Blockchain GDPR Data Compliance through Container Utilisation

Can Blockchain be used for secure data storage after the implementation of GDPR? Can this be achieved using data containerisation?

Mitch J. Durso

A thesis submitted in fulfilment of the requirements of the University of Lincoln for the degree of MSc by Research.

College of Sciences

School of Computer Science

UNIVERSITY OF LINCOLN

March 2021

Acknowledgements

I wish to show my appreciation to those closest to me both in a professional manner and personal. Without the care, help and love of my family and my partner through the late nights and long drives, the many hours spent on the road to see family around everything. Also, a thank you to Morgan H. for the time spent helping me with the wording when I needed to bounce ideas off someone. Along with providing the second set of eyes for reviewing my work. I would not have been able to put in the time and dedication that this research needed. Even with my desire to explore the topic and subject sometimes you cannot go it alone.

Thank you.

Definitions

1.0 Containerisation: Containerisation, when discussed throughout this research, is the process in which data, in any shape or form, is handled and stored within its own designated or identified area, complimented with respective identification logging. Similar to the act of physical objects being placed into a container or box, and identifiably tagged.

Abstract

This research presents an investigation of the utilization of data containerization on a Blockchain network, as this will allow for the compliance of GDPR. Blockchain is a form of technology that is cryptographically protected and is therefore an immutable system by design. This is achieved through cryptographical calculation of entered data from which a hash is generated this can then be used like a signature as the same content has to be present to gain the same hash again. A Hash is a long string of letters and numbers. This hash is then stored inside a 'block' of information within the Blockchain's ledger. These blocks are all linked together through the storing of a previous block's hash in the newest entry, this is to ensure that the blocks of information can be cryptographically checked and guarantees that the data has not been tampered with or modified. Blockchain's flaw lies in the fact that any inputted data cannot be removed, otherwise, the chain is broken. This is a clear breach of GDPR if user data is stored inside it; GDPR is a regulation that affects the handling, collection and management of user data and information. The specific article which affects Blockchain in relation to GDPR is Article 17, which states all members of the EU, and UK (United Kingdom) as the law was ratified into UK law, have the right to erasure and to be forgotten. This causes issues for Blockchain, as data cannot be modified or removed from an existing Blockchain network, without invalidating the Blockchain. This is due to Blockchain's design, as an incorrect entry, or more specifically, a cryptographic hash, on the ledger invalidates the Blockchain. Containerisation solves this issue by storing data that needs to be removable inside of a separate

storage method. For the purposes of this research, these files are being stored on the disk, metaphorical to being stored on a docker system or FTP server. Docker is an open-source software, which allows the user to pack, provision and run virtualized application containers on an operating system. It contains dependencies needed to execute code within containers, allowing containers to move between a docker environment and OS. It uses resource isolation in the OS kernel, allowing the common operating system to be contained and run multiple times. It runs containers of docker images, which contain the dependencies needed to execute within a container but should not be mistaken for a Virtual Machine (VM), as a VM operates differently, encapsulating an entire OS, and being run through dedicated hardware resources on the machine.

An FTP server is used to facilitate file transfers over the internet, FTP standing for File Transfer Protocol. Files are either uploaded to or downloaded from an FTP server. In context of this paper, containerisation allows for the data to be stored outside of the ledger, substituting the original data with container or file identifiers on the Blockchain. This allows the data to be deleted, without affecting the Blockchain's cryptographic chain. The containerised data is protected by utilizing RSA public key encryption. This is to ensure data off the Blockchain, cannot be modified without destroying the original information or data. The encrypted data is also split up into chunks, to ensure that the data has an even lower chance of being decrypted or lost, as the Blockchain network, which is private, holds the reconstructive information for these chunks. These 3 systems, Encryption & Chunking, Blockchain Ledger and Storage Method, together allow for a Blockchain system that can continue to be used as an immutable storage method, whilst allowing the extraction of data as per a user's request, in compliance with GDPR, specifically, Article 17.

Contents

1.0 Introduction

1.1 Research Question

The academic research questions I will be investigating is: “Can Blockchain be used for secure data storage after the implementation of GDPR? Can this be achieved using data containerisation?” This will consist of 2 broad focal points of research, GDPR and Blockchain Technology. Respectively, this research therefore must answer 2 main questions; “Can Blockchain be used for secure data storage after the implementation of GDPR” and “Can this be achieved using data containerisation?” These two questions must be answered by extensive research and development to achieve new blockchain architecture that is compliant with GDPR.

1.2 Research Hypotheses

It is my hypothesis that this research will result in a positive outcome to both research questions, due to the flexibility of technology and it’s continued shifting nature; I believe it will be possible to find a workaround to the limitations imposed by both Blockchain and GDPR, achieved through the proposed method of containerisation (Definitions; 1.0).

1.3 Research Aims and Objectives

It is my aim throughout this research to develop a method of Blockchain data storage whilst complying with GDPR, achieved by utilizing data containerization. I also wish to further develop my knowledge in the fields of data storage, GDPR and Blockchain Technology. It would be my hope that this research can also begin a line of questioning towards current policies, practices that help lay the foundation for further research into similar fields and policies.

1.4 Research Outline

Due to beginning of GDPR's public development in early 2016, the internet and media were heavily discussing and reporting on how GDPR might change the public's use of the internet and the ways in which both physical, and digital, information was handled, permanently. Using the internet to host many of my own projects and communities, I wanted to further investigate how GDPR would affect my work online, as well as the wider implications it may have on businesses and end-users alike.

Throughout this research, it is proposed to create a method of Blockchain storage with higher levels of security, and more importantly, accountability, than current SQL based databases. This must also comply with GDPR's data erasure laws; by utilizing Blockchain's functionality of chaining inputted data together, we can create a database that has traceability like no other, however, this method has obvious concerns due to its chaining system; it is not possible to delete user data within current and existing Blockchain architectures, due to its ledger system. This will be further explained in Section 2.3, "Blockchain Background". By using data containerization, it is hoped that a database-like system can be developed that allows user data to be extracted and deleted, whilst maintaining an elevated level of security, accountability, and reliability.

The research will utilize an "append-only structure" at its core nature to keep the notion of security and ensure immutability. This will ensure that the Blockchain system remains efficient and upholds what makes it ideal for data storage: it's security. The artefact will achieve this by using a containerised storage medium, in which data can be removed from, without damaging the Blockchain's integral chain or ledger. This ensures that newer additions to the ledger, and the original data, cannot be damaged or cause the ledger to need to be rebuilt.

There are 5 main factors to consider when creating this artefact so that it complies with GDPR. These 5 factors are:

- Server Handling and Data Storage Method
- Data Encryption
- File Separation and Upload and Download/Deletion Requests
- UI and Application
- Blockchain System

These 5 factors will be discussed during Section 3.0, “Literature Review,” woven and considered throughout multiple contrasting, and echoing, pieces of literature. These 5 factors will then be further developed throughout Section 5.0, “Design, Development and Evaluation.” These factors will be heavily considered during the design process and will be implemented during the development phase of the artefact. This will ensure a GDPR compliant artefact which successfully addresses the outlined research question. Accomplishment of these 5 factors will be then discussed within Section 6.0, “Results.” These 5 main factors will also undergo minor title alterations within Section 5.1, “Design Process” to adapt from research titles to development requirements.

2.0 Research & Artefact Background

2.1 GDPR Background

General Data Protection Regulation (GDPR) (European Parliament and of the Council, 2016) is a regulation in UK and European law that further addresses data protection and privacy. The law is designed to protect data subjects’ rights and information.

Personal data is any data relating to a data subject or owner. The European Commission states: “personal data means any information relating to an identified or identifiable natural person (‘data subject’); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person.” (European Parliament and of the Council, 2016)

This regulation was implemented to reform the European Union’s (EU) outdated 1995 Data Protection Law (European Parliament and of the Council, 1995), in the hopes of strengthening online privacy and data collection overall. It was obvious in recent times that the development of technology, as well as the online space, that previous legislation was no longer comprehensive enough to cover the necessary ground, limitations and factors for data collection and online privacy.

The regulation was implemented, as before GDPR, data holders had less obligation to disclose and give users their data, upon the data subject’s request (European Parliament and of the Council, 2016), as well as less regulation on how user data should be handled and collected. This left possibilities for companies to exploit data subject’s information and data without their consent, including selling, sharing, and directly using their data for directed advertising.

The General Data Protection Regulation (European Parliament and of the Council, 2016) has shaped the collection of data across the globe and is affecting all member states of the European Union (EU), as well many businesses outside of it too. Originally proposed in 2012, GDPR legislation was put into effect on the 25th of May 2018 (European Parliament and of the Council, 2016) and as an agreed part of Brexit, GDPR was ratified into UK law as the Data Protection Act 2018 (United Kingdom Government, 2018).

The collection of user data has shifted into a much more secure, monitored and user-orientated environment, due to both the implementation of these regulations as well as the ways that user data can now be collected; GDPR has reshaped industry over the last 2 – 3 years as compliance with this regulation is paramount, and is a legal requirement for any company that is based within the European Union; or is based outside of the European Union but wishes to trade with businesses, or cater for citizens, who are within the European Union. The organisations outside the EU that wish to interact with organisations or end-users within the EU, must abide by these laws and regulations to ensure the fair use and collection of European user's data.

The penalties for those who do not comply with GDPR can be substantial fines or imprisonment; these fines can be based upon a fixed maximum amount; however, it is at the discretion of the court. (European Parliament and of the Council, 2016). Fines can also be based on company earning percentages, either a maximum fine amount of €10,000,000 or 2% of a company's total turnover from their previous financial year, whichever is higher (Information Commissioner's Office, 2020). This is to ensure that repercussions are equal to company size. If the infringement is big enough, imprisonment can be taken into consideration for those in management and control over user data, however, this is in extreme circumstances.

Certain articles within the GDPR legislation have implemented new factors that never used to be considered necessary. One example of this would be Article 17 (European Parliament and of the Council, 2016), which states that everyone within the European Union has the right for all their data to be erased and for the data subject to be forgotten upon request. However, this article raises questions about the application of GDPR (European Parliament and of the Council, 2016). For example, the legislation could be 'bent' or interpreted to manipulate what is considered to be user data.

The legislation (European Parliament and of the Council, 2016) also states that in certain circumstances you are required to tell other organisations about the erasure of a user's personal data, either by request or 'digital clean-up', though it does not strictly specify that these other organisations must also carry out the erasure on their end. This too is an obvious point of interpretation within the legislation and highlights the ambiguity that makes GDPR so difficult to comply with.

On the other hand, this lack of absolute certainty gives plenty of room for companies to bend the legislation to their wants, needs or desires to support them, and their intentions, rather than the data subject, which is what could be argued the legislation (European Parliament and of the Council, 2016) was created for - the protection of an entity's data, and the rightful control over it.

2.2 The Data Problem

"Today, data is a valuable asset in our economy" (Zyskind et al., 2015) as stated by Guy Zyskind, Oz Nathan, and Alex 'Sandy' Pentland; researchers in Blockchain's usage to protect Personal Data.

Zyskind et al. point out the fact that applications are constantly collecting personal data of which the user has no specific knowledge or control, assuming most services are: "honest-but-curious." (Zyskind et al., 2015) Meaning the data collected is within regulation but the service owner's collection may not be as honest with the user as the end-user would prefer.

This idea of applications storing, collecting, and using user data for their business' or corporation's benefit is not an unknown fact. Data holders and collectors often have more information about you, than even your own government, proven by Apple, the FBI, and Dr Unal Tatar. (Tatar et al., 2020)

Tatar et al. conducted research that mentioned a lawsuit involving Apple and the FBI, relating to user data, and its sensitivity. (Tatar et al., 2020) In this case, the FBI wanted access to a citizen's Apple device, which belonged to Syed Farook, who was responsible for the shootings in San Bernardino in December, which left 14 people dead. They hoped that the data could give them insight into the events leading up to the shooting since they considered the incident to be a terrorist attack. The court granted permission for the device to be opened, and subsequently, Apple could charge the FBI for this service (Boutrous et al., 2016) (Wilkinson and Allen Chiu, 2016).

These investigations represent the complexity of the society most of us live in today. Our society allows your mobile phone company to collect such valuable information on you, that even the Federal Bureau of Investigation is willing to bring the said company to court. Resulting in the FBI being charged by an independent company just to gain access to a phone, that is already in the FBI's possession.

It should be noted that during and before these investigations, it was highly expected for companies to include a back-door entry system for various legal uses. This system gave access to data within the system without standard direct access. This back-door access could be granted under fair legal reasoning, however, was not mandatory; due to this, it can be assured that a system such as a Blockchain network can legally comply with data access regulations, as this backdoor is not a legally enforceable expectation.

The possibilities that a Blockchain system could provide are both positive and negative. Extremely tight security comes with risks; information cannot be monitored by the government or data holder, and what is shared or stored could be illegal content. Although that level of privacy is daunting, the lack of accessibility for data holders is a potentially good thing for GDPR compliance, as sensitive information cannot easily be shared or leaked to those who are not supposed to have it. The research done by (Tatar et al., 2020) represents

the complexity of GDPR compliance and outlines what this research focuses on; creating a data storage solution, whilst addressing the inconsistency between the way the law is structured, and how technologies operate.

2.3 Blockchain Background

The secondary focal point, and main technology used throughout this research, is Blockchain. Blockchain can be used in a multitude of ways, its most common being used in cryptocurrency. Cryptocurrency is a form of virtual currency that is recorded, using Blockchain. Each transaction that cryptocurrency is used for is entered and recorded onto a defined type of Blockchain dependant on said cryptocurrency's architecture.

However, cryptocurrency is not Blockchain's only potential application. Blockchain systems are admired for their decentralisation and traceability due to do their architecture and have been used when compiling sales data, tracking digital payments to content creators, and recording online political votes, as well as growing government and private usage for health records and personal information.

Blockchain allows for the secure recording of information, without the ability to modify the data within it; when creating data networks that must be carefully administrated, Blockchain is an extremely secure option due to its cryptographic nature and has a broad variety of potential application, as discussed prior.

It is important to understand the basics of Blockchain and its potential application, to fully understand the importance this technology holds throughout the current industry.

"Blockchain is a shared, immutable ledger for recording the history of transactions. It fosters a new generation of transactional applications that help establish accountability and transparency. Blockchain provides an unmatched level of accountability for how data is

managed based on its tamper-resistant data store and its consensus mechanism used to modify the data. Basically, Blockchain data is protected by design." (Compert et al., 2018)

Blockchain is a log of records that are set up in such a way that they are immutable because it stores a hash, for example, "e826bea354b01e9ec5c5333..." of the entered data or record within the next entry. This means that when a previous entry is modified or deleted, it will no longer match the stored hash stored within the next block on the Blockchain, and therefore the record is no longer accurate. See figure 1.

It is purposely designed like this as the blocks of data within the chain, each have their hash calculated and then stored within the next block for optimal security. As well as transparency as it is not possible to modify existing data that has been entered into a Blockchain system, without completely invalidating it. If any data within the previous block was changed or deleted, the chain itself becomes incorrect from this point onwards and would no longer be cryptographically correct as the hashes cannot be verified backwards or forwards. See Figure 1.

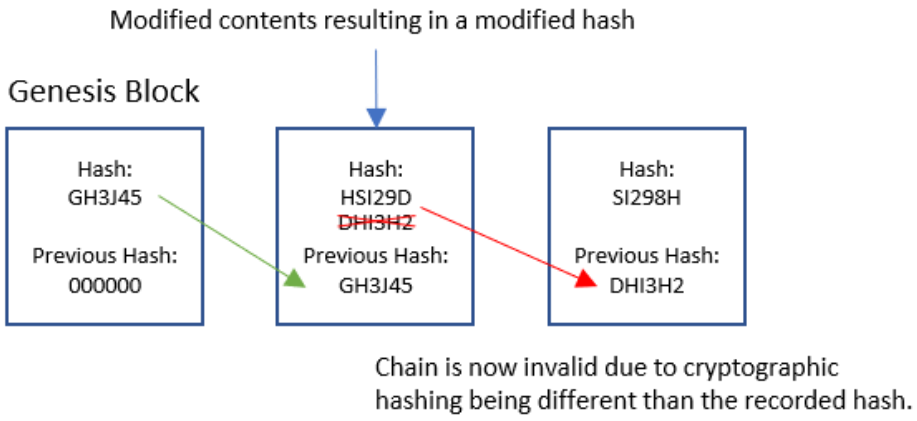

Figure 1 – How a chain is invalidated upon the contents being modified.

This means the Blockchain cannot be edited or continued without permanently invalidating and breaking the chain due to its cryptographic nature. In this case, the chain would need to

be rebuilt to continue accurately storing information from the point in which it was originally invalidated.

In short, Blockchain uses a block chaining system to ensure data cannot be edited. It uses interlocking hashes and blocks to verify each block's data, and if the data within any block is edited, the entire chain beyond that point becomes invalidated.

A shopping receipt can be used as a form of example when explaining Blockchain's concept and method of storing information: When you shop at the supermarket, each item on a receipt can be taken as a metaphor for an entry on a Blockchain's ledger, and relates to a specific item you have purchased, and put into your final cart. Items can be removed from the final cart, but they cannot be removed from an already printed receipt. In a metaphorical Blockchain environment, two carts must be considered. The first is an initial cart of items that are yet to be printed to the receipt and the second, a cart of items that have already been printed to the receipt and purchased. If an item was moved from the initial cart to the final cart, also considered as purchasing the item, and subsequently printed to the receipt, that would be considered as a standard transaction. If an item was removed from the final cart, and another item was purchased, and added to the receipt afterwards, it would be the equivalent of invalidating the Blockchain's Ledger, as the final cart's items would no longer tally with the receipt, as a purchased item had been discarded or removed. This is only a metaphorical example of Blockchain and does not take into consideration the linking procedure of Blockchain, however, this gives a basic explanation to the principle in which Blockchain operates. All items that have been printed to the receipt must be in the final cart, or the Blockchain is invalidated. In this example, your data or files are represented by the items in your carts, and the receipt represents the Blockchain's ledger.

Typical storage methods that allow data to be removed or modified, for example MySQL, MongoDB, SQLite, have no chaining system, this means that the data is potentially

vulnerable if not secured appropriately, as it can be easily edited or removed. These non-chaining systems work well for systems that store records such as stock quantities but has potential security flaws if storing personal information, as it can be manipulated without traceability or accountability.

Depending on the requirements and individual setup, a Blockchain can be a centralised system, which allows private access only, a decentralized system, which allows public access only, or a distributed system, which can allow either public access or private access. A centralised system is where participants are given specific rights to access or management depending on their granted permission. This centralized system connects all parties to one central node, to which their access can be decided by the Blockchain's controlling entity. This type of system can also be known as a permissioned Blockchain. A decentralized system differs from this, as its access and management requests are forwarded to one of the many main nodes on the network. These main nodes sync with coexisting nodes but do not always synchronise system wide. This system type can be difficult to maintain due to incorrect data syncing and overwriting. A distributed network follows a similar architecture to the decentralized system, however, does not use any type of main node. This means all nodes within the system relate to no central node. This also means that all computational resources are split evenly throughout the network. This results in higher levels of data syncing, with less chance of incorrect data overwriting, whilst maintaining equal rights across the network.

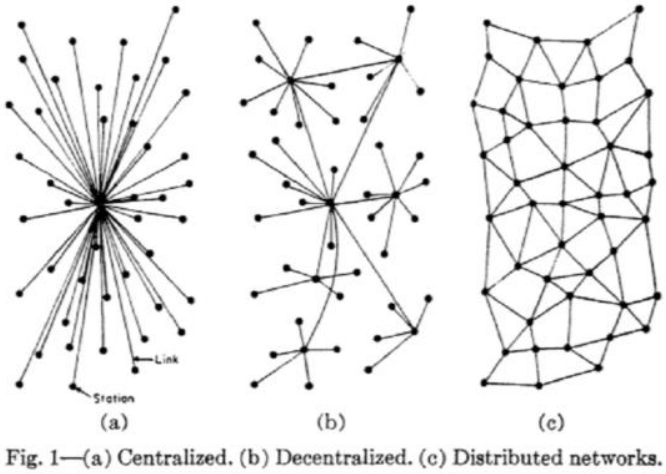

Fig. 1—(a) Centralized. (b) Decentralized. (c) Distributed networks.

Figure 2 – Network Types (a. Centralized, b. Decentralized, c. Distributed networks.) (Baran, 1964)

A Blockchain system could be implemented to maintain control over a service's data storage or for other reasons, such as removing the anonymity of the network participants to ensure accountability. A private system is best used for securing confidential data, because the stored data should not be accessible to outside sources, except those who have been granted the correct access, and right, to the stored information or data, such as the data owner or manager. A public system is publicly viewable to anyone on or off the network, however, these three Blockchain architectures can be combined and modified to create hybrid Blockchain architectures. A hybrid blockchain is a combination of both public and private entities. These entities are managed by using a public blockchain in correlation to a private network. Hybrid blockchain architecture is not open to everyone on the network, however, still provides desired Blockchain characteristics such as integrity, transparency, and security. Another example of this is a Consortium Blockchain architecture, also known as a Federated Blockchain. A Consortium Blockchain consists of multiple organizations governing the platform and is based upon a permissioned architecture. A Consortium Blockchain is built upon a centralized architecture, however, allows for multiple organizations to make decisions within the platform.

3.0 Literature Review

3.1 Introduction

This research has two main parts for consideration, GDPR (European Parliament and of the Council, 2016) and Blockchain technology. It is important to understand both Blockchain and GDPR (European Parliament and of the Council, 2016) are individual elements but also to consider the relationship between them.

Specific articles, such as Article 17 in GDPR, have created issues for Blockchain when relating to the storage of personal information. (European Parliament and of the Council, 2016). Article 17 of GDPR addresses the deletion process for personal information, something that Blockchain was not designed to facilitate due to its architecture. This was not an issue when Blockchain technology was originally conceptualised and developed due to differing legislation at the time, (European Parliament and of the Council, 1995) however, it is now one of its largest issues when attempting to use the technology as a form of storage for personal information, or personal data. This is a cause of great concern for Blockchain and its future, due to the introduction of the new regulations implemented and passed as law in 2018 (European Commission, 2018). As a result of these new rights, Blockchain technology is in a tricky situation, where the technology itself needs to be modified to comply with the newly introduced regulations.

When modifying Blockchain's architecture, many characteristics need to be considered when attempting to create a system that can both facilitate GDPR whilst maintaining Blockchain's security benefits. These characteristics will be investigated and developed throughout the Analysis of Existing Literature.

3.2 Analysis of Existing Literature

In 2018, IBM published an article (Compert et al., 2018) outlining issues relating to GDPR and Blockchain. The paper discussed the changes which GDPR would bring, as well as what it meant for those outside the EU. The paper discussed the accountability and transparency that Blockchain can provide in a GDPR environment and the benefits, as well as disadvantages it could bring.

The paper (Compert et al., 2018) states the rights of EU data subjects concerning Article 17. This mostly consists of the broad idea of GDPR, and what it means for both the user and data holder, as well as the difficulty that "Article 17: Right to erasure ('right to be forgotten')" (European Parliament and of the Council, 2016) could cause for Blockchain due to Blockchain's nature.

IBM's research (Compert et al., 2018) highlights the efforts being made by companies that are collaborating and trying to implement Blockchain as their method of data storage. Their main examples of these were in the areas of "food safety" and "global trade" (Compert et al., 2018). It should be noted that both examples, provided by IBM, were using IBM based Blockchain software (Compert et al., 2018).

IBM highlights the advantages that Blockchain offers relating to companies trying to comply with GDPR (Compert et al., 2018), whilst holding delicate user data or information. This includes the potential for easier access for data consumers if so desired and if correct permission is granted. The example they use for this is the "Estonian eHealth Foundation" describing their implementation of Blockchain as a revolution to the healthcare system (Compert et al., 2018). This reference was from a completely unbiased position as the Estonian eHealth Foundation use a Keyless Signature Infrastructure Blockchain system (Reede, 2020) designed in Estonia without IBM's input. The research IBM proposes (Compert et al., 2018) shows the heightened levels of accountability that a Blockchain system can

provide and the increased compliance of GDPR due to Blockchain's nature; however, IBM does not introduce any new form of system or direct advice as to how Blockchain should be adjusted to fit situations that Blockchain is not GDPR compliant with. IBM failed to provide any new Blockchain architecture as they intended to simply show how the current form of Blockchain can be used to comply with GDPR. Chowdhury et al. fill this gap by proposing altered Blockchain architecture, which will be discussed in the following paragraphs, (Chowdhury et al., 2018), whilst discussing data storage, management and shareability under the new GDPR legislation.

Chowdhury et al.'s discussion brings light to the recent changes concerning data subjects and consumer relationships, and just how much information is now available to those who have the desire to see their personal data. Chowdhury et al. highlight the importance of correct data management (Chowdhury et al., 2018) as before GDPR 's implementation, data holders had less obligation to display and give users their information on request (European Parliament and of the Council, 1995), meaning large scale changes must be made across many industries; services that hold personal information now must comply with these newer and stricter regulations.

To combat these issues relating to data management, Chowdhury et al. suggest methods of storage that have been newly implemented by different industries (Chowdhury et al., 2018), however, with modifications to implement GDPR compliance when using Blockchain's technology. Specifically, the authors suggest cloud-based storage solutions for encrypted personal data. The system which Chowdhury et al. suggests is based upon a Blockchain system that can be integrated alongside a Personal Data Storage (PDS) system (Chowdhury et al., 2018). (An example of a PDS system would be Dropbox). Zyskind et al. also discuss the protection of personal data by utilizing Blockchain, however, Zyskind et al.'s research (Zyskind et al., 2015) highlights a similar perspective to Chowdhury et al. (Chowdhury et al.,

2018) on the usability of Blockchain; suggesting Blockchain can be used as a “Notarization Service for Data Sharing with Personal Data Store.” (Zyskind et al., 2015)

Chowdhury et al. discuss how Personal Data Storage (PDS) is an extension of the Personal Health Records system, but that a PDS based system allows for more versatility (Chowdhury et al., 2018). The benefit of this system is the ability for an external service, such as a hospital or university, to access these files on a permission-based system, and that users would be allowed to selectively share sets of data with other users. These benefits of accessibility are why Zyskind et al. also adopted a PDS based system (Zyskind et al., 2015), however, there is no mention of Article 17 (European Parliament and of the Council, 2016), or the deletion of personal data and these questions seem to have gone unanswered.

A PDS system is a service that allows an individual to manage and store their data in a traditionally secured manner. A PDS system gives a user a central point of control, and data can be sorted in multiple external repositories.

Both Chowdhury et al. and Zyskind et al. point out the growing concern that is: user privacy, as well as the lack of control users have over their own data (Chowdhury et al., 2018,) (Zyskind et al., 2015). Zyskind et al. point out that standard, non-Blockchain off-chain storage systems are no longer reliable (Zyskind et al., 2015). This is due to too much transparency and easy accessibility for companies to retrieve information from data centres, which is echoed by IBM (Compert et al., 2018) when they discussed the modern use of Blockchain technology in non-cryptocurrency environments.

To combat the issues surrounding GDPR and current off-chain storage systems, both Chowdhury et al. and Zyskind et al. suggest a multi-identifier system, consisting of assigning IDs to the data-subject, data-custodian, and data resource (Chowdhury et al., 2018) (Zyskind et al., 2015). Chowdhury et al. suggest generating a copy of a document in question,

generating a hash of the data-subject ID, data-custodian ID, and data resource, and then using those IDs to create an interlocking Blockchain (Chowdhury et al., 2018). These three separate entities, all involved in data management, ownership or handling, combine to work as an "ecosystem".(Zyskind et al., 2015)

These ecosystems are based upon a multi-transaction service, consisting of two types. The first of these being an access transaction, which would be used to access the Blockchain, for access control management. The second type of transaction would be used to access the Blockchain's data storage, where the user data could be retrieved. It is intended that the 'access control management' transaction would be operated by the user, and the 'data storage' transaction would be used by the service requesting the user data.

Chowdhury et al. and Zyskind et al. develop extremely similar Blockchain architectures, using transactional-based services to gain control management or data retrieval, and giving the end-user the ability to share these data retrieval transactions (Chowdhury et al., 2018,) (Zyskind et al., 2015). This would mean the user could change the permission granted to a service at any time by using an access control management transaction. This work does not specify how the user would be able to delete their data, even after the service no longer had access to the stored information. This would allow the service to comply with GDPR, however, does not outline how the Blockchain storage provider complies with GDPR. If this data was to be shared, questions about its storage location go somewhat unanswered. The only information provided is that the data subject's information would be stored on a PDS based system, however, who controls this system is not specified by Chowdhury et al. or Zyskind et al. and leave questions relating to the compliance of Article 17: Right to Erasure (European Parliament and of the Council, 2016).

M.Alessi, A.Camillò, E.Giangreco, M.Matera, S.Pino, and D.Storelli also released a related publication (Alessi et al., 2019), attempting to create a PDS system that works alongside

Blockchain; however, this system is designed and based on Ethereum (Alessi et al., 2019). They too adopted a system of transactional access where their system was split into blocks, both private and public. Their system was somewhat modified to suit an Ethereum based Blockchain, however, their fundamental principles agree with both Chowdhury et al. and Zyskind et al. in the management of service access and transactional relationships (Chowdhury et al., 2018) (Zyskind et al., 2015).

The nature of this proposed solution means the end-user would have the ability to request an organisation to stop using their data at the user's discretion; this system, however, relies on the user understanding how to give or remove said access to their data, and more importantly, knowing that they have the right to do so. This could cause potential issues when trying to implement this method of storage into daily applications, due to the assumption that the end user's GDPR knowledge is of a relevant level, and that they would have the understanding that this form of data storage was being implemented and used. Alterations could be made to this system by changing the purpose of the data identifiers used by the file management and storing system. A system could be adapted to link individual files to the Blockchain's ledger by file identification, facilitating an Off-Chain Data Storage system; the main opposition to this Off-Chain data system relates to the potential security risks of using an Off-Chain (non-Blockchain) Data Storage medium. A potential solution to this issue is data encryption, which Chowdhury et al. mentioned as a form of data protection for this suggested Blockchain Architecture (Chowdhury et al., 2018).

Chowdhury et al.'s system also aims to address verification issues when a user is trying to access their own information. The system Chowdhury proposes is based on IDs for each piece of the data relationship; "a hash of the data-subject id, data-custodian id and data resource" (Chowdhury et al., 2018) to allow a user input for each id relating to the data retrieval process. This causes a lot more difficulty for the end-user, as they must write down

or remember their set of IDs; a potentially challenging task if sharing substantial amounts of documents and resources relating to the data subject's confidential information.

This system also leaves more potential security risks when trying to authenticate and verify the user. Chowdhury et al. suggest using a symmetric encryption method to solve this issue (Chowdhury et al., 2018). This type of encryption uses a single key, which would not be publicly viewable, to both encrypt and decrypt the stored files, however, Chowdhury et al. do not specify if one private key would be used system-wide, or if each user would be defined their own key, however, it is implied that it would be one key per user, as they state, "Then it is encrypted by the public key of the data subject." (Chowdhury et al., 2018)

If this system intended to apply one key per data subject, user authentication would have to be implemented in the traditional sense, to verify the identity of the data accessor, and retrieve their singular private key. Chowdhury et al. fail to provide the system of User Authentication, and therefore do not acknowledge a potential flaw in their security; standard user authentication or login would consist of a username or email and a password; however, this is potentially insecure method of data protection, and as Chowdhury et al. fail to provide their own system of user authentication, it is assumed that this method would be implemented as it is the most common throughout the industry.

The main reason that a password method of authentication is so vulnerable is due to user error: many people find it too difficult to remember complex passwords, or they do not see the value in trying to remember a complex password and decide to use an easy to remember password; smaller, less complex passwords, consisting of less special characters, numbers, or capitals. It is one of the main reasons the most common password on the entire internet is: "123456", proven to be cracked in less than 1 second by a computer, according to NordPass. (NordPass, 2020). Logically this suggests there must be another, more secure

method of data protection, as it has been proven password creation can be often predicted and a cause of great security concern when securing important user information or data.

A potential alternative method of encryption is a Public Key Cryptography system, as per Dr Lee's suggestion (Lee, 2017) when developing their Blockchain-based storage system. Dr Lee suggests that a public-key cryptography method can be used to aid identity and security within a Blockchain system, and that the public ledger serves to record these public identifiers (Lee, 2017). A public-key cryptography system consists of two keys, a public encrypting key, and a private decrypting key. The public key is used to encrypt the data, and the private key is used to decrypt the data. This is also known as asymmetric encryption, as the two keys, both public and private, are not the same, or 'symmetrical'. The study introduces the concept of mutual authentication in which the two keys are compared and is one of the most secure methods of encryption that could be found when researching the topic. This method of encryption could potentially solve the issue that Chowdhury et al. fail to provide the answers for, as the user requesting access to the data stored on the off-chain storage system must have access to, and provide, a one-of-a-kind private key, that was linked to the original encryption that the public key provided. Comparing these two keys allows for the encryption and decryption of data using a singular method, without the use of potentially hackable passwords.

The main questions around this system arise when considering the real-world application; how exactly would this form of data encryption be implemented in a sustainable method? There are issues around what happens if a user loses their key, or if the key was stolen. Dr Lee fails to address these questions, and there is an apparent lack of a key-recovery system that upholds the desired security with this application. A potential solution could be Biometrics, as this would allow the user to gain access to their keys through a naturally

secure method, such as fingerprint or facial recognition. Biometrics, however, may also have identity theft implications, or sudden identity change implications.

This method of Biometrics could also be adapted by Chowdhury et al. in their use of symmetric encryption (Chowdhury et al., 2018); however, asymmetric is more secure than the symmetric by design, and Dr Lee's method of Public Key Cryptography encryption is more applicable (Lee, 2017) than the symmetric encryption suggested by Chowdhury et al (Chowdhury et al., 2018).

Chowdhury et al.'s secondary proposed method is a Consortium based Blockchain (Chowdhury et al., 2018), which was not addressed or commented upon by Zyskind et al. as they limited their research to only one proposed solution (Zyskind et al., 2015). This Consortium Blockchain introduces a different styled PDS based system that would use a style of Blockchain called "permissioned Blockchain" (Chowdhury et al., 2018). This type of Blockchain is designed for access control, allowing certain actions to be performed on the Blockchain by assigned individuals. This type of Blockchain system would solve the issue of a user having to verify a transaction, as their assigned ID or account would give them access to the system, if the permissions are assigned, whilst keeping their identity hidden if so desired. Using this type of Blockchain system would also allow traceability and accountability, to ensure staff and data managers were not mismanaging user data or information, but this system also causes multiple concerns relating to GDPR (European Parliament and of the Council, 2016).

The first concern is the complications relating to the deletion of data that is directly on a Consortium Blockchain; this is where a containerized system: the storage of data off-chain from the ledger, used in conjunction with Blockchain would solve the complications surrounding GDPR (European Parliament and of the Council, 2016) and the rights to erasure.

The second concern relates to audibility: “in blockchain, a transaction is an activity which changes the state of the current Blockchain” (Chowdhury et al., 2018). This is important to note, as reading or querying the data within the Blockchain does not count as a transaction, which means a data consumer could query the Blockchain, gaining access to information, without making a change to the Blockchain itself and subsequently their actions not being logged. This could also create the potential for information leaks without any form of traceability; a sincere concern when using a method of storage that is intended to be highly secure. Alongside these issues, is the potential for sharing information and data without the user’s consent or knowledge. IBM also voiced their concerns relating to this potential security vulnerability when using a system such as this (Compert et al., 2018).

If a “vulnerable application” (Compert et al., 2018) was connected to the Blockchain, its vulnerabilities could be broken down and used to access the information on the Blockchain. A suggested solution to this issue is to divide the transactions into both public and private and encrypt the data before entering the Blockchain, as suggested by Chowdhury et al. (Chowdhury et al., 2018) when they address this issue. The use of this Blockchain architecture would ensure the security of the private network as the vulnerable application would only provide access to the public Blockchain, rather than the internal private Blockchain, provided correct management was provided for both Blockchains.

If the Blockchain’s data was queried it would be stored as a private transaction and would be encrypted by the symmetric public key that Chowdhury et al. implemented (Chowdhury et al., 2018). This would mean only the data subject could see who has accessed their data, leaving less chance of data targeting. For example, if a country’s president used an application that was later found to have a security risk or to have been compromised, the President, could not be targeted by viewing the Blockchain transactions as there would be no connection or identifier; when the query transactions are hidden, weak links in the

Blockchain's system are much more difficult to find. This too was suggested, and implemented, by M. Alessi et al. in their application of Ethereum based Blockchain (Alessi et al., 2019) as a data storage medium, however, this method could be adapted for even higher levels of security.

An alternative method to addressing the issue is to split the entered data, and encrypt it chunk by chunk, using asymmetric, Public Key Cryptography. This would constitute a form of containerisation. This solution would consist of hiding the reconstructive instructions for the now split-up data, inside of a private distributed Blockchain ledger. This would be retrieved by using the uploaded file's identifier. This too raises obvious concerns as to how data consumers would receive the data subject's information; to send data from the data subject to the consumer, the data subject would encrypt their data using the data consumers public key, which would have to be made available to them. Upon a data subject uploading their information using the data consumer's public key, the data subject would assign a password to the file and then be given a file identifier. These two pieces of information would be sent to the data consumer, who could then download the data subject's file by entering their private decrypting key, file identifier and password. An automated system for this could be designed, however, this issue is raised due to lack of innovation when sharing user data in Chowdhury et al., Zyskind et al., M. Alessi et al. and IBM's research (Chowdhury et al., 2018) (Zyskind et al., 2015) (Alessi et al., 2019) (Compert et al., 2018).

These research papers demonstrate the power of Blockchain and data management; if it was possible to directly affect what data can or cannot be seen by an entity or service, without using a 'permission' system within their own application, it would be revolutionary for the data privacy world and demonstrates the major potential for modern technology, such as Blockchain. It also demonstrates that compliance with GDPR can be met by adapting current technology, however, Chowdhury et al., Zyskind et al., M. Alessi et al., and Dr Lee's research

papers (Chowdhury et al., 2018) (Zyskind et al., 2015) (Alessi et al., 2019) (Lee, 2017) leave gaps for potential in Blockchain usability and application, and do not address every need of the data subject. This is due to the nature of the PDS based systems that are suggested. Although Chowdhury et al. outline many integral parts of a complicated Blockchain system (Chowdhury et al., 2018), and in my opinion, suggest the best established potential Blockchain architecture, their system has limitations and does not cover potentially forgotten areas, such as storing user data without the intentions of sharing it through a smart contract. A smart contract is a self-executing contract with the terms of the contract run by a computer program or a transaction protocol. Smart contracts are computer programs or transaction protocols designed to control and document legally relevant events when a contract or an agreement is signed. They operate in a similar way to an “escro” service, in that it requires both parties or conditions to be met or to accept before it will release the subject it was created for, for example, buying a house. One party enters the housing deed, and the other enters the agreed money, both parties then accept the smart contract. This smart contract is then automatically executed and both parties receive their agreed part of the trade.

3.3 Final Comments

The general conclusion is that Blockchain is still in its infancy when it comes to applications other than cryptocurrency, however, that does not mean it does not have any potential for use and implementation in many other areas of society. There is already research being done towards its use for medical records & medicine such as the Institute of Innovative Research, Tokyo Institute of Technology based in Yokohama Japan (Tith et al., 2020) & also, by Universidade do Vale do Rio dos Sinos (UNISINOS), Brazil (Mayer et al., 2020). There are also efforts being made towards its potential use in the construction and food industry, which were briefly spoken about at an IEEE convention in London (Dr Cyril Onwubiko, 2020).

Many of the articles and research papers that I viewed, were similar to Chowdhury et al.'s research in that they were based around designing a system with smart contract integration. The system which Chowdhury et al. propose looks at the use of Blockchain from a service perspective (Chowdhury et al., 2018); showing the benefits of being able to use a storage system that information is stored upon, and that can be securely "shared with hospitals, universities, schools, etc. "(Chowdhury et al., 2018).

This system works extremely well for its intended purpose of data sharing, evident by the fact that similar theoretical ideas were mentioned by Zyskind et al., M. Alessi et al., Dr Lee, and IBM, (Zyskind et al., 2015) (Alessi et al., 2019) (Lee, 2017) (Compert et al., 2018) focusing upon service-based relationships with data subjects and data consumers; proving its fundamental usability, however, what these papers fail to address is the relationship between the data subject and their data. Throughout their research, their focus is the sharing of data for the advancement of the data consumer and their compliance with GDPR legislation (European Parliament and of the Council, 2016), rather than the data subject themselves.

These services are designed and intended for the sharing of information, whereas this research aims to integrate more than that, the use of an immutable, secure, and encrypted Blockchain system for the storage of user-submitted data; this research focuses on allowing the user to input their own data, for their own storage, rather than an intended purpose of sharing it, though this is easily achievable by using another entity's public key, but also having direct control over its holding and deletion to comply with GDPR's Article 17 (European Parliament and of the Council, 2016), arguably Blockchain's main opposition due to its immutable nature.

In short, there are many potential applications for Blockchain within the current society, economy, and industry. I do not think it will be too long before Blockchain will be

implemented in many areas of modern life and we will not even be aware of it, however, it will be there.

3.4 Literature Review Summary

To summarise a couple of conclusions and key points that I found notable throughout this literature review:

- GDPR was introduced in 2018 (European Parliament and of the Council, 2016) in an attempt to address and update the data protection and privacy laws for EU citizen's data, and in turn: regulates personal data holders outside and within the European Union.
- The introduction of GDPR (European Parliament and of the Council, 2016) has reshaped the collection of personal, user data and, in turn, has had a major effect on the types of data collection methods that can be used.
- Blockchain data storage is one type of data storage that has been affected, as in its current state, it is a non-editable type of data and record storage.
- Blockchain technology has an extremely versatile usage parameter, and can be implemented into many different industries, however, is still in its infancy concerning versatile implementation and use.
- Every person within the European Union has the right to erasure, also known as the right, to be forgotten.
- Blockchain uses a ledger that cannot be edited and therefore whatever data is stored inside the ledger, must not be personal, identifiable, or specific to any user or entity under GDPR compliance (European Parliament and of the Council, 2016).

- A solution to an un-editable Blockchain is to store the data within an external file or storage medium database which would create the ability to mimic an editable ledger's record.
- Chowdhury et al., Zyskind et al. and M. Alessi et al. all suggest using a multi-identification system for the sharing of personal data through smart contracts (Chowdhury et al., 2018) (Zyskind et al., 2015) (Alessi et al., 2019).
- These services use two types of transaction for data access and management but fail to address the compliance of Article 17 (European Parliament and of the Council, 2016) when considering the PDS system, Chowdhury et al., Zyskind et al. and M. Alessi et al. introduce to store their Blockchain's data, off-chain.
- They (Chowdhury et al., Zyskind et al. and M. Alessi et al.) mention the revoking of data access for the data consumer and do not address the deletion of data as per the data subject's request.
- To combat these unanswered questions, the data stored would be located on an off-chain storage medium, and on the Blockchain, would be an identifier for the data that is stored, allowing the data to be removed, without affecting the integrity of the Blockchain itself, with an implemented system for data deletion requests.
- Another method of avoiding certain GDPR (European Parliament and of the Council, 2016) issues, concerning user identification, would be implementing a system in which uses a public-key cryptography system for identity and security.
- The possibility of future user authentication, such as biometrics, could solve the current issues surrounding user verification and make the access of Blockchain data much easier, specifically relating to the implementation of a dual-transaction

Blockchain recording system. Biometrics, however, may also have identity theft implications, or sudden identity change implications.

4.0 Research Methodology

4.1 Introduction

The study of GDPR (European Parliament and of the Council, 2016) and its complexity when in relation to Blockchain being used as a form of user-data storage, is vital to the research. To this end, an extensive literature survey and review has been conducted, and identification of the main challenges that must be overcome in order to remain compliant with GDPR, whilst using Blockchain as a secure method of personal-data storage. It is these challenges imposed by GDPR (European Parliament and of the Council, 2016) that shaped the research methodology most. This research investigated how user data could be stored, whilst maintaining the added security of Blockchain, instead of using a standard data storage system.

Academic articles and papers started to be released from 2018 onwards, analysing GDPR's overall potential impact on society and the internet, such as (Tatar et al., 2020) and (Compert et al., 2018). This is heavily related to both large and small businesses, as well as their conduct, which led me to investigate the potential ways to create a secure and editable storage medium for what is likely, extremely sensitive information.

It was because of the newly imposed laws and regulation, alongside the impression that Blockchain technology had on me, that I decided to research this topic further. The information on the subject was limited at first but the changes from GDPR became increasingly apparent throughout the affected industries, and as a result, I was able to access more information relating to the topic.

For each of the 5 challenges, imposed by the 5 main focal points, consideration was given during the investigation and research section, by developing a conceptional solution founded on the information and research acquired at that point. Further research was

conducted by finding existing testing or research that had been conducted on said topic by a reliable third-party source, or, if sources could not be found, it outlined sections of investigation and research that had to be tested, during the development phase. I will go into even further details regarding my hypothetical solutions later.

The security factors, regulation details and capabilities of Blockchain had to be heavily researched, to conclude that Blockchain could be used as a viable method of data storage in correlation with GDPR (European Parliament and of the Council, 2016), if the Blockchain system was not suitable to be implemented and utilized, research into encryption methods and other integral parts of a working data storage system, would have been redundant. Specifically relating to GDPR's limitations over user-data collection and holding. The right to erasure was also one of the main focal points when researching the complexity of personal information storage.

Considering Blockchain's architecture, of cryptographically linking blocks of data together, a solution had to be conceptualized through research and intuition, to ensure the blocks within the Blockchain did not contain the user's file identifiers. Method of access to these files then had to be investigated, as well as user's password security, method of data encryption, and the reconstruction of the user's files.

It was apparent to me at the time of research, that this conceptual system could allow the user to delete and download their own data, provided I understood the different type of Blockchain systems, found viable methods of encryption, and most importantly, if I could develop a form of data splitting and identification, that kept the main Blockchain intact whilst upholding it's intended security and immutability. Research and investigation was an opportunity to rule out or answer hypothetical questions, discover potentially unasked questions, and distinguish the potential methods of development for the artefact.

4.2 Data Collection

To gain a better insight into the potential problems that may arise or hinder the artefact, the overall research was approached topic by topic following the main 5 research headings: GDPR, The Blockchain's Security, User Data Access, File Separation, Data Security and Encryption. This was conducted by analysing existing data, and the research methods used throughout other research papers, to give a fundamental idea of the process of development and research into similar topics to this.

One of the issues that arose from the research conducted, was in relation to GDPR, specifically, the thin line of interpretation when trying to define what is considered user data, for example, a hash sum of the data used for integrity. This was due to the lack of exact reference within the GDPR legislation, (European Parliament and of the Council, 2016) and as a result, easily caused confusion. As such, the information had to be interpreted to uphold GDPR's unclear requirements and expectations; research into this GDPR concluded no definitive answer.

The method in which data was collected had to be considered before investigation could begin. The student resources provided by the University were the first and most obvious source of information, however, to ensure more thorough research was conducted, other methods had to be considered.

Prototyping and experimental design was one of the less common methods of research conducted, specifically, data encryption and data splitting. Testing was conducted by prototype to gain an understanding of potential issues that may arise in the areas of data encryption and splitting, as this was the least publicly researched topics of the artefact. Most of the information used was retrieved from sources such as the University's library. This library consisted of a wide range of literature, but the most beneficial sources, were online versions of existing academic papers that were within the student library resources.

Another potential method of data gathering that could have been considered, was investigation into the operation of a cryptocurrency, and how it's Blockchain system handles information, such as a Bitcoin, Tether or Binance Coin. However, it was decided that due to the lack of similarity between the existing architecture of financial Blockchain systems, and planned artefact architecture for completion of this research; this method of data gathering would not be beneficial to the design or research process.

It should be noted this method of data collection would have differed from the research undertook into Ethereum Blockchain based systems, as their focus was on non-financial Blockchain use constructed upon a cryptocurrency Blockchain architecture. Since this research is orientated around file storage with compliance of GDPR, it was necessary to return to the basic principle and architecture of Blockchain in order to accommodate the functions required by GDPR, such as data erasure.

Research towards different Software Development Life Cycle (SDLC) methodologies also had to be conducted and evaluated; Agile, Lean, Waterfall, Iterative, Spiral, and DevOps were all considered, however, due to the nature of the artefact, it's unusual governing regulations, blockchain technology and multicomponent research/design factors to consider, alongside my personal work preferences from practice within the field; I voted against using a SDLC methodology, as I didn't believe it would be beneficial to the design process of the artefact and did not suit my coding practices. My approach and practices towards the design and development of the artefact, are discussed within Section 5.0, "Design, Development and Evaluation."

To decide what information was useful for the artefact and research, a range of studies and papers were gathered. These were based on topics that would potentially benefit the research, chosen based on article title, description, article tags and abstract keywords. The documents were first read without taking notes or pinpointing any information that was

within the documents. If they were likely to be constructive towards the research, based on the researcher's content and comparison to my own research question, the sources would be saved to both a physical hard drive and online storage or backup. Once the initial list of papers was comprised, the papers were reread. To ensure the information was constructive to the investigation process, multiple concepts were made of the viable solutions to achieve GDPR compliance when utilising Blockchain. These viable and conceptual solutions were then compared to the saved third-party research content and founding principles; their reasoning for their research, methods of practice, and requirements of their artefacts to fulfil their research expectations. This information was then given an in-depth analysis. This was to investigate if the chosen studies would advance the development, understanding or potential solutions for my own artefact, or if the research material would impose modifications to any of the existing conceptual builds of the potential artefact. Depending on whether the information was related and helpful, the papers were then compiled into a second list of documentation, which was used throughout the literature review. Once this was done, the finalized list of references and documents were reread, and the most essential and valuable information was extracted and put into bullet points for review and guidance during the development process. These points were especially beneficial when developing the artefact as they created a direct reference of the artefact's intended aims, potential issues, and/or limitations.

Overall, approximately 25 - 35 studies and papers were found and read, all relating to similar topics that would potentially benefit the research and artefact. Approximately 40 – 50% of the papers read were not beneficial.

The research relating to the Blockchain system was one of the two main focal points of the literature review. Blockchain was a method of storage that I had a basic understanding of but not on an architectural level. To gain insight into this, online resources, as well as

research papers, helped to highlight the complexity of the system, how to build one, and how to properly utilize it. The most beneficial information collected from the literature review related to the issues surrounding data encryption. There were few methods of data encryption that would work for the artefact, except an assertive study done by Lee, their paper titled “BIDaaS: Blockchain Based ID as a Service” (Lee, 2017). This paper was extremely vital to the artefact’s research as their work helped to develop a possible solution using their proposed method of encryption, that would benefit the user information’s security and encryption.

The second focal point of the data collection was surrounding GDPR. The legal legislation was challenging to comprehend, so to best absorb the information, online videos explaining the legislation, in a more easily understood manner, assisted in processing the information. The documents were then reread and compared to the notes that had been taken from viewing the online videos.

This method of research gave a comprehensive understanding of literature in a more simplified manner.

The most beneficial information collected confirming the interpretability of GDPR as well as the possibilities Blockchain provides for widespread industry use, was the paper by IBM (Compert et al., 2018). This was extremely vital to the artefact’s research as their work helped to establish a founding principle on how to interpret GDPR as well as confirming the difficulties the regulation inflicts through its open-ended nature.

4.3 Research Methodology Conclusion

The artefact’s research primarily focused on collecting qualitative data at first, as quantitative data was most useful in representing the research’s limits and capabilities, rather than its potential flaws which could easily be amended in the Alpha Testing phase of a working artefact and did not need to be specifically researched. This artefact was

experimental in nature, and statistical data was not necessary for the initial development of the artefact.

The analysis of other literature allowed for initial hypotheses to develop and encouraged experimentation. These initial hypotheses were then tested to construct a final hypothesis based upon other's existing research, alongside experimentation.

With the help of various articles, such Blockchain as a Notarization Service for Data Sharing with Personal Data Store (Chowdhury et al., 2018) Decentralizing Privacy: Using Blockchain to Protect Personal Data (Zyskind et al., 2015), Blockchain and GDPR (Compert et al., 2018), and BIDaaS: Blockchain Based ID As a Service (Lee, 2017), the artefact's main factors for consideration can be broken down into five factors. These five main factors will ensure the creation of a suitable Blockchain system to fulfil the intended compliance with GDPR. These five factors outline what must be done in order to deem the artefact successful. These five factors are:

- GDPR
- The Blockchain's Security
- User Data Access
- File Separation or Chunking/Data Security
- Encryption

These five factors' must be considered to ensure the personal information Blockchain storage system is effective. A breakdown of these five factors will follow:

<u>GDPR:</u>

The most impactful factor to this research is GDPR (European Parliament and of the Council, 2016). The article which mandates personal information erasure, Article 13, had to be researched and heavily analysed, as it is one of the most important parts of this research due to the standard nature of Blockchain.

What GDPR (European Parliament and of the Council, 2016) attempts to do, in relation to the user and the control of their own data, also had to be researched, addressed, and carefully outlined. It is vital to comply with these regulations when attempting to consider the artefact's outcome as a success. GDPR also relates to Blockchain technology so heavily due to Blockchain's nature and its traits when relating to security and immutability. It is paramount to research a method in which Blockchain could be maintained as a secure method of storage, whilst maintaining compliance with modern regulations and laws, such as GDPR (European Parliament and of the Council, 2016).

The Blockchain's Security:

One of Blockchain's most important factors and characteristics is its lack of editability and immutability.

To fully comply with the issues surrounding GDPR's data erasure regulations and Blockchain's immutability, research had to be into the different types of Blockchain architectures and the ways in which Blockchain could be modified and altered, to allow the development of a deletion process upon request. Without this implementation, companies using a Blockchain system, based on this research's architecture, could be left open to being heavily fined for breaching GDPR, or other similar regulations (Information Commissioner's Office, 2020).

For this element of the artefact to be deemed successful, a method had to be researched that allows a secure access method to retrieve or delete the user data, be it sensitive or not,

whilst minimalizing user-stress and complication when trying to access said data on the Blockchain.

User Data Access:

As stated above, the most impactful factor in this research and artefact, is GDPR.

To remain compliant with GDPR's attempts to keep the user in control of their own data and give them the right to removal from any data storage, existing methods of user data access to request an erasure of this sensitive information, had to be researched. This would allow the data owner the right and ability to request deletion of their data through a simplistic and user-friendly method of access, however, this had to include a method of requestion which ensures security in order to avoid the destruction of user data without the permission of the data owner. The ability of sharing this data also had to be considered, and it was preferable to find a method of data access that would allow the data owner to share their information, without having to provide any sensitive or personal information, such as a password. If access could be easily obtained by others who do not own the information or data or have the right to view it, it would create large concern for any business that used this type of Blockchain architecture.

A method for not just easy, but also secure, user access also had to be researched; a password-based system will be used during research as a control measurement for security, as they have been found to have quite major security flaws, yet, are consistently used for user access to personal information.

For this element of the artefact to be deemed successful, a method had to be researched that provides a secure access method of retrieval or deletion for the user data, be it sensitive or not, whilst minimalizing user stress and complication when utilizing the artefact.

File Separation / Data Security:

Due to the necessity of a system architecture that allows data from the Blockchain to be removed, research had to be conducted into potential ways this could be done. Building the artefact's architecture based upon an off-chain storage method with file identifiers on the blockchain's ledger was my initial concept; research had to be done into ways of increasing file security through both containerisation and file encryption. If a file was containerised into multiple different chunks and spread across different servers, it would be almost impossible to reconstruct them without appropriate reconstruction access. Research into using a separate off-chain storage method for chunk storage had to be conducted, whilst allowing file traceability, access, and reconstruction. If the off-chain storage method did not uphold a heightened level of security, Blockchain's fundamental security will be redundant.

A file chunking system had to be researched that allowed the conversion of a single file, into multiple unreadable files, with private identifiers. This inherently meant that research must include a reconstruction system which could not be publicly accessed or used. Preferably with the option to increase/decrease file separation amounts.

Encryption:

The secondary research path alongside file separation, was achieving a secure file encryption system.

Both file separation and file encryption working in unison would result in an adequate level of security for the off-chain storage system, due to the likely situation that many users would fail to split their file into additional chunks. If a user did not understand the mechanics behind the file separation system and wanted the quickest and easiest method of uploading their documents, using a smaller number of chunks is the obvious choice, meaning an additional service had to be researched to ensure adequate security.

Research had to ensure a method of encrypting was either found or noted for development, that would allow the files to be encrypted after they were split into chunks and before being uploaded to a server, allowing for the most secure method of storage and eliminating the chance of a “man in the middle attack” when paired with the file chunking system that was to be developed; A man in the middle attack, is a type of cyber-attack in which hackers intercept an existing conversation or data transfer in real-time, disguising themselves as a legitimate part of the transfer, copying the transfer to their machines.

Research also had to be conducted into the benefits of HTTPs, to ensure the best method of data transfer was being used to help avoid any unnecessary security flaws.

This research would allow the user to choose the less secure file separation settings: lower separation amounts, whilst keeping their information as safe and secure as possible.

4.4 Research Methodology Analysis

The conclusions made in the literature review, and summarized within it, are valid, on-topic and factually correct, however, the articles stated in the literature review could have been replaced with more beneficial forms of research. For example, I attended an IEEE event in London in 2020 (Dr Cyril Onwubiko, 2020). I found this event to be extremely valuable to me on a personal level, regarding my understanding of the research topics. This event benefited the research methodology as I had a base understanding of the topics at hand. I was able to gain a lot more insight into experimental ideas that are currently being developed and was able to interact with those behind the development within their fields. However, this is an issue, as it is not possible to reference information collected at an event through conversation.

I believe that if the existing articles for the literature review were more alike to the information received from the IEEE event, more time could have been dedicated to finding

this information, which was not possible when attempting to receive existing articles on a topic such as this; If the research conducted had been citable through real-life practises and conversational guidance by members of the industry, more specific questions surrounding potential issues could have been answered, such as the issues regarding encryption types, methods and reliability. This research would also have greatly benefited from a group approach, due to the ability to do continuous and more extensive research, on a faster basis over the same time frame.

5.0 Design, Development and Evaluation

5.1 Design Process

To create a working, viable artefact, I had to combine my research, viable solutions, and artefact limitations to formulate a working design that achieved optimal results. This was done over a course of approximately a month.

I decided to use JavaScript and NodeJS throughout the artefact due to its cross-platform uses and due to its growing adoption in server-side applications.

There were also many means of support available for this language, for example, online forums, existing projects, GitHub, to name a few. These sources of information helped to troubleshoot bugs within the code and find solutions for any implications that may arise. I decided to write a certain proportion of the artefact using standard JavaScript as JavaScript is most used as a web browser programming language for the client-side application, being better suited for User Interfaces; NodeJS, which is a modified version of JavaScript, was used for the server-side as it is a version of JavaScript that has been modified and ported for the purpose server-side applications.

During this design process, there were five main factors I had to consider when creating a working system that would be needed for the desired outcome. These five main factors were mentioned prior, however, will now be discussed in a more detailed manner:

- Server Handling and Data Storage Method
- Data Encryption
- File Separation and Upload and Data Download/Deletion Requests
- UI and Application
- Blockchain System

Server Handling and Data Storage:

The Server Handling operation was the most difficult part of the design system to find a definitive version for. This was due to the complexity of having to manage multiple servers at the same time. The original intention was to create a system that would allow all the servers to work together, however, this was difficult to create without the servers overwriting or overruling each other. This is due to no server receiving any updates to what another is doing as they all would be of equal responsibility and power. As a result, two servers could use the same file identifier at the same time, or two server servers could write to the ledger at the same time which would result in an incorrect chain. This is a similar result to what would happen if you modified an entry on the ledger.

The initial plan for the client, as shown in Figure 3 below, outlines a simplistic version of the file management system.

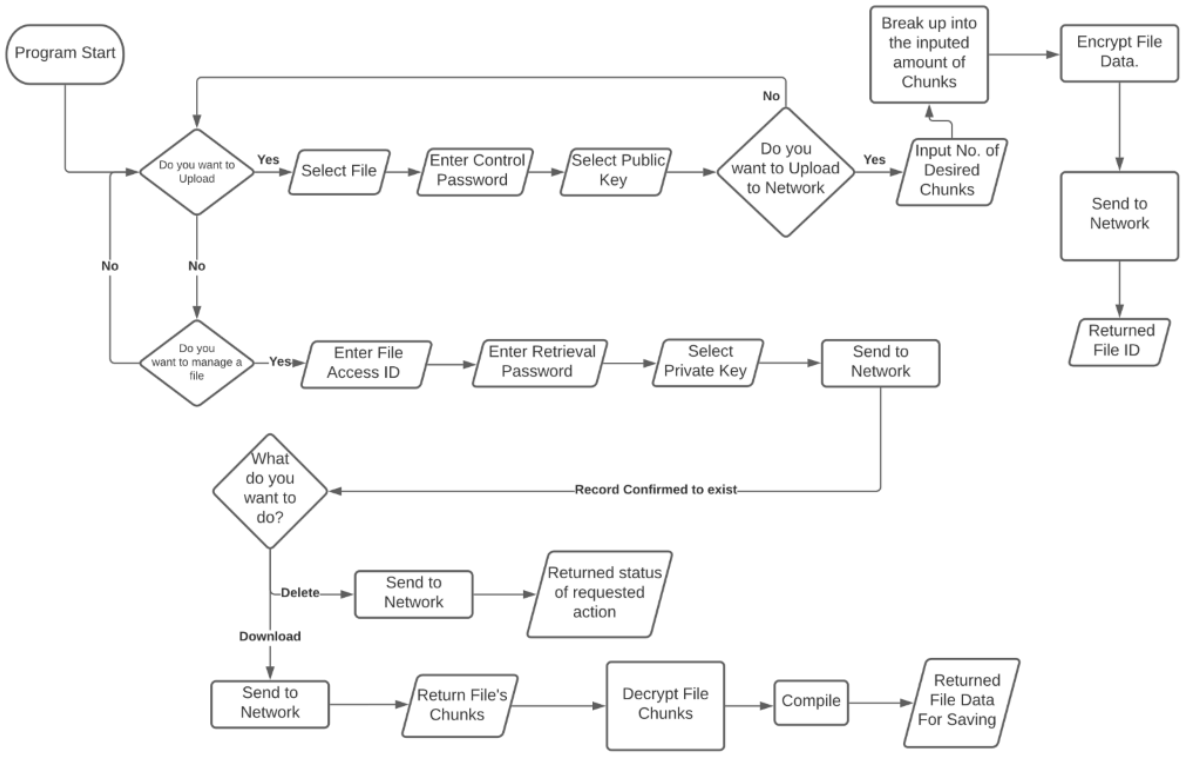

Figure 3 – Client Program Flow Chart

Upon starting the artefact, the user would be presented with a user interface giving them two choices; Download and Upload. If the user wanted to upload a file, they would be required to input the public key that they generated and the file they wish to upload. The artefact would then request the number of chunks they desired to break their file up into, and finally to set a password, used when accessing the file in the future. Once their details were inputted, the upload process would begin. It would be at this point that the system began breaking up their file into the specified number of chunks and encrypting it. Once this process was complete, the now 'chunked' file would be uploaded to their artefact's server, to which a file identifier would be returned to the user for access and management.

The download process would follow a similar process. The user would be required to select their private key, enter a password for their file, and finally an already provided file identifier. If the user clicks continue, the data: the password for their file, and their file identifier, would be submitted to the network. This would enable the system to check whether the file ID existed, and if so, would proceed to check if the existing password stored, correctly matched the password that the user submitted. If the password was incorrect, the system would refuse the user access to the file; If the password was correct, the user would be presented with two options: the ability to download their file, or to delete their file. If the user requested to delete their file, the server would delete all files associated with the user-submitted file ID, meaning all chunks of encrypted data would be removed permanently from the server's disk. If the user requested to download their file, the system would read all file chunks from the disk that matched the user-submitted ID, and deliver them back to the client, so they could be reconstructed and decrypted. The reconstructive process would decrypt all data chunks associated with the user-submitted ID, and then recompile them to recreate the user's original file. This file could then be saved to the user's download destination.

This original design, in which the servers were set up to communicate with other servers was later revised to use a Primary-Secondary server configuration, in which the servers elect a nominated server that is assigned for write requests to go through. This decision was made as the previous design did not allow for the incremental hashing of each block that was entered into the Blockchain, and issues arose relating to the client program.

The lack of client-side error handling caused issues when attempting to debug the artefact, and similar issues arose surrounding the server architecture. The lack of server-side error handling caused large issues during a network failure, or if a server went offline. These issues were discovered during the development phase, and so another design had to be drafted.

The updated server software is designed to follow an event-driven nature (Figure 4), this is to reduce any stress on the server's hardware and to increase overall performance. If this system is not updated to its current state, the system's efficiency could be greatly reduced, due to lost performance in 'sleeping' the process. This temporarily halts the application's process until the sleep value expired, before checking again for any updates or new HTTP messages. This would result in a slower system alongside more system errors relating to overwriting or block conflicts within the Blockchain due to further performance reductions, potentially leading to further latency between servers. This caused each server node to be constantly checked in an attempt to add the next block of data to the Blockchain and resulted in strain to the server, and a speed reduction to the system. This overwriting occurred where each server failed to match against another causing undue strain.

The artefact's working theory was slightly adapted to facilitate this Primary Secondary Server architecture: When a user submits a data download or data deletion request, the request will be executed on any of the Secondary servers, assigned by the Primary Server. This will reduce stress on the Primary Server, as its resources are being prioritized for writing

requests and managing the Secondary servers, due to their importance in the working artefact. See Figure 4 and 5:

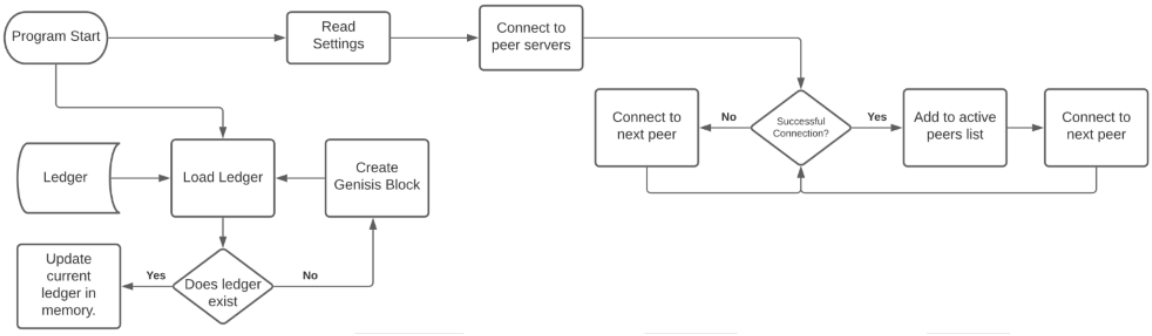

Figure 4 - Server Program Flow Chart

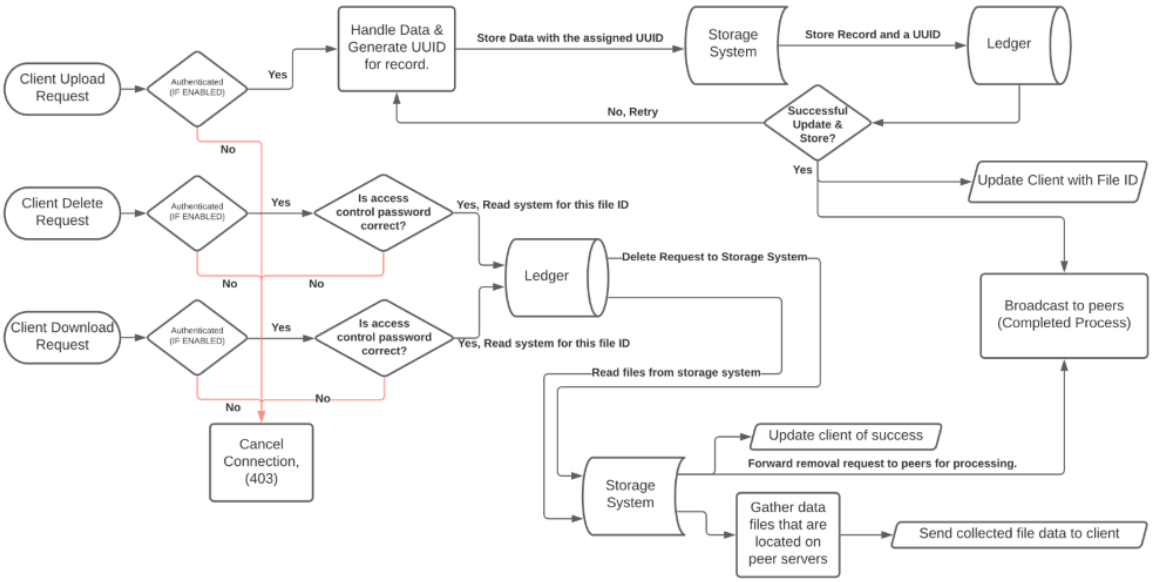

Figure 5 – Client Request (Server Program Flow Chart)

All user information uploaded using the service will be stored on servers under strict regulation. This is to reduce further GDPR complication relating to data handlers and the

issues surrounding GDPR's interpretability. It also assures that control over the stored data is managed by trusted individuals, such as a GDPR Data Protection Officer. This eliminates the risk of third-party tampering with the system and its security; when considering personal information, this is a particularly crucial factor.

Data Encryption:

The main difficulty when evaluating the system's file encryption related to the amount of hardware usage that a Public Key Cryptography system uses. To encrypt the user's file(s), a public key must be used, however, due to the complexity of the public and private key system, more computational resources are required.

The main issue with this system arises when substantial amounts of data must be encrypted, generating a public key encryption for a file over the size of 4,000 bytes is incredibly slow and puts too much strain on hardware resources.

I was unable to find any other method of encryption/decryption that provided adequate security for such sensitive information, so a workaround had to be designed to solve this issue.

The method of encryption will remain as Public Key Cryptography, however, instead of encrypting each file chunk individually before uploading them, as each chunk would most likely be over 4,000 bytes in size, a method was developed of splitting each chunk of information into a piece of data to encrypt and relink together will be used. This will be discussed in the File Separation and Upload section.

Encrypting lesser amounts of data using a public key cryptography system is an extremely reliable method of data security as only the user's single private key can decrypt the file(s). This system also ensures that if an attacker gathered all the user's chunks, without mixing them up with other file's chunks, and reconstruct them into the correct order, they would

still not be able to access the file's information, due to the public cryptography's incredibly secure encryption.

File Separation and Upload:

The file separation system was one of the most computationally taxing parts of the artefact. To separate the user's file into individual, reconstructable parts, a mathematical formula and file processing system had to be developed, this is discussed below.

To begin, I designed a system that can convert the data to an alphanumerical string, achieved by turning the data and information collected from the user's files into a JSON format. JSON stands for JavaScript Object Notation. It is a lightweight format for storing and transporting data. The object is then converted into a string, which holds the information in a more appropriate manner for 'chunking' it.

Once the file is converted into a string, it can then split into a user-defined number of chunks. Once this is done, the strings can be stored as standard files, with a custom extension of ".fc" which stood for file chunk.

The names of these files can then be stored in order, on a centralized Blockchain for development and testing purposes. However, issues arose during the development stage in relation to data encryption when using this method, so the existing data chunks had to be split again into "Primary Chunks" and "Secondary Chunks" to allow for 4,000-byte encryption, as discussed prior.

To fix this issue, a redesign of the system had to be designed to allow 4,000-byte encryption. This consisted of multiple levels of chunking that can be identified as "Primary Chunks" and "Secondary Chunks."

To easily understand what would identify as the main system's chunking system; the chunks that the data owner has control over, and the separated information that would coexist within these main chunks. These are the parts of data that are split for the second time and have a maximum of 4,000 bytes in size; identifiers will be set for both, as either a "Primary Chunk" (the main system's chunking system) or a "Secondary Chunk" (the user's data that is split into a maximum size of 4,00 bytes).

The "Secondary Chunks," within the "Primary Chunks," will be encrypted using the user's public key, meaning the "Secondary Chunks" will only be decryptable only if the user is using their personal private key. Originally, I had intended for the "Secondary Chunks" to be identified for linking, and each "Secondary Chunk" within the "Primary Chunk" would be assembled in the chain to the previous one, however, this original method also ensures that the file's decryption worked in a 'backwards linear motion' when compared to the encryption. This means the data within the user's file can be reconstructed in the correct order without a chaining system. I found this through the development process, as the "Secondary Chunk" chaining system was unnecessary as the compiling process works in a linear, single chunk movement. This means that the files can be decrypted in a backwards linear movement without any identification or linking system.

<u>UI and Application:</u>

The user interface will be built using Electron. Electron is an open-source software framework developed and maintained by GitHub. I decided to use Electron as it allows me to develop a desktop GUI application using the same language as most of the artefact, JavaScript.

I searched for another platform that would allow similar frameworks to use, however, this was best suited for the task.

The Blockchain System:

When designing the Blockchain System, NodeJS will be used to keep the artefact written in a single language. This also works best for ease of development as the language is intended for this style of application. The Blockchain architecture will be a private distributed Blockchain system, as this is the best option for the security of sensitive user information but also redundancy of the data. The distributed Blockchain system assures that the data within the Blockchain is private and cannot be accessed by the public. If the Blockchain system is developed using a public decentralized, public centralized, public distributed or a private centralized system, for which access had been retrieved, the only thing needed to reconstruct the file's chunks would be the user's private key. This key could potentially be stolen from a user's hard drive in the event of a robbery or could be stolen from an online backup in the event of a data breach/cyber-security attack, and a distributed Blockchain will resolve this issue.

5.2 Development Process

I decided to create the artefact using NodeJS version 10.16.3 on my own primary computer. All the other devices tested were running NodeJS version 12.18.4. This language was used because of its cross-platform combability, as well as a large amount of documentation, information, and packages/modules available for the resource.

Windows 10 Pro 64-bit was used for all the development of the artefact as this was my primary OS, however, the artefact has also been successfully tested on the following operating systems as well:

- Windows Server 2016
- Ubuntu 20.04 (Linux)

Though this is a small portion of the available operating systems that could be used, the artefact is compatible on most desktop operating systems due to NodeJS being tested on many operating systems and architectures, including ARM, which is typically used in mobile devices, potentially allowing for a mobile-based app variation which could be developed.

Development and testing of prototypes were conducted on localhost. Split-up and encrypted data were saved directly to the disk of my personal computer, in place of a dedicated server. This localhost approach was chosen for its convenience and efficiency. Data handling, key generation, and data retrieval were all conducted on localhost, utilizing a personal hard drive, as all program executions could be translated with minimal effort, but could unnecessarily increase the time spent during the repetitive task of testing and development; the program could be run on a cloud, or dedicated, server configuration, however, configuring this system was not seen as a valuable use of development time, due to the program's cross-compatibility from localhost to server configuration. It is cross-compatibility is inherent, due to the coding language and network used. For application debugging and testing, a dedicated server was used with no adjustment.

The first development of the artefact consisted of creating a system in which users could upload their files to a secure platform. This meant that once the data had been uploaded, it was only retrievable by the data owner themselves. The advantage of this was that the owner always remained in full control of their information, however, there were unfortunate issues with this build of the artefact. This build came with one major flaw, there was an exception to those who had access to the data, this flaw lay in the data storage medium: Access was possible through those who were physically holding the data within their server/storage medium. It should be noted that even with this flaw, due to the data being split into separate chunks, the data holder would have great difficulty if they attempted to reconstruct the data. The only way the data could be reconstructed, if the

original development build was used, was if there was access to both the centralized Blockchain and data storage medium. This build did not include the public key cryptography system.

To solve this issue, and to ensure that the data holders themselves could not reconstruct the user information/data, it was decided that a strict method of encryption was needed. This led to further research which was added to the literature review and built upon existing work done with additional information relating to a public key cryptography method that was later implemented into the development build.

An updated method of encryption was so vital towards the security of the information, as, in the event of a double data breach, one in the centralized Blockchain (See Appendix 1.1) and the data server/storage medium, the user files, after being split could be downloaded by an attacker due to the data server being stored within the main server and identified by the data identifiers (See Appendix 1.2) stored within Blockchain. Allowing them to then be reconstructed by matching the chunks to the Blockchain's information (See Appendix 1.3), this would result in all the personal information and files within the server being stolen from the data holder, and lead to the software being held liable for a lack of secure data encryption.

It was at this point, that it was decided to introduce a public key cryptography system (See Appendix 1.4). This caused more issues surrounding the system; if the data was encrypted and then stored using a single server concept for this data operation, software efficiency, redundancy, and security, could become a large issue. To solve this, an architecture was developed to ensure that each server available was used, keeping the system updated and concise. This was achieved by utilizing a synchronizing data loop; Message Queuing Telemetry Transport (MQTT) ensured that all systems that ran the individual data servers, were kept updated. This too was to avoid any issues surrounding data overwriting, and in

the event of a Primary Server being attacked, it also reduced the likelihood of a fatal data breach. An attacker would need access to multiple servers across multiple locations to reconstruct the data.

To break down the artefact's methodology of handling user requests: if the system receives a request from a user, the assigned handling server (See Appendix 1.5) (the currently elected Primary Server) determines if the user's instructions are a read or a write request, and follow a determined procedure:

If a user submits a read request, an attempt to download or view their information/data, the artefact will fulfil the request by compiling the information using the Blockchain's reconstructive instructions. This data would be then reassembled by one of the Secondary servers (See Appendix 1.6).

If a user requests a write request, an attempt to upload their information to the artefact's data servers, the artefact will pass this to the elected Primary Server to prevent issues of overwriting, write conflicts and other issues associated with multiple applications attempting to write the possible same file/material/location. The Primary Server then decides where all the uploaded user information/data is going to be written and stored, which is then synchronized by sharing the same request with all Secondary servers on the network, ensuring all server remain up to date. This avoids two servers being nominated at the same time for a write request and duplicating information/data, as well as avoiding a server being chosen to write information to their storage twice, overwriting one of the requests with another (See Appendix 1.7).

The Primary Server is elected by all participating server instances on the network, which is randomly allocated based on a vote, as stated prior. This voting system will only run in the event of the current Primary Server becomes unresponsive for a set period, which was

decided to be anything longer than two seconds. This gives the Primary Server more than enough time to respond in the event of a common latency issue, if it does not respond, it is assumed offline. When a new Primary Server is required, one or more servers will make a request for all servers to conduct a 'Primary Server vote,' and a new server will be elected as the Primary Server.

One further issue with this proposed method relates to the service start-up. The service could not conduct a Primary Server vote if the service was not already running. To solve this, a method was developed that automatically ran when the network first started; it was decided that the existing electoral system would be overridden when the first instance of the network ran, and instead of multiple servers voting for a Primary Server, the first server to launch and run is automatically selected as the Primary instance (See Appendix 1.8). This server handles any user write requests until said server was overridden by a Primary Server vote.

One other potential issue is if a Primary Server suddenly goes offline whilst uploading or in the process of completing a write request. In this scenario, the Primary Server would be the only server up to date but would be unable to push any data to the other instances on the network as it is no longer operational. This could be resolved by delaying the ‘write successful’ notification, ensuring that the client is told that the upload/write was unsuccessful if the write request was not sent out to all Secondary instances on the network. The user would be informed that they must retry uploading their file/data and would ensure no data upload was marked as completed, when it was not, however, this system was not implemented as this scenario only needs to be considered if the system was being implemented on a public, commercial or enterprise level. It is also unlikely that a Primary Server would go offline during an upload/write request and does not need to be fixed to

prove the artefact's founding principles; the main requirements of the artefact to fulfil the 5 aforementioned main factors and conclude a successful research portfolio.

Overall, the adaptation to a Primary-Secondary configuration (See Appendix 1.9) helped maintain the system's predictability; achieved by configuring the server to wait for a client's request before attempting to process anything on the server's side, for example an upload or download. This server was also setup up to communicate with the other servers, to maintain redundancy in the event of a failure.

The final issue during the development process was deciding how best to handle the data and download/deletion requests. There were multiple options available, such as TCP or other socket-based connections, however, I decided that a HTTPs system (See Appendix 1.10) was the best option to use due to its widespread usability. Using this system meant that the network could be easily adapted for use within a web browser client. In addition to this, it also made it easier to handle incoming requests from any user. Each request could be assigned to a different path of the URL for the system to handle. This also meant that if an IP changed through a load balancer or reverse proxy, the system could easily be migrated to a new host/server or instance, overall improving the efficiency and reliability of the network.

These implementations were mostly tested individually, the operation of the entire artefact, however, needed to be assessed and debugged. I did this through a combination of minor testing and an engineering debugging approach (See Appendix 4.0). Two of these tests failed, checking the system for incorrect file IDs, and inputting the incorrect password or decryption key. These operations do not affect the security, deletion, storage or containerisation operation of the artefact, and so, are only UI bugs, in which the process hangs on a blank screen until the user presses ESC. These bugs were deemed to not affect the successful operation of the artefact.

6.0 Results

6.1 Results

The artefact was tested 3000 times by uploading files to the service (See Appendix 2.0). These uploads consisted of multiple file sizes ranging from 1-kilobyte to 1-megabyte, and considered reading of the file, chunking of the file, and encrypting of the file.

Of these 3000 tests, no files failed to upload or split into their correct chunks. The average upload time for a 1-kilobyte file was 1303 milliseconds, taken from 1000 uploads of 1-kilobyte text files, and there was an average upload time of 8049 milliseconds when using a 1-megabyte file, once again, this statistic was taken from 1000 text-file uploads as an average. The average processing time for a 1-kilobyte file was 7 milliseconds and the average processing time for a 1-megabyte file was 3093 milliseconds. 1000 uploads were also done using 500-kilobyte text files; it took an average of 4720 milliseconds to upload, and 1677 milliseconds to processing these files.

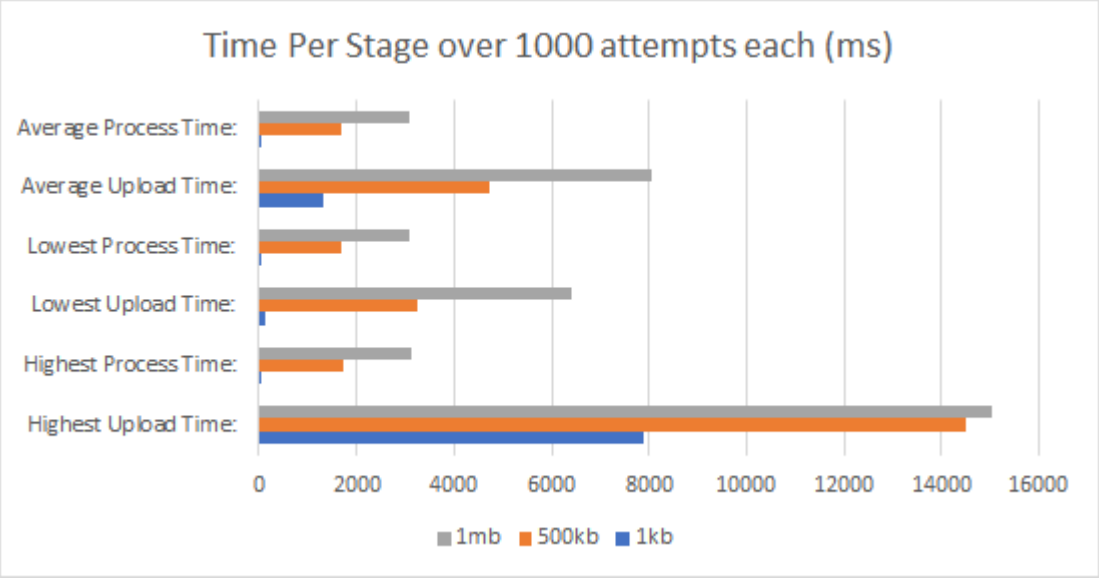

Figure 6 – Processing time per category and file size.

This data relates to:

- Server Handling and Data Storage Method
- Data Encryption
- File Separation and Upload + Data Download/Deletion Requests

6.2 User Feedback

A group of fourteen individuals were given access to the artefact to test over the course of a week. These fourteen individuals were then given a questionnaire to answer as honestly as possible.

This questionnaire consisted of ten questions, excluding basic information about the individual, 9 of which related to the artefact's functionality. The first question asked was if the individual had used the application. All fourteen answered with “yes”. The second question asked for their name/username for the purpose of traceability. The individuals were then asked: “Overall how would you describe your experience when using the application?” with a scale of 0 – 5 (Bad to Good). The average score retrieved from fourteen answers was: 3.79 out of 5; 1 individual recorded a score of 3, stating: “The UI is not very clear / user friendly enough” and another individual recorded a score of 2, stating: “Needs loading bar.” The implementation of a loading bar was considered, however, deemed unnecessary for testing purposes.

The individuals were then asked: “How would you describe the process of uploading a file?”, once again with a scale of 0 – 5 (Bad to Good). The average score retrieved from fourteen answers was: 4.36 out of 5.

This concluded the uploading system used for the artefact, and then the fourteen individuals were asked about the downloading process.

The fourteen individuals were asked: “How would you described the process when trying to download your file.” The average score of the 14 recorded answers was: 4.21 out of 5, with one outlier recording a score of 1, due to a bug that was discovered from the collection of the user-feedback: Specific User feedback: “Didn't work, error on line 440, missing }”. The error thrown back to the individual was fixed, which also amended a file separation/chunking bug that was reported by two other individuals.

When asked: “In your opinion, how could the overall usability of the application be improved?”, three individuals commented on faster downloading times, three individuals commented on progress indicators/bars and three individuals commented on better explanation on how to use the software. Individuals were then given a breakdown of how the Blockchain system operated on a fundamental level and were asked: “What were your thoughts in relation to the application and process, and how it handles data?”, none of which reported any issues, and generally agreed the software was “secure”.

Finally, the fourteen individuals were asked: “If you have the technical know-how what were your thoughts on the applications design and security?”. Of the fourteen, nine answered the question; The longest recorded answer was: “I checked the developer console with CTRL + SHIFT + I. From what I saw, it looked like the file was encrypted and decryption was working correctly. I have full trust in the security of the Blockchain system.” and the shortest was: “Needs loading bar.” Only two complaints/recommendations were recorded from that question: “there isn't enough information around the user on how it works,” relating to the ease of operation, and “Needs loading bar.”

All questions and answers, in full, can be found in the Appendix (See Appendix 2.0 & Appendix 2.1)

7.0 Final Conclusion

In conclusion, the research and artefact were successful in every fundamental aspect. The artefact's results show promising results and evidence that the artefact was successful in creating an editable Blockchain system, that would comply with GDPR, through the process of user feedback, direct testing, and ensuring the compliance of GPDR throughout.

The collected user feedback outlined a few reproducible bugs that the artefact consisted of, but also gave clear feedback on the positive aspects that were achieved when considering the artefact's founding principles.

The main focal points from the user feedback were:

1. The User interface: This was remarked upon multiple times throughout user feedback. It was a common feeling that improvements could be made to the application's layout, to make it more "user friendly" and professional. Remarks were made that the UI was rather 'barebones' and could use further development, such as showing more error information and more specific error handling on the user's side, such as during the upload process. This feedback primarily relates to cosmetic aspects of the artefact, and does not affect the artefact's founding principles, but is constructive feedback that would enhance the user experience.

2. Debugging of the artefact's existing issue: An example of feedback that helped to debug the artefact and some of its founding principles, related to when a bug was found surrounding a missing brace bracket at the end of one of the lines of code; when reviewing the code, the brace bracket was present, so it was not clear as to why the issue arose, however, it was fixed thanks to user testing. Preliminary investigations showed that some data was sporadically missing during the download/decryption process. This issue was resolved by later discovering there was a counting issue from 0 rather than 1, which also resolved the

brace bracket issue. These bugs no longer have any effect on the artefact's efficiency and reliability.

3. System efficiency: Another finding from the user feedback relates to low system speeds when uploading, downloading, and decrypting. The system could be further optimized to avoid points of low speed when receiving the user's file ID, as well as when downloading & decrypting; a solution to the decryption stage, of the retrieval process, would be the ability to have the file decryption parallelised. This would decrease the waiting time for the user, as more of the file could be decrypted at the same time, however, this does not relate to the fundamental principles of the artefact and is not an issue when attempting to create a 'proof of concept.'

These three issues were the main focal points of the qualitative data that was suggested, I would consider the simple bugs and design flaws that arose, unrelated to the artefact 's founding principles, and do not affect the artefact and research's conclusive results.

The results collected from user input had a majority feeling of security when using the system's storage and file encryption. No traces, excluding a string of characters used as a file identifier, are left on the servers relating to any user information or data, once the deletion protocol has been carried out. The entire system can run without the use of any third-party services, although the system could be configured to use an external, commercial, storage system if so desired.

Compliance with GDPR was one of the main key points regarding the artefact. As previously mentioned in this thesis and by others, GDPR has plenty of room for interpretation and manipulation. This is something that the researcher can only outline, when attempting to find complete compliance, without help from lawyers and/or members of the justice system. The question of whether a hash of user data would still count as personal data is still open for debate and will probably take a major court case in the future to decide if it is or is not,

but for now, it is assumed that it is not as there is no identifiable data associated with the hash. In which case, the system created and proposed here allows the user to remove their data upon request, meaning that the artefact is GDPR compliant as per Article 17. As a result, we can conclude the artefact is GDPR compliant based on the limitations that apply, such as the right to be forgotten (European Communities, 2016). Therefore: the artefact is successful in the removal of user data upon request, and Blockchain can be used as a viable method of storage within the EU.

The artefact's main founding principles that were found to work correctly, and in line with what was expected, were:

- Server Handling and Data Storage Method
- Data Encryption
- File Separation and Upload and Download/Deletion Requests
- UI and Application
- Blockchain System

1. Server Handling and Data Storage Method: File Data Storage; The file data storage worked correctly for the user, splitting their file into their desired chunks, once bug fixes had been implemented, and storing the information on the artefact's servers with the correct identifiers being inserted to the Blockchain for reconstruction.

2. Data Encryption: The use of the public key cryptography system can only be evaluated as successful; the files were correctly encrypted into their 4000-byte secondary chunks, within their primary chunks, and when downloaded: were correctly decrypted, after minor bug fixes, resulting in a secure and working data encryption system.

3. File Separation & Upload File Separation and Upload and Download/Deletion Requests: File Download/Deletion Protocol; The file download/deletion protocol worked correctly for the user, identifying their file's chunks by reconstructing the information from the Blockchain, sending a reconstructed file to their download queue, or deleting the information from the artefact's servers upon the user's request and input of a correct password and private key.

4. The UI and Application: This functionality of the application is mainly discussed throughout the user feedback, however; although the application had remarks relating to its user interface, every aspect of the application worked correctly. The public key cryptography system was able to generate, read, encrypt, and decrypt the users' files, although there were initially some issues surrounding the reconstruction of the file and information. This was solved and it was discovered not to be due to the cryptography system; it was an issue in the math behind the file splitting and as such resulted in missing data during reconstruction, this issue was observed once during initial development though I could not replicate the issue. However, user testing outlined that the issue was more random occurrence and upon further investigation the issue was determined to be a minor math error that resulted in the loop being one iteration short in order to complete the data reassembly. This was a minor bug in the file splitting process. The ability to upload the user's file worked correctly with the only limitation relating to file size and upload speed; an issue that could be resolved with further development, but that is not necessary to conclude a working concept. The user was able to use their file identifier correctly to retrieve their file, and the system consisted of minimal bugs that were not related to the artefact's key principles.

5. The Blockchain's System: Upon the deletion of the user's information from the artefact's servers, the Blockchain's file identifier, user password or essential Blockchain information did not get removed, and therefore the Blockchain and its ledger maintain their integrity.

The key aspects taken from this research are the interpretability of implemented regulation, and the reality of ever-shifting technology. No matter what law is implemented, there will always be a method of workaround to achieve whatever goal is desired. Existing technologies can be changed, laws can be argued and there will always be a solution to the problems at hand. Though what must be sacrificed to find such a solution is to be concluded by those who are willing to question it.

The key limitation of my research relates to a lack of existing examples and testing when attempting to alter the fundamentals of Blockchain technology. This research sets an example of the possibilities that Blockchain technology can provide and the creativity that can be achieved for all existing forms of software, hardware, and technology alike. Limitations are to be questioned and no task is unachievable with the correct dedication, determination, and creativity.

To conclude, “Can this be achieved by the use of data containerisation?” Yes. The containerisation aspect has been done and complies with the requirements of the user and implementation. Even though the likes of docker, Amazon S3 or FTP were not used, they could easily be implemented, if so desired. I believe my artefact and research fulfilled its intended purpose of Blockchain being used for secure data storage after the implementation of GDPR. A system can and has now been, developed that allows the use of this extremely secure technology, whilst remaining compliant with the implementation of recent law and regulations, specifically, GDPR (United Kingdom Government, 2018).

References

Alessi, M., Camillò, A., Giangreco, E., Matera, M., Pino, S. and Storelli, D. (2019) A Decentralized Personal Data Store based on Ethereum: Towards GDPR Compliance. *Journal of communications software and systems,* 15 (2) 79-88. Available from https://hrcak.srce.hr/220954.

Baran, P. (1964) *On Distributed Communications Networks.* Available from http://web.stanford.edu/class/cs244/papers/DistributedCommunicationsNetworks.pdf.

Boutrous, T., Hanna, N., Vandevelde, E., Crutcher, L., Dunn, G., Olson, T., Zwillinger, M. and Landis, J. (2016) *Apple Inc's motion to vacate order compelling Apple Inc. to assist agents in search, and opposition to Government's motion to compel assistance.* California. Available from https://assets.documentcloud.org/documents/2722199/5-15-MJ-00451-SP-USA-v-Black-Lexus-IS300.pdf.

Chowdhury, M.J.M., Colman, A., Kabir, M.A., Han, J. and Sarda, P. (Aug 2018) Blockchain as a Notarization Service for Data Sharing with Personal Data Store. In: Anonymous IEEE. 1330-1335 Available from https://ieeexplore.ieee.org/document/8456052.

Compert, C., Luinetti, M. and Portier, B. (2018) *Blockchain and GDPR.* IBM Security. Available from https://iapp.org/media/pdf/resource_center/blockchain_and_gdpr.pdf.

Dr Cyril Onwubiko, (2020) IEEE United Kingdom & Ireland Blockchain Launch | 2020. London, Available from https://www.ieee-ukandireland.org/event/ieee-united-kingdom-ireland-blockchain-launch-2020/.

European Parliament and of the Council. (2016) *2016/679 (General Data Protection Regulation).* Available from https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679.

European Parliament and of the Council. (1995) *Directive 95/46/EC*. Available from https://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:31995L0046:en:PDF.

Gewirtz, D. (2018) Volume, velocity, and variety: Understanding the three V's of big data. [Online] Available from https://www.zdnet.com/article/volume-velocity-and-variety-understanding-the-three-vs-of-big-data/.

Information Commissioner's Office. (2020) What are the GDPR Fines? - GDPR.eu. [Online] Available from https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-law-enforcement-processing/penalties/.

Lee, J. (2017) BIDaaS: Blockchain Based ID As a Service. *IEEE access,* 6 2274-2278. Available from https://ieeexplore.ieee.org/document/8187625.

Mayer, A.H., da Costa, C.A. and Righi, R.d.R. (2020) Electronic health records in a Blockchain: A systematic review. *Health informatics journal,* 26 (2) 1273-1288. Available from https://journals.sagepub.com/doi/full/10.1177/1460458219866350.

Nakamoto, S. Bitcoin: A Peer-to-Peer Electronic Cash System. Available from www.bitcoin.org.

NordPass. (2020) Most common passwords of 2020 | NordPass. [Online] Available from https://nordpass.com/most-common-passwords-list/.

Vagata, P, K. W. (2014) Scaling the Facebook data warehouse to 300 PB. [Online] Available from https://engineering.fb.com/2014/04/10/core-data/scaling-the-facebook-data-warehouse-to-300-pb/.

Tatar, U., Gokce, Y. and Nussbaum, B. (2020) Law versus technology: Blockchain, GDPR, and tough tradeoffs. *The computer law and security report,* 38 105454. Available from http://dx.doi.org/10.1016/j.clsr.2020.105454.

Tith, D., Lee, J., Suzuki, H., Wijesundara, W M A B, Taira, N., Obi, T. and Ohyama, N. (2020) Application of Blockchain to Maintaining Patient Records in Electronic Health Record for Enhanced Privacy, Scalability, and Availability. *Healthcare informatics research,* 26 (1) 3-12. Available from https://www.ncbi.nlm.nih.gov/pubmed/32082695.

United Kingdom Government. (2018) *Data Protection Act 2018.* London:. Available from https://www.legislation.gov.uk/ukpga/2018/12/pdfs/ukpga_20180012_en.pdf.

Van Reede, M. (2020) *Evaluating the practicality of using blockchain technology in different use cases in the healthcare sector.* Available from https://www.cs.ru.nl/bachelors-theses/2020/Mischa_van_Reede___4557816___Evaluating_the_practicality_of_using_blockchain_technology_in_different_use_cases_in_the_healthcare_sector.pdf.

Wilkinson, T. and Chiu, A. (2016) *Order Compelling Apple, Inc. to Assist Agents in Search.* California. Available from https://assets.documentcloud.org/documents/2722199/5-15-MJ-00451-SP-USA-v-Black-Lexus-IS300.pdf.

Zyskind, G., Nathan, O. and Pentland, A. (May 2015) Decentralizing Privacy: Using Blockchain to Protect Personal Data. In: Anonymous IEEE. 180-184 Available from https://ieeexplore.ieee.org/document/7163223.

Appendix

1. Artefact Code: https://github.com/Starystars67/Blockchain-FileStorage

    1.1. Centralized Blockchain System: https://github.com/Starystars67/Containerised-Blockchain

    1.2. Data Identifiers: https://github.com/Starystars67/Blockchain-FileStorage/blob/main/Server/modules/network.js#L122

    1.3. File Reconstruction: https://github.com/Starystars67/Blockchain-FileStorage/blob/main/Server/modules/network.js#L188

    1.4. Public Key Cryptography System: https://github.com/Starystars67/Blockchain-FileStorage/blob/main/Client/script.js#L234

1.5. Server Voting Procedure: https://github.com/Starystars67/Containerised-Blockchain/blob/master/Server/modules/network.js#L728

1.6. Download Procedure: https://github.com/Starystars67/Containerised-Blockchain/blob/cd37601d07c1d55dd0126a175b5da171c7306ec1/Server/modules/network.js#L188

1.7. Upload Procedure: https://github.com/Starystars67/Containerised-Blockchain/blob/master/Server/modules/network.js#L108

1.8. Server Start-Up Procedure (Primary Voting): https://github.com/Starystars67/Containerised-Blockchain/blob/master/Server/modules/network.js#L619

1.9. Primary/Secondary Server System: https://github.com/Starystars67/Containerised-Blockchain/blob/master/Server/modules/network.js#L560

1.10. HTTPs Protocol System: https://github.com/Starystars67/Blockchain-FileStorage/blob/main/Server/modules/network.js#L31

2. Feedback Form: https://docs.google.com/forms/d/e/1FAIpQLSdlgVdFv8S9WqJIBO4sARAX1cCmn7Ec_dtlJk4WFc7nbPg3sg/viewform?usp=sf_link

    2.1. Feedback Answers: https://docs.google.com/spreadsheets/d/1BLL7FxhVgh9z9UU5PBDcLN0bVSPwkgXUoIGOB19dYK0/edit?usp=sharing

3. Feedback Results: https://docs.google.com/spreadsheets/d/1BLL7FxhVgh9z9UU5PBDcLN0bVSPwkgXUoIGOB19dYK0/edit?usp=sharing

4. Engineer Test Approach: https://universityoflincoln-my.sharepoint.com/:x:/g/personal/16610086_students_lincoln_ac_uk/EQoxXH-LZ3ZNm-aXT4U5N0MBmI1azMLal1mds75Kb4-DdQ?e=HxfyfA