Technology
e-Health	Electronic Health
GPS	Global Positioning System
HIPAA	Health Insurance Portability and Accountability Act
iOS	iPhone Operating System
IoT	Internet of Things
IP	Internet Protocol
JSON	JavaScript Object Notation
KSI	Keyless Signature Infrastructure
MSP	Membership Service Provider
OS	Operating System
OSINT	Open Source Intelligence
PKI	Public Key Infrastructure
TLS	Transport Layer Security
US	United States
UX	User Experience
XML	Extensible Markup Language

Table of contents

List of figures

1 Introduction

In recent years, wearable devices, especially smartwatches and fitness trackers, are emerging and we are seeing a large increase in the number of people who is owning such devices. Wearable devices are seen in many daily usages, for example tracking fitness activities and health monitoring, checking notifications, using it as a watch and many more. However, majority of the owners are using these devices to keep track of their activities and constantly monitoring their health. The advancement of wearable device technology, complemented with many health monitoring features, is allowing doctors to constantly monitor and collect health data from patients and receive medical check-ups remotely. Additionally, health institutions are also collecting data for use of health research and even the device manufacturers and software developers are interested in collecting these data for improving their hardware and software. These data are constantly being collected from participants who are willing to share their data from healthcare applications by entities who are interested in collecting for research purposes.

Despite the increase in number of people using fitness trackers, smartwatches and smartphones to record their healthcare activities there are some concerns. Many users are worried about sharing their data between manufacturers and software developers due to the privacy issue and do not have control over what data are being shared and with which entities. Although, there have been no solutions discussed for a new method of data collection from healthcare applications. The purpose of the collection of healthcare applications' data from wearables and smartphones is to use these data to improve algorithms for software, hardware and for research of a new disease. Which enables device manufacturers and software developers to deliver improved products for a better experience and accurate healthcare and activity monitoring.

To make possible where people can feel confident sharing their data from healthcare applications between the entities that are interested in collecting data for improving hardware and software, the author will propose a new method of data collection. As a solution to the issues mentioned, the author's aim for this thesis is to propose a new

method of data collection from healthcare applications with the focus on confidentiality and integrity of data. The goals of this thesis are to design an infrastructure diagram model with the implementation of Hyperledger Fabric for data collection from user health applications by entities who are interested in collecting data. With the created proposed diagram of the infrastructure, the author will deeply analyze the data sharing flow between the healthcare applications, and device manufacturers and software developers. A flowchart for data flow will be also created by the author for ease of understanding. Moreover, a deep analysis of the components that are involved in the infrastructure, its purpose and functionality for implementing confidentiality and integrity of data will be done. Furthermore, the infrastructure for the data collection is aimed for data collectors to be able to collect reliable data from the participants who are authorized to contribute for research and improvements of healthcare applications purpose. Therefore, the proposed infrastructure diagram for data collection will not just focus from data sharers' perspective, however also from the perspective of data collector. Lastly, the author will analyze the use of this infrastructure in other part of healthcare scenarios, not just for healthcare applications, however, how machine learning can make use of collected data and referring use case with medical equipment. However, in this thesis the author will not discuss data collection from medical equipment, since access to the devices are not possible and data collection from these devices are outside of the scope.

As an overview of the thesis, in Chapter 1, the author will discuss the usage of wearable devices, the aims and goals of this thesis. In Chapter 2, background research for use of healthcare applications and its privacy issues will be discussed. In Chapter 3, the methodology for the thesis will be discussed, giving brief ideas of what the author will conduct to contribute to this thesis. In Chapter 4, a discussion of the solution for the improved data collection from healthcare applications that is proposed by the author will be made. In Chapter 5, a discussion of importance of big data, data security and privacy will be done. In Chapter 6, an analysis for future work that needs to be further conducted will be discussed. Lastly in Chapter 7, as a conclusion, the author will mention what has been achieved and discuss if the aim has been met from the research of this thesis.

In this thesis, the scope for wearable devices is only limited to smartwatches and fitness trackers. The author is limiting to only these devices, since these devices are directly connected to the Internet and have ability to share data from the device itself directly between entities who are interested in collecting healthcare application data.

2 Background

In this chapter, the author will be conducting background research on healthcare applications. The research covers, changes in statistical usage of healthcare applications over the past few years, privacy issues with current healthcare applications data collection and briefly looks into how Estonia has implemented DLT in e-Healthcare system.

2.1 Use of IoTs and Wearables in Healthcare

Nowadays many healthcare data are being handled electronically as technology rapidly keeps developing. For example, the use of IoTs and wearables are allowing data management to be done digitally and automatically rather than having data in paper format and reduced manual work, such as the use of machine learning to further conduct health research [1]. According to a research, data that was collected back in 2019 November, 19% of Americans are using healthcare wearables and 15% have used them before. Additionally, 19% of Americans are preserving their health data digitally on application and 13% have used it before [2]. According to research done by Gartner back in 2019, smartwatch spending was $12,412 million worldwide in 2018 [3]. Research that was conducted in 2021 by Gartner suggests that in 2019, $18,501 million worldwide and in 2020 it was $21,758 million [4]. Additionally, Deloitte is suggesting that wearable device usage will keep growing and higher demand worldwide [5]. As the research suggests there is an increase in the use of wearables for healthcare, which suggests that there is a transition in healthcare to digitalization from paper format. As many people are also transitioning to digitally recording their healthcare and activity records, hospitals need to adapt to the change and remotely monitor patients' healthcare.

Moreover, since COVID-19 has begun, there are changes seen how people have started using wearables, there is an increase in the use of wearables on regular basis by the people who had the device already. Other 46% of people have been keep using as usual since before the pandemic has started, on the other hand there is also an indication of a decrease in usage of 21%. However, there are interests seen in owning wearables from those who have not owned one yet, the main purpose is for the health monitoring. Where there is an

increase in the number of people who are interested in owning a wearable compared to before COVID-19 started, the increase was from 24% to 27% [6, p9]. Although, not just wearable devices, health and fitness applications usage on smartphones have increased throughout the period of COVID-19 pandemic. The research done by Deloitte shows that there is increase of between 11% and 14% of people who have started monitoring their health using smartphones. Moreover, there are many interests in keep monitoring their health using smartphones even the COVID-19 pandemic finishes [6, p10].

As the research shows, the use of wearables and smartphones are increasing in healthcare since the pandemic has started and people are willing to keep using wearable healthcare technologies. Which suggests that there is a need for an awareness of how the use of technologies such as wearables and smartphones will change medical care in the near future. At the same time, since the increase in demand for use of wearable devices and healthcare software, there must be a consideration of how these data can be handled carefully between device manufacturers, application developers and health institutions for developing and improving hardware, and software for research. Since these health data that have been collected from wearables contain many sensitive data, especially when shared between healthcare institutions and doctors. Data sent from health applications can contain such as heart rate, sleep duration, chronical disease data, blood oxygen level and physiological data which are very sensitive and must not be shared with non-authorized entities. Therefore, there is a need for a new method for handling data between data sharing participants, and device manufacturers and software developers for use of data from healthcare applications for research purposes of improving its products.

There is a rise in a number of interests in owning wearables since COVID-19 started as it has been suggested by Delloite's research, for the purpose of monitoring individuals' health and activities from their wrist for convenience. Moreover, wearables can help detecting and giving the wearer an early warning of disease symptoms by collecting and constantly monitoring health [7]. According to the study of Tejaswini Mishra et al, data collection from wearables of heart rate, steps and other physiological data, enable to inform the users an early symptom warning by continuous monitoring of health. The research has been conducted by collecting a large number of healthcare data from participants' wearable devices with those who have a normal health conditions, with COVID-19 symptom, and other health conditions. With those data collected, it has enabled Tejaswini Mishra et al to construct a model for detecting COVID-19 symptoms

by examining data with application of statistical analysis using cumulative distribution. [8] Similar application can be applied for research on different disease symptoms to give a warning for early signs of symptoms. This enables both the patients and doctors aware of health condition by constantly monitoring before it gets worse and too late.

Furthermore, companies such as Apple and Google have started studying health, such as heart study, illness detection, DNA analysis and many more with the use of healthcare applications. These companies partner up with medical institutions to study health using health data collected from wearables and smartphones [9] [10]. One of the example cases is, Mintu P Turakhia et al, a research group from Stanford Medicine and Apple have worked together back in 2019 to conduct heart studies using wearables. Using health data collected from participants' wearables, Mintu P Turakhia et al have conducted research in identifying patterns of irregular rate of pulse to identify the risk of getting a stroke. This research involved 419,093 volunteers for the collection of various data from people with different health conditions. [11]

As the research suggests, to improve healthcare applications features such as predicting health and medical conditions, device manufacturers and software developers will have to collect healthcare and activity data from various sources. The same applies to health institutions that are researching healthcare technology. To collect these data, they are collected from healthcare applications of users. This enables manufacturers and health institutions to conduct further research and improve its hardware and software, also algorithms for detecting early signs of symptoms. As the emergence in increasing use of wearables, especially smartwatches and fitness trackers are seen recently, it is visible how these devices are attracting many people to get hands-on with it to take the advantage of tracking their healthcare and activity records. However, there is a problem from the privacy perspective, where users who are willing to participate to share their healthcare and activity records are not feeling confident sharing data, which applies to the data confidentiality. However, there is another issue from the perspective of data collectors. Once the data has been collected, the data should not be changed, no unauthorized users should be able to access, which applies to the data integrity. Since the data that has been collected earlier can be used for machine learning purposes for developing algorithms for detecting symptoms and hardware.

2.2 Privacy Concerns in Health Record Usage

As mentioned in Section 2.1, the usage of wearables for healthcare purpose is increasing, what this suggest is that more people are starting to record their health conditions from wearable devices. Although, there are concerns with data collection from these devices, which is privacy issues, how devices are handling data, data sharing between third parties and what is being collected. According to Deloitte, which the statistics that were collected in the US alone, 60% of wearable device users are not concerned with privacy of data collected through wearable devices, [6] [12] this includes data such as heart rate, GPS data of where they have gone, steps per day, sleep duration, disease conditions and much more. Relatively, many users are not concerned over how their data is being used and handled, which suggests, that there is not much awareness of how these data are being used or can be a cause of threat on them. On the other hand, 40% of those who own wearable devices are concerned with privacy and data [12]. Additionally, among those who own wearable devices, 26% of the owners are subscribing to a health report service, these services which creates individual health report based on data collected through their wearable devices. However, 60 percent of those who are subscribing to such services are worried about how the data is being handled by those service providers [12]. This shows that, there are users who own a wearable device have an awareness and understand the risks of their privacy and personal data for sharing with third party services. Although, when it comes to the device itself collecting data, awareness of how data is being handled by these devices decreases, many people in the US have answered as they are not concerned. What this suggest is that wearable device users are not aware of how manufacturers are handling health data that have been collected through their devices.

For instance, on 30 June 2021, according to Jeremiah Fowler et al [13], the team found out that the data that has been collected through smartwatches and fitness trackers had been breached from a database with no password protection. Which led 61 million user data across world being exposed. Through the investigation, the data breach was originating from a service that offers a unified API solution to access health data from various wearables' applications, the application name called GetHealth. The data that was exposed had contained various health and private data collected from wearable devices:

- Fitness application information

- Activity records consisting of heart rate, blood oxygen level, sleep statistics, GPS tracking data, walking and running speed, weight
- User profile information: First and last name, username, date of birthday, gender
- Manufacturer of the device and OS
- Some other private information collected through GetHealth [13]

The data above can expose sensitive information about the users. For example, if the data is to be used for monitoring patients' health data over time for examining their health based on history, exposed data allows people with an intention to disturb activities of organizations or other's work to amend data. This leads to compromised data integrity, as a result, studies done by doctors and health researchers will not be reliable to be used and valid record of history. Additionally, the author has mentioned in Section 2.1, some organizations are using health data collected through wearables that are used for machine learning for research, such as for predicting the health of patients and its wearable users. As data gets amended, it will also not be available to be used for training models for machine learning, since the data collected is not valid anymore. Machine learning requires a large resource of valid and reliable collection of data. As the data is reliable, the model for predicting data will become reliable. However, if there are false data or any kind of outliers, these data need to be deleted during the training of model, otherwise including faulty information will cause inconsistency and unreliable prediction model.

In case data collected from healthcare applications gets to the hands of unauthorized personnel, it can be used in various ways, such as phishing attacks based on the data that was collected through the exposed user data. Since the data may contain a name, email, and other essential information, which enables to send a phishing email. Additionally, based on exposed health data that has been collected through users' wearable devices, it can be used to re-identify the original user by further collecting and analyzing the patterns from various sources such as social media, to perform OSINT. [13]

Moreover, statistical data that was collected back in 2019, in the US, 40% have answered that they are happy to share health data with the manufacturer of the wearable device. On the other hand, roughly only 20% of consumers from elsewhere apart from the US have answered that they are happy to share data with a manufacturer of the wearable device

[14]. Although there is a population difference depending on the country and purpose for using health applications may be different, although overall number of people who are willing to share their health data is low. This suggest that there are many wearable device users who are concerned of privacy, how their data are going to be used and handled. As users do not feel confident sharing data, this causing device manufacturers, software developers and health institutions are not able to collect enough data for research purposes from a wide range of users with different backgrounds.

In figure 1 below, there are images for preference for data sharing options for healthcare applications from smartphones and wearable devices, which the screenshots were taken from Apple iOS and Android devices. As figure 1 is showing, on the left, is a screenshot from Apple iOS device. Apple is mentioning that health data collected from individuals are used for improving its products and services. Additionally, in their "About Improve health and Activity & Privacy…" [15] it is mentioned that collected data that has been shared by users maybe shared between health and medical services, also stored information between third-party providers. Moreover, in "About Improve Health Records & Privacy…" [16] they are stating that during the transfer of information and data storage in the database, data are in the state of encrypted format. Looking at the Google Fit on the other hand in figure 1, the pictures attached in middle and right, it is not showing any information how Google is handling collected data. Although, users have the opportunity to whether they do not want to connect Google Fit with other applications on the phone, which also exist in the Apple health application.

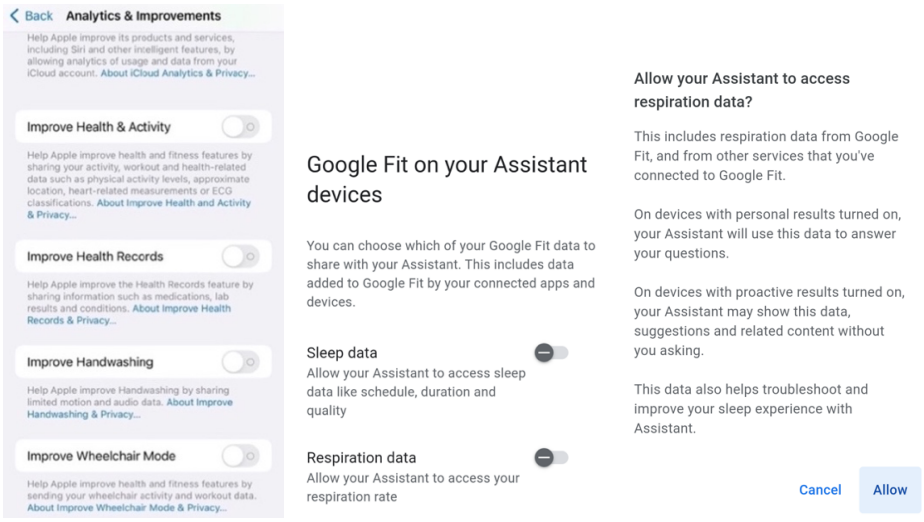


Figure 1. Healthcare sharing preference

Although, it is not clear whether the data is stored securely and during the transmission from device to their database. Since the author has taken up the case, where health data that has been collected from smartwatches and fitness trackers have been breached, affecting a large number of users putting into threats. Moreover, the manufacturer of the device has mentioned that they may share data between third parties, however, the users who are contributing do not know, which third parties have their data. Which causing users who are willing to contribute for improving the hardware and software of healthcare applications to not feel confident sharing data, due to no understanding of who has read access to the data, since it is confidential data. Furthermore, the data they are sharing is not for advertisement purposes, the main aim is to contribution toward improved hardware and software for healthcare and activity recordings. It is concerning for users that are contributing for improving the healthcare application, instead, if their health data gets breached due to a lack of security measures taken by third parties, which leads data can be used to analyze the pattern to re-identify the original person and unauthorized personnel to have access to confidential data.

Although, sharing data between device manufacturers may have some threats. Since the users may not know how the data are being handled even though it is mentioned that data are encrypted and only used for improving their products, for instance, data can be used for advertisement purposes. Additionally, the data collector may share data between third parties to conduct research together to improve hardware and software without consent.

2.3 Distributed Ledger Technology in Healthcare

One of the types of DLT is a blockchain [17], where every block holds information, such as transaction details, and they are chained up together to form one chain. The blocks that are chained up together are distributed among other participants who are participating in the network. To preserve historical data, the blocks are chained together to form one chain, only new records can be added to the chain, therefore it allows to keep the data integrity by avoiding any changes to be made in the information that has been added earlier. Since the record, which is a ledger, is shared among the participants in the network, it is decentralized, not only one person has a ledger, but people who are all participating have the same ledger to enable the data availability. In blockchain the first block is called the genesis block, which has no link to any previous block [18]. Each block

holds certain information, which consists of the two sections, block header and block data, which contains the following in each section:

- Block header
    - Block location number
    - Hash value of the previous block in header
    - Hash value of the block itself
    - A timestamp of the transaction
    - Block size
    - Nonce, this may or may not be included, use case for solving a hash of a block
- Block data
    - Details of transactions and events
    - Other necessary data [19]

The information stored in the block header section keep records of how each block is connected to one another. If new information comes in, new block is created and chained together after the final block in the chain. The purpose of DLT is, by referring to how blockchain is constructed, one of them is “integrity.” Integrity is to ensure that the data is consistent and protected against any unauthorized changes to be made in data. Since each block accommodates the hash of its block and the hash of its previous block in the header, it makes sure that the sequential order of the chain of the blocks is correct and connected. Otherwise, if an attacker modifies the data in a block, the hash value of the block will change, which causes the chain of the block to be broken. Since the blocks will not be able to be chained up together, as the block that is chained behind the block that has been modified will have different hash value from the previous block in its header, which breaks the connection. Even a small change made in the data causes in completely different hash value, which DLT focuses on tamper-proof record. Additionally, the decentralized model makes it harder to make changes in ledger structure, since all the

nodes that are participating in the network have a copy of same structure. Therefore, if ledger gets modified, it can be compared between other nodes, enabling consistency of the structure [20]. This makes sure integrity and availability of data from the CIA triad.

Second purpose of DLT is to keep availability [21, p7]. Availability is to ensure that the data is reliable and accessible promptly. Since DLT is decentralized, therefore everyone who is participating in the network has a copy of the same ledger. For instance, if new information comes in and ledger gets updated, ledger will also be updated on other nodes, which makes sure that the information is up to date. Additionally, as the record is kept on multiple nodes that are participating in-network, even if one of the nodes goes down, data is still available, [22] which avoids from a single point of failure. This is possible with the point of view of decentralized technology.

One of the examples of DLT usage in healthcare is e-Health service in Estonia. The electronic health record is being used in Estonia to keep healthcare records accessible from different healthcare services of each patient from a single place. To keep electronic healthcare records tamper-proof and preserve privacy from any kind of attacks, Estonia is making use of KSI blockchain which has been developed by Guardtime in Estonia [23]. KSI blockchain focuses on data integrity, which enables to ensure that data is reliable and consistent, therefore both the doctors and the patient will see the same latest data of their health records. Moreover, to preserve the history of healthcare records. To preserve privacy for the data stored in the ledger, PKI is used to ensure only authorized users can access the private data.

Estonia has started to implement KSI blockchain since 2008 to test out the technology and by 2012, they have started to use in live. Since then, Estonia has implemented use of KSI blockchain in many registries that is used by government, not only in healthcare, for instance, population registry, property registry, business registry and many more [24]. Additionally, as a solution for COVID-19 vaccination certificates, KSI blockchain is also being used to ensure data integrity [25], therefore the certificates cannot be amended and keep the validity of period by having a timestamp.

3 Methodology

Speaking of emergence in number of usages in wearable devices and smartphones to keep health records, there must be a need for consideration of improving data collection by manufacturers and developers. From the perspective of integrity and confidentiality of data, there is a need for a solution with a new initiative way to securely handle health records. The research to improve data collection of healthcare data from smartwatches and fitness trackers with the use of DLT is constructed with the implementation of the author's theory knowledge. Which are required during the data collection from healthcare applications from smartwatches and fitness trackers. To support the author's implementation, external resources, scientific reports and sources available online will be used. The outcome of this thesis is to create a prototype diagram of infrastructure will be created.

Overall aim of this thesis is to:

- Analyze different types of available distributed ledger technology [26][27] and introduce Hyperledger Fabric DLT framework [28]
- The author will create a proposed infrastructure diagram model of how Hyperledger Fabric DLT framework can be implemented for data collection from healthcare applications by organizations with the aid of research papers [41]
- Analysis of the proposed infrastructure and its functionalities and implementation of data collection
- Briefly analyze the application of data collected through proposed infrastructure for improving healthcare applications.

4 Solution

In this chapter the author will propose a solution for a new method of data collection from healthcare applications with the implementation of Hyperledger Fabric framework with a focus of confidentiality and integrity of data.

4.1 Available Blockchain Technologies

There are public and private blockchains, the author will briefly describe each blockchain, how they are different from each other. By understanding the difference, the author will further implement blockchain for data collection from healthcare applications with a focus on data confidentiality and integrity.

4.1.1 Public Blockchain Technology

The use of public blockchain is such as cryptocurrency, Bitcoin. Public blockchain is publicly open for anyone to join the network, participants do not require permissions or identify themselves as they are part of the member [26, p153]. Since anyone can join the network, there is no central authority who manages the network, moreover, anyone can read and make a new transaction to update the blockchain [27, p2]. As every new transaction is made, new block will be added to the very end of the blockchain, as it is mentioned in Chapter 2, Section 2.3. Once after the transaction has been made in public blockchain, verification must happen, for example if money is to be sent to another user, it must check if the sender has enough money in their account. To verify this, users who are participating in the network, in other word, “miners” will validate it, then appended to the end of the blockchain [29, p3].

4.1.2 Private Blockchain Technology

The use of private blockchain is where the access to the network needs to be restricted to certain users, they must be trusted to be a member of the blockchain. Unlike public blockchain network, private blockchain is not open to anyone, it may be created by a company for only specific members to be able to join or clients who have been approved to join the network [26, p153]. In private blockchain network, it can be restricted to who

can read and write data, whether only to certain or all participants in the network. Once a transaction is made, the validation of block can be done by certain nodes that is responsible for validating [27, p2]. Private blockchain allows restrictions on who can join the network, which prevents from data to be publicly available and who can write data in blockchain. Which private blockchain is suitable for use within the organization purpose.

4.2 The Proposed Technology: Hyperledger Fabric

The author's proposal for data sharing from participants' healthcare applications to device manufacturers and software developers, is to implement Hyperledger Fabric. It is a framework for private and permissioned DLT [28], it defines access privilege for each participant to restrict read and write data in blockchain [29, A22]. Therefore, even if the participants in the network are approved to join, not everyone has permission to read and write inside blockchain, which offers greater control of the data confidentiality. Since healthcare and activity records collected from smartwatches and fitness trackers are confidential, storing this kind of data in public blockchain will enable anyone to see. However, using private and permissioned DLT, Hyperledger Fabric framework enables users to share data between certain data collection entities with the confidence of data confidentiality and integrity. Since, Hyperledger Fabric has not been applied in the use of data collections by device manufacturers and software developers. Therefore, the author will propose the implementation of the technology for improving data collection from healthcare applications in this thesis.

To understand Hyperledger Fabric, it is one of the private DLTs, however this is a framework, it aids for the development of applications where it requires implementation of private DLT functionality. Hyperledger Fabric is an open source project and the architecture can be designed according to the needs of solution for an application, these components will be discussed later in the chapter, Section 4.5 Components. Moreover, since it is a private DLT, only the members can interact inside the blockchain network. However, Hyperledger Fabric allows further restrictions inside the network, where only permitted members inside the network can read and write ledger. Confidential data can be only shared between permitted members in the network, therefore not all the members inside the network can read nor write in ledger. [28] Further implementation of the

Hyperledger Fabric for proposed data collection from healthcare application will be introduced later in the chapter.

4.3 Implementation

This thesis is to implement a proposed model diagram for collecting healthcare and activity records from smartwatches and fitness trackers between users, and device manufacturers and software developers. To understand the whole picture of the implementation, proposed infrastructure diagram from a new method of data collection from healthcare applications, which has been created by the author is attached below, figure 2. The components and its descriptions which are involved to create infrastructure with the use of Hyperledger Fabric is described later in the chapter. They are implemented accordingly to meet the goal of this thesis, which is improving data collection from healthcare applications by device manufacturers and software developers with the focus of data confidentiality and integrity. Which enables to develop trust between the users and data collectors for creating better health applications.

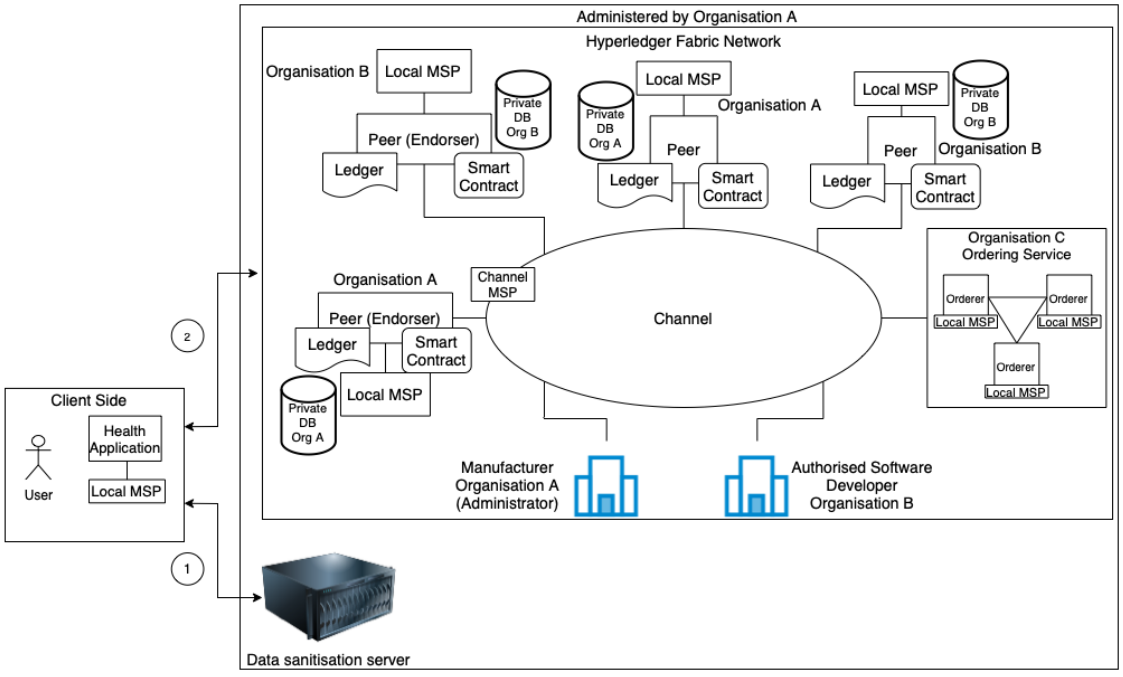


Figure 2. Data collection with Hyperledger Fabric diagram

4.4 Components

4.4.1 Health Application

The main source where the data is collected is the health application. From the health applications, users who are willing to share their data will be sent to third party entities, device manufacturers and software developers. It is the point where data will be collected from a user and sent out from.

4.4.2 Local and Channel MSP

Every participant on the channel has an identity. The identities which are held by individual participants who are involved in the Hyperledger Fabric channel network performs an important role. Since Hyperledger Fabric channel is a permissioned blockchain, the participants who are involved are authorized and identified what permissions they have. To identify, individual participants' principals are identified in their digital identity for defining the permissions they are granted. These identities are issued by CAs [29]. The participants are the following according to figure 2 diagram created by the author:

- User application
- Peer and ordering nodes
- Organizations

The MSP has a responsibility to provide what roles every member in the channel has, by holding the identity that was issued by CA. With the use of PKI, MSP allows identity to be verified on the channel between nodes during the transaction [31], which allows them to identify themselves that they are authorized participants on the channel. During the issuance of certificates to the members who are participating in the channel, certificates are issued from intermediate CA rather than the root CA [32]. This is due to the case if the root CA has been compromised by an attacker, they have ability to create certificates for anyone. Therefore, to prevent this, intermediate CA is used to create certificates for the participants. The intermediate CAs do not have the ability to self-sign its certificate, therefore, the root CA will have to issue certificates for the intermediate CAs. This

enables to mitigate attackers from creating certificates even if the intermediate CA gets compromised.

There are two types of MSPs, one of them is local MSP, which must exist on every client, peer and orderer node, this is in order to hold their certificate for identity and group the participants into an organization. It is necessary since it defines which peer and orderer nodes have what kind of permissions on the channel and to identify its peer administrator [31]. The local MSP on the client allows them to identify that they are a member of the network and is authorized to make transactions to send data from their health applications. The local MSP is only applied on the node level, locally itself. The second type of MSP, channel MSP identifies which participants are part of the channel and holds MSPs of the organizations who are part of the channel. All the participants of the channel have the same configuration of channel MSP, therefore it allows them to identify if a peer who is making a transaction is authorized in the channel [31]. Moreover, in figure 2, the proposed infrastructure has three organizations currently, Org 1, Org 2 and ordering service. These organizations have their MSP defined and their members/peer nodes are linked to an MSP that is managed by the organization that they belong to. Therefore, it allows identifying which members/peer nodes belong to which organization and who has the administrative privilege, a node which overlooks the members on the channel for their organization [31]. Otherwise, if all the nodes are in the same organization, there will be only one MSP, which there will be no control over which organization has access to specific data in the ledger, private data will be visible by all the nodes. The channel MSP is applied on the channel level, which they are applied to all the nodes who are part of the channel, which allows all the nodes to have the latest information of who is authorized to be part of the channel, permissions and role of every node.

As the members who are involved in the data collection are authorized and identified to be part of the channel, this makes sure data confidentiality, which prevents sensitive data to be available to anyone and eavesdropped, only by the authorized members. Moreover, members can be grouped into organizations to control access to data. From the perspective of data collectors, such as device manufacturers and software developers, it allows them to collect reliable data. As the data can be only sent from the clients who are authorized, part of the member of the channel, otherwise, the transaction will be rejected. Therefore, it enables both the confidentiality and reliability of data by collecting data from the sources that are authorized.

4.4.3 Data Sanitisation Server

The term data sanitization is the process of deleting data that is not necessary [33]. The data sanitization is required to delete all the sensitive data that is unnecessary to be held by third parties in the process of developing health applications and its hardware. For instance IP address, GPS location, username, real name and many more of these data are not necessary during the process of research and development. Since device manufacturers and software developers requires only certain information that is related to healthcare and activity data and other few technical statistics data, such as hardware and application that the user is using. Other data that can be used to re-identify the actual user, such information mentioned earlier needs to be deleted before that is going to be stored by organizations. This enables to ensure there is data confidentiality in place from the perspective of the users. Any sensitive data that is unnecessary is deleted to ensure it is not made available to be seen and stored by third parties.

The purpose behind why there is a need for a dedicated server that does data sanitization is due to the lack of computational power available in smartwatches and fitness trackers. Wrist worn wearables have limited computational power, and they are not as powerful as smartphones and computers. Before sending to the data sanitization server, data can be locally on the user smartwatch or fitness tracker sanitized first, as the user can select what healthcare and activity data to send. However, smartwatches and fitness trackers only have limited computational power and limited battery, performing many tasks on user side will not be effective and UX will not be satisfied. Therefore, rather than only locally sanitizing data, sending data to a server side, the task can be done in a shorter period as it has got more computational power and consumes less power on the client's device.

4.4.4 Peer Nodes

The peer nodes which are connected to the channel of the blockchain network, have roles to keep hold of smart contracts and ledgers. In figure 2, there are multiple peer nodes, and they are owned by specific organizations. There are two organizations in the infrastructure and two peer nodes are owned by each of them. There are two types of peers, committing peer, which is responsible for validating the transaction and updating and holding the ledger. Secondly, the endorsement peer, which receives a proposal of the transaction to simulate the transaction to create the following, first set is readset, it is the hashed keys that were read, and second set is writeset, it is the new key and value that needs to be

written to the ledger. Once the transaction simulation has been validated, the endorsement peers will send back readset and writeset back to the user along its signature [34, p6] [35]. The endorsement peers are selected according to an endorsement policy which will be explained in Section 4.4.7. The endorsement peer node hold ledger as well. Figure 2 explicitly identifying "endorsement" node for ease of explanation for the transaction flow. There are multiple peer nodes owned by organizations to ensure consistency, in case data gets modified by an unauthorized user, nodes can check each other for the correct data that must be stored in the ledger ensuring data integrity. Further detailed explanation of ledger and smart contract will be done in Section 4.4.7.

4.4.5 Orderer Nodes and Ordering Service

The main concept of implementing Hyperledger Fabric framework in blockchain for the purpose of collecting data from healthcare applications from users by device manufacturers and software developers is, since it is a permissioned blockchain. As the author has described earlier, only authorized participants can join and have different access privileges in the network. The orderer node manages ordering of transactions and packing transactions into blocks in order, which these processes are specifically done by orderer nodes by the process called ordering service [36]. Therefore, unlike permissionless blockchain, with Hyperledger Fabric, only certain nodes have ability to consensus transactions, as it is a permissioned blockchain.

To implement ordering service, the author has proposed to use Raft to design the infrastructure, although there can be implementation using Solo. However, there is a drawback with the Solo implementation, it is not fault tolerant. What this means is since Solo order service only runs on one node, therefore, in case the node goes down, there is no backup node, which can cause a single point of failure as transactions cannot be validated. Additionally, Solo is used for development environment [34, p9]. Therefore, from the consideration of these two points, Solo ordering service is not suitable during the collection of data from healthcare applications, where there will be transactions one after another every seconds. On the other hand, Raft is a crash fault tolerant ordering service, there are multiple orderer nodes, which allows even if one node fails, there are still working nodes that can order the transactions. In figure 2, there are 3 orderer nodes, in Raft nodes can turn into one of the following:

- Leader: A node which is responsible for sending logs, which is command or instruction, to follower node. All the followers must have same log entries as the leader node.

- Follower: At the first state, before electing the leader, all the nodes are in the state of follower. Once the leader has been assigned, all the nodes are in follower states. They must have the same log entries as what the leader node has.

- Candidate: A candidate state is, before turning from follower state to leader state, follower turns into candidate, then becomes a leader [37].

4.4.6 Ledger and Database

A ledger is where all the transactional data are recorded. Inside the ledger there are two components, a world state and a blockchain. In figure 2, where the author has designed an infrastructure, every peer has a ledger, although world state and blockchain are not shown, however, every peer holds both components inside its ledger.

The world state is a database where it holds the latest values, which enables ease of access to retrieve the latest value, rather than going through every block in the blockchain to retrieve the hash of the previous and last block in the chain. The record inside the world state is made up of pair of a hashed key and a value. As every new transaction is made, it will update the world state.

On the other hand, the blockchain is a chain of blocks that is linked in a sequential order of changes that were made in the world state. In Chapter 2 Section 2.3, the author has briefly talked about blockchain. However, the author will deeply explain its components, block header, block data and block metadata further in details.

Inside the block header, there are three components. Block number, which indicates the position of the block, starting from 0, which is assigned to the genesis block and increase by 1 as new block is appended. Secondly, hash of all the transaction in the current block and lastly the hash of the previous block header [38]. The header ensures data integrity since every block is linked to each other using a hash, otherwise, if the data is modified, the hash will change leading the link between blocks to be destroyed.

Inside the block data, where transactions are held, there are the following attributes, header, signature, proposal response and endorsements. Header section which stores metadata, such as smart contract that has been applied for this transaction, smart contract will be explained in Section 4.4.7. Signature section stores the signature that has been applied by the participant's health application, which is essential to check the integrity of data. During the transaction of data, it makes sure if the data has not been changed by an unauthorized user. Which is another essential for validity of data and building the trust between the users and third parties who are interested in collecting data. Proposal section is responsible for holding data that is going to be applied into smart contract on endorsing peer nodes. Further details for smart contract will be introduced in Section 4.4.7. Response section stores result of values from running a smart contract based on the proposal. Endorsements section holds transactions that have been approved from organization peer nodes that are needed to meet the reequipment of endorsement policy [38]. Endorsement policy will be described in Section 4.4.7.

Finally, block metadata, holds essential information to ensure data confidentiality and integrity, it has certificate and signature of the block creator, since the block needs to be verified by the nodes who are part of the network [38].

The main components block header and block metadata enables to ensure the data integrity and confidentiality. The block header links the blocks using the hash of the previous block, if the previous block data gets modified, the chain will break, this applies to the data integrity. Moreover, it makes sure of data confidentiality, as the nodes that are part of the network can verify the creator of the block, if the certificate and signature are not recognized by other peer nodes in the channel, it suggests that the block creator is unauthorized and is not part of the network. This enables to understand that data could have been modified by an unauthorized party, also breached data.

The proposed model for the data collection from healthcare applications uses private data collection. This enables healthcare and activity records, which are confidential data, to be only shared between chosen organizations by the user in the network, rather than to all the organizations. To store private data, instead of storing on the ledger, they are stored in a database of peer nodes of authorized organizations that the user has chosen to share with [39]. Using private data collection, instead of storing actual data on the ledger, the hash of key and value are stored, in the database, healthcare and activity data are stored.

Moreover, every data stored in database has salt added, therefore, attackers will not be possible to find the same hash value of data that is stored in ledger. [40] The hash values of private data are stored in the ledger of every peer node, which enables to ensure data integrity. In case the hash value in the ledger does not match the hash of the data stored in the database, it suggests that the data has been modified. This can be recovered by retrieving private data stored in other peer nodes from same organization to maintain consistency. For the database, CouchDB is used on peer nodes since it allows to store data in JSON format [45]. The author has extracted raw health data collected on Apple Watch to have a look at the content, part of the data is attached in Appendix 2, Figure 3. The data was exported as XML format, although it is easily convertible to JSON format, which led to implement CouchDB on peer nodes to store health data collected from users. Furthermore, with JSON format, it enables to store continuous data from activities such as walking and running, in array format as one document in CouchDB. The format for how the data are going to be stored in database is proposed by the author, both non-continuous and continuous data are available in Appendix 2.

In the proposed model, following data are stored in the ledger in hashed format:

- Key: Unique user identification number assigned by the application on client
- Value: Actual health data shared from healthcare applications

Proposed data types are stored in the database of the authorized organization's peer nodes:

- userId: Unique user identification number that was assigned by an application for the participants who are data sharing, stored as in string
- sourceName: Name of the source where the data come from, such as application name, data type for this is a string
- device: Model of the device, OS and its version, stored as in string
- applicationVersion: Version of the application, stored as in string
- value: Result for the measurement(s) of health data, stored in an array
- type: Type of measurement, such as heart rate, blood oxygen level, glucose level and other measuring values, stored as in string

- unit: Unit for the value, such as count per minute, percentage, stored as in string

- time: The time when the measurement was done, formatted in YYYY-MM-DD HH:MM:SS +0000, stored as in datetimeoffset which specifies the time zone

The above data above are proposed by the author based on health data collected on Apple Watch, raw data can be checked in Appendix 2 figure 4, additionally there are samples of how health data will be stored in the peer nodes' database proposed by the author. One for non-continuous data such as heart rate taken at a regular interval and another is continuous data for walking and running activities.

4.4.7 Smart Contract

Smart contract manages the access privilege to the data in the blockchain, between the users and third-party entities [41, p7], which enables to maintain data confidentiality in the blockchain, as only authorized users have access. The implementation of a smart contract in the proposed model for the data collection from healthcare applications by device manufacturers and software developers can be as following rules:

- First smart contract, for sharing data from client side, a participant can specify which organizations have access to private data and are willing to be collected for research purposes, additionally choose what data to be collected. Therefore, the users can limit data collected and accessed by certain organizations. In the health application, the user has the opportunity to select data they would like to share, to confirm, the user will have to consent to share data in the application. Once the smart contract is run, the data owner is the participant, this can be defined by identity which is assigned by the local MSP on every individual participant's application. Therefore, they have rights over control to share data with other organizations from the same channel.

- Second smart contract, for a user querying own health data from peer node database, allowing them to see what data have been collected and shared with authorized entities. A user can query data from the application by sending a read request along their identity assigned by local MSP, then this smart contract will run on peer node. Based on identity, peer node will send back health data that user has shared and has authority to access.

- Third smart contract is for organizations for querying data from a peer node's database, data that have been collected from users to share their healthcare and activity data. With this smart contract, data can only be queried by nodes from the peer node database of the same organization that they are belonging to. To verify that the organization node is authorized for querying data and is belonging to the same organization, the identity assigned by local MSP on the node is checked if they are valid when this smart contract is run. If it is valid, peer node will send health data to the organization node that requires to read data stored in a private database.

Moreover, endorsement policy must be defined by the user from application, the endorsement process is to make sure if a transaction sent by a client to an endorsement peer is valid during the execution of a smart contract. If the endorsement is valid, approval and transaction with the endorsement peer's signature will be sent back to the client before the transaction is written to the ledger. An endorsement policy defines by which organization transaction must be endorsed, which is specified by the user [35]. In the proposed infrastructure, in the healthcare application a user will have the option to select sharing their data between Org 1 or Org 2 or both, before sending a proposal of transaction. Therefore, health data will be shared and the endorsement process to be done by only certain organizations, which makes sure data confidentiality by restricting sharing of data, additionally the validity of data before being recorded in the ledger and database.

Furthermore, the private data collection definition proposed by the author can be checked in Appendix 3. Private data collection definition defines a policy for endorsement of the transaction, how many peer nodes must receive health data, what are the private data, who can read and write data and which organization must add a signature for the endorsement to be valid. The definition has been created based on Hyperledger Fabric documentation [42] to suit the requirement of the proposed infrastructure proposal.

4.4.8 Channel

The channel allows communication to happen between clients and Hyperledger Fabric blockchain network, where peer nodes and ordering service is located. In the channel, there is always at least one peer node, order node and organization connected. As different organizations and peer nodes join, they share a ledger which is the same across all the peer nodes, enables consistency across the network [43] and ensures data integrity. In

case data gets modified by an unauthorized user, peer nodes in the network have same copy of the ledger, therefore it can be compared once it has gotten tampered and can copy the ledger. As figure 2 shows, the proposed infrastructure diagram by the author shows, an application from client side, peer nodes from “Organization A” and “Organization B,” ordering service and organization branches are connected to the Hyperledger Fabric blockchain channel. Moreover, there can be more than one channel depending on different security policy placement in the infrastructure.

With the author's proposed model, there is only one channel in place, as the scenario of this network is initially the network was created by “Organization A,” which is the device manufacturer and has the network administrative rights. Later, the “Organization B,” which is the software developer company, who agreed to join the “Organization A” network to start collecting data from their health application available on devices manufactured by “Organization A.” This network model allows sharing of the same ledger between “Organization A” and “Organization B” peer nodes. For privacy, data confidentiality, a private data collection mechanism is used. With private data collection, to share private data between the authorized peer nodes, data will be sent peer-to-peer using a protocol called gossip [44]. The private data that was sent between peer nodes of the authorized organization, the data will be stored in a database for private data, as it was mentioned in the “ledger” section. In the ledger of every peer, a hash of the private data will be stored. This is to ensure to keep a record of the transaction has been validated.

Although, multiple channels can be set up, where each organization have their own channel to isolate data collection from different organizations [39]. However, the author is not implementing multiple channels, rather using a private data collection mechanism. In the proposed infrastructure, there is only one network administrator, which is the device manufacturer. The device manufacturer has the right to invite other entities such as a software developer to the network for data collection from their own healthcare application or for research purposes between multiple organizations. However, adding channels leads to more administrative work to be done [39] by the network administrator, since multiple channels must be overseen. Moreover, if a user is willing to share data with multiple organizations for data collection, having only one channel with private data collection will enable users requiring send data to only a single channel, rather than to multiple. The users still have the opportunity to restrict data collection by the organizations who are in the channel.

4.4.9 Organizations

Organizations are the data collectors. An organization can be a device manufacturer, software developer or health institution. In the proposed model, administrator of the network is the device manufacturer who initially set up. Then software developers and health institutions can partner up together and join the network for data collection. Organizations have responsibility to manage their peer nodes in the channel, including adding and removing of peer nodes.

4.5 Data Flow

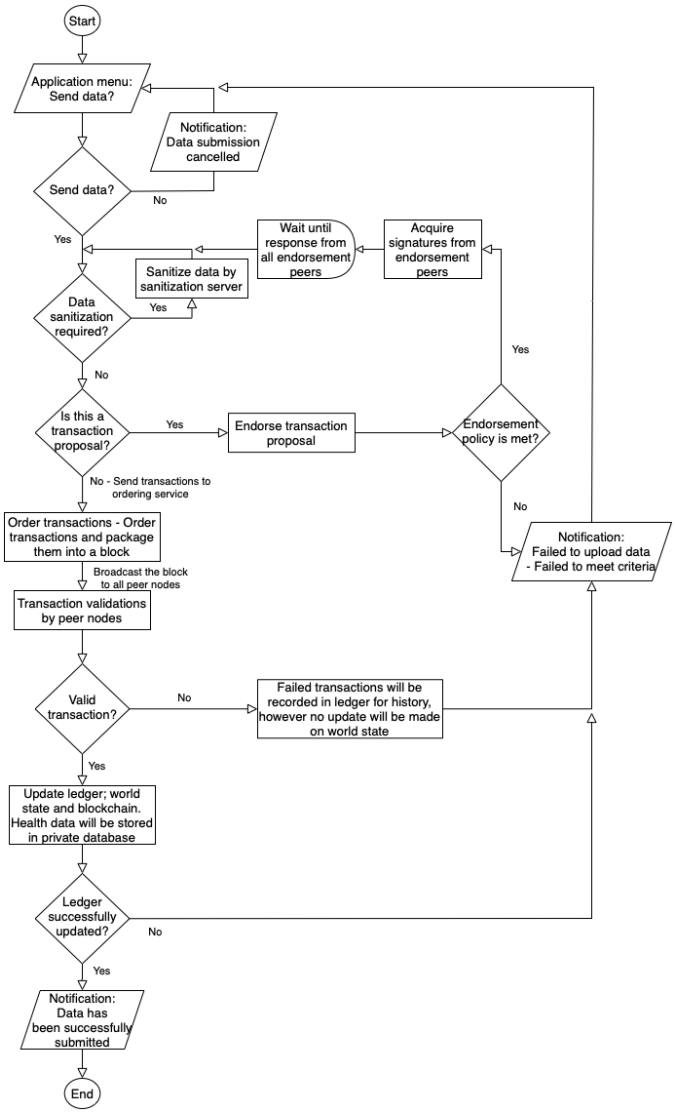


Figure 3. Transaction flow diagram based on the author's proposal

4.5.1 Data Sanitisation

First step of before data is going to be added to the ledger and stored in database of authorized organization peer nodes, data that is going to be appended needs to be sanitized, therefore the data is going to be sent to the "Data sanitization server" at the first stage. This allows to anonymize the data by deleting the data that can be used to re-identify the original user. Before the data is going to be sent to the data sanitization server, users will be notified in the application what data to share, they have choice to opt in or out and final confirmation before sharing.

During the transfer of data between client side and "Data sanitization server" side, TLS communication will be used. The use of TLS is to ensure that the communication between two peer nodes, in this case client and server, is to safeguard [46] from attacks such as data being eavesdropped.

During the data sanitation process, the server side will only read the data that has been received from the client side and delete all unnecessary sensitive data. Therefore, no data will be stored on the server. Once the sanitization has been finished, the data will be sent back to the client side for the next process.

4.5.2 Data Endorsement

Once data has been sanitized and sent back from "Data sanitization server," the next step is to make a proposal for the data transaction [43]. At this stage, a user will first select which organizations they would like to share data with, between device manufacturer Org A or software developer Org B or both. Once the proposal has been sent to the selected endorsement peers according to endorsement policy and private data collection definition in Appendix 3, peer nodes that received the transaction will execute a smart contract which is defined in Section 4.4.7. At the same time, client will send salt inside the private data. The health data received on endorsement peers will be temporarily stored in a place called transient of the peer node. The health data that is stored in the transient area will be broadcasted between authorized organization peer nodes according to the minimum number it needs to be shared with, which is specified in private data collection definition in Appendix 3. At this point, no data is still written to the database and ledger yet.

As a response to the user, the endorsement peers will send back the hashed keys and values of health data that is going to be written to the ledger. All the endorsement peers

who have endorsed the transaction must send back a response to the user. Moreover, for the validity of the endorsement that it has been processed by the certain organization peer nodes on the channel which was specified by the user, signature of the endorsement peer nodes will be added to the response from every endorsement peer. If an endorsement policy does not meet the criteria, the transaction proposal will fail and this will be notified to the user in the healthcare application.

4.5.3 Ordering

In this step, transaction and responses from endorsement peers will now be sent to ordering nodes from the client application along user's signature, although plain healthcare and activity records will not be sent this time. At this state, only keys and values in a hashed format, and endorsement peer and client's signatures are sent by a client's healthcare application to the ordering nodes. Ordering nodes will order all the delivered transactions from users in the priority of first come first served [36]. These ordered transactions will be packed into a block, the ordered transactions in a block can never be amended. The block will be then distributed across all the peer nodes on the channel.

4.5.4 Validation and Ledger Update

At this step, peer nodes will first validate the blocks that have been delivered by the ordering service. The peer nodes of the organizations who have been permitted to access healthcare and activity data will check if data exists on their transient storage that has been shared during the transaction proposal at the data endorsement step in Section 4.5.2. If the permitted organization peer nodes do not have data, then they will request other peer nodes from same organization for data, which will be shared using gossip protocol. Otherwise, if all the permitted organizations have private data in the transient storage, it will validate if the data matches the values in the transaction in the block that has been distributed. This ensures data integrity during the transaction. Additionally, it will check if the endorsement policy has been satisfied during the endorsement of data in Section 4.5.2, if it does not meet, the validation will fail and a notification will be sent to the user in the application that it has failed sharing data. This activity will be recorded on the blockchain for history purposes, however no update will be made in the world state. Block that has been successfully validated will be recorded on the ledger, [43] both on world state and blockchain. Data stored in transient storage will be stored in the database where

all private data are recorded, which will be done only on the permitted organization peer nodes. The data will be stored in JSON format in CouchDB, the example format can be checked in Appendix 2 and an explanation of values are described in Section 4.4.6. The data in transient storage will be deleted completely. Finally, the successful transaction will be notified to the user that the data has been successfully shared between an organization.

4.6 Application

The data collection from healthcare application infrastructure model design proposed by the author can be implemented in different use cases by the device manufacturers and software developers. Although the main use case of collected healthcare data are used for research purpose to improve the device and software for the reliable use of healthcare and activity monitoring from user's wrist.

As the author has taken up in Section 2.1, researchers have made use of health data collected from participants' wearable devices to conduct research for detecting early symptoms of disease to minimize the risk before it gets too late. The collected healthcare and activity data using proposed data collection method by the author may be used for machine learning purpose by the entities who have collected the data. The entities who have collected the data can invite a member, a server that is specialized computing for finding patterns of a disease symptom by performing machine learning, with the collection of big data from participants. By performing machine learning on various data collected from people with different health conditions, it may be able to identify patterns with a specific disease from multiple data [47, p4]. With the aid of the patterns found from the data collected, a new algorithm for healthcare applications can be developed for every type of disease, further improvements on hardware can be made as well. As the data that are collected are reliable with data integrity, reliable algorithms for software can be created also to aid the development of devices.

However, it is not just limited for creating algorithms for software, algorithm can be further distributed for medical equipment that are being used in hospitals. Which similarly to smartwatches and fitness trackers, by deploying the algorithm for medical equipment, it can be used to predict patients' health condition and identify disease in the earlier stage before it gets worse.

5 Discussion

In this chapter the author is going to discuss the importance of data collection and preservation for improving healthcare software and hardware. Additionally, security and privacy for protection of data that have been considered during the creation of proposal modal of infrastructure will be discussed.

5.1 Big Data

The process of machine learning, it requires large datasets to train models for creation of algorithms for healthcare applications. Although to collect and preserve reliable data, to avoid any changes to be made in the data, there must be history preservation functionality. Therefore, the implementation of Hyperledger Fabric, it allows to store data in a database and hash of data to be stored in blockchain for history preservation and data integrity. In case data gets amended, the hash value will not match to the corresponding hash that is inside blockchain, however original data can be retrieved from other peer nodes from same organization. Which allows to mitigate from any false data to be included during the training of machine learning model that have been changed by an unauthorized user. This enables to create a reliable algorithm for healthcare software. Although, not just for development of healthcare applications where it requires reliable data. Even with the development of new healthcare devices require large data set to conduct testing of its new devices similarly to how software is developed.

5.2 Data Security

To meet the basic requirements of HIPAA compliance for data security guidelines, [48] firstly, CIA triad has been taken into consideration during the proposal of infrastructure model. Confidentiality is referred in Section 5.3. To meet integrity, a hash of data is stored in the blockchain, therefore if data changes inside the database, the hash value will change causing the chain of blocks to break. This ensures if data have not been modified that are held inside the database. Secondly, to meet the availability criteria, database is distributed across multiple nodes within the same organization. However, ledgers are distributed across all the nodes who are inside the same Hyperledger Fabric channel and consistency is maintained. Therefore, if data in one

node gets amended, valid data can be retrieved from other nodes from same organization. Moreover, to ensure only authorized authority have access to certain data, every participant have MSP, identifying whether they have access to network and data, and its role. This manages the restriction to access the network where data are held and transactions are happening. Data sharing policy is implemented by private data collection method, users have ability to restrict which entities can collect data. Therefore, not all organizations who are in the network will receive healthcare data, data cannot be stored nor be accessed from other organizations' databases. Furthermore, since data cannot be stored permanently by organizations, it needs to be deleted at some point. The implementation of Hyperledger Fabric allows to set for how long data can be stored inside the database and its hash value inside the ledger. This can be defined by "blockToLive" also defined in Appendix 3. Once the defined number of blocks have been added to the blockchain, data will be deleted permanently. Organizations should not store more than 6 years of data. This enables the proposal data collection method to meet the requirement for period, data that can be held by organizations. During the communication between the user and data collection network, TLS is used to encrypt the communication.

5.3 Data Privacy

One of the aspects of data protection is privacy, during the collection of health data from applications, the author has proposed anonymization of data by removing any data that can be used to re-identify the original user. With consideration of HIPAA compliance, privacy rules for use of health data for research purposes, [49] the following listed example parameters will be removed before data are going to be stored by data collection entities, which is done at data sanitization processing:

- Names – Including username and email address
- GPS location – Only limited to postal code
- Date – Which are related to the user, such as birth of date and date of diagnosis
- Phone number
- IP address

Any kind of unique identifiable data that can be traced back to the original user must be removed. Moreover, basic sanitization of data will be done on the user's application side before it is sent to the sanitization server and users have the opportunity to select which

data they would like to share before sending. This allows to limit data collection to only what users are willing to share.

6 Future work

In the thesis, the author has created a prototype infrastructure model for the data collection from health applications and entities that are interested in collecting data for research purposes. The theoretical understandings have been defined, understood and a model has been created to meet the requirements for data confidentiality and integrity. Although further future work must be done to create a working infrastructure to perform the transaction. However, during this research, due to the limited allocated time, the working infrastructure was not set up.

Moreover, the scope for this research was only limited to smartwatches and fitness trackers, this was considered specifically due to the emerging number of uses with these devices. However, in future research, consideration for collecting data sources from various wearable IoT devices and medical equipment that are used in hospitals can be further taken in into account for collection of data. Therefore, the data collection source is not only limited to smartwatches and fitness trackers that is used commercially and can be used by anyone, but also from different sources where it is used in a critical environment. Which further may lead to different results from the data that has been collected from smartwatches and fitness trackers, and medical equipment.

7 Conclusion

Overall, the author has analyzed the importance of collecting healthcare and activity data from smartwatch and fitness tracker users in the era where there is emergence in use of such devices for monitoring and recording health. However, there is a privacy issue where users do not feel confident sharing their data. On the other hand, device manufacturers and software developers are required to collect valid data from users.

The goal of this thesis was to improve data collection from healthcare applications, where the users can feel confident sharing data with the entities who are interested in collecting data for research purposes. Also the participants to have control over their shared data, one entity cannot share data with another entity, there must be consent from the user side. To achieve this, suitable DLT was analyzed for data collection. With the use of Hyperledger Fabric DLT framework, the author has created a proposed infrastructure design of how data sharing can be done between the participants, and device manufacturers and software developers. Alongside the creation of the infrastructure diagram, components involved in the infrastructure and its data transaction have been analyzed deeply with the author's proposed theories.

The goal to create a new method for data collection from participants' healthcare applications by device manufacturers and software developers, where people can feel confident, has been developed and achieved. To achieve the development goal, public research papers and online sources and documentation for Hyperledger Fabric were looked into it carefully for the development solution of a new way of data collection from healthcare applications.

For future development, this infrastructure can be used as a prototype model for further development and implementation of the collection of data from healthcare applications. Although, it is still necessary to conduct further research, since this is still a prototype design of the infrastructure for how data collections from healthcare applications can be done.

References

[1] Abhinav Sharma, Robert A. Harrington, Mark B. McClellan, Mintu P. Turakhia, Zubin J. Eapen, Steven Steinhubl, James R. Mault, Maulik D. Majmudar, Lothar Roessig, Karen J. Chandross, Eric M. Green, Bakul Patel, Andrew Hamer, Jeffrey Olgin, John S. Rumsfeld, Matthew T. Roe, Eric D. Peterson, Using Digital Health Technology to Better Generate Evidence and Deliver Evidence-Based Care, Journal of the American College of Cardiology, Volume 71, Issue 23, 2018, Pages 2680-2690, ISSN 0735-1097, https://doi.org/10.1016/j.jacc.2018.03.523.

[2] "One in Five U.S. Adults Use Health Apps, Wearable Trackers" [Online]. Available: https://news.gallup.com/poll/269096/one-five-adults-health-apps-wearable-trackers.aspx (Accessed: 20 February 2022)

[3] "Gartner Says Global End-User Spending on Wearable Devices to Total $52 Billion in 2020" [Online]. Available: https://www.gartner.com/en/newsroom/press-releases/2019-10-30-gartner-says-global-end-user-spending-on-wearable-dev (Accessed: 20 February 2022)

[4] "Gartner Forecasts Global Spending on Wearable Devices to Total $81.5 Billion in 2021" [Online]. Available: https://www.gartner.com/en/newsroom/press-releases/2021-01-11-gartner-forecasts-global-spending-on-wearable-devices-to-total-81-5-billion-in-2021(Accessed: 20 February 2022)

[5] "Wearable technology in health care: Getting better all the time" [Online]. Available: https://www2.deloitte.com/global/en/insights/industry/technology/technology-media-and-telecom-predictions/2022/wearable-technology-healthcare.html (Accessed: 21 February 2022)

[6] "How the pandemic has stress tested the crowded digital home" [Online]. Available: https://www2.deloitte.com/content/dam/insights/articles/6978_TMT-Connectivity-and-mobile-trends/DI_TMT-Connectivity-and-mobile-trends.pdf (Accessed: 20 February 2022)

[7] "Smartwatches can help detect COVID-19 days before symptoms appear" [Online]. Available: https://www.cbsnews.com/news/covid-symptoms-smart-watch/ (Accessed: 5 March 2022)

[8] Mishra, T., Wang, M., Metwally, A.A. et al. Pre-symptomatic detection of COVID-19 from smartwatch data. Nat Biomed Eng 4, 1208–1220 (2020). https://doi.org/10.1038/s41551-020-00640-6

[9] "Apple Heart Study" [Online]. Available: https://med.stanford.edu/appleheartstudy.html (Accessed: 7 March 2022)

[10] "Google Health" [Online]. Available: https://health.google (Accessed: 7 March 2022)

[11] Turakhia MP, Desai M, Hedlin H, et al. Rationale and design of a large-scale, app-based study to identify cardiac arrhythmias using a smartwatch: The Apple Heart Study. Am Heart J. 2019;207:66-75. doi:10.1016/j.ahj.2018.09.002

[12] "Why consumers-and doctors-are wary about wearable data" [Online]. Available: https://www2.deloitte.com/us/en/insights/industry/technology/wearable-technology-healthcare-data.html (Accessed: 21 February 2022)

[13] “Fitness Tracker Data Breach Exposed 61 Million Records and User Data Online” [Online]. Available https://www.websiteplanet.com/blog/gethealth-leak-report/ (Accessed: 22 February 2022)

[14] “A consumer-centered future of health” [Online]. Available: https://www2.deloitte.com/content/dam/Deloitte/global/Documents/Life-Sciences-Health-Care/gx-global-health-care-trends-survey.pdf (Accessed: 22 Febraury 2022)

[15] “Improve Health and Activit & Privacy” [Online]. Available: https://www.apple.com/legal/privacy/data/en/improve-health-activity/ (Accessed: 8 March 2022)

[16] “Improve Health Records & Privacy” [Online]. Available: https://www.apple.com/legal/privacy/data/en/improve-health-records/ (Accessed: 8 March 2022)

[17] Jennifer Li, Mohamad Kassem Applications of distributed ledger technology (DLT) and Blockchain-enabled smart contracts in construction Automation in Construction Volume 132 2021 103955 ISSN 0926-5805 https://doi.org/10.1016/j.autcon.2021.103955

[18] “Blockchain technology” [Online]. Available: https://www.guru99.com/blockchain-tutorial.html (Accessed: 9 March 2022)

[19] Yaga, D. , Mell, P. , Roby, N. and Scarfone, K. (2018), Blockchain Technology Overview, NIST Interagency/Internal Report (NISTIR), National Institute of Standards and Technology, Gaithersburg, MD, [online], https://doi.org/10.6028/NIST.IR.8202 (Accessed March 13, 2022)

[20] “Blockchain & Cyber Security” [Online]. Available: https://www2.deloitte.com/content/dam/Deloitte/tr/Documents/technology-media-telecommunications/Blockchain-and-Cyber.pdf (Accessed: 9 March 2022)

[21] Fernández, Antonio, Chryssis Georgiou, Kishori M. Konwar and Nicolas C. Nicolaou. “Formalizing and Implementing Distributed Ledger Objects.” NETYS (2018).

[22] “Decentralisation: a multidisciplinary perspective” [Online]. Available: https://policyreview.info/concepts/decentralisation (Accessed: 11 March 2022)

[23] “e-Health Record” [Online]. Available: https://e-estonia.com/solutions/healthcare/e-health-records/ (Accessed: 15 March 2022)

[24] “Guardtime KSI Protecting Estonian Digital State” [Online]. Available: https://showroom.demos.guardtime.com/1-ksi-stack.html (Accessed: 15 March 2022)

[25] “WHO Digital COVID-19 VaccinationInfrastructure” [Online]. Available: https://guardtime.com/blog/who-digital-covid-19-vaccination-infrastructure (Accessed: 15 March 2022)

[26] Handbook of Blockchain, Digital Finance, and Inclusion, Volume 2 : ChinaTech, Mobile Security, and Distributed Ledger, edited by Kuo Chuen, David Lee, and Robert H. Deng, Elsevier Science & Technology, 2017. ProQuest Ebook Central, https://ebookcentral.proquest.com/lib/tuee/detail.action?docID=4939397. (Accessed: 17 March 2022)

[27] Dominique Guegan. Public Blockchain versus Private blockhain. 2017. ⟨halshs-01524440⟩ (Accessed: 17 March 2022)

[28] “Hyperledger Fabric Introduction” [Online]. Available: https://hyperledger-fabric.readthedocs.io/en/latest/whatis.html (Accessed: 18 March 2022)

[29] Manlu Liu, Kean Wu, Jennifer Jie Xu; How Will Blockchain Technology Impact Auditing and Accounting: Permissionless versus Permissioned Blockchain. Current Issues in Auditing 1 September 2019; 13 (2): A19–A29. doi: https://doi.org/10.2308/ciia-52540

[30] “Identity” [Online]. Available: https://hyperledger-fabric.readthedocs.io/en/latest/identity/identity.html (Accessed: 18 March 2022)

[31] “Membership Service Provider” [Online]. Available: https://hyperledger-fabric.readthedocs.io/en/latest/membership/membership.html (Accessed: 18 March 2022)

[32] Souppaya, M. , Haag, W. , Akram, M. , Barker, W. , Clatterbuck, R. , Everhart, B. , Johnson, B. , Kapasouris, A. , Lam, D. , Pleasant, B. , Raguso, M. , Symington, S. , Turner, P. , Wilson, C. and Dodson, D. (2020), Securing Web Transactions TLS Server Certificate Management, Special Publication (NIST SP), National Institute of Standards and Technology, Gaithersburg, MD, [online], https://doi.org/10.6028/NIST.SP.1800-16 (Accessed March 20, 2022)

[33] Regenscheid, A. , Feldman, L. and Witte, G. (2015), NIST Special Publication 800-88, Revision 1: Guidelines for Media Sanitization, ITL Bulletin, National Institute of Standards and Technology, Gaithersburg, MD, [online], https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=917935 (Accessed March 13, 2022)

[34] Elli Androulaki, Artem Barger, Vita Bortnikov, Christian Cachin, Konstantinos Christidis, Angelo De Caro, David Enyeart, Christopher Ferris, Gennady Laventman, Yacov Manevich, Srinivasan Muralidharan, Chet Murthy, Binh Nguyen, Manish Sethi, Gari Singh, Keith Smith, Alessandro Sorniotti, Chrysoula Stathakopoulou, Marko Vukolić, Sharon Weed Cocco, and Jason Yellick. 2018. Hyperledger fabric: a distributed operating system for permissioned blockchains. In Proceedings of the Thirteenth EuroSys Conference(EuroSys '18). Association for Computing Machinery, New York, NY, USA, Article 30, 1–15. DOI:https://doi.org/10.1145/3190508.3190538

[35] “Fabric Gateway” [Online]. Available: https://hyperledger-fabric.readthedocs.io/en/latest/gateway.html (Accessed: 26 March 2022)

[36] “Ordering Service” [Online]. Available: https://hyperledger-fabric.readthedocs.io/en/latest/orderer/ordering_service.html (Accessed: 30 March 2022)

[37] Hu, Junjie & Liu, Ke. (2020). Raft consensus mechanism and the applications. Journal of Physics: Conference Series. 1544. 012079. 10.1088/1742-6596/1544/1/012079. (Accessed: 31 March 2022)

[38] “Ledger” [Online]. Available: https://hyperledger-fabric.readthedocs.io/en/latest/ledger/ledger.html (Accessed: 4 Apirl 2022)

[39] “Private Data” [Online]. Available: https://hyperledger-fabric.readthedocs.io/en/latest/private-data/private-data.html#what-is-private-data (Accessed: 7 April 2022)

[40] “Architecture for Private Data” [Online]. Available: https://hyperledger-fabric.readthedocs.io/en/latest/private-data-arch.html (Accessed: 7 April 2022)

[41] Wang Q, Qin S. A Hyperledger Fabric-Based System Framework for Healthcare Data Management. Applied Sciences. 2021; 11(24):11693. https://doi.org/10.3390/app112411693

[42] “Using Private Data in Fabric” [Online]. Available: https://hyperledger-fabric.readthedocs.io/en/latest/private_data_tutorial.html (Accessed: 7 April 2022)

[43] “Peers” [Online]. Available: https://hyperledger-fabric.readthedocs.io/en/latest/peers/peers.html (Accessed: 11 Apirl)

[44] “Gossip data dissemination protocol” [Online]. Available: https://hyperledger-fabric.readthedocs.io/en/latest/gossip.html (Accessed: 7 April 2022)

[45] “CouchDB Introduction” [Online]. Available: https://docs.couchdb.org/en/stable/intro/index.html (Accessed: 13 April 2022)

[46] McKay, K. and Cooper, D. (2019), Guidelines for the Selection, Configuration, and Use of Transport Layer Security (TLS) Implementations, Special Publication (NIST SP), National Institute of Standards and Technology, Gaithersburg, MD, [online], https://doi.org/10.6028/NIST.SP.800-52r2 (Accessed March 13, 2022)

[47] S. Vijaya Kumar M.Sc., M.Phil, M.M.M. 2017. Applications of Big Data Analytics and Machine Learning Techniques in Health Care Sectors. International Journal of Engineering and Computer Science. 6, 7 (Jul. 2017).

[48] 45 CFR Part 164 Subpart C https://www.ecfr.gov/current/title-45/subtitle-A/subchapter-C/part-164/subpart-C (Accessed: 13 March 2022)

[49] “45 CFR § 164.514 - Other requirements relating to uses and disclosures of protected health information” [Online]. Available: https://www.law.cornell.edu/cfr/text/45/164.514#a (Accessed: 14 March 2022)

Appendix 1 – Non-exclusive licence for reproduction and publication of a graduation thesis[1]

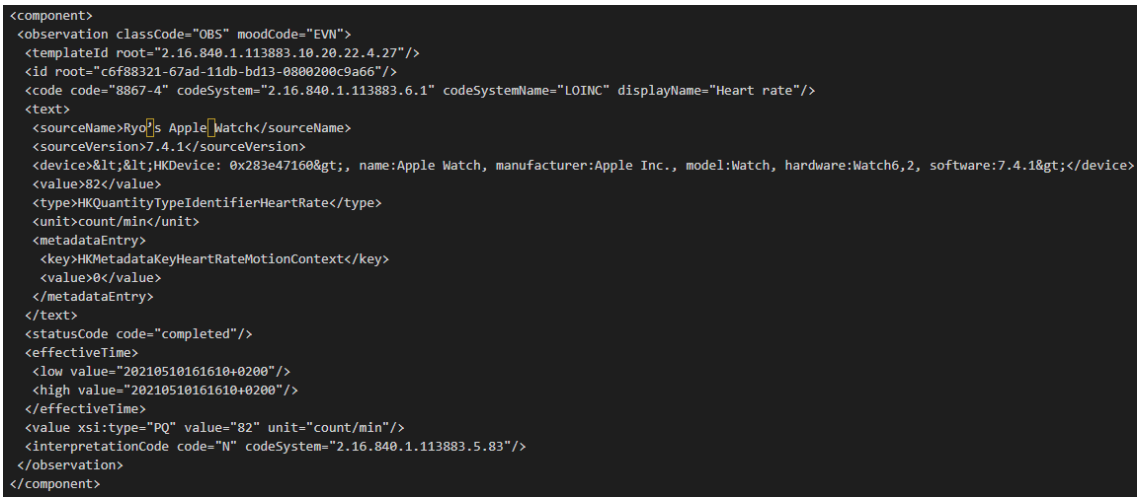
I Ryo Shiraishi

1. Grant Tallinn University of Technology free licence (non-exclusive licence) for my thesis “Improving Data Collection from Healthcare Applications”, supervised by Kaido Kikkas
   1.1. to be reproduced for the purposes of preservation and electronic publication of the graduation thesis, incl. to be entered in the digital collection of the library of Tallinn University of Technology until expiry of the term of copyright;
   1.2. to be published via the web of Tallinn University of Technology, incl. to be entered in the digital collection of the library of Tallinn University of Technology until expiry of the term of copyright.
2. I am aware that the author also retains the rights specified in clause 1 of the non-exclusive licence.
3. I confirm that granting the non-exclusive licence does not infringe other persons' intellectual property rights, the rights arising from the Personal Data Protection Act or rights arising from other legislation.

12.05.2022

1 The non-exclusive licence is not valid during the validity of access restriction indicated in the student's application for restriction on access to the graduation thesis that has been signed by the school's dean, except in case of the university's right to reproduce the thesis for preservation purposes only. If a graduation thesis is based on the joint creative activity of two or more persons and the co-author(s) has/have not granted, by the set deadline, the student defending his/her graduation thesis consent to reproduce and publish the graduation thesis in compliance with clauses 1.1 and 1.2 of the non-exclusive licence, the non-exclusive license shall not be valid for the period.

Appendix 2 – Sample data stored in database



```
<component>
 <observation classCode="OBS" moodCode="EVN">
  <templateId root="2.16.840.1.113883.10.20.22.4.27"/>
  <id root="c6f88321-67ad-11db-bd13-0800200c9a66"/>
  <code code="8867-4" codeSystem="2.16.840.1.113883.6.1" codeSystemName="LOINC" displayName="Heart rate"/>
  <text>
   <sourceName>Ryo's Apple Watch</sourceName>
   <sourceVersion>7.4.1</sourceVersion>
   <device>&lt;&lt;HKDevice: 0x283e47160&gt;, name:Apple Watch, manufacturer:Apple Inc., model:Watch, hardware:Watch6,2, software:7.4.1&gt;</device>
   <value>82</value>
   <type>HKQuantityTypeIdentifierHeartRate</type>
   <unit>count/min</unit>
   <metadataEntry>
    <key>HKMetadataKeyHeartRateMotionContext</key>
    <value>0</value>
   </metadataEntry>
  </text>
  <statusCode code="completed"/>
  <effectiveTime>
   <low value="20210510161610+0200"/>
   <high value="20210510161610+0200"/>
  </effectiveTime>
  <value xsi:type="PQ" value="82" unit="count/min"/>
  <interpretationCode code="N" codeSystem="2.16.840.1.113883.5.83"/>
 </observation>
</component>
```

Figure 4. data collected from Apple Watch Series 6 Watch OS 8.5

Non-continuous data, such as heart rate taken at a regular interval:

```
{
    "userId": "03414fdf-75d9-4de8-9dcc-b91cdb1fe5a5",
    "sourceName": "Health Application",
    "device": "name:Apple Watch, manufacturer:Apple Inc., model:Watch,
hardware:Watch6,2, operatingSystem:7.4;",
    "applicationVersion": "1.0.5",
    "value": [
        {
            "rate": 82
        }
    ],
    "type": "HeartRate",
    "unit": "beat/minute",
    "time": "2022-03-24 15:30:16 +0200"
}
```

Continuous data, such as data collection from running and walking activities:


```
{
    "userId": "03414fdf-75d9-4de8-9dcc-b91cdb1fe5a5",
    "sourceName": "Running/Walking Application",
    "device": "name:Apple Watch, manufacturer:Apple Inc., model:Watch,
hardware:Watch6,2, operatingSystem:7.4;",
    "applicationVersion": "2.0.8",
    "value": [
        {
            "ele": 25.668412022-03-243,
            "time": "T05:29:11Z",
            "speed": 1.709691,
            "lat": 59.415981,
            "long": 24.722648,
            "HR": 79
        },
        {
            "ele": 26.485037,
            "time": "2022-03-24T05:29:40Z",
            "speed": 1.186919,
            "lat": 59.416055,
            "long": 24.722751,
            "HR": 82
          }
    ],
    "type": "Walk",
    "unit": "",
    "time": "2022-03-24 17:50:35 +0200"
}
```

Appendix 3 – Private Data Collection Definition

```
[
 {
    "name": "collectionPrivateOrgAll",
    "policy": "OR('Org1MSP.member', 'Org2MSP.member')",
    "requiredPeerCount": 2,
    "maxPeerCount": 4,
    "blockToLive":20000,
    "memberOnlyRead": true,
    "memberOnlyWrite": true,
    "endorsementPolicy": {
        "signaturePolicy": "OR('Org1MSP.member', 'Org2MSP.member')"
    }
 },
 {
    "name": "collectionPrivateOrgOne",
    "policy": "OR('Org1MSP.member')",
    "requiredPeerCount": 2,
    "maxPeerCount": 2,
    "blockToLive":20000,
    "memberOnlyRead": true,
    "memberOnlyWrite": true,
    "endorsementPolicy": {
        "signaturePolicy": "OR('Org1MSP.member')"
    }
 },
 {
    "name": "collectionPrivateOrgTwo",
    "policy": "OR('Org2MSP.member')",
    "requiredPeerCount": 2,
    "maxPeerCount": 2,
    "blockToLive":20000,
    "memberOnlyRead": true,
    "memberOnlyWrite":true,
    "endorsementPolicy": {
      "signaturePolicy": "OR('Org2MSP.member')"
    }
 }
]
```