ATHABASCA UNIVERITY

SAFEGUARDING PATIENTS' HEALTHCARE DATA:
INTRODUCING DATA PRIVACY BROKER

BY

SAMSON MIHIRETTE

A THESIS

SUBMITTED TO THE FACULTY OF GRADUATE STUDIES

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE IN INFORMATION SYSTEMS

FACULTY OF SCIENCE AND TECHNOLOGY

ATHABASCA, ALBERTA

DECEMBER, 2021

Approval of Thesis

The undersigned certify that they have read the thesis entitled

**SAFEGUARDING PATIENTS' HEALTHCARE DATA:
INTRODUCING DATA PRIVACY BROKER**

Submitted by

Samson Mihirette

In partial fulfillment of the requirements for the degree of

Master of Science in Information Systems

The thesis examination committee certifies that the thesis
and the oral examination is approved

Supervisor:
Dr. Qing Tan
Athabasca University

Committee Member:
Dr. Dunwei Wen
Athabasca University

External Examiner:
Dr. Richard Johnstone
Information Consultants

December 13, 2021

Dedication

This project is dedicated to the memory of my father who always wanted me and my siblings to be lifelong learners and encouraged us to keep learning and contribute our best to the community. It is also dedicated to my mother whose continuous support and encouragement has got me through the process.

Acknowledgments

First, I would like to express my sincere gratitude to my professor and supervisor, Dr. Qing Tan, whose guidance, support, encouragement, and direction enabled me to develop an understanding of the subject from the very beginning through the end.

Finally, I like to thank and offer my regards to my colleagues, friends, and family and all who supported me in any respect throughout the period when working on this thesis. My wife Aynalem and my two daughters Bethel and Eden always encouraged me throughout the years of my studies and working on my thesis.

Abstract

To strengthen the healthcare data privacy protecting techniques and ensure the transparency of healthcare data exchange between healthcare stakeholders, various types of data privacy-preserving methods are introduced continuously, and this elevates the privacy concern of ultimate data owners. This thesis highlights privacy concerns and introduces techniques and research directions towards data privacy on healthcare data in Healthcare Information Systems. The thesis uses the power of Shapley Additive exPlanations (SHAP) machine learning algorithms to identify critical data elements that can put personal privacy at risk within a dataset and proposes a patient-centric healthcare information system architecture with a data broker based on what the industries refer to as Cloud Access Security Broker (CASB.) The proposed data privacy broker is inspired by the CASB and middleware integration applications used for financial and administrative purposes by industries. The privacy broker leverages application programming interface services and integration middleware in safeguarding healthcare data privacy.

Keywords: CASB, Data Privacy, Data Privacy Broker, Healthcare, SHAP

Table of Contents

List of Tables

List of Figures and Illustrations

Chapter 1 – Introduction

Data privacy has a deep historical root and is discussed and explained among philosophers, sociologists, psychologists, and legal scholars. All those disciplines agree that there should be a limit on authorities, government, or private sector companies on the use of citizens' personal data. Citizens have the right to know who can access their private or personal information and for what purpose. Data privacy has been a big concern in this digital era. There are now many laws, regulations, policies, agreements, and guidelines across the world to address the concern and to secure the data privacy of citizens. Technological innovations, social media, and centralized information platforms such as cloud computing have contributed to the urgent need for protecting data privacy. Healthcare data is one of the most sensitive and private information of citizens that is available in those centralized information systems.

A good recent example of the concern on healthcare data privacy is the COVID-19 track and trace strategy using citizens' personal data. Some potential elements of this could be used elsewhere. According to some articles, up to 60 percent of COVID information transmissions happen before someone is aware, they are infected (O'Connell, n.d.). For numerous reasons such as COVID-19, the system is normalizing mass surveillance of citizens therefore keeping citizens' personal data secured will continue to be a pressing issue and hence our research in this thesis contributes architectural solution to safeguard healthcare data privacy and introduces a state-of-the-art patient-centric solution that makes the awareness of the movement of data transparent and timely. Although leveraging every possible angle in controlling pandemics is beneficial towards the safety and health of the public and for effective health policy decisions during outbreaks, citizens' healthcare data needs to be accessed in an ethical and secured manner.

Furthermore, as more and more organizations are migrating their IT workloads to the cloud infrastructure, they are trying to find ways and means on hardening the security and privacy of data and comply with their organizations’ IT security guiding principles.

Data privacy is recognized as a fundamental human right and a number of jurisdictions have enacted privacy and data protection laws and organizations are obliged to be governed under all relevant international human rights treaties.

Research Background

As technology develops, Internet of Things (IoT) technologies, smart medical devices, sensors, and wearable devices are used extensively for healthcare-related purposes. The use of healthcare data has been vital for the medical research community, health policymakers to improve the quality of healthcare information systems. Therefore, while leveraging the healthcare data for the good of society, there is also a responsibility of using the healthcare data in compliance with the data privacy regulations and all the stakeholders should have a transparent mechanism of monitoring the use of their healthcare information.

Furthermore, the use of medical devices such as the Internet of Health Things (IoHT) forces the use of a central cloud-based medical information system and therefore healthcare data privacy will continue to be a concern and the research community needs to continue to innovate patient-centric data privacy techniques and methodologies. Hence, the techniques are applicable for the data residing on both on-premises and cloud platforms.

Due to the nature of the data stored in the healthcare information system and the sensitive nature of data – the data goes to the very core of a human being therefore when storing those types of data in a central system or a cloud platform, a very careful and secured method should be implemented. In the cloud platform where there is less control over remote storage data, the existing security measures are inadequate for the e-healthcare systems. When the healthcare providers publish their data, they need to make sure that all the key sensitive identifiers such as name, address, health card identification number, and social insurance number need to be preserved to maintain the privacy of data.

As compared to financial records, personal medical records can be damaging in nature if they end up in the wrong hands. Personal medical records such as date of birth, address, medical history cannot be altered by the owner of the data if they are stolen.

As a result, there is a need for a robust solution for identifying data elements in a dataset that are crucial for protecting data privacy and a transparent solution that can help monitor and notify stakeholders on the usage of those private data in a timely manner.

Research Purpose

The purpose of this research is to enhance the healthcare information system by including a more transparent data privacy-preserving component into the system architecture to ensure that there is a data privacy-aware mechanism in using personal healthcare data. The research also uses machine learning algorithmic techniques to identify and classify privacy-sensitive data elements within healthcare data. In short, the purpose of this research is to create an architecture component that provides monitoring mechanisms on healthcare data exchange among all stakeholders.

Although some data anonymization methods such as data masking, encryption, and tokenization can mitigate the risk of compromising data privacy in the cloud platform, cloud customers, patients and all healthcare data stakeholders should have a way to monitor and control their own healthcare data at any given time. Patients should have the right to know who can access their personal information in a timely manner and cloud service providers should be transparent in providing the information to their customers. As healthcare data (Electronic Health Records – HER) needs to be modified or updated very frequently, to ensure proper access control, a number of technologies and research studies have been found on strong authentication, the query set size restrictions, inference control techniques, and more robust anonymization techniques.

The following are characteristics of Healthcare Information Systems that make healthcare data special and highly sensitive in terms of privacy.

- Healthcare data (medical records) contains intimate details about people's physical and mental health information.
- Enhancing Healthcare data privacy protection promotes effective communication between physician and patients.
- Healthcare data promotes a good relationship between healthcare providers and patient.
- Healthcare information systems with strong healthcare data privacy protection are an important part of quality care.
- Enhanced autonomy.
- Preventing economic harm, embarrassment, and discrimination.
- In some societies, it is a matter of individual rights, personal choice, and a private sphere protected from intrusion.

Because of these characteristics of the healthcare data and the rapid growth of technology, digitization of healthcare data and medical research, it is imperative to conduct rigorous research within the area of secured data exchange between healthcare providers, healthcare officials, patients, researchers, and all other stakeholders. In protecting the privacy of data, a substantial amount of research has been conducted around data encryption and data anonymization. There are also some research materials around the use of encryption keys (public key and private key) for exchanging data among different healthcare units (Sharma et al., 2018).

Some vendors in the industry have introduced the concept of Cloud Access Security Broker (CASB) to address the visibility and policy enforcements into the cloud environment and make sure that the cloud environment is more secured. However, CASB has not been studied extensively by the research community (*What Is a CASB (Cloud Access Security Broker)?*, n.d.).

Regardless of where data is stored on-premises or cloud platforms it needs to comply with data privacy guidelines and principles. For example, recently there are some public concerns on handling public health information such as - controlling the COVID-19 pandemic has amplified the urgency for effective healthcare data management within the government and healthcare officials in getting accurate information in a timely manner so that decisions can be made swiftly to contain outbreaks, and to make proper decisions and effective policies. The campaign has already begun extensively around the use of cell phone data for identifying and tracking patients affected by the virus. On the other hand, citizens are seeking a reliable method of managing and taking ownership of their own data on healthcare information systems, and users need to know who can access the data for what purpose.

In this research, in helping create a balanced approach between securing healthcare data privacy and approved use of healthcare data for the purpose of analytics by healthcare officials, (balance data privacy with utility) an extensive study has been conducted in the benefit and an additional middleware has been introduced into the healthcare information system based on service-oriented architecture (SOA) built-in with the use of Application Programming Interface (API). A high-level architecture component of middleware on healthcare data privacy broker was developed including the techniques from CASB to provide a solution in protecting people's privacy from healthcare data.

Research Question

Due to healthcare data's sensitive nature, there is a great concern about healthcare data privacy from the public. Therefore, this research strives to provide a solution to the concern on how to protect healthcare data from authorized or unauthorized access. If the data can be accessed without considering data owners' privacy, it could lead to an unexpected impact on the data owner. For example, if employers may use an employee's healthcare data for employment administration purposes, it could impact the employee's position, promotion, or employment benefit, if insurance agents can also exploit the healthcare data of their clients for their own benefits.

As Yuval Noah Harari said - "If corporations and governments start harvesting our biometric data en masse they can . . . not just predict our feelings but also manipulate our feelings and sell us anything they want – be it a product or a politician."(Harari, 2020) Therefore, we need to be mindful on the privacy of citizens' healthcare data.

The two main research questions this thesis addresses:

1. *What is the status of the patient-centric healthcare information system in terms of data privacy protection?*
2. *What are the existing data privacy-preserving solutions in the healthcare information system, and can they are enhanced by introducing a patient-centric data privacy broker?*

In solving the question of finding a patient-centric solution, we are presenting data tracking tools or a robust architecture model to help implement ongoing data privacy and security policies in the healthcare information technology infrastructure. This solution will help patients or any stakeholder in the healthcare domain to update, delete, or insert healthcare information at any given time. Then, we propose a robust transparent solution that will prevent healthcare data privacy breaches. The solution will also provide patients or any stakeholders a proactive awareness of all healthcare data movement or access to sensitive information in the healthcare information system. Our research aims to enhance healthcare data privacy-protecting methods in the healthcare information system.

Research Hypotheses

The research hypotheses are:

1. *We can use machine learning algorithms to identify and model influencing data elements for the purpose of data privacy protection.*
2. *Centralizing data privacy protection processes by introducing the data privacy broker into the healthcare information system can be an effective way to enhance personal privacy protection.*

Having a bulletproof method of protecting healthcare data privacy breaches is not always guaranteed. By verifying the research hypotheses, we will propose a data privacy broker as a solution that could provide a transparent and robust way of controlling the movement of healthcare data within the healthcare infrastructure.

Research Objective

These are four objectives of the research in this thesis:

1) *To answer the research question.*
2) *To understand the existing healthcare information system in respect to data privacy and identify the areas that require attention.*

   We use machine learning algorithms to identify and classify data elements in a dataset that requires attention for safeguarding privacy; to identify data elements within a dataset that are crucial for exposing the personal content.
3) *To design architecture to ensure there is transparent data communication between healthcare stakeholders within the healthcare information system.*
4) *To verify the research hypotheses.*

We introduced an architecture including a component that monitors healthcare data movement among stakeholders and notifies data accesses for protected and sensitive data to all stakeholders including the patient.

To successfully achieve the research objectives, we will answer the research question and be able to verify the research hypotheses.

In our research, we contributed a healthcare information system with architecture that includes a data privacy broker element in the healthcare domain to narrow the gap between healthcare data privacy and the effective and secured use of the data.

To get a balance between data privacy and the use of healthcare data for protecting the public and scientific research might be still challenging even with a complex data privacy protection method. There is always a trade-off between healthcare data privacy and utility.

In our research, we studied the ongoing crucial requirement of finding a robust patient-centric solution to the healthcare information system regarding data privacy protections. We are contributing to ensuring that healthcare data is protected and used responsibly by healthcare providers, researchers, and policymakers.

Research Methodology

In our research, we conducted a thorough literature review with both exploratory and case studies. We delivered a solution that can help mitigate the risk of data privacy breaches, and we also studied the digital healthcare information system of Ontario - the eHealth Ontario. Athabasca University's Online Library has been used extensively in the research for gathering peer-reviewed journal papers and conference materials within the 2014 and 2021 timeframe. Materials on the internet and white papers have been used to understand the architecture of the current healthcare information system in the industry and government organizations. We also studied machine learning algorithms to identify data element contributions to machine learning prediction models and then included the algorithm to the data privacy protection. We used open-source health datasets to validate algorithms.

In the preliminary and exploratory phases of our study, we used a mixed research methodology (qualitative and quantitative research methodology), studied the healthcare dataset and understand the privacy aspect within the Canadian perspective. We used data analysis approaches empowered by machine learning algorithms to make sense of the data elements in datasets with respect to data privacy. In the explanatory phase, we conducted an extensive study within academic research materials, industry practices and areas that require additional future work.

In the case studies portion, we studied and analyzed the current and target state architecture of the Ontario healthcare online system (e-Health Ontario.)

Research Significance and Contributions

The main contributions of this thesis are:

1) Identify and classify model or research output influencing data elements in dataset that require attention for safeguarding privacy with the use of SHAP machine learning algorithms. These data elements within a dataset are crucial for exposing the personal content and influencing the prediction. We used machine learning algorithms to classify data elements based on the degree of each element's potential contribution to exposing the privacy of the entire dataset or a record within the dataset.
2) The introduction of the state-of-the-art novel Healthcare Privacy Data Broker component as a middleware in the healthcare information system to ensure there is a proper and transparent patient-centric data privacy issue prevention in the exchange of data between systems and applications. This novel Healthcare Data

Privacy Broker is also integrated with all the technologies that are currently used for cloud security – Cloud Access Security Broker (CASB).

In introducing the healthcare information system architecture with the Healthcare Data Privacy Broker, we explained and documented the objectives of the novel component, functionalities, and the capabilities of each component. We also provided algorithms that can be used in the systems in the infrastructure (both in the healthcare applications, repositories, and the healthcare data privacy broker.

Research Scope and Limitations

The scope of this work is limited to the data privacy issues in the healthcare information system. The research work focused on machine learning algorithms to identify crucial data elements within a dataset that contributes highly to models and have the potential for protecting the privacy of a dataset. The research leveraged a state-of-the-art healthcare data broker and a middleware process integration system, but the work is not focused on building a healthcare application.

Organization of the Thesis

This thesis is organized as follows: In Chapter 2, the literature review and related works are presented. The literature review portion is categorized by different data privacy methods the academic research community studied, encryption, data anonymization, data perturbation and authentication, blockchain approach, differential privacy, and fog computing methods. The major concepts or technologies of data privacy methods are summarized. In the literature review, two major integration architectures are also presented – Amazon Web Service and e-Health Ontario architecture views.

Chapter 2 also discusses the relevant works done by many researchers in the field of privacy and security of healthcare data through extensive literature review. It looks at problems currently being experienced in these fields and summarizes the findings from this research.

Chapter 3 focuses on the research in terms of architecture design. It develops the methodology and instrumentation of the study. Chapter 4 discusses the healthcare data privacy broker. Chapter 5 analyzes the use of machine learning algorithms in data classification and provides the high-level design of the healthcare infrastructure by including the data privacy broker. Chapter 6 provides trends, and challenges around healthcare data privacy-protecting techniques and suggests areas in which further work can be focused. And chapter 7 provides conclusions and recommendations.

Table 1

Terminology

Term	Definition	Explanation
SOA	Service-Oriented Architecture	The Modern and innovative way used for exchange data between applications
REST	Representational state transfer	Web service API that provides interoperability between applications
Process Integration	A middleware that integrates applications	A middleware that can be used for exchange data
e-Health Ontario	Ontario online healthcare infrastructure	
API	Application Programming Interface	Used highly for exchanging data between applications
AWS	Amazon Web Services	Cloud Service Provider and provides other online services.
EHR	Electronic Health Records	
SOAP	Simple Object Access Protocol	messaging protocol specification for exchanging structured information in the implementation of web services
SFTP	Secure File Transfer Protocol	A protocol used to exchange data between applications securely
CASB	Cloud Access Security Broker	On-prem to cloud solution package consists of security mechanisms.
Privacy	Medical privacy or health privacy is the action of maintaining the security and confidentiality of patient data or records.	

Term	Definition	Explanation
Data privacy	Data that belongs to an individual and the individual is willing to share to the people he/she wants. Data privacy is about data that has a footprint on an individual personal information	
Data privacy broker	A device that makes sure that when private data is accessed by an authorized or unauthorized party, it will make sure to monitor and notify all stakeholders.	
Healthcare data	Healthcare data is any data related to health conditions, cause of disease, clinical information, etc. It can also be administrative-related data about a patient.	
Healthcare Dataset	A healthcare dataset is a list of records about patients and their information such as type of disease, the reason for death and other administrative, medical, and clinical information.	
Data record	The data record is a record of an individual patient, in database terms it is a row in a database table.	
Data element	Data element is one or more attribute of a record. In database terms, it is a field or attributes in of a record or group of records. A field in a database table.	

Chapter 2 – Literature Review and Related Work

In our research and literature review process within the time of 2014 and 2021, a significant number of research papers focus on the following main topics or methods in mitigating the risk of healthcare data breaches and creating a healthcare information system with a strong data privacy management system:

- Encryption and decryption methods (AES-256 or RSA)
- Encryption using cryptographic algorithms
- Homomorphic encryption
- Data anonymization methods
- Masking, de-identification, randomization, pseudonymization and manipulation
- Multi-factor authentication
- Service level or data-sharing agreement between stakeholders
- Adopting blockchain technology for healthcare information system
- Data perturbation methods
- The trust relationship between stakeholders
- Differential privacy methods
- Fog computing architecture
- Security middleware
- Detect privacy violations and eliminate inferences.
- Proactive privacy measures – removing private data during extraction for Big Data.
- Hardware-based protection of data - Trusted Execution Environment

We used various academic journals and conducted systematic reviews to identify the latest advancement published and previous research with respect to healthcare data privacy issues and recommendations proposed to protect data on-premises and in the cloud platform including third-party cloud environments (interconnected cloud environments) also known as cloud federations. In my research, I have conducted research on data privacy issues in both traditional relational database management systems and Big Data.

With the goal of integrating healthcare mobile devices with cloud computing, the research community mainly focuses on how to effectively distribute data across multiple data centers on the cloud and develop efficient and effective algorithms that allow better reliable communication between the cloud and mobile devices.

There are a substantial number of research papers discussing the subject of healthcare data and options in mitigating the risk of data breaches and data privacy protections. The extreme importance of data privacy concerns in data management and Big Data across cloud data centers is also recognized among IT practitioners and academicians in various literature and information technology forums in various angles.

A number of academic papers greatly emphasized the level of healthcare data privacy importance due to the distributed nature of data across multiple regions in the form of virtual machines, storage and computing components and indicates both cloud service providers and customers are mutually responsible for safeguarding data in the cloud by providing the usual security measures that are used on on-premises infrastructures.

Research Encryption and Decryption Method

In their research paper, Sagar Sharma, Keke, Chen and Amit Sheth discussed the modern healthcare system and its progress towards practical privacy-preserving analytics for IoT and cloud-based healthcare systems. The researchers based their study on kHealth – a personalized digital healthcare information system that is developed and tested for disease monitoring. They discussed potential trade-offs among privacy, efficiency, and model quality. The researchers propose solutions based on the cost of each method. The suggested methods are privacy-preserving computation on untrusted platforms – fully homomorphic encryption (FHE) and Yao's garbled circuit (GC together with oblivious transfer (OT). After the recommendation, they concluded that it might not be possible to implement the best healthcare modelling algorithms in privacy standards because of numerous trade-offs and restrictions and further concluded that modern healthcare frameworks including kHealth can overcome privacy-preserving and yet be practical if the right privacy building blocks and the application scenarios are identified and implemented. The researchers agree that preserving data from adversary parties in a healthcare information system without affecting data utility, model learning, and data sharing are challenging. The researchers indicate that ideal data privacy is almost unachievable throughout the storage, processing, and communication phases. They discussed the potential of AES-256 encryption methods to protect data from potential adversaries that could seriously jeopardize computations and services in providing useful analytics data. Although simpler anonymization techniques can help make the use of data use from a modelling or analytics perspective, it is somehow inadequate against data privacy breaches (Sharma et al., 2018). The researchers also discussed the use of third-party methods such as homomorphic encryption schemes, data perturbation, and differential privacy techniques to mitigate data privacy risks. Finally, the researchers further discussed the need for additional methods on what needs to be done to continue to reduce the risk of data breaches by monitoring the data consumption by all healthcare data stakeholders and thus their proposal needs to be supplemented by a more robust data privacy enforcer architecture.

Data Anonymization Method

Researchers Benjamin Eze, Craig Kuziemsky, Liam Peyton in their research paper, "Operationalizing Privacy Compliance for Cloud-Hosted Sharing of Healthcare Data"(Eze et al., 2018), have showen how_Data Sharing Agreement (DSA)_can be used to operationalize privacy compliance for a cloud-hosted surveillance and performance management infrastructure by leveraging selective anonymization approach based on patients concept and some types of governance to protect data privacy. In their paper, the researchers emphasized anonymization should be an integral part of healthcare data sharing among stakeholders along with a proper form of patients' consent and governance; however, the researcher emphasizes novice techniques that help ensure the privacy of healthcare data is protected.

Researchers Kingsford Kissi Mireku, Fengli Zhang and Komlan Gbongli, discuss the importance of healthcare data privacy in their research paper "Patient Knowledge and Data Privacy in Healthcare Records System," which focuses on the consent of patients when accessing healthcare data. They used Pearson Correlation and linear regression analysis techniques to determine the relationship of the knowledge with data privacy. The researchers concluded that patients could play a role to mitigate the risk of monitoring and eavesdropping. On patient knowledge and data privacy in healthcare records system(Kissi Mireku et al., 2017). Then the researchers suggest integrating social forces and factors as an enabler of healthcare record protection and conclude the requirement for extensive research to examine big data, data privacy and healthcare records. In their research, they referred, "anonymization algorithms such as k-Anonymity, l-diversity, t-closeness mostly deal with the removal of identifiable information in data mining is widely accepted in privacy preservation but unfortunately, the anonymization algorithms do not completely preserve privacy." (Kissi Mireku et al., 2017) And at the same time, a high degree of anonymization might make the data not fully usable.

Researchers, Harsh Kepware Pail and Ravi Seshadri examined_how patients' knowledge about data privacy, Big Data in the healthcare record system contributes to the overall protection of healthcare records. These researchers agree the healthcare industry harnesses the power of Big Data, security, and privacy. They also discussed the impact of the Internet of Things (IoT) and their ability to provide real-time monitoring and expedited access management system (Kupwade Patil & Seshadri, 2014) and agreed that resource-exhaustive operations while preserving privacy is a challenge. They also discussed the use of distributed processing across cloud platforms and leverage collective intelligence however there is no proposal on what technique should be used to tighten healthcare data privacy. Their main goal in the research thesis is to show the relationship of reactive and proactive healthcare data to healthcare costs and the economy in general.

Researchers, Silvina, Pitsillides, Rossbory, Vinov, Legay, Blackledge and Huand, in their research paper, “The SERUMS tool-chain: Ensuring Security and Privacy of Medical Data in Smart Patient-Centric Healthcare Systems” agree in order to minimize cost and embrace technology to provide a great response time for patients’ diagnostics and treatments that the future of healthcare system is highly distributed and decentralized home environment and this forces the sharing of healthcare data among different stakeholders and therefore there is a threat of data privacy. To mitigate privacy breach risk, the researchers presented a design and proof of concept of a Securing Medical Data in Smart Patient-Centric Healthcare Systems (SERUMS) toolchain for accessing, storing, communicating, and analyzing highly confidential medical data. The researchers included Blockchain technology, data masking, data encryption, data fabrication, a noise-adding mechanism for differential privacy and secure authentication components in their SERUMS tool-chain architecture. They also included a use case to support their proposal(Janjic et al., 2019). The research paper is very extensive, comprehensive, and sound. If the solution provided in the paper is supplemented with a patient-centric tracking solution it will further tighten the data privacy and transparency of the healthcare infrastructure.

Research Data Perturbation Method and Authentication

In their paper “A Privacy Framework in Cloud Computing for Healthcare Data,” researchers Kundalwal, Singh and Chatterjee - to prevent impersonation attacks, introduced data perturbation based on privacy techniques for healthcare data in the cloud platform(Kundalwal et al., 2018). In their paper, they tried to look at two angles: 1) randomizing the result of the query by using a hybrid data perturbation approach. 2) a rule-based approach to ensure if the record is accessed by a legitimate user to resist impersonation attack in the cloud database. In their research, they proposed data perturbation and rule-based authentication.

In their research paper, "PHeDHA: Protecting Healthcare Data in Health Information Exchanges with active Data Bundles", Fadheel, Salih and Lilien propose a Health Information Exchanges (HIEs) called PHeDHA that provides privacy and security protection for patient data during transmission between multiple healthcare providers with security policy enforcement engine virtual machine(Fadheel et al., 2018). In the paper, the researchers explained how their proposed framework protects the transmission of healthcare data.

Elmisery, Rho and Botvich, in their paper "A Fog Based Middleware for Automated Compliance With OECD Privacy Principles on the Internet of Healthcare Things", proposed the idea of a personal gateway that acts as an intermediate node (called fog nodes) between the IoHT devices and the cloud-based healthcare services to preserve the privacy of end-users data.(Elmisery et al., 2016) The researchers presented a two-stage concealment process, which provides complete privacy control to patients. Although the researchers presented a middleware that handles data privacy which is helpful to have control between the device and the cloud platform, the framework still needs to include other robust data privacy control techniques to ensure a patient-centric data privacy technique. The researchers, at the end of their research, noted the need for a future research agenda on utilizing game theory for a better formulation of a group of IoHT devices, sequential vital measurements release and the impact on data privacy.

Li and Pino, in their research work, "D&D: A Distributed and Disposable Approach to Privacy Preserving Data Analytics in User-Centric Healthcare"(Z. Li & Pino, 2019) acknowledge that privacy-preserving measures, such as Data Encryption, Data Perturbation, and De-identification have been considered inadequate to completely address the diverse privacy challenges in big healthcare data analytics. In their research, they proposed a strategy of data analytics routine in a distributed and disposable manner.

In their research work, "An efficient access privacy protocol for healthcare patient information system", Sharif, Li, Ullah, Haq and Alam, proposed a two-factor authentication protocol that can be used by both patients and healthcare providers to access healthcare data and ensure the data access is controlled and protected(Sharif et al., 2019). Although the two-factor authentication method is useful for information communication between two parties in a decentralized healthcare information system, in addition to the two-factor authentication method a more robust and complex information system is required in a centralized and modern healthcare information system in order to better serve healthcare patients, preserve data privacies and help healthcare providers to leverage the healthcare data and come up with a better statistical model, healthcare research to serve and protect the community at large.

Ghafour, Ghodous and Bonnett, in their research paper, "Privacy-Preserving Data Integration Across Autonomous cloud Services"(Ghafour et al., 2015), acknowledged the need for healthcare data privacy and proposed a service-oriented model (solution) that enforces privacy policies for services involved in answering a query by applying access control models and data anonymization algorithms. The model allows the execution of aggregates (compositions) of data without revealing extra information. This model gives some data privacy however if the healthcare providers and researchers want to leverage healthcare data in the information system, on top of aggregate healthcare data, a solution should be provided to ensure the improvement of the healthcare research work for a better healthcare system.

In their research paper, "Privacy as a Service: Protecting the Individual in Healthcare Data Processing, Su, Maarala, Li and Honko, argued health applications involve multiple data sources, individuals, and services that work against guarantees that an individual's personal data will not be used without consent and proposed privacy-centred architecture that integrates security, a trust-query framework that enables the provision of user consent as a service. Their architecture recommendation is based on privacy-as-a-service (PRIAAS) and MyData principles augmented with General Data Privacy Regulation (GDPR). They explained the User-Managed Access protocol over personal data, content, and services based on the OAuth 2.0 protocol(Su et al., 2016). The architecture mainly emphasizes consent control between the data source and the MyData operators and it also leverages the API technology. This is very extensive research but mainly around consents but in addition to this method, a more robust method is required to ensure patients' data is protected and there is a robust information system that is protected and helps the healthcare information system for the betterment of the society.

Kato Mivule, in his research work "Targeted Data Swapping and K-Means Clustering for Healthcare Data Privacy and Usability,"(Mivule, n.d.) proposed the use of mathematical targeted data swapping and K-Methods to hide sensitive healthcare data however the method hides very important information and might make the data distorted and not useful for healthcare research or other purposes.

Dingh and Jangra, in their research work, "Healthcare Data Privacy Measures to Cure & Care cloud Uncertainties"(Singh et al., 2017), discussed the data privacy issues in terms of architectural measures, technique-based and conventional healthcare data privacy-preserving schemes. They discussed ZigBee e-health monitoring application, secured storage, secure data sharing, and secure computation including conventional encryption such as AES and ECIES. These encryption technologies, if they are supplemented by a transparent data privacy tracking technology, will improve the data privacy protections in the healthcare information system.

Research Adoption of the Blockchain Approach

In their research work – "Studies Using a Universal Exchange Language Solution for Application of a Blockchain Approach in Healthcare."(Robson & Srinidhi Boray, 2016), Barry Robson and Srinidhi Boray studied the blockchain model used primarily for financial data structure and adopt it for electronic health records. There is a good potential to leverage blockchain technology for healthcare data privacy however there are several challenges for the main reason that healthcare information is not at the level of financial information in terms of standard therefore there is a problem of interoperability and integration.

The research work: "Blockchain: A Panacea for Healthcare cloud-based Data Security and Privacy?" by Esposito and De Sentis recognize that the convenience of the cloud infrastructure is making the healthcare information system shift to the cloud and recognize the limitations of the cloud infrastructure in terms of security and privacy and the researchers highlight the potential of the use of Blockchain technology to protect the healthcare data hosted in the cloud(Esposito et al., 2018). They also presented the challenges of adopting Blockchain technology to the healthcare information system. Although the Blockchain is serving in protecting the financial information system, it has its own challenges and limitations. The major benefits of blockchain technology are not third-party brokers, data owners have full control of data and any change to the blockchain is visible to the entire blockchain network, however, there is a challenge in altering or deleting data that is already in the chain.

In their research paper, "A Review on the State-of-the-Art Privacy-Preserving Approaches in the e-Health clouds", Abbas and Kan(Abbas & Khan, 2014), after recognizing the concern of healthcare data privacy, proposed healthcare information system design by first classifying healthcare data into cryptographic, non-cryptographic proposed a design depending on the usage of the data. They also discussed the challenges their proposal consists of. They presented a comparative analysis on data privacy-protecting approaches and concluded that, despite all the efforts to enhance healthcare data privacy techniques, there are still issues and require more attention within the area of security and privacy of data in the cloud environment. They mentioned, although there is some protection within the healthcare data from privacy threats, the provenance of healthcare data might reveal sensitive patient data to unauthorized parties and raised an important research question around the research on healthcare data monitoring, tracking and accountability.

In their research paper, "Privacy-Preserving Scoring of Tree Ensembles: A Novel Framework for AI in Healthcare", Fritchman and Saminathan explained the use of "privacy preserving machine learning (PPML)"(Fritchman et al., 2018). The researchers provide a secure multiparty computation (SMP) technique that can be used to protect the potential privacy problems that come because of Big Data and AI. The technique proposed in the research paper is more on how to mitigate the risk of healthcare data privacy when using ML for collecting data, but the research doesn't include or propose a robust method for mitigating data privacy monitoring and accountability.

In their research paper, "SPOC: A Secure and Privacy-Preserving Opportunistic Computing Framework for Mobile-Healthcare Emergency,"(Lu et al., 2013) Lu and Lin, proposed a secure and privacy-preserving opportunistic computing framework called SPOC, for Mobile-Healthcare Emergency with the objective of minimizing the disclosure of personal data during mobile healthcare emergency situations. Their proposal SPOC framework consists of three parts: system initialization, user-centric privacy access control and analysis of opportunistic computing in a mobile healthcare emergency. Within these three parts, they included the idea of trust authorities, mathematical algorithms for patient location when a healthcare emergency is required, and opportunistic analysis for arrival time based on the location of the patient.

Puppala, He, You and Chen, in their research work, "Data Security and Privacy Management in Healthcare Applications and Clinical Data Warehouse Environment"(Puppala et al., 2016), studied a case study in the Houston Methodist church and proposed a security and privacy model that consists two components: the enterprise data warehouse (EDW) and a software intelligence and analytics (SIA) layer and they conducted their research to indicate that patient privacy is protected best by implementing a mix of technologies such as de-identification of records and restriction of data access. Their approach is mainly on traditional access management control, security schemes and healthcare application policies. However, although the approach maintains the traditional security and privacy standards, it lacks to contribute additional novel healthcare data privacy methodology.

Differential Privacy Method

Alnemari, Romanowski and Raj, in their research work, “An Adaptive Differential Privacy Algorithm for Range Queries over Healthcare Data”(Alnemari et al., 2017), presented a mathematical differential privacy algorithm to explain the vector partitioning into small ranges for healthcare data queries improves data privacy. The paper contributes to improving vector partitioning algorithms and providing a consideration of sensitivity before providing query results. Although the research substantially contributed to protecting data privacy by integrating a differential privacy technique, the method needs to be completed with more data privacy algorithms to mitigate privacy risks.

In their research paper, Privacy-preserving Smart IoT-based Healthcare Big Data Storage and Self-adaptive Access Control System, Yang, Zheng and Guo conducted research on healthcare data privacy and proposed a secured system with a two-fold access privilege, in the emergency application, patient’s historical medical data and using password-based access mechanisms. Although their proposal introduces a new way of accessing patients’ data by integrating the traditional encryption mechanisms, it still needs to be supplemented by a new patient-centric healthcare data privacy framework to further restrict and monitor the movement of healthcare data.

In their research paper, “No Free Lunch in Data Privacy,” Kifer and Machanavajjhala argue that differential privacy is a powerful tool for providing privacy-preserving noisy query answers over statistical databases.(Kifer & Machanavajjhala, 2011) They explained how differential privacy helps in answering queries over sensitive datasets without compromising the privacy of personal records in the information system. They also mention that it is not possible to guarantee privacy and utility without making assumptions about the data-generating mechanisms and presented their argument on techniques and algorithms for inferences in mathematical theorems. These researchers explained in detail how differential privacy helps in hiding sensitive or protected records but if the tools of differential privacy techniques are integrated with a more privacy framework or architecture, there will be an optimal trade-off between data privacy and utility within the healthcare information system.

After recognizing the security threats to the healthcare-related data, Xu, Wei and Wu, in their research work, "Privacy-preserving data integrity verification by using lightweight streaming authenticated data structures for healthcare cyber-physical system," proposed a privacy-preserving data integrity verification model by using lightweight streaming authenticated data structures for Health-Cyber Physical Security (CPS.)

Tracking the data privacy concerns of smart healthcare and remote devices, Laarje, Fatiha and Bouhorm, conducted research on IoTs based wireless devices in their research paper, "Protecting E-healthcare Data Privacy for Internet of Things Based Wireless Body Area Network," (Rghioui et al., 2015) presented data communication using security keys for symmetric cryptography in order to ensure the privacy of healthcare devices or sensors in the context of IoT. In using security keys, their proposal mitigates the risk of unauthorized access to healthcare data however for the central healthcare system their model did not include a patient-aware framework to make the entire information system transparent.

Fog Computing-based Method

In the research paper, "A Security Model for Preserving the Privacy of Medical Big Data in a Healthcare Cloud Using a Fog Computing Facility With Pairing-Based Cryptography,"(Al Hamid et al., 2017) the researchers explained the importance of big data for keeping healthcare-related data such as x-rays, ultrasounds, CT scans and MRI reports in a cloud platform. And the use of Big Data will continue to grow however there is a big concern of data theft. Therefore, they proposed a tri-party one-round authenticated key based on the bilinear pairing cryptography that can generate a key among the healthcare data stakeholders. The researchers introduce a fog computing algorithm to add a security mechanism to the infrastructure. Their approach is great in creating a middleware between information system components but to ensure that the data communication between stakeholders is tracked, the architecture still needs to be supplemented with additional security and data privacy algorithms or framework.

In the research paper on “Fog Computing for Smart Healthcare data Analytics: An Urgent Necessary”, by Khaloufi, Abouelmehdi and Beni-Hassane(Elmisery et al., 2016), the researchers explained the use of Fog Computing an extension to cloud computing for the performance reasons implemented between smart devices and the cloud infrastructure for a healthcare information system. They also discuss the importance; challenges of data privacy and security including Fog gateways and emphasize the concern around patient privacy and the Fog Computing architecture creates additional concern for preserving data privacy.

Privacy Around Big Data

In the research paper, “Security Solutions For Big Data Analytics in Healthcare,”(Rao et al., 2015) having realized the importance of a secured mechanism in keeping the privacy of patient’s healthcare data, the researchers proposed a secured solution for big data pertaining to healthcare. In their proposal, they included the traditional methods such as data de-identification, security models at the application level, and the creation of a dynamic recommendation engine that recommends sensitive data and vulnerabilities. Their approach still complements the healthcare information system in terms of keeping security and data privacy, but the research does not include a modern innovation and integrated approach in keeping the healthcare information system impenetrable in keeping the privacy of information.

To ensure a secured healthcare information system, researchers proposed a middleware in the e-healthcare system, in their research paper “A Support Middleware Solution for e-Healthcare System Security.”(Bruce et al., 2014) Although this research’s goal is to propose a middleware solution for security it only covered mainly authentication mechanisms and did not cover the healthcare data privacy framework.

Daniels, Rose and Farkas proposed a framework to detect privacy violations and implement removal of undesired inference from data, in their paper – “Protecting Patients’ Data: An Efficient Method for Health Data Privacy.”(Daniels et al., 2018) They presented a sound framework to remove data inferences that leads to the threat of data privacy however, as suggested by the researchers, the framework needs to be enhanced with a more patient-centric method for visualization.

In the research paper, “Customized privacy-preserving for inherent data and latent data,”(He et al., 2017) the researchers introduced novel inherent data privacy and latent data privacy to combat data against powerful third-party users from data inference attacks. The researchers use mathematical techniques and algorithms to propose a data sanitization strategy so that there is a trade-off between utility and privacy. Although their proposal gives a trade-off, further research is required to ensure there is an additional method in making healthcare data useful by ensuring data privacy protection.

In the research paper, “Sociological Aspects of Big Data Privacy,”(Yuxuan et al., 2020) the researchers analyzed the sociological aspect of data privacy in relation to Big data – sociology theories such as anomie theory, culture conflict theory, social control theory, and social exchange theory to understand the privacy of big data. The researchers explained the data privacy concerns in detail in sociological terms and proposed technical solutions within the area of anonymity, data encryption and differential privacy. Although the research consists of theoretical implications of data privacy breaches, it did not propose a novel data privacy solution.

Further to privacy around traditional database management systems, data privacy issues are also discussed extensively among the research community in respect to Big Data. Therefore, a large amount of healthcare data such as x-rays, medical results and others leverages Big Data for healthcare data storage. I have also read a number of research papers about data privacy on big data in my research.

José Farnesio Huesca Barril and Dr. Qing Tan, in their research paper " Integrating privacy in architecture design of student information system for big data analytics"(Barril & Qing Tan, 2017), conducted research on the privacy of data when integrating a hybrid information system using on-premises and cloud infrastructure for relational database and Big Data within Educational Data Mining (EDM) domain for a school integrated information system. In their research, they explained that most EMD tools do not have integrated privacy in their architecture design and proposed a design that includes a privacy component. Their proposal focuses on a proactive approach in the data collection phase and includes the idea of stripping Personal Identifiable Information (PII) during data extraction to protect privacy. The approach of stripping PII on for EDM can also be adapted to healthcare data privacy however we still need to keep the balance of privacy and the use of the healthcare data for medical research and health policy decision making.

Application Architecture Validation

Validating architecture can be performed in multiple methods depending on the complexity of the proposal. In my literature review, I went through a number of research materials that are written on architecture validation.

“Architecture-Based Testing and System Validation”(Tekinerdogan et al., 2011), a paper that summarizes architecture-based testing and system validation presented at the 9th working IEEE/IFIP conference on software architecture.

“Design principles in Test Suite Architecture”,(Nishi, 2015) explains notations or concepts for test architecture such as Unified Model Language (UML) test profile and NGT. The paper also discusses on Based on the notations, research on modelling is necessary next. They include design principles, patterns, frameworks, tools, methodologies etc. In their research, they showed the ten design principles in test architecture: coupling, cohesion, maintainability,

automatability, circumstance consistency, development consistency, decidability, design direction, design positiveness and execution velocity consistency.

“IEEE Standard for System, Software, and Hardware Verification and Validation”(*IEEE Standard for System, Software, and Hardware Verification and Validation*, n.d.), in this paper the researchers discussed the importance of verification and validation (V&V) in determining the development of software products of a given activity that conform to the requirements of the activity. V&V life cycle process requirements are specified for different integrity levels. The scope of V&V processes encompasses systems, software, and hardware, and it includes their interfaces.

“Model-Driven Validation Of System Architectures,”(Pfluger et al., 2011) in this research paper, the researchers use the model-driven process approach to validate a given architecture that meets the requirement.

“Study of C4ISR Architecture Simulation Validation with UML and Object-based Petri Nets,”(Xiaohui Bai, 2008) this research explains the dynamic behavior of a system that can be displayed by a simulation model, and whether the relation of time-order, information flow and the interface are reasonable, or the system function satisfies the military requirement, and the change of system state is correct, can be checked evidently. The paper also provides a case in the air defense system using the simulation validation method proposed by it.

"System Architecture Validation with UML"(Pflüger et al., n.d.), this research paper uses an approach that defines a model-driven process for the architect to validate system architecture against system requirements based on UML. It supports the architect in designing the architecture and in analyzing the impacts of requirements changes.

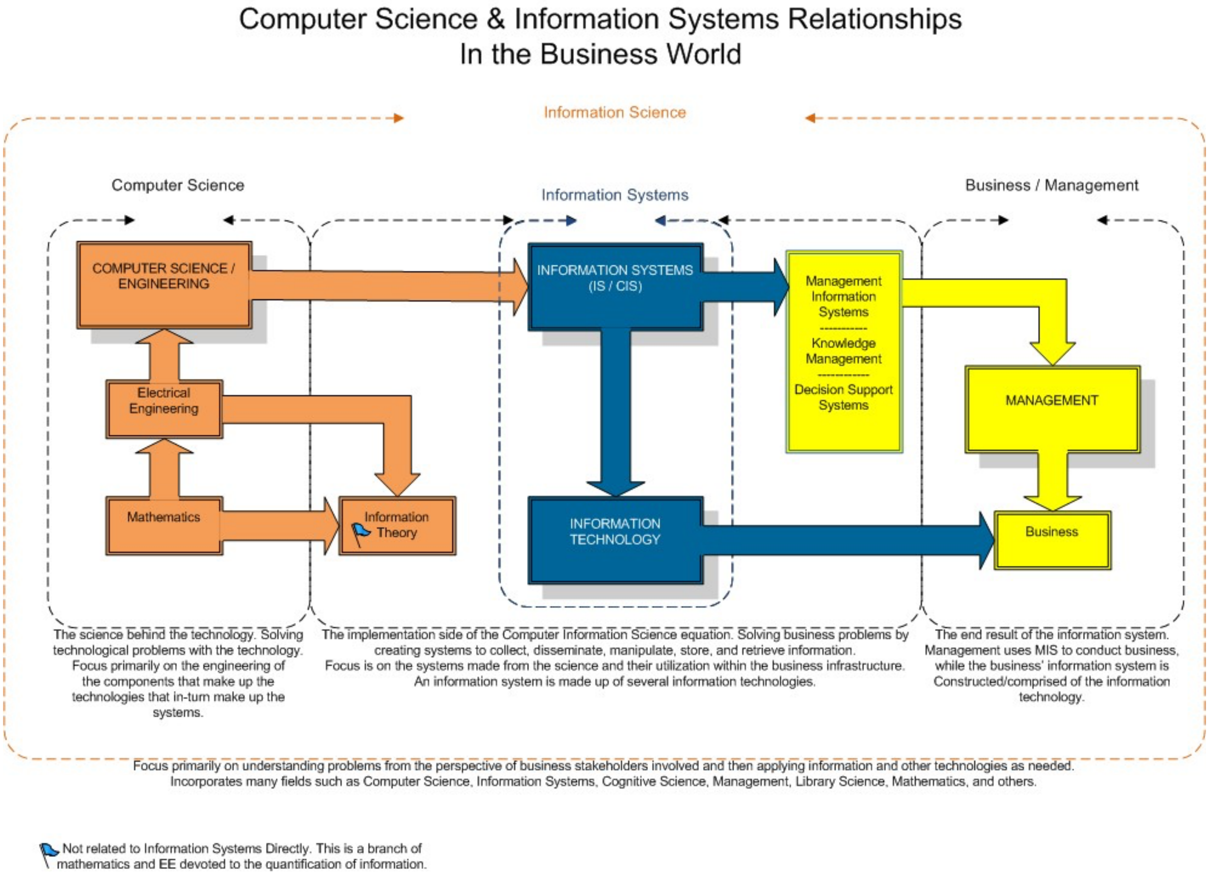
"Automatic and Continuous Software Architecture Validation," (Goldstein & Segall, 2015), in this research paper the researchers propose and implement a solution for automatic detection of architectural violations in software artifacts. The solution, which utilizes several predefined and user-defined patterns, does not require prior knowledge of the system or its intended architecture.

In our research, we have also conducted application architecture deliverables in the public service to understand the current practice around fundamentals of architectures in different development methodologies (Waterfall and Agile.)

The following figure explains the relationship between mathematical models and information systems, and this can be used in building software that fulfills requirements.

Figure 1

Computer Science and Information Systems Relationships



The science behind the technology. Solving technological problems with the technology. Focus primarily on the engineering of the components that make up the technologies that in-turn make up the systems.

The implementation side of the Computer Information Science equation. Solving business problems by creating systems to collect, disseminate, manipulate, store, and retrieve information. Focus is on the systems made from the science and their utilization within the business infrastructure. An information system is made up of several information technologies.

The end result of the information system. Management uses MIS to conduct business, while the business' information system is Constructed/comprised of the information technology.

Focus primarily on understanding problems from the perspective of business stakeholders involved and then applying information and other technologies as needed. Incorporates many fields such as Computer Science, Information Systems, Cognitive Science, Management, Library Science, Mathematics, and others.

Not related to Information Systems Directly. This is a branch of mathematics and EE devoted to the quantification of information.

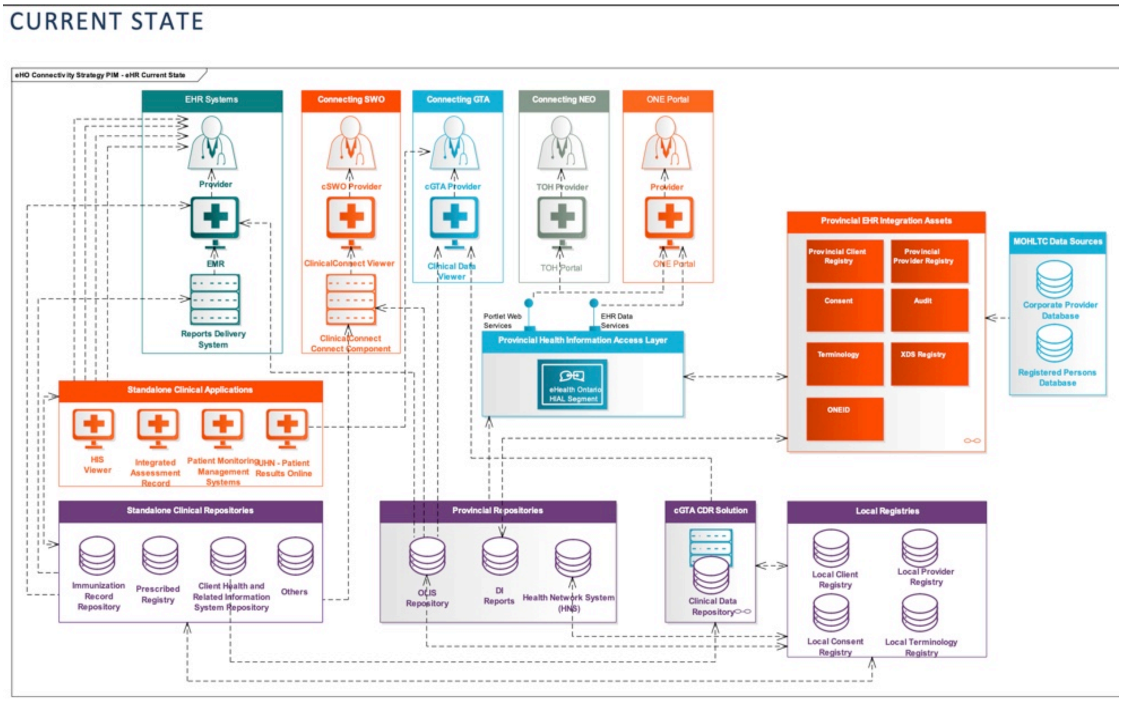
Note: details on the relationship between computer science and information system in the industry.(Azimi, 2019)

E-Health Ontario Related Work

The Ontario provincial government, in its effort to modernize the healthcare infrastructure, has set up a digital healthcare infrastructure called e-Health Ontario to manage patient care using electronic health records (EHR) for Ontario's 13.6 million residents. The healthcare infrastructure has improved the quality and access to healthcare. In my research in electronic healthcare infrastructure, I conducted an in-depth analysis of the architecture of eHealth Ontario.(*EHR_Connectivity_Strategy_Summary-En.Pdf*, n.d.) The infrastructure has proper data access monitoring, client registry, provider and consent registries, audit services for privacy and security for all transactions. The infrastructure is a very robust healthcare information system, and the government is also planning to improve the infrastructure and I closely studied the final target state architecture. Although the solution has a significant number of functionalities, privacy detecting measures, it can also be improved further in terms of protecting patients' records and maximizing the use of healthcare data for public health policy measures and medical research. Here is the current state architecture of the Ontario eHealth connectivity. The architecture has integration components, and the government is also planning to enhance the architecture in the future; however, to ensure the data privacy of patients, the architecture needs to be enhanced and supplemented with architecture and a platform that has a more transparent and patient-centric mechanism

Figure 2

Reference Architecture – Current State EHR Assets and How EHR Information is Viewed


CURRENT STATE
eHO Connectivity Strategy PIM - eHR Current State
EHR Systems
Provider
EMR
Reports Delivery System
Connecting SWO
cSWO Provider
ClinicalConnect Viewer
ClinicalConnect Connect Component
Connecting GTA
cGTA Provider
Clinical Data Viewer
Connecting NEO
TOH Provider
TOH Portal
ONE Portal
Provider
ONE Portal
Portlet Web Services
EHR Data Services
Provincial Health Information Access Layer
eHealth Ontario HIAL Segment
Provincial EHR Integration Assets
Provincial Client Registry
Provincial Provider Registry
Consent
Audit
Terminology
XDS Registry
ONEID
MOHLTC Data Sources
Corporate Provider Database
Registered Persons Database
Standalone Clinical Applications
HIS Viewer
Integrated Assessment Record
Patient Monitoring Management Systems
UHN - Patient Results Online
Standalone Clinical Repositories
Immunization Record Repository
Prescribed Registry
Client Health and Related Information System Repository
Others
Provincial Repositories
OLIS Repository
DI Reports
Health Network System (HNS)
cGTA CDR Solution
Clinical Data Repository
Local Registries
Local Client Registry
Local Provider Registry
Local Consent Registry
Local Terminology Registry

Note: components within the current state architecture of the E-Health Ontario. (*EHR Connectivity Strategy | EHealth Ontario | It's Working For You*, n.d.)

Trusted Execution Environment

The increasing popularity of connected devices and the prevalence of technologies, such as cloud, mobile computing, and the Internet of Things (IoT), have strained existing security capabilities and exposed gaps in data security (Lowans, 2020). Organizations that handle Personally Identifiable Information (PII) must mitigate threats that target the confidentiality and integrity of either the application, data on transit across the network, data at rest on multiple storage devices or the data in system memory (The Confidential Computing Consortium, 2021). As a result, Gartner predicts, by 2025, 50% of large organizations will adopt privacy-enhancing computation (PEC) for processing data in untrusted environments and multiparty data analytics use cases (Gartner, 2020). Of the several PEC techniques, Trusted Execution Environment is the only one that relies on hardware to accomplish its privacy-enhancing goal.

A Trusted Execution Environment (TEE), or Secure Enclave as they are sometimes known, is an environment with special hardware modules that allow for a secure area inside the device. It runs in parallel with the operating system, in an isolated environment. Input is passed into the TEE and computation is performed within the TEE ('secure world'), thereby protected from the rest of the untrusted system ('normal world'). These secure and isolated environments protect content confidentiality and integrity, preventing unauthorized (*Confidential Computing Consortium - Open Source Community*, n.d.) access to, or modification of, applications and data while in use.

While protecting sensitive data poses significant architecture, governance, and technology challenges, using a TEE may provide a starting point for an alternative means of enhancing security from the lowest level. However, a TEE is not plug-and-play; it is a technically challenging mechanism that should be reserved for the highest-risk use cases (Lowans, 2020). Nonetheless, it is certainly harder to steal secrets from inside a secured TEE than from the unsecured 'normal world.' Thus, when a TEE is used as a secondary defense mechanism or to protect low-level secrets, it makes a lot of sense. In these cases, it makes the attacker's job harder, and that is always a good thing (Lindell, 2020).

The term 'confidential computing' is often used synonymously with TEE. While they are related, as per the Confidential Computing Consortium (CCC)(*Confidential Computing Consortium - Open Source Community*, n.d.), confidential computing is the protection of data in use by performing the computation in a hardware-based TEE. Of note, the definition of confidential computing is independent of topographical location (no mention of cloud, a user's device, etc.), processors (a regular processor or a separate one), and whether encryption or some other isolation technique is used (or not used).

Importantly, a TEE […] can be applied anywhere including public cloud servers, on-premises servers, gateways, IoT devices, edge deployments, user devices [(e.g., Smartphones or watches)]. It is also not limited to being done by any processor, since trusted processing might also be in various other places such as a graphic processing unit or network interface card; neither is it limited to solutions that use encryption (The Confidential Computing Consortium, 2021). Edge computing is distributed computing technology where information processing is located close to the edge, which is where things and people produce and/or consume that information.

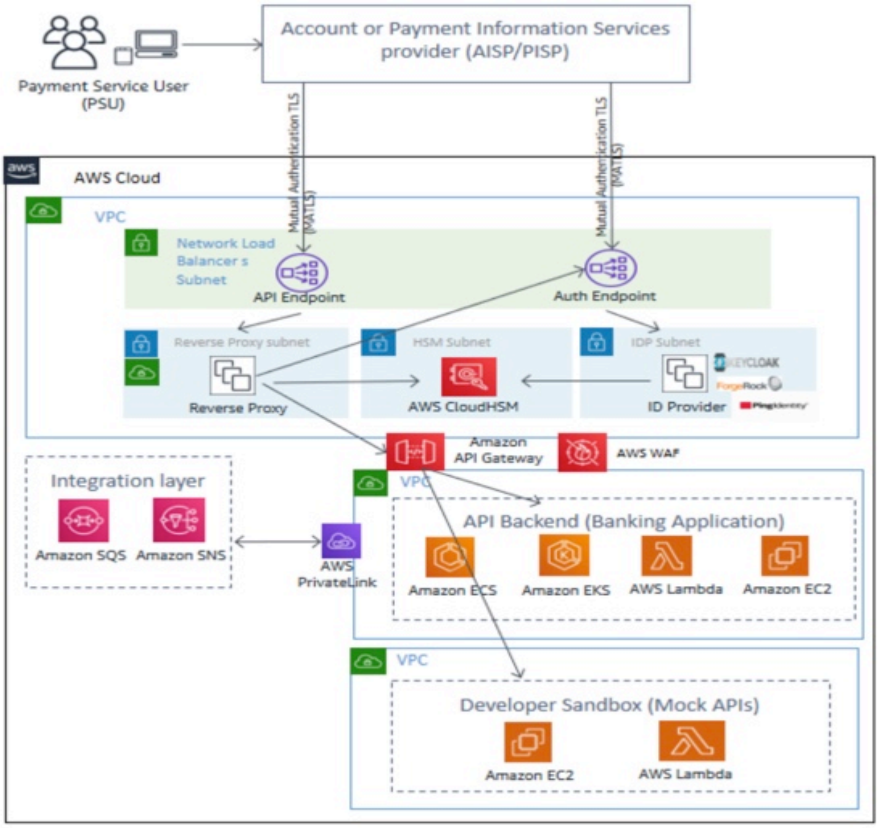
TEEs are provided by solutions such as Intel's Software Guard eXtensions (SGX) or ARM's TrustZone; via hardware vendor Software Development Kits (SDKs); or with abstraction layers (e.g., Google's Asylo) that eliminate the requirement to code explicitly for a TEE.
Many cloud vendors (e.g., Alibaba, Microsoft, IBM, and Oracle) are now providing TEE capabilities as a dedicated low-level service aligned with their computation offerings. However, the specifications offered by cloud vendors should be considered carefully (Fritsch, Bartley, & Ni, 2020).

Amazon Web Services – Related Architecture

The financial, manufacturing and service sectors leverage the Application Programming Interface (API) technology on service-oriented architecture (SOA) platform to exchange data between applications in an integrated and secure manner. To provide an example, here is a reference architecture for Amazon Web Services (AWS) accounts payable system, a company that provides on-demand cloud computing platforms and APIs to individuals, companies, and governments.

Figure 3

Reference Architecture AWS



Note: This AWS reference architecture provides the integration of different information system components using APIs. (*T*he AWS Security Reference Architecture - AWS Prescriptive Guidance, n.d.)

As shown in the above reference architecture (*AWS Well-Architected Framework Financial Services Industry Lens*, n.d.), Amazon Web Services’ accounts payment information service, AWS is leveraging the API technology, API gateway, reverse proxy and the integration layer component to ensure there is a very efficient integration of resources. Within this integration architecture, there are internal privacy mechanisms that clearly monitor the movement of data across different components in the infrastructure. This API integration method and architecture can also be leveraged, adapted, and enhanced for other disciplines such as a healthcare information system.

In the research paper, “Privacy-aware Big Data Warehouse Architecture,”(Navuluri et al., 2016) Mukkamala and Ahmed argue that there is a lack of comprehensive system-wide solutions and that privacy continues to be a major concern in the big data era. In their research, they describe a privacy-aware architecture for big data warehouses that provides controlled data access to users based on data owner-specific privacy policies – Embedded Privacy Agreement (EPA). A data owner defines EPA at the dataset level (database table, relation, file). Their proposal architecture introduced a middleware called warehouse server that represents a logical entity that stores the data coming from data owners and provides data to clients by respecting privacy policies. They have proposed a novel architecture model that introduces a middleware to control the privacy of the data movement among all the stakeholders. Their proposal contributes a new privacy model, but it needs to be enhanced and adapted to the healthcare information system to further ensure data privacy in the healthcare infrastructure.

Frej, Dichter, and Guptan, in the research paper, “Comparison of Privacy-Preserving Models based on a Third-Party Auditor in Cloud Computing”(Frej et al., 2019), discussed and Third Party Auditor (TPA) privacy-preserving models proposed their own Light-Weight Accountable Privacy-Preserving model (LAPP). They also discussed cloud infrastructure security vulnerabilities with respect to lack of trust – authentication and loss of control –

privacy. Their research is mainly within the area of cloud auditing and reporting vulnerabilities and discussing the strength and limitations of each TPA and presenting comparative analysis. The paper presented a novel model by providing encryption for secure authentication and secret sharing mechanisms. Although this model has added a security mechanism to the infrastructure, a more robust data privacy technique is needed to make sure there is a patient-centric information system in place.

Federated Learning

Federated machine learning is designed based on a machine learning framework that allows a collective model to be constructed by using data distributed across repositories or databases owned by different organizations or devices. The use of data and models in building across organizations and devices while meeting applicable privacy, security and regulatory requirements is provided within federated learning. Industries and research communities struggle to improve the efficacy of machine learning models to strengthen the weak model training practices.

Chapter 3 – Research Design and Descriptions

After conducting in-depth research on healthcare data privacy-preserving architecture, models, and algorithms, it seems imperative to continue the innovation of more complex data privacy- preserving solutions so that the sensitive and personal information of citizens should fall under all the ethical standards. Therefore, to contribute state-of-the-art architecture to the healthcare information system, we have conducted research on the following architecture component to be included in the healthcare infrastructure. We provided a Secured Data Privacy broker as a middleware and integration component in the healthcare information system infrastructure. This novel architecture contribution enhances the existing data privacy-preserving techniques recommended by researchers - techniques such as encryption, anonymization, differential privacy, and others.

Overall Design

In our research thesis, we also included other methods such as differential privacy, classification of data and encryption methods within the architecture. The foundation of this novel architecture is based on a service-oriented architecture that utilizes the application programming interfaces (API) extensively. These APIs will be used for communication between two systems and their databases.

We also conducted a study on the use of API, forward and reverse proxies and designed a middleware integration solution that will help enhance the healthcare data privacy capability in relation to the cloud environment. This integrated architecture is implemented and will be used among healthcare providers, cloud providers, patients, and other healthcare data stakeholders. Our research contribution includes how healthcare clients access their healthcare information on the cloud, how healthcare officials access patients' records in the cloud environment for online processing, analytics, and reporting purposes.

In our in-depth research, we conducted studies and provided methods and mechanisms around the following main items:

- Efficient methods of identifying data elements within data sets that are key for data privacy protection. Once those data elements are identified, classification of data privacy will follow. Then based on the privacy classification, effective and transparent data privacy awareness techniques follow.
- The secure health data broker notifies users or patients of any access to personal data by any of the parties using regular reports. Depending on the nature of the data, the system can also be configured to seek user approvals before data is provided to external parties.
- The secure health data broker provides a list of external bodies for which the data is provided.
- If a patient wants to delete or modify a health record from the healthcare data repository, it can be done through the repository. Verification or any query by the patient can be conducted through the broker.
- Data mapping is completed in the broker to mask or hide any private data the patient doesn't want to share.
- The healthcare data repository can only supply data to the data broker.
- Audit and access logs are captured at the secured data broker system and timely reports will be given to the data owner on-demand or on a scheduled basis.
- Roles and responsibilities should be clearly defined and isolated between the healthcare data repository and the secure data broker in the cloud environment. For security and privacy, it is preferable to have these components in their own cloud platform. Segregation of duties is important for better security and privacy of data.

- The secured health data broker is built on a service-oriented architecture (SOA) and uses secured adaptors such as RESTful services, web services, SFTP for all inbound and outbound messages.
- A very clear Service level agreement (SLA) between different components of the healthcare information systems needs to be prepared and agreed upon by the major stakeholders. For example, between the parties who are responsible for maintaining the secure data broker and the data repository.

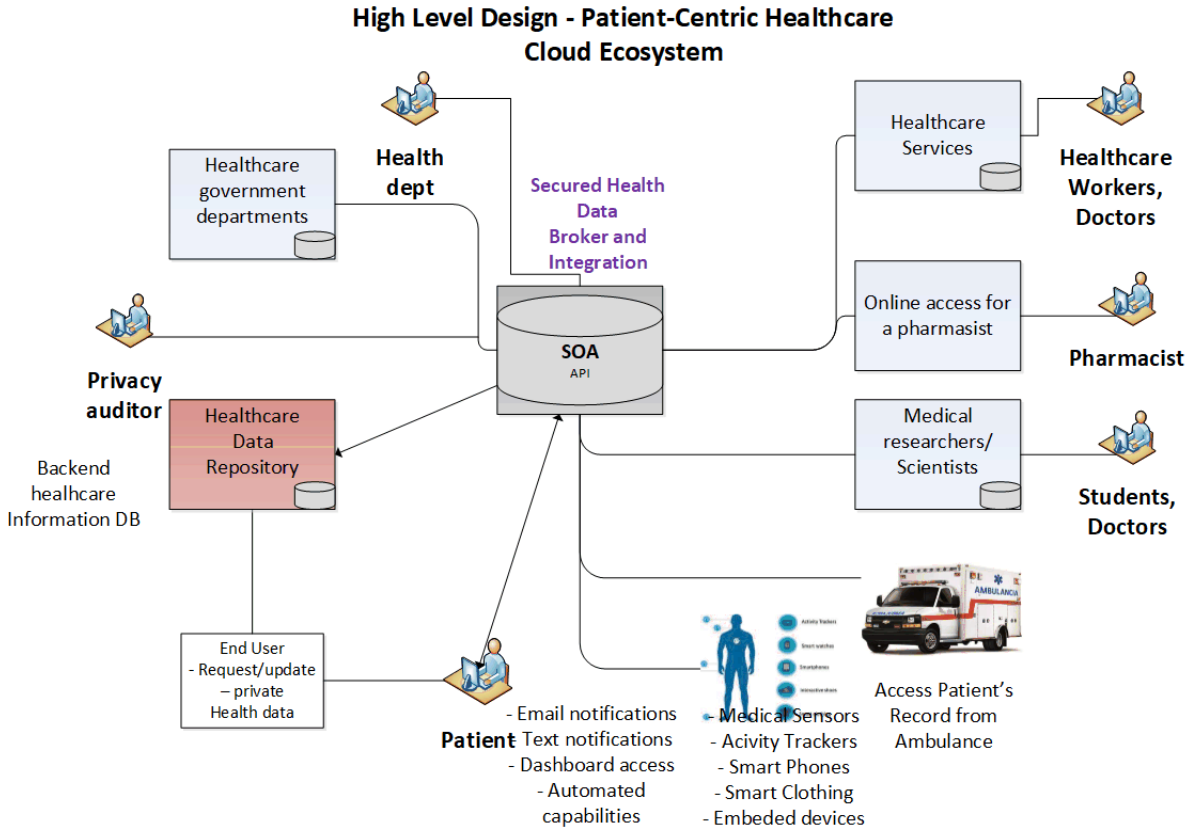
Apart from introducing a state-of-the-art secured data broker, as an integration middleware, I have conducted a study focusing on a technology that can help ensure data privacy in the healthcare information system in the cloud computing platform. In my recommendation, I covered the capabilities of Cloud Access Security Brokers (CASB)(*What Is a CASB (Cloud Access Security Broker)?*, n.d.) to support and enhance the data privacy architecture for the healthcare information system.

High-Level Architecture

The following is a high-level and simplified summary of information system and architecture models in our research. All communication or in-transit data between cloud environment components is encrypted. Data can only be accessed from the secured data broker, not from the healthcare repository.

Figure 4

High-Level Architecture for Healthcare Information System with Secured Health Data Broker


High Level Design - Patient-Centric Healthcare Cloud Ecosystem
Health dept
Healthcare government departments
Secured Health Data Broker and Integration
Healthcare Services
Healthcare Workers, Doctors
Online access for a pharmasist
Pharmacist
SOA
API
Privacy auditor
Healthcare Data Repository
Backend healhcare Information DB
Medical researchers/ Scientists
Students, Doctors
End User - Request/update – private Health data
Patient
- Email notifications
Text notifications
- Dashboard access
- Automated capabilities
- Medical Sensors
- Acivity Trackers
- Smart Phones
- Smart Clothing
- Embeded devices
Access Patient’s Record from Ambulance

Note: A sample conceptual view of a healthcare information system with secured healthcare data broker and integration component.

Component’s Capabilities in the Design

The centralized secure data broker, a system based on service-oriented architecture, consists of many capabilities within the area of data movement monitoring, integration, firewall, and IT security mechanisms. Here are some of the high-level capabilities of the Secure Data Broker component in the infrastructure.

- Synchronous and asynchronous integration with data repositories or any healthcare application.
- Web service interface between any frontend and backend healthcare applications.

- Central system to control and monitor data flows between different applications.
- A system that triggers a notification to data owners for approval or awareness.
- Transform or map data for efficient integration.
- Provide a run-time environment for exchanging data between systems and interface monitoring capabilities.
- Execute integration workflows with a series of steps, such as to request and approval.
- Connecting systems using secured communication protocols such as SFTP, SOAP, HTTPS, mail, and others.
- Capture audit, access, and system logs.

In the following chapters we will discuss a couple of state-of-the-art contributions to enhance the privacy protection processes for healthcare information systems:

- Identify and classify model or research output influencing data elements in the dataset that require attention for safeguarding privacy with the use of machine learning algorithms. These data elements within a dataset are crucial for exposing the personal content, we used machine learning algorithms to classify data elements based on the degree of each element's potential contribution to exposing the privacy of the entire dataset or a record within the dataset.

- architecture design to include a data privacy broker in the healthcare information system landscape.

Chapter 4 – Healthcare Information System Architecture With Data Privacy Broker

For several years software architecture was represented in terms of boxes and lines however in the 1990s software architecture started to be an integral part of the software development life cycle. A very good software architecture design can lead to a very excellent software product. A software product that started with a poor architecture can result in a very disastrous result from the withdrawal of a project to substantial financial loss. There are several types of software architecture: objected-oriented architecture, agent-oriented architecture, and service-oriented architecture.

Today's information technology software architecture tries to solve the following three main questions:

- What are the needs?
- Is it achievable?
- How does it work?

To come up with an effective healthcare information system architecture that includes a data privacy broker, we conducted research on the steps of building the design, different components of architecture disciples: application architecture, data architecture and technology architecture.

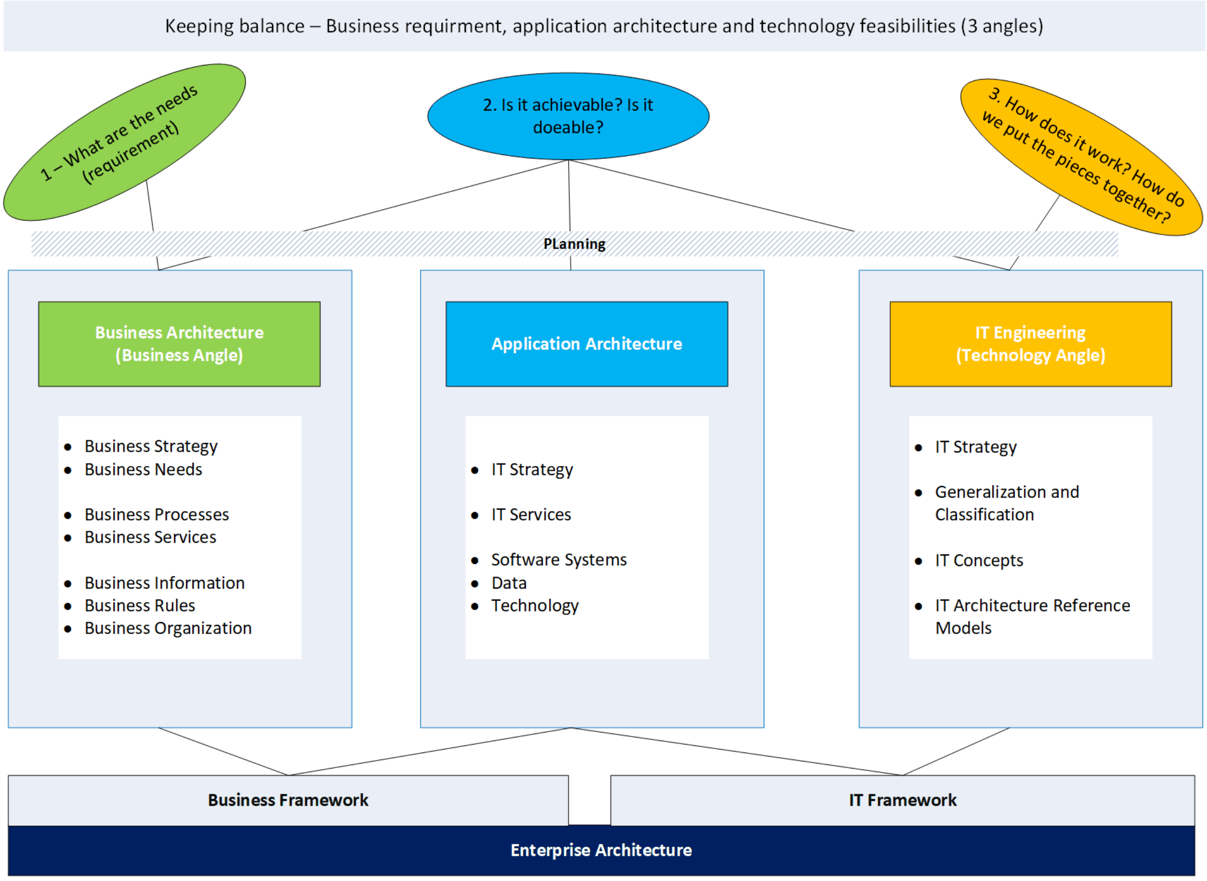
The application architect is responsible for gathering the business requirement and providing architectural solutions that meet the demand.

A data architect is responsible for efficient data flow, and data integrity within the information system.

The technology architect is responsible to provide the technology and tools required to implement the recommended architecture.

Figure 5

Business Requirement, Application Architecture and Technology Capabilities


Keeping balance – Business requirment, application architecture and technology feasibilities (3 angles)
1 – What are the needs (requirement)
2. Is it achievable? Is it doeable?
3. How does it work? How do we put the pieces together?
PLanning
Business Architecture (Business Angle)
Business Strategy
Business Needs
Business Processes
Business Services
Business Information
Business Rules
Business Organization
Application Architecture
IT Strategy
IT Services
Software Systems
Data
Technology
IT Engineering (Technology Angle)
IT Strategy
Generalization and Classification
IT Concepts
IT Architecture Reference Models
Business Framework
IT Framework
Enterprise Architecture

Note: a sample explanation on industries on business requirement, application architecture and technical feasibly based on enterprise architecture principles.

The effective architecture ensures that the IT solution we build:

- Are suited to the purpose for which they are intended
- Fit gracefully in the information technology platform
- Are structurally sound
- Comply with codes, regulations, and standards
- Are defined using the applicable framework, terminology, and classification
- Are sustainable through their expected lifespan

- Are user friendly with all user experience principles (aesthetically pleasing and elegant)

Building the right software architecture requires collaboration between different disciplines: technology architects, data architects, application architects, security architects, database administrators, infrastructure architects and business architects.

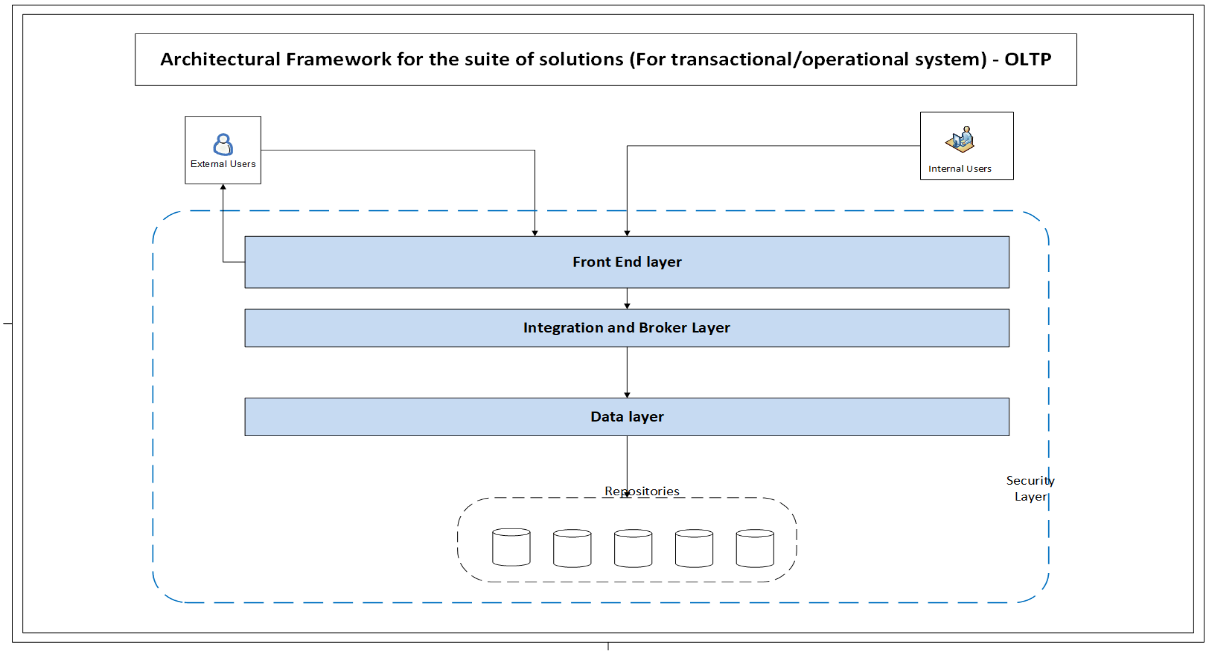
In designing the architecture for solutions, in my thesis, I will split the architecture components including 1) Business Operation (BO) used mainly for Online Transactional Processing (OLTP) systems: architecture for the transactional or operational system – ex. data entry and online processing and 2) Business Analytics (BA) used mainly for Online Analytical Processing (OLAP) systems: architecture for analytics and reporting functionality of the system – ex. data modelling, management report on a dashboard, audit report and forecasting.

Data Privacy Broker Architecture

Data Privacy Broker architecture – Conceptual for Transactional System

Figure 6

Conceptual Representation of the Architecture

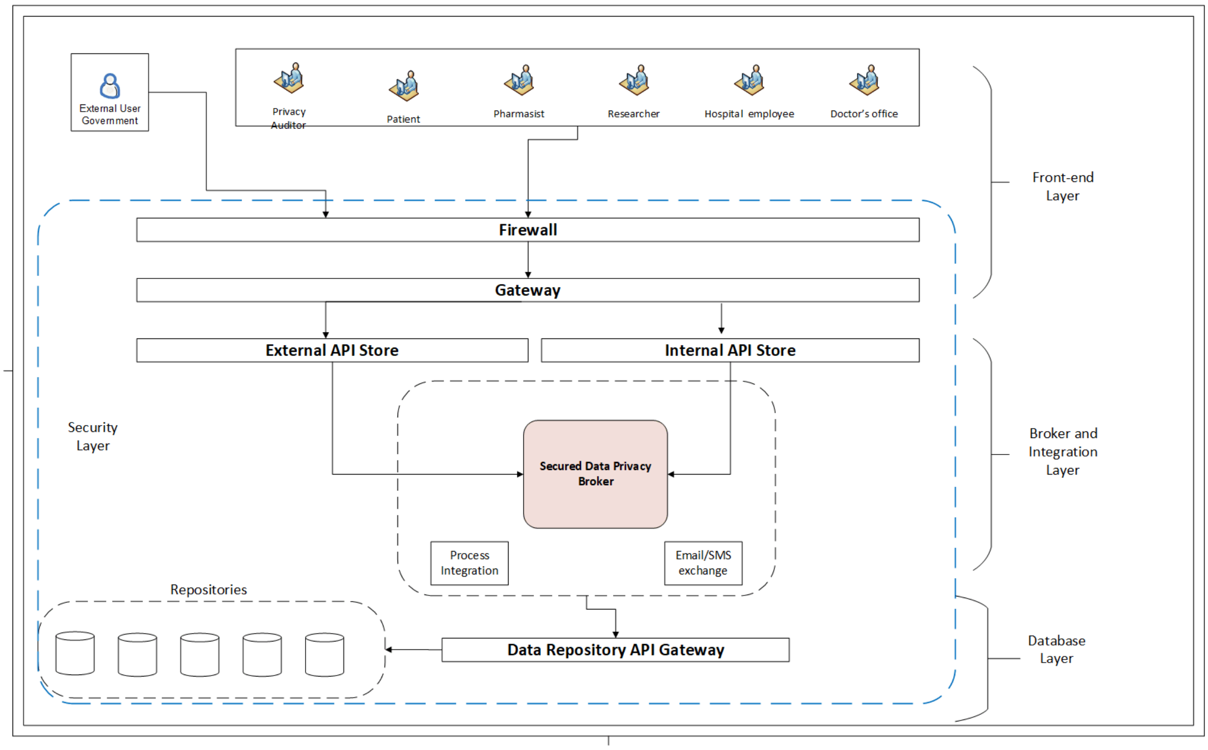

Architectural Framework for the suite of solutions (For transactional/operational system) - OLTP
External Users
Internal Users
Front End layer
Integration and Broker Layer
Data layer
Repositories
Security Layer

Note: a sample conception high level diagram based on three-layer architecture.

Data Privacy Broker Architecture – High-Level Architecture Infrastructure View for Transactional System.

Figure 7

IT Components in the Three Layers of the Architecture



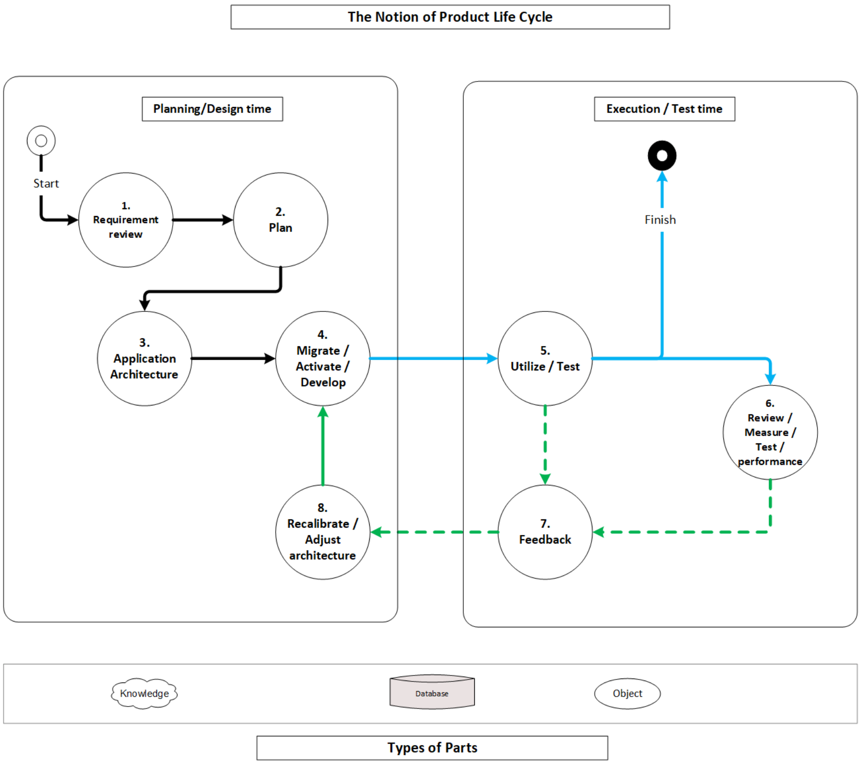
Note: expanded sample to demonstrate the three layers architecture.

Architecture Implementation Process

A holistic approach to software product life cycle and its validation depends on the software development methodology such as waterfall, agile methodologies

Figure 8

Different Steps in the Product Life Cycle


The Notion of Product Life Cycle
Planning/Design time
Start
1. Requirement review
2. Plan
3. Application Architecture
4. Migrate / Activate / Develop
8. Recalibrate / Adjust architecture
Execution / Test time
Finish
5. Utilize / Test
6. Review / Measure / Test / performance
7. Feedback
Knowledge
Database
Object
Types of Parts

Characteristics of Product Life Cycle

These product life cycle phases have different sequences for different lifecycles methedologies:
1) Waterfall – The Define, Build, release and migrate are sequential
2) Agile – there are multiple and repetitive Define, Build Release phases mulitple times.

Waterfaill is more traditional of the two methodologies with stricter process steps. Agile is more adaptive and allows for flexibility during the development processes.

The architecture framework or design will go through more testing and requirement confirmation in multiple phases during the agile methodology and during the test or feedback phase in the waterfall methodology.

Note: a sample demonstration of a product life cycle based on industry practices.

Agile and Water fall software development life cycle approaches:

Agile:

- Architects define the architectural vision, aligned to the projects' goals and the organization's long-term strategic objectives.
- Architects work on epics, user stories, and tasks to provide iterative artifacts.
- Architects need to take the lead in choosing the right tools and technologies.
- Architects should encourage comments and suggestions and should not be protective of ideas.
- Architects need to make iterative design decisions with the project team.

Waterfall:

- Architects define the architectural vision, aligned to the project's goals and the organization's strategic objectives.
- Architects design and create solutions and record them in architecture artifacts.
- Architecture artifacts are extensive and made available before the project's development phase.
- Architects mostly design with predefined rules, guidelines, and standards; only making modifications to baseline standards when required.
- Architecture artefacts have no room for design or scope change.

The data privacy broker will be supported by a number of encryption methods, k-anonymity, l-diversity. I will explain some of these methods in this chapter.

Analysis on Privacy Characterization and Quantification in Data Sharing or Publishing

With the help of inexpensive computing power, vast amounts of storage space and due to the simplicity and accessibility of digitized (electronic) data, a significant volume of data is available over the internet in multiple digital formats. This data is being leveraged for different purposes such as scientific research, market research, statistical modelling, and training. For example, healthcare data is used for scientific research on finding medicines or vaccinations, analysis on healthcare-related data can help forecast the outbreak of a pandemic. Although there is increasing pressure to share healthcare data, there is also a concern of releasing sensitive information that can cause several unwanted results. There are some cases where aggregate of data is used to share information while suppressing data privacy however there is a substantial increase in demand in sharing data and maximizing the use of data for scientific research and analysis.

Several data privacy protecting techniques have been proposed, studied, and privacy-preserving data publishing (PPDP) techniques have been proposed in scientific research. In this chapter, we will discuss some of the well-known PPDP techniques such as k-anonymity, l-diversity, and t-closeness in terms of privacy characterization. In most data privacy protecting techniques, the most obvious personal identifiers will be removed however by joining the remaining data, there is a possibility of identifying, inferring, or exposing the original data.

a) 𝓀-anonymity

One of the popular data anonymization techniques is k-anonymization. For a given data, to satisfy a k-anonymity principle, every record in the table should not be distinguishable from at least k-1 other records in the same table with respect to every set of quasi-identifier attribute and these kinds of tables are called k-anonymous tables.

Equivalence class refers to a set of records that are indistinguishable from each other with respect to certain "identifying" attributes.

Table 1 shows medical records from a fictitious healthcare dataset. The table contains no uniquely identifying attributes like name, healthcare number, social insurance number or driving license number. This example divides the attributes into two groups: 1) the sensitive attributes (medical conditions) and 2) the non-sensitive attributes (postal code, age, salary, and race.) An attribute is considered sensitive if the data in that attribute is exposed it might lead the record to be identified in its fullest or partially. Let collection of attributes {number, postal code, age, salary, race} be a quasi-identifier for this dataset.

$$f\ (Pi) = set\ of\ \{postal\ code, age, salary, race\}$$

To demonstrate how the *𝓀-anonymity* and ℓ-diversity works in protecting the privacy of data, we created the following sample tables of data.

Table 2

Original Sample Data

	Postal Code	Age	Salary	Place of birth	Disease
1	K2S 0G1	67	250,000	America	Lung disease
2	K2S 6T2	35	150,000	Asia	Heart disease
3	K2S 7GS	45	130,000	Canada	Liver disease
4	K6T R5Y	25	100,000	Africa	Cancer
5	K7G H6G	29	25,000	Asia	Cancer
6	K7Y G5S	39	75,000	Africa	Heart disease
7	K2S 0H1	45	98,000	America	Liver disease

Table 3

A 3-Anonymous Version of the Original Sample Data

	Postal Code	Age	Salary	Place of birth	Disease
1	K2S*	>60	> 200,00	America	Lung disease
2	K2S*	3*	>=100,000	Canada	Heart disease
3	K2S*	4*	>=100,000	Asia	Liver disease
4	K6T*	2*	>=100,000	Africa	Cancer
5	K7G*	2*	<=100,000	America	Cancer
6	K7Y*	3*	<=100,000	Africa	Heart disease
7	K2S*	4*	<=100,000	Canada	Liver disease

Table 2 represents the original data with a list of patients, their postal code, age, salary, race, and disease. In this example, the attribute - disease is sensitive information because the information of certain diseases for any individual is very personal and can have an unwanted negative result on the individual's employment or health insurance.

Table 3 shows the anonymized table with the method of 3-anonymous version (*𝓀-anonymity*.) Suppose a person knows an African person who is in his 30's, then it is very easy to find out that the patient has heart disease.

If record linkage, data matching or entity relationship is possible in a dataset, then the privacy of the record on the sensitive data will be possible. Names, birthdays, or other specific information on a record can expose the sensitive information in a table if they are linked with other outside information. For example, partial knowledge of some data outside the table can help link the person to identify the exact sensitive data in the table.

In their paper, "Protecting Privacy when disclosing information: 𝓀-anonymity and Its Enforcement through Generalization and Suppression"(Samarati & Sweeney, n.d.), the researchers illustrate how k-anonymity can be applied when using generalization and suppression techniques in releasing the data by not suppressing the data more than needed.

To address the limitations of the 𝓀-anonymity, ℓ-diversity has been introduced as a stronger notion of privacy.

a) ℓ-diversity

ℓ-diversity is a form of group-based anonymization that is used to preserve privacy in data sets by reducing the granularity of a data representation. This reduction is a trade-off that results in some loss of effectiveness of data management or mining algorithms in order to gain some privacy. ("*L*-Diversity," 2020)

(The ℓ-diversity Principle) An equivalence class is said to have l-diversity if there are at least one "well-represented" value for the sensitive attribute. A table is said to have l-diversity if every equivalence class of the table has ℓ-diversity.(N. Li et al., n.d.)

Machanavajjhala et al (Machanavajjhala et al., 2006) has given a number of explanations for the term "well-represented" in the l-diversity principle. l-diversity introduces an important addition to k-anonymity against attribute identification, but it also has its own weaknesses that we can discuss further.

When the dataset has recorded with the result of a medical lab test (positive or negative) for a list of patients with certain diseases and if the dataset has 99% positive and 1% negative, then it is possible for a given patient's record to be inferred with the help of external data. In this situation, the degree of sensitivity is different when there are records with 99% positive and 1% negative. In a dataset such as this, there will be very easy attribute disclosure.

The main principle of ℓ-diversity is to prevent data leakage or disclosure of information on a specific record by creating groups and ensuring diversity of sensitive values in each group. However, this method still does not prevent information disclosure if the groups have close values similar to records in Table 3b. Therefore, depending on the closeness of the values in the group ℓ-diversity can be exposed to a similarity attack.

Table 4

Original Healthcare Sample Data

	Postal Code	Age	Salary	Place of birth	Disease
1	K2S 0G1	27	250,000	America	Lung disease
2	K2S 6T2	25	150,000	Canada	Heart disease
3	K2S 7GS	25	130,000	Asia	Liver disease
4	K2S R5Y	45	100,000	Africa	Cancer
5	K2S R6G	59	25,000	America	Cancer
6	K2S R5S	39	75,000	Africa	Heart disease
7	K2S 0H1	35	98,000	Canada	Liver disease
8	K2S 0L1	37	109,000	Africa	Stomach cancer
9	K2S 9K1	39	200,000	Asia	Stomach cancer

Table 5

A 3-Diverse Version of Table 4

	Postal Code	Age	Salary	Place of birth	Disease
1	K2S ***	27	250,000	America	Lung disease
2	K2S ***	25	150,000	Canada	Heart disease
3	K2S ***	25	130,000	Asia	Liver disease
4	K2S R**	> 40	100,000	Africa	Cancer
5	K2S R**	> 40	25,000	America	Cancer
6	K2S R**	> 40	75,000	Africa	Heart disease
7	K2S ***	3*	98,000	Canada	Liver disease
8	K2S ***	3*	109,000	Africa	Stomach cancer
9	K2S ***	3*	200,000	Asia	Stomach cancer

Both k-anonymity and ℓ-diversity principles are showing limitations in their anonymization or data masking techniques on privacy characterization; therefore, research needs to continue to ensure there are advanced technologies to protect the privacy of sensitive data by using emerging and evolving technologies.

Figure 9 shows the relationship between direct identifier, indirect identifies and privacy disclosure. The more identifiers are exposed the higher it is to expose the original records in a dataset.

$$f\ (Pi) = f(Di)\ \mathrm{U}\ f(Ii)$$

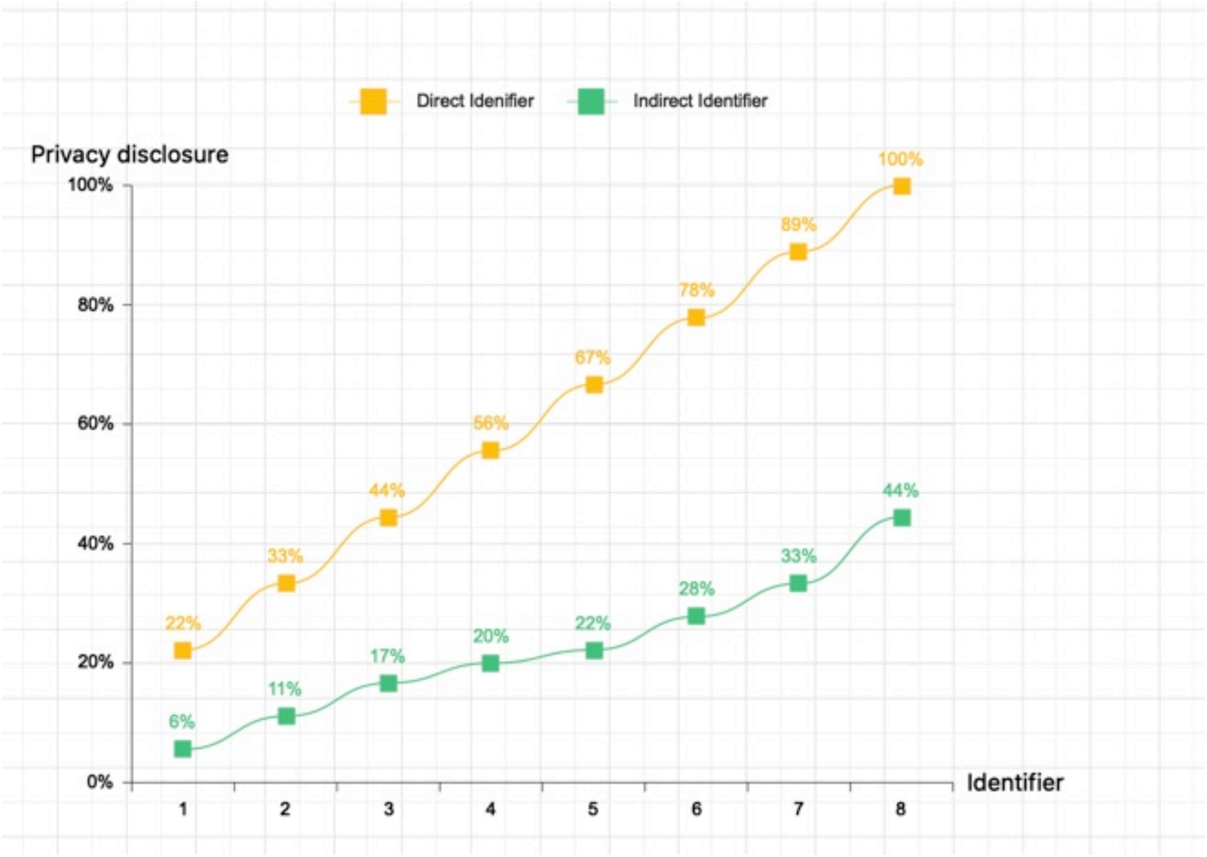
Where $f(Pi)$= Privacy disclosure on a record in a dataset

$f(Di)$= Direct Identifier on an attribute in a record

$f(Ii)$ = Indirect Identifier on an attribute in a record

Figure 9

The Estimated Relationship Between Privacy Disclosure and Identifiers Based on a Sample Data



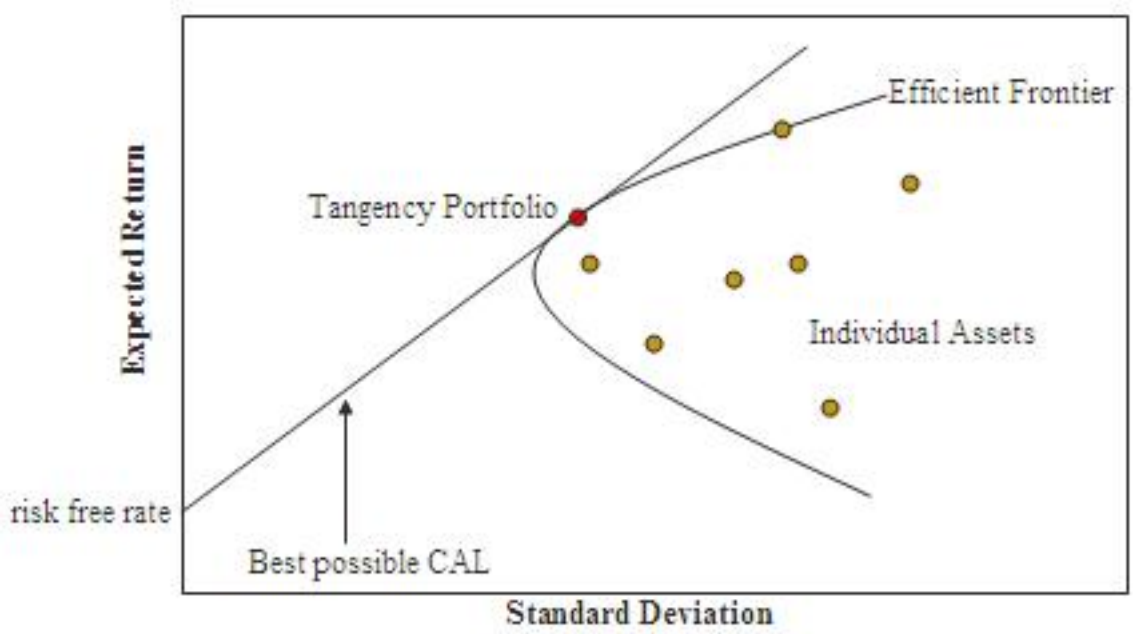
Note: a diagram based on a mock data to show relationship.

In recent years, a significant amount of research has been published to develop and ensure strong data privacy protection methods, algorithms, and ideas and to share datasets for scientific research purposes. There is always a balance between data privacy and the utility of the data for research and public policy purposes. Therefore, in addition to all the data anonymization techniques, the research community needs to develop transparent methods to notify and alert data owners when data is used for those research and public policy useful purposes. There are always trade-offs between data privacy and utility.

The privacy and utility of data can be compared with the concept of the Modern Portfolio Theory concept in the financial investment domain. When investing, one should consider the volatility of the market, and risk. The efficient frontier is the set of portfolios that offer the highest expected return for a defined level of risk or the lowest risk for a given level of expected return. ("Efficient Frontier," 2020)This concept is captured in the following diagram.

Figure 10

Efficiency Frontier



Note: privacy vs utility ("Efficient Frontier," 2020)

Chapter 5 – Architecture Implementation, Validation, and Discussion

The first part of the research focused on finding ways to identify potential data elements within datasets that have significant contributions in exposing the privacy of a record in a dataset - elements such as a full name of an individual combined with address information within healthcare data. In our research, after experimenting with multiple open-source healthcare datasets, we identified the potential use of machine learning algorithms to help classify or categorize potential data elements that need to be protected to safeguard data privacy.

Once we understand the relative importance of elements towards the contribution of data privacy using machine learning algorithms and the Canadian data privacy laws and guidelines, the second part of our research focused on finding architectural components on protecting the privacy of healthcare data within healthcare information systems.

Personal information is data about an "identifiable individual". It is information that on its own or combined with other pieces of data, can identify *you* as an individual.(Canada, 2014)

Classification of Data Using Machine Learning Algorithms

Machine learning (ML) is a domain within the field of artificial intelligence that leverages algorithms that can provide accurate predictions or models without explicit programming. The fundamental premise of machine learning is to prepare algorithms that can get data and use statistical analysis to predict an outcome while updating the prediction as new data is found. In this research, by using ML algorithms, we have demonstrated how to identify the relative importance of each variable in the dataset to the outcome of the model or prediction on healthcare data.

Machine learning explainability and interpretability are being considered as tools for explaining predictions to ML model users or end customers. There are several open-source toolkits, providing explainable AI functionality. In our research, we have used Shapley Values

(SHAP values,) ML tools to explain the importance of a variable or attribute of an element or field within a dataset towards the model and eventually to the privacy of the information. Machine learning helps gain insight into specific problems. In this research, we used Shapley Values to understand the influence of elements in a dataset in exposing the privacy of records within the dataset. Once the influence of each data element in terms of Shapley values is known, the next step is to ensure that we provide a systematic and integrated approach to protect the privacy of the data and eventually provide techniques and tools to let the stakeholders of the data-aware whenever the data is accessed by authorized or unauthorized parties.

SHAP, which stands for Shapley Additive exPlanations, is an interpretability method based on Shapley values and was introduced by Lundberg and Lee (2017)(Lundberg & Lee, 2017) to explain individual predictions of any machine learning model.

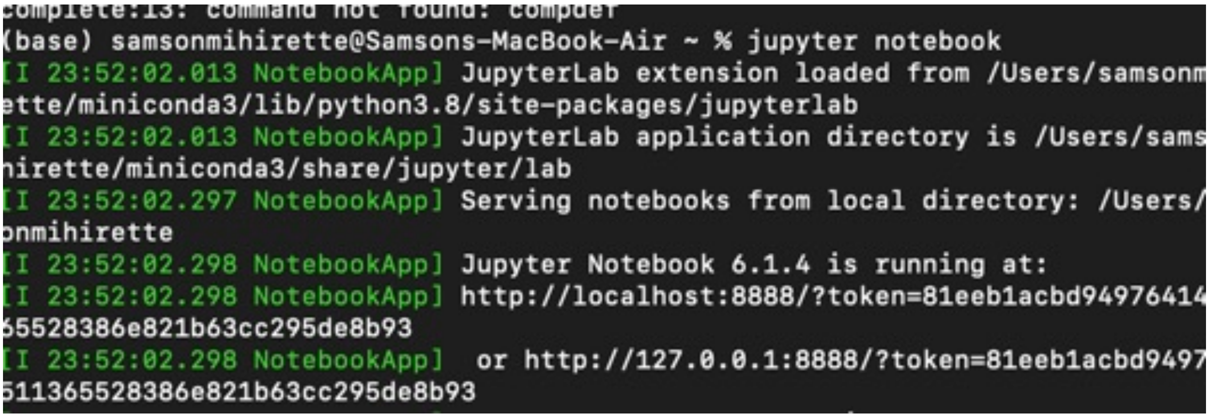
Shapley Values (SHAP), a method used in game theory to determine how much each player in a collaborative game has contributed to the successful result of the game. The Shapley value is the average of all marginal contributions to all possible coalitions or combinations. Therefore, in our research, the use of Shapley values is used to measure how much each data element within a healthcare dataset contributes to the final model. This is particularly useful to find out which elements contribute and by how much to the result of the model. Understanding that value helps us protect the data set if those elements have significant potential to expose the privacy of the data.

To understand, interpret patterns, and data elements within datasets, there are several integrated development environments (IDEs) open-source applications that are used by the research and developer communities. In our research, we used Jupyter Notebook IDE as it is supporting multiple programming languages.

In this research thesis, the model is performed using data science tools for ML prediction and modelling tools (to what degree negative or positive) - SHAP and CATBOOST. To invoke the development tool "jupyter notebook" is executed from the command line as shown in the screenshot below:

Figure 11

How to Invoke Jupyter in a Command Line



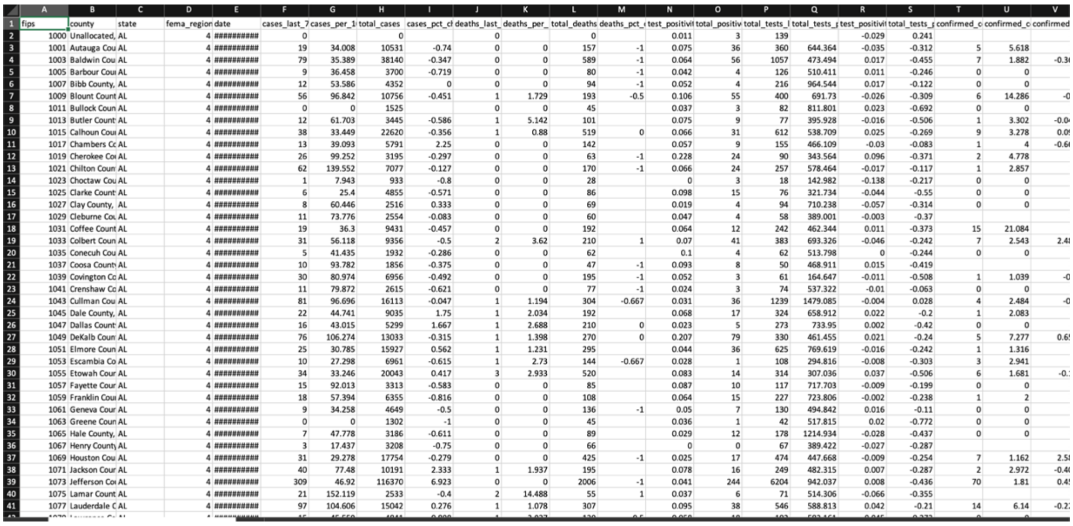
```
complete:13: command not found: compdef
(base) samsonmihirette@Samsons-MacBook-Air ~ % jupyter notebook
[I 23:52:02.013 NotebookApp] JupyterLab extension loaded from /Users/samsonm
ette/miniconda3/lib/python3.8/site-packages/jupyterlab
[I 23:52:02.013 NotebookApp] JupyterLab application directory is /Users/sams
hirette/miniconda3/share/jupyter/lab
[I 23:52:02.297 NotebookApp] Serving notebooks from local directory: /Users/
onmihirette
[I 23:52:02.298 NotebookApp] Jupyter Notebook 6.1.4 is running at:
[I 23:52:02.298 NotebookApp] http://localhost:8888/?token=81eeb1acbd94976414
65528386e821b63cc295de8b93
[I 23:52:02.298 NotebookApp]  or http://127.0.0.1:8888/?token=81eeb1acbd9497
511365528386e821b63cc295de8b93
```

Note: a screenshot to show the command using a command line interface.

Here is a portion of the healthcare dataset used for identifying SHAP Value of each element the data set contains more than 75,000 records:

Figure 12

Example of an Opensource Healthcare Dataset



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V
1	fips	county	state	fema_region	date	cases_last_7	cases_per_1	total_cases	cases_pct_cl	deaths_last_	deaths_per_	total_deaths	deaths_pct_	test_positivit	total_positiv	total_tests_l	total_tests_	test_positivit	total_tests_	confirmed_c	confirmed_c	confirmed
2	1000	Unallocated,	AL	4	##########	0		0		0		0		0.011	3	139		-0.029	0.241			
3	1001	Autauga Cou	AL	4	##########	19	34.008	10531	-0.74	0	0	157	-1	0.075	36	360	644.364	-0.035	-0.312	5	5.618	
4	1003	Baldwin Cou	AL	4	##########	79	35.389	38140	-0.347	0	0	589	-1	0.064	56	1057	473.494	0.017	-0.455	7	1.882	-0.3
5	1005	Barbour Cou	AL	4	##########	9	36.458	3700	-0.719	0	0	80	-1	0.042	4	126	510.411	0.011	-0.246	0	0	
6	1007	Bibb County,	AL	4	##########	12	53.586	4352	0	0	0	94	-1	0.052	4	216	964.544	0.017	-0.122	0	0	
7	1009	Blount Count	AL	4	##########	56	96.842	10756	-0.451	1	1.729	193	-0.5	0.106	55	400	691.73	-0.026	-0.309	6	14.286	-0
8	1011	Bullock Coun	AL	4	##########	0	0	1525		0	0	45		0.037	3	82	811.801	0.023	-0.692	0	0	
9	1013	Butler Count	AL	4	##########	12	61.703	3445	-0.586	1	5.142	101		0.075	9	77	395.928	-0.016	-0.506	1	3.302	-0.0
10	1015	Calhoun Cou	AL	4	##########	38	33.449	22620	-0.356	1	0.88	519	0	0.066	31	612	538.709	0.025	-0.269	9	3.278	0.0
11	1017	Chambers Cc	AL	4	##########	13	39.093	5791	2.25	0	0	142		0.057	9	155	466.109	-0.03	-0.083	1	4	-0.6
12	1019	Cherokee Co	AL	4	##########	26	99.252	3195	-0.297	0	0	63	-1	0.228	24	90	343.564	0.096	-0.371	2	4.778	
13	1021	Chilton Coun	AL	4	##########	62	139.552	7077	-0.127	0	0	170	-1	0.066	24	257	578.464	-0.017	-0.117	1	2.857	
14	1023	Choctaw Cou	AL	4	##########	1	7.943	933	-0.8	0	0	28		0	3	18	142.982	-0.138	-0.217	0	0	
15	1025	Clarke Count	AL	4	##########	6	25.4	4855	-0.571	0	0	86		0.098	15	76	321.734	-0.044	-0.55	0	0	
16	1027	Clay County,	AL	4	##########	8	60.446	2516	0.333	0	0	69		0.019	4	94	710.238	-0.057	-0.314	0	0	
17	1029	Cleburne Cou	AL	4	##########	11	73.776	2554	-0.083	0	0	60		0.047	4	58	389.001	-0.003	-0.37			
18	1031	Coffee Count	AL	4	##########	19	36.3	9431	-0.457	0	0	192		0.064	12	242	462.344	0.011	-0.373	15	21.084	
19	1033	Colbert Coun	AL	4	##########	31	56.118	9356	-0.5	2	3.62	210	1	0.07	41	383	693.326	-0.046	-0.242	7	2.543	2.4
20	1035	Conecuh Cou	AL	4	##########	5	41.435	1932	-0.286	0	0	62		0.1	4	62	513.798	0	-0.244	0	0	
21	1037	Coosa County	AL	4	##########	10	93.782	1856	-0.375	0	0	47	-1	0.093	8	50	468.911	0.015	-0.419			
22	1039	Covington Co	AL	4	##########	30	80.974	6956	-0.492	0	0	195	-1	0.052	3	61	164.647	-0.011	-0.508	1	1.039	-0
23	1041	Crenshaw Co	AL	4	##########	11	79.872	2615	-0.621	0	0	77	-1	0.024	3	74	537.322	-0.01	-0.063	0	0	
24	1043	Cullman Cou	AL	4	##########	81	96.696	16113	-0.047	1	1.194	304	-0.667	0.031	36	1239	1479.085	-0.004	0.028	4	2.484	-0
25	1045	Dale County,	AL	4	##########	22	44.741	9035	1.75	1	2.034	192		0.068	17	324	658.912	0.022	-0.2	1	2.083	
26	1047	Dallas Count	AL	4	##########	16	43.015	5299	1.667	1	2.688	210	0	0.023	5	273	733.95	0.002	-0.42	0	0	
27	1049	DeKalb Coun	AL	4	##########	76	106.274	13033	-0.315	1	1.398	270	0	0.207	79	330	461.455	0.021	-0.24	5	7.277	0.6
28	1051	Elmore Coun	AL	4	##########	25	30.785	15927	0.562	1	1.231	295		0.044	36	625	769.619	-0.016	-0.242	1	1.316	
29	1053	Escambia Co	AL	4	##########	10	27.298	6961	-0.615	1	2.73	144	-0.667	0.028	1	108	294.816	-0.008	-0.303	3	2.941	
30	1055	Etowah Cour	AL	4	##########	34	33.246	20043	0.417	3	2.933	520		0.083	14	314	307.036	0.037	-0.506	6	1.681	-0.
31	1057	Fayette Cour	AL	4	##########	15	92.013	3313	-0.583	0	0	85		0.087	10	117	717.703	-0.009	-0.199	0	0	
32	1059	Franklin Cou	AL	4	##########	18	57.394	6355	-0.816	0	0	108		0.064	15	227	723.806	-0.002	-0.238	1	2	
33	1061	Geneva Cour	AL	4	##########	9	34.258	4649	-0.5	0	0	136	-1	0.05	7	130	494.842	0.016	-0.11	0	0	
34	1063	Greene Coun	AL	4	##########	0	0	1302	-1	0	0	45		0.036	1	42	517.815	0.02	-0.772	0	0	
35	1065	Hale County,	AL	4	##########	7	47.778	3186	-0.611	0	0	89		0.029	12	178	1214.934	-0.028	-0.437	0	0	
36	1067	Henry County	AL	4	##########	3	17.437	3208	-0.75	0	0	66		0	0	67	389.422	-0.027	-0.287			
37	1069	Houston Cou	AL	4	##########	31	29.278	17754	-0.279	0	0	425	-1	0.025	17	474	447.668	-0.009	-0.254	7	1.162	2.5
38	1071	Jackson Cour	AL	4	##########	40	77.48	10191	2.333	1	1.937	195		0.078	16	249	482.315	0.007	-0.287	2	2.972	-0.4
39	1073	Jefferson Co	AL	4	##########	309	46.92	116370	6.923	0	0	2006	-1	0.041	244	6204	942.037	0.008	-0.436	70	1.81	0.4
40	1075	Lamar Count	AL	4	##########	21	152.119	2533	-0.4	2	14.488	55	1	0.037	6	71	514.306	-0.066	-0.355			
41	1077	Lauderdale C	AL	4	##########	97	104.606	15042	0.276	1	1.078	307		0.095	38	546	588.813	0.042	-0.21	14	6.14	-0.2

Note: a dataset used for the experiment *(COVID-19 Community Vulnerability Crosswalk - Crosswalk by Census Tract | HealthData.Gov*, n.d.)

I have used multiple experiments using various data sources to verify the result of the SHAP algorithm. The healthcare data source used in this Thesis is from the "COVID-19 Community Vulnerability Crosswalk – Crosswalk by Census Tract" provided at the US healthcare research opensource data set.(*COVID-19 Community Vulnerability Crosswalk - Crosswalk by Census Tract | HealthData.Gov*, n.d.)

In our analysis, we are determining by how much each element contributed to the model.

Therefore, the Shapley Values provides value on how much each of the elements in the dataset contributes to the prediction or model. For example, how much a 'postal code' data element (from a record that contains name, postal code, types of medication) contributes to the difference between a single prediction and the average prediction when training the model. So, each element in the dataset is responsible for the prediction therefore we would like to know the amount of contribution each one of them has for the final prediction.

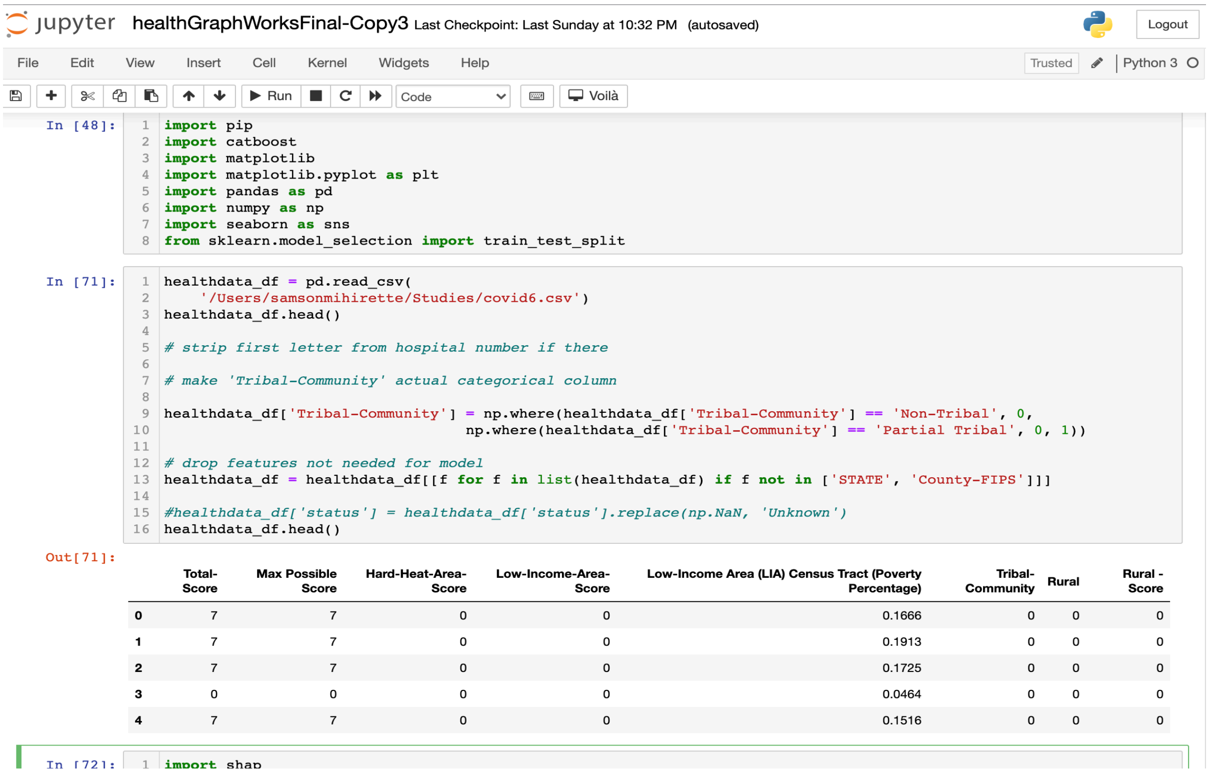
When determining Shapley Values techniques, the following major steps are taken:

- Step 1 – Find random permutation of elements in the dataset.
- Step 2 – Get random samples from the dataset.
- Step 3 – For vectors using datasets (a very key mathematical step.)
- Step 4 – Record the difference between actual and average and return to step 1.

To analyze the model, we have conducted a few steps to make the application read the data set to read and produce models – Install, configure, and import PIP, MATPLOTLIB, CATBOOST, SHAP (Load executable software libraries into the memory.) CATBOOST(*CatBoostClassifier*, n.d.) is used to handle the categories of data elements.

Figure 13

Example of Codes Using Jupyter Notebook – Reading Datafile


jupyter
healthGraphWorksFinal-Copy3 Last Checkpoint: Last Sunday at 10:32 PM (autosaved)
Logout
File
Edit
View
Insert
Cell
Kernel
Widgets
Help
Trusted
Python 3
Run
Code
Voilà

```
In [48]:
import pip
import catboost
import matplotlib
import matplotlib.pyplot as plt
import pandas as pd
import numpy as np
import seaborn as sns
from sklearn.model_selection import train_test_split
```

```
In [71]:
healthdata_df = pd.read_csv(
    '/Users/samsonmihirette/Studies/covid6.csv')
healthdata_df.head()

# strip first letter from hospital number if there

# make 'Tribal-Community' actual categorical column

healthdata_df['Tribal-Community'] = np.where(healthdata_df['Tribal-Community'] == 'Non-Tribal', 0,
                                    np.where(healthdata_df['Tribal-Community'] == 'Partial Tribal', 0, 1))

# drop features not needed for model
healthdata_df = healthdata_df[[f for f in list(healthdata_df) if f not in ['STATE', 'County-FIPS']]]

#healthdata_df['status'] = healthdata_df['status'].replace(np.NaN, 'Unknown')
healthdata_df.head()
```

```
Out[71]:
     Total-  Max Possible  Hard-Heat-Area-  Low-Income-Area-  Low-Income Area (LIA) Census Tract (Poverty     Tribal-  Rural  Rural -
     Score          Score            Score             Score                               Percentage)  Community          Score
0        7              7                0                 0                                    0.1666          0      0        0
1        7              7                0                 0                                    0.1913          0      0        0
2        7              7                0                 0                                    0.1725          0      0        0
3        0              0                0                 0                                    0.0464          0      0        0
4        7              7                0                 0                                    0.1516          0      0        0
```

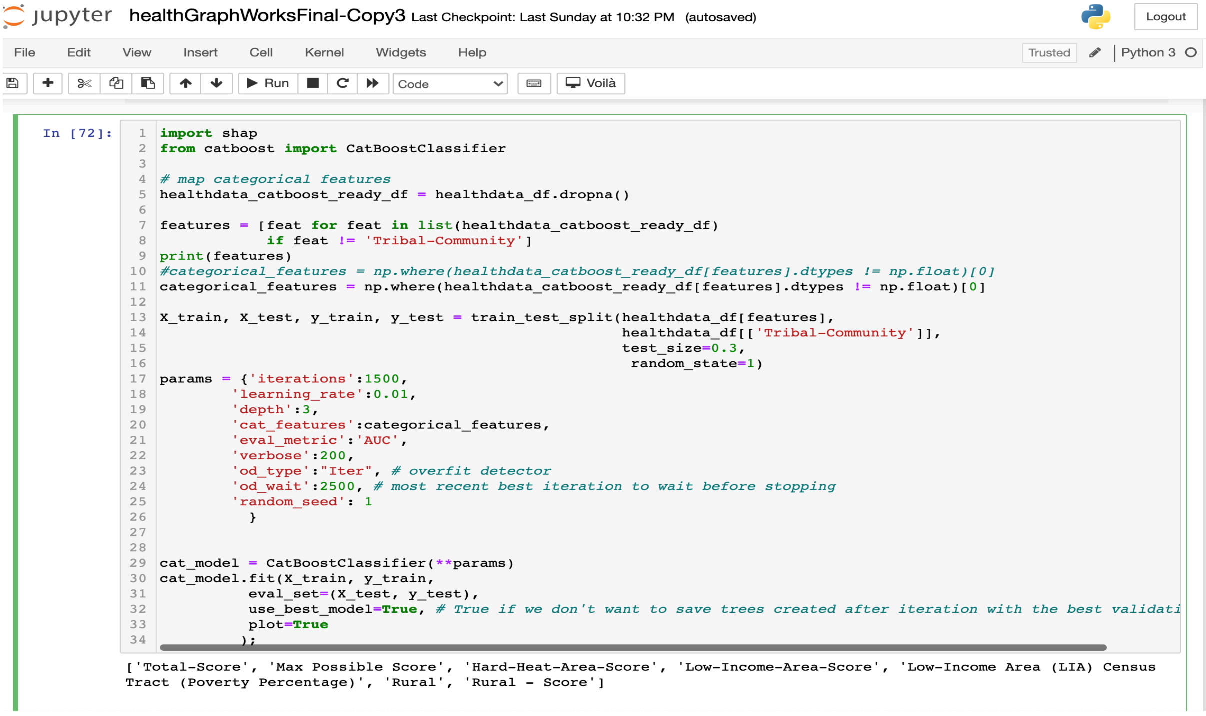
```
In [72]:
import shap
```

Note: shows python codes on how to import modules and read datafiles.

Cleaning up a dataset for statistical modelling (preparing the dataset by taking only the important variables in the dataset.) In this example, we took only the following variables – ['Total-Score', 'Max Possible Score', 'Hard-Heat-Area-Score', 'Low-Income-Area-Score', 'Low-Income Area (LIA) Census Tract (Poverty Percentage)', 'Rural', 'Rural - Score']

Figure 14

Trimming and Preparing Critical Data Elements



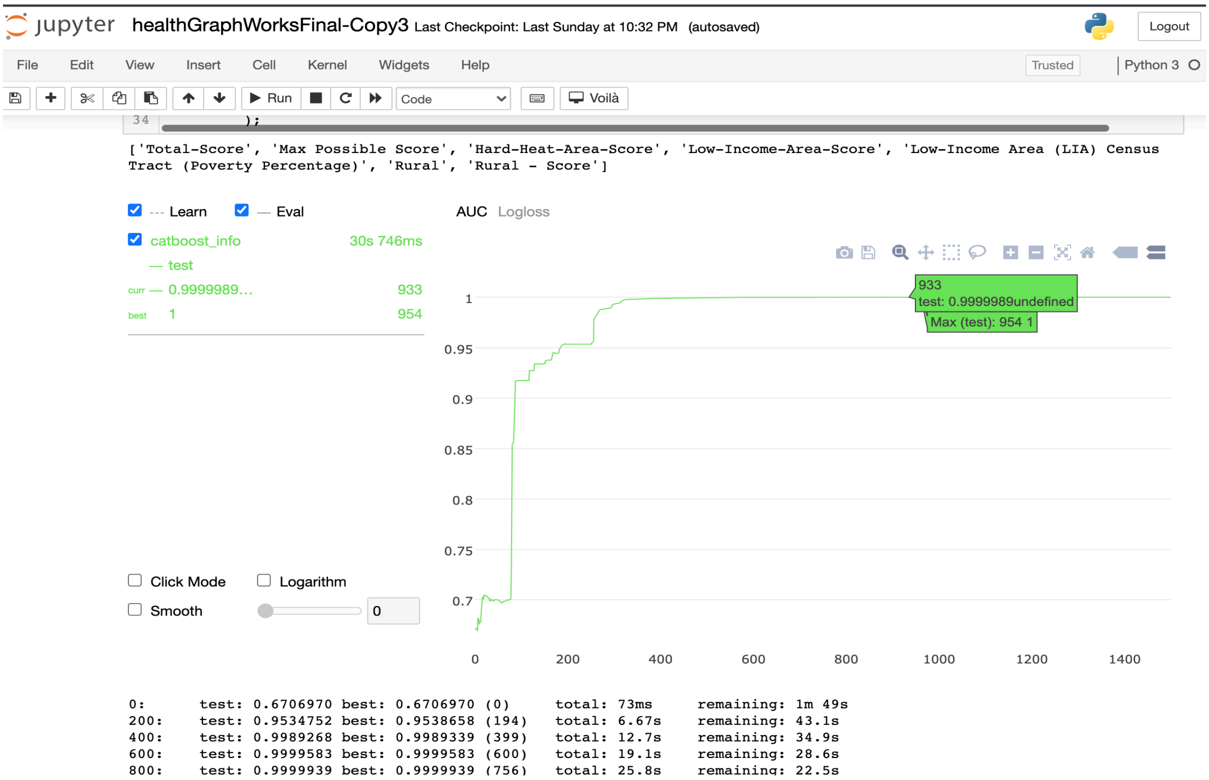
```
import shap
from catboost import CatBoostClassifier

# map categorical features
healthdata_catboost_ready_df = healthdata_df.dropna()

features = [feat for feat in list(healthdata_catboost_ready_df)
            if feat != 'Tribal-Community']
print(features)
#categorical_features = np.where(healthdata_catboost_ready_df[features].dtypes != np.float)[0]
categorical_features = np.where(healthdata_catboost_ready_df[features].dtypes != np.float)[0]

X_train, X_test, y_train, y_test = train_test_split(healthdata_df[features],
                                                    healthdata_df[['Tribal-Community']],
                                                    test_size=0.3,
                                                     random_state=1)
params = {'iterations':1500,
        'learning_rate':0.01,
        'depth':3,
        'cat_features':categorical_features,
        'eval_metric':'AUC',
        'verbose':200,
        'od_type':"Iter", # overfit detector
        'od_wait':2500, # most recent best iteration to wait before stopping
        'random_seed': 1
          }


cat_model = CatBoostClassifier(**params)
cat_model.fit(X_train, y_train,
          eval_set=(X_test, y_test),
          use_best_model=True, # True if we don't want to save trees created after iteration with the best validati
          plot=True
         );
```

```
['Total-Score', 'Max Possible Score', 'Hard-Heat-Area-Score', 'Low-Income-Area-Score', 'Low-Income Area (LIA) Census
Tract (Poverty Percentage)', 'Rural', 'Rural - Score']
```

Note: piece of code to demonstrate how to prepare the data for further analysis and enter useful parameter values for SHAP.

Figure 15

The Process of Training the ML Algorithm



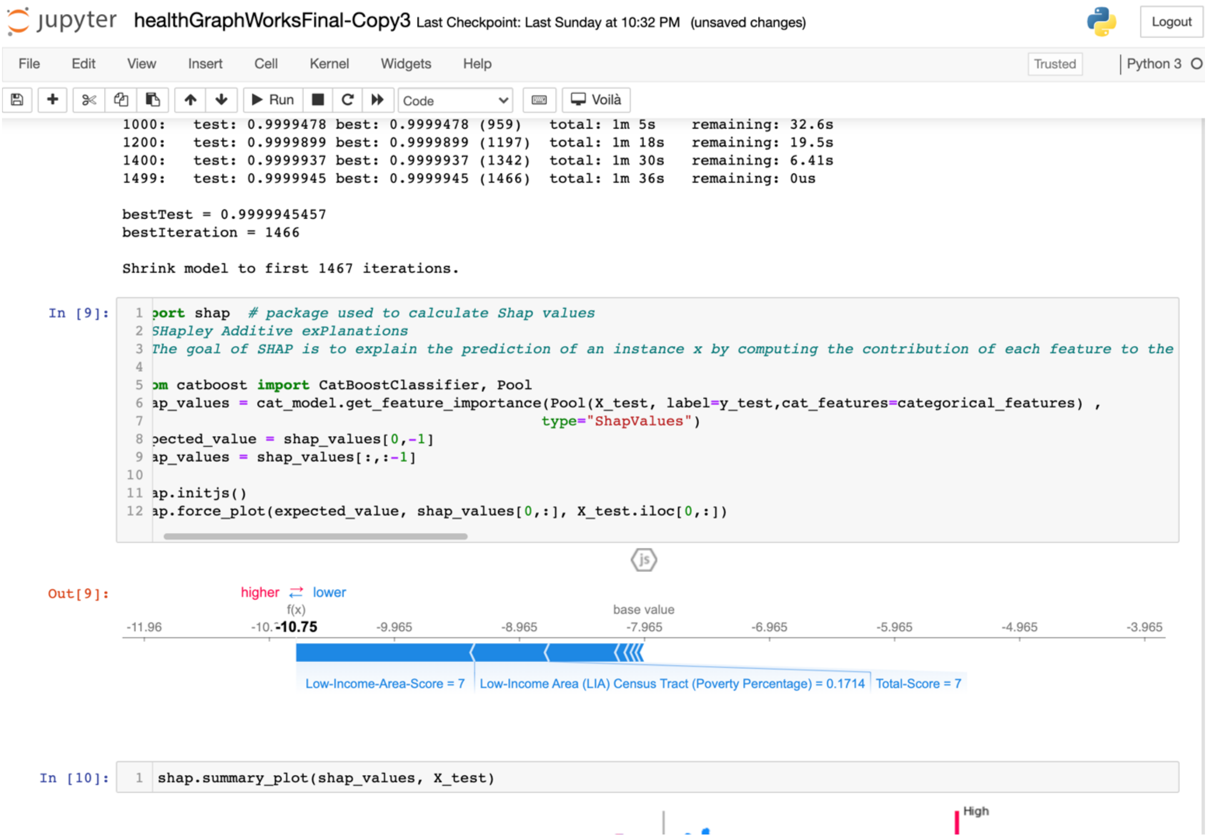
```
               );
['Total-Score', 'Max Possible Score', 'Hard-Heat-Area-Score', 'Low-Income-Area-Score', 'Low-Income Area (LIA) Census
Tract (Poverty Percentage)', 'Rural', 'Rural - Score']
```

```
0:      test: 0.6706970 best: 0.6706970 (0)     total: 73ms     remaining: 1m 49s
200:    test: 0.9534752 best: 0.9538658 (194)   total: 6.67s    remaining: 43.1s
400:    test: 0.9989268 best: 0.9989339 (399)   total: 12.7s    remaining: 34.9s
600:    test: 0.9999583 best: 0.9999583 (600)   total: 19.1s    remaining: 28.6s
800:    test: 0.9999939 best: 0.9999939 (756)   total: 25.8s    remaining: 22.5s
```

Note: when running the code, it scans each record in the dataset to understand patterns.

Figure 16

Example of the SHAP Values for Some Data Elements



jupyter
healthGraphWorksFinal-Copy3 Last Checkpoint: Last Sunday at 10:32 PM (unsaved changes)
Logout
File
Edit
View
Insert
Cell
Kernel
Widgets
Help
Trusted
Python 3
Run
Code
Voilà

```
1000:	test: 0.9999478 best: 0.9999478 (959)	total: 1m 5s	remaining: 32.6s
1200:	test: 0.9999899 best: 0.9999899 (1197)	total: 1m 18s	remaining: 19.5s
1400:	test: 0.9999937 best: 0.9999937 (1342)	total: 1m 30s	remaining: 6.41s
1499:	test: 0.9999945 best: 0.9999945 (1466)	total: 1m 36s	remaining: 0us

bestTest = 0.9999945457
bestIteration = 1466

Shrink model to first 1467 iterations.
```

In [9]:

```
port shap  # package used to calculate Shap values
SHapley Additive exPlanations
The goal of SHAP is to explain the prediction of an instance x by computing the contribution of each feature to the

om catboost import CatBoostClassifier, Pool
ap_values = cat_model.get_feature_importance(Pool(X_test, label=y_test,cat_features=categorical_features) ,
                                             type="ShapValues")
pected_value = shap_values[0,-1]
ap_values = shap_values[:,:-1]

ap.initjs()
ap.force_plot(expected_value, shap_values[0,:], X_test.iloc[0,:])
```

Out[9]:

In [10]:

```
shap.summary_plot(shap_values, X_test)
```

Note: preparing the result of the algorithm for a summary.

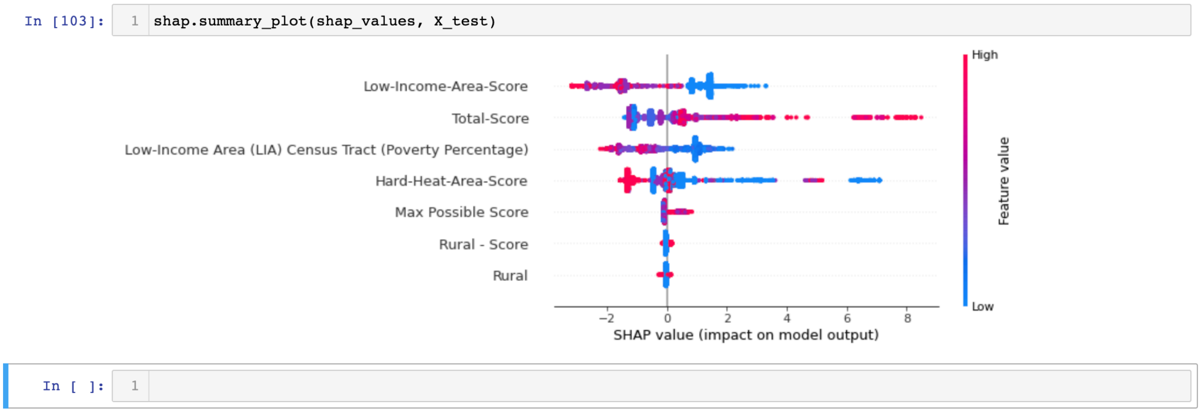
This plotting technique of SHAP simplifies the model to understand the influence of each element in the model. It is a regression model based on classifiers, non-linear. In this dataset, the race is the most important variable that influenced the model.

The following diagram also depicts the SHAP values for given health-based open-source data set. For example, race (a dataset element) has the highest SHAP value for the model. Based on this model, several healthcare decisions can be made. For example, new public health policy decisions can be made based on neighbourhood or postal code information. As shown in the model, the postal code has a substantial contribution to the model. At this point, the question we need to ask is on privacy protection measures – are the stakeholders aware of the decision? Does this model expose the data privacy of stakeholders (patients, healthcare providers and others?) Therefore, these SHAP values can help enhance the data privacy protection measures in terms of data classification. In our research, these values help to build algorithms for a robust data privacy broker. Whenever elements with a substantial value in determining the outcome (model) are used by authorities or any other public policy organizations, data privacy brokers will help document and notify the appropriate stakeholders.

To summarize, in this dataset, after ‘Low-Income_area-Score’, the ‘Total-Score’ is the most important variable. So, in combination with other elements in the dataset ‘Total-Score’ related data has a high SHAP value (high in red and low in blue) with a significant impact on the model. Therefore, the following SHAP summary plot diagram provides information for the selected variables and the impact of each variable on the model.

Figure 17

SHAP Values for Each Player (Data Element)



```
In [103]: shap.summary_plot(shap_values, X_test)
```

```
In [ ]:
```

Note: a SHAP summary that shows 3 characteristics for each plot in the diagram.

The y-axis indicates – the variable name, in order of importance from top to bottom. SHAP value can show how much each predictor contributes, either positively or negatively to the target variable. This is like the variable importance plot but it is able to show the positive or negative relationship for each variable with the target.

This plot is made of all the dots in the train data. It demonstrates the following information:

- Feature importance: The horizontal location shows whether the effect of that value is associated with a higher or lower prediction.
- Impact: The horizontal location shows whether that variable is high (in red) or low (in blue) for that observation.
- Original value: Color shows whether that variable is high (in red) or low (in blue) for that observation.
- Correlation: A high level of the "Total-Score" content has a high and positive impact on the quality rating. The "high" comes from the red color, and the "positive" impact is shown on the X-axis.

Data Privacy Broker Within the Healthcare Information System Landscapes

Existing technologies are evolving at a rapid rate, and new technologies are emerging to accelerate and leverage cloud computing, artificial intelligence, and edge computing, to name a few. There is no end to this technological growth and digital transformation; therefore, the role of IT application architects is significant in designing an optimal landscape for each business or digital requirement in determining the correct use of application technologies and architecture. In preparing the healthcare information systems infrastructure, I have conducted research on several architecture designs and solutions within the three major domains of architecture: application, data, and technology.

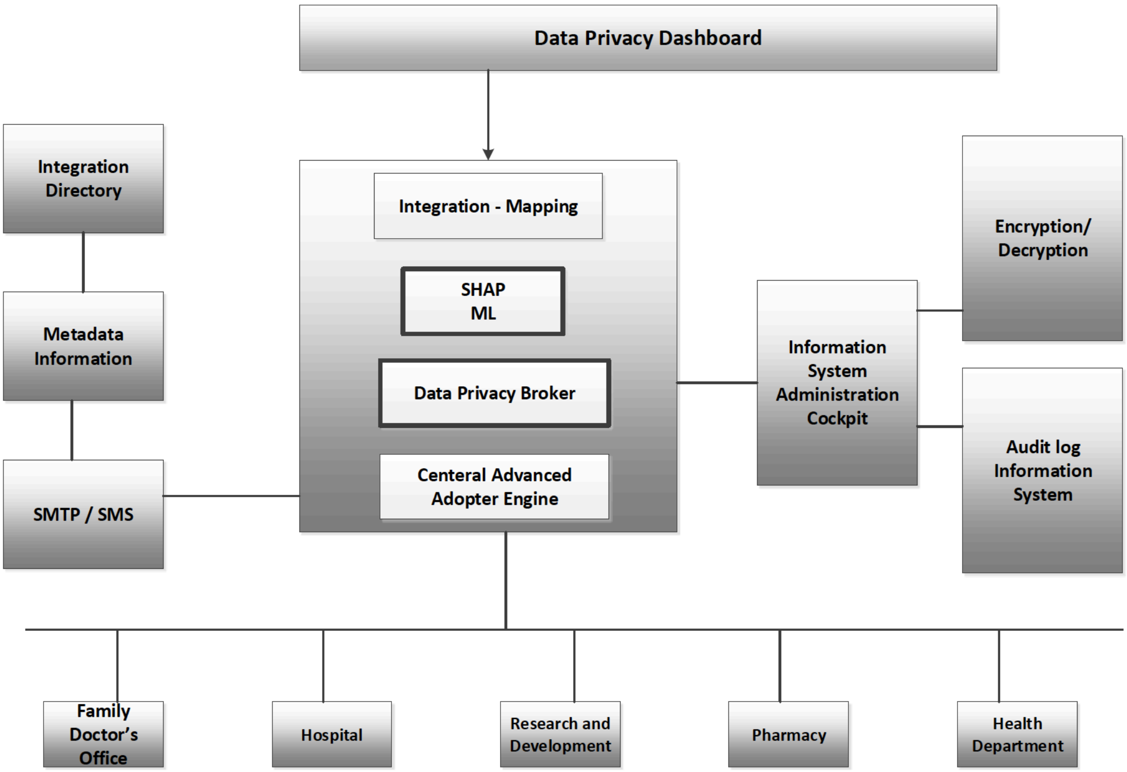
In the following paragraphs, we will explain the role of the data privacy broker within the healthcare information system architecture. This data privacy broker plays a significant role in providing a transparent mechanism when exchanging healthcare information from one stakeholder to the other, mainly by providing easy and timely access to all major stakeholders of the healthcare data including the patient or the customer who is primarily impacted if the data is exposed to other entities. It is debatable to conclude the owner of healthcare information - for example, records of a given patient, the doctors' office, however, patients will always be involved or impacted when we discuss healthcare-related records. We are not against the use of healthcare data for research or any other policymaking purposes; however, when using personal data which is considered private for a patient, all appropriate stakeholders should be notified and provide awareness on who accessed the data and for what purpose. In this architecture, we are introducing a data privacy broker in the healthcare information system.

Application Architecture

To discuss infrastructure components of healthcare architecture, we will use the following high-level architecture design diagram. There are multiple ways of implementing and integrating the data privacy broker within the healthcare information system, in this diagram we are introducing the data privacy broker as a middleware central integration component.

Figure 18

Healthcare Information System Including the Data Privacy Broker



Note: a sample information system with SHAP machine learning and the Data Privacy Broker.

Integration – Mapping

The purpose of the integration mapping component of the middleware in the architecture:

- Take data format and translate it to a new format to make sure there exists a compatible messaging or data understanding between two or more parties.
- Data element or database table field filtering to make sure the destination application received the requested and authorized data only.
- Scripting functionalities can be included if robust messaging or communication is needed. Scripts such as XSL or CSS can be used to package the message.
- Sender and receiver addresses, communication methods are included.

Data Privacy Broker

The Data Privacy Broker is used mainly to ensure there is a transparent mechanism when data is communicated from the sender to all receivers for all highly classified data in terms of privacy. The Data integration uses the SMTP, or SMS or a MAIL server to communicate the action to the stakeholders. All activities can also be displayed using the Data Privacy Dashboard as shown in the figure above.

- Data privacy elements identified by the Machine Learning algorithm will be captured.
- Data elements are identified as protected by the privacy of laws (ex. Canadian privacy laws.)

Central Advanced Adapter

- Adapter Engine contains a set of communication formats or adapters such as SOAP, HTTP, SFTP, and FTP

Integration Directory

Integration directory provides answers to all these questions of the end-to-end integration scenario,

- The message sender system.
- Are there multiple message receivers?
- Should the message receiver system be derived dynamically from the message content?
- What is the mapping program to transform the sender message to the target message format?
- How should the adapters be defined?

Metadata information repository

Metadata repository contains useful information that can be leveraged to enable security and governance, data recommendations, and user telemetry. This repository stores technical and operational metadata that allows the logical mapping of physical data stores to a more application-neutral model. This neutral model assists in representing common keys with different names. Look-up tables to translate keys between systems can also be part of the metadata repository. The translation from a neutral term to a specific data detail enables query planning and execution. Metadata also forms the foundation of data fabric capabilities, which offer AI/ML recommendation engines for various purposes, including data recommendations and performance optimization recommendations.

SMTP / SMS services:

Simple Mail Transfer Protocol (SMTP) is used for sending and receiving emails and SMS is a text messaging solution used on smartphones or any text messaging application. Both SMTP/SMS services will be used to notify important information to the patient or any stakeholder that need-to-know data access by any party.

Encryption/Decryption:

The encryption/decryption component of the healthcare infrastructure is responsible for any type of high-level secured communication between applications.

Audit Log information system:

All access logs will be captured by the audit log capturing component of the infrastructure. Who requested the private data, the date and time the data is accessed, and other information is captured?

Data Privacy Dashboard:

This is an element of the healthcare information system that provides users and all other healthcare stakeholders front-end services. As a point of entry portal system for any authorized user including the patient, it provides basic information such as data access activity log for a given period.

Data Architecture

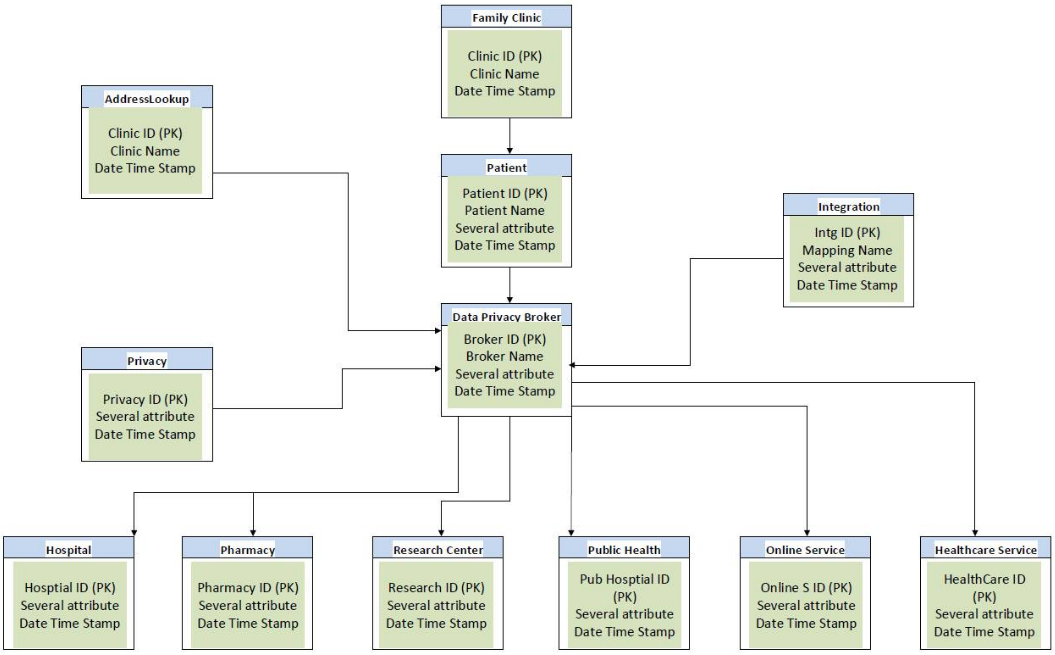
Data architecture is one major component within information technology architecture that mainly deals with the physical and logical flow of data for a given architecture therefore it is very important for ensuring better data management and governance initiatives for a given solution architecture. As one of the IT management frameworks, The Open Group Architecture Framework (TOGAF), data architecture is a domain within the TOGAF that defines the organization’s data storage, management, and maintenance, including logical and physical data models.(White, 2018)

Data Modelling High-Level Description

Data modelling is a group of techniques used to explain the kinds of information that are important in an information system. The following conceptual data model diagram represents the relationships between various components in a healthcare information system. Depending on the number of stakeholders and information required, the diagram can be very complex; however, in this model, we are showing the role of the data privacy broker and its relationship with the main components in the healthcare information system.

Figure 19

High-Level Data Modelling Including Data Privacy Model Integration



Note: a sample data modeling to demonstrate important data elements for integration with the data privacy broker.

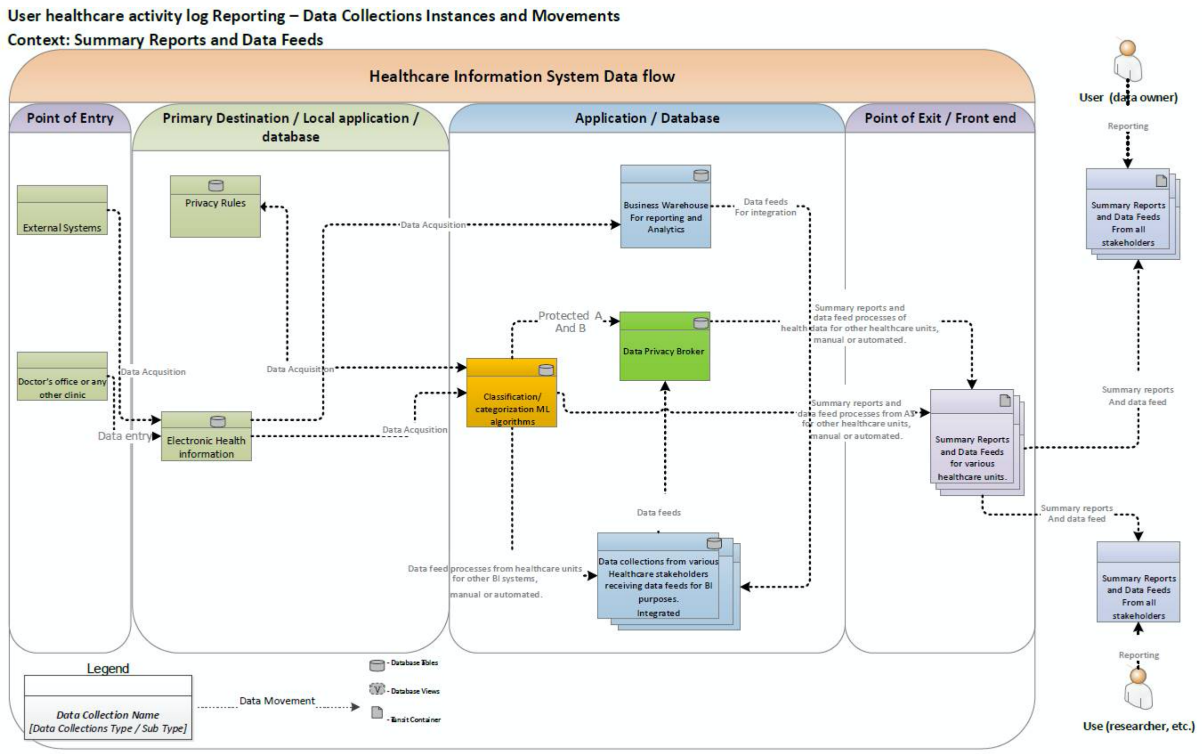
Data Architecture View

The data architecture views described in this section cover various data-related topics including data collections, data movement, major data entities, and reporting.

The following target architecture data movement diagram illustrates the data movement in more detail:

Figure 20

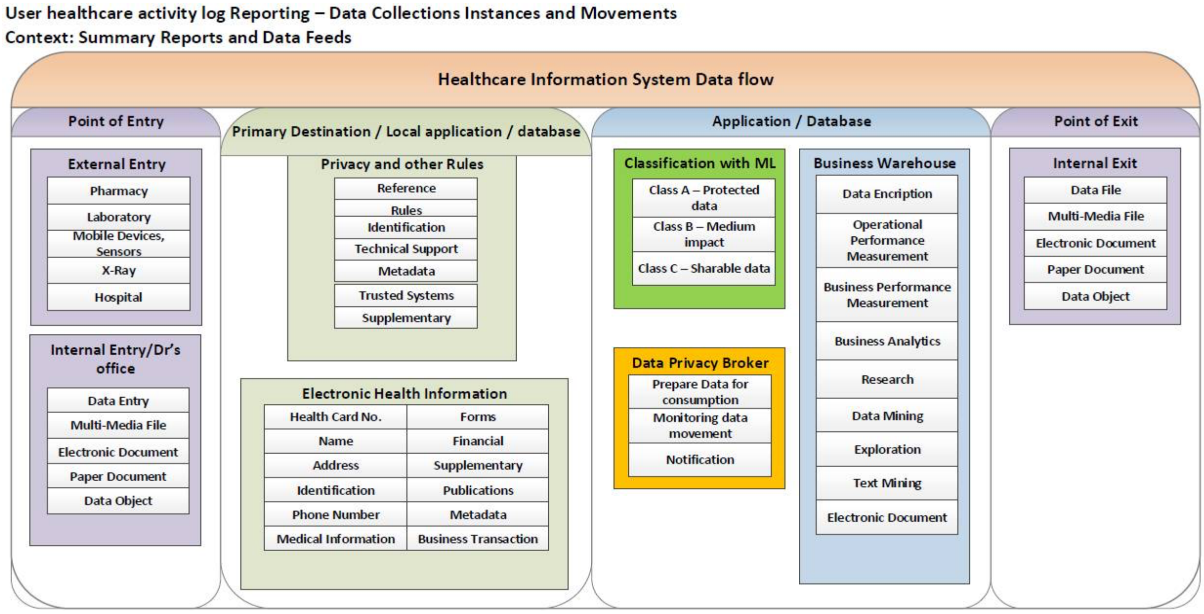
Data Collection Instances and Movements


User healthcare activity log Reporting – Data Collections Instances and Movements
Context: Summary Reports and Data Feeds
Healthcare Information System Data flow
Point of Entry
Primary Destination / Local application / database
Application / Database
Point of Exit / Front end
External Systems
Doctor's office or any other clinic
Data Acqusition
Data entry
Privacy Rules
Data Acquisition
Electronic Health information
Data Acqusition
Data Acqusition
Business Warehouse For reporting and Analytics
Data feeds For integration
Protected A And B
Data Privacy Broker
Classification/ categorization ML algorithms
Summary reports and data feed processes of health data for other healthcare units, manual or automated.
Summary reports and data feed processes from AS for other healthcare units, manual or automated.
Data feeds
Data feed processes from healthcare units for other BI systems, manual or automated.
Data collections from various Healthcare stakeholders receiving data feeds for BI purposes. Integrated
Summary Reports and Data Feeds for various healthcare units.
User (data owner)
Reporting
Summary Reports and Data Feeds From all stakeholders
Summary reports And data feed
Summary reports And data feed
Summary Reports and Data Feeds From all stakeholders
Reporting
Use (researcher, etc.)
Legend
Data Collection Name [Data Collections Type / Sub Type]
Data Movement
- Database Tables
- Database Views
- Transit Container

Note: Demonstrating the details of the healthcare information system data flow

Figure 21

Details of the Data Collection



Note: Demonstrating data collection components at higher level.

Point of Entry (Internal and External Sources)

Internal Sources:

- Data entry by clinic receptionist over the phone, online or at the doctor's office. Information such as name, last name, age, health care information of the patient.
- prescriptions, insurance, laboratory, pharmacy, and other administrative data.
- Doctor's note during patient's consultation. Clinical data such as decision-support information, medical imaging, physician written notes.

External Sources:

- Clinical data, diagnostic imaging, laboratory results and other health-related information from laboratories.
- Machine-generated or sensor data such as data from devices monitoring health data and healthcare IoTs.
- Information from pharmacies, hospitals, or other sources about patients' health information.
- Data is extracted from any healthcare unit in many formats for reporting purposes. Data can be submitted as a flat file or scanned invoices.

Primary Destination

- Electronic healthcare applications reside locally on the computers in the doctor's office.
- Access and identification information repository related to integration to other systems in healthcare information system.
- Scanners and other devices are used to capture healthcare information.

Secondary Destination

- Centralized healthcare information system infrastructure, healthcare research centers, universities, government agencies, hospitals, and other third-party healthcare information stakeholders. These stakeholders collect and leverage information such as patient identification and checking for recurring visits, medical admission, nursing admission, appointing and visitor information.

Data Classification

- The machine learning algorithm categorizes the data into three classifications:
    - Class A – Data that should be protected and need to pass through the data privacy broker so that all stakeholders are aware when it is accessed by anyone.
    - Class B – Data with a medium impact on privacy issues but it needs to pass through the data privacy broker if it is not encrypted.
    - Class C – Data with no impact on privacy therefore these data are sharable with other stakeholders. No need to let Class C data pass through a data privacy broker because if the data is exposed to any stakeholder, it will not expose personal information.

Data Privacy Broker

- Data movement to a patient in the form of notification or when accessed by the user through a data privacy dashboard. Central healthcare portal system with user login capability providing all healthcare access log information.
- Any data which is vetted by the machine learning algorithm as protected data or 'Class A' passes through the data privacy broker.
- Data that is noted as protected data will not go through the data privacy broker.
- The data privacy broker also provides a middleware component therefore it is a health integration access layer that uses API services.

Chapter 6 – Opportunity, Challenges, and Trends

As the future of the healthcare system continues to be transformed to a new level to reach patients at any location such as the patients' home, workplace, on the road, at doctor's office, the researchers and innovators will continue to innovate several devices with respect of healthcare monitoring and other devices for the purpose of easily identifying patient healthcare conditions in any given time to eventually improve patient healthcare conditions. These innovations will continue to create research opportunities for information system professionals to deliver healthcare mobile devices, data protection mechanisms, application integration technologies within cloud platforms and on-premises healthcare information systems. The increasing heterogeneity, complexity, and distributed nature of healthcare solutions dictate the continued transformation of the healthcare information system. The trend within these healthcare innovations is to improve personalized patient experience, quality, and safety of treatments. Those innovations include virtual healthcare technologies also called telehealth or telemedicine to reach out to patients remotely using emerging and evolving technologies. These technologies are using the cloud platform substantially for ease and centralized access to healthcare information. Those trends in technology also bring opportunities and challenges to the information technology research communities in the areas of healthcare data privacy.

Data privacy concerns and delivering a robust solution to protect data will continue to be a hot topic as technology evolves and new methods of electronic data communications and storage media are created. Various types of electronic data are stored and transmitted over the internet across multiple devices and over the cloud platform. From these data, healthcare data continues to be very personal, sensitive, and important that needs to be protected and secured. Patients or anyone who owns the healthcare data will not want to expose the data to their employers or insurance agents due to a fear of employment termination, insurance premium impacts or any other negative health-related policy implications. Therefore, as information technology professionals, we will continue to conduct research and provide transparent solutions for data access to stakeholders including the patient to ensure there are data protection components and techniques within the information system. There are plenty of opportunities, and challenges to protect individual-level healthcare information from misuse and maximize social benefit, data, and analytics. The technological innovation trend will also continue in the areas of artificial intelligence, service-oriented architecture and application integration using APIs. In general, the challenges and the trends of information technology are within the following privacy- enhancing technique areas – the use of artificial intelligence, application, and data integration (data sharing,) enhancing data privacy protection around healthcare Internet of Things (IoTs), data-centric design for n-tier architecture, and the use of healthcare data for research and healthcare policy decisions by healthcare public officials while preserving data privacy.

As discussed in this research, machine learning algorithms can be used to identify data elements and use the result for data classification to feed the data privacy broker or apply encryption, differential privacy, homomorphic encryption, or anonymization algorithms. Opportunities will continue to exist on leveraging and integrating machine learning algorithms in various areas of healthcare infrastructure. At the same time healthcare information systems continue to be involved in collecting biometric information, the use of healthcare IoTs and cloud infrastructure, there are multiple areas in which we continue to have challenges. Therefore, machine learning algorithms, in general, data science disciplines including artificial intelligence, can be used to resolve data privacy, performance problems and enhance the monitoring of all data movement in the healthcare information system. There are plenty of research opportunities around healthcare data privacy at every level, in different layers in the information system. Data communication between, the internet of things and the Fog computing (a form of distributed computing that offers services at the network edge) layer, and the cloud computing layer. Research and development opportunities on providing security, monitoring, enforcing privacy preservations, and the creation of awareness will continue to be important and therefore artificial intelligence can help mitigate those areas. As technology advances, the processing power of devices will continue to grow, and storage will not be expensive to handle massive heterogeneous data within the cloud platform for both structured and unstructured data. Therefore, there are opportunities to continue to improve the use of AI algorithms within all types of data including big data to identify privacy loopholes, and risks associated with it.

Since it is not practical or effective to work in a silo, there is always a need for data integration for analytics or transactional systems. There are several research opportunities within the areas of healthcare application and data integrations. As technology grows, more IoTs or transforms the healthcare devices, application integration continues to be an area of opportunity and challenges. Application interfaces between healthcare solutions, healthcare devices, wearable fitness devices, and cloud platforms will be required for effective and secured data communications. For this, there is a trend, opportunities, and advancement within the domain of service-oriented architecture, APIs, web services, continued work in designing wrappers and plugins based on the concept of service component architecture (SCA). There are various research papers within the area of healthcare data ingestion to the cloud platform and several methods have been identified to secure data within the cloud platform. Cloud-based healthcare data ingestion and storage must be carefully planned and executed in order to ensure predictable outcomes and prevent harm.(Tao et al., 2016) The continued improvement of healthcare monitoring devices requires the improvement and advancement of network communication methods such as application plugins, REST APIs, and web service communication technologies. The other potential area for healthcare data privacy research opportunity for secure healthcare data communication within the service-oriented architecture is the use of cryptography and blockchain technology. There is a potential research work in adopting blockchain framework. Using blockchain technology to secure patient records. A blockchain is a decentralized and distributed database that validates, records, timestamps, and maintains all transactions in a network of computers available only to authenticated participants.(Salahuddin et al., 2017) With further research, the blockchain-based electronic healthcare systems can contribute in mitigating by adding security and privacy capabilities.(Hossein et al., 2019)

When transferring data from one application to the other through the data privacy broker, unexpected negative performance impacts, and data latency issues can arise due to the volume of data transferring across networks. However, this performance problem can be mitigated by using distributed data privacy brokers. Moreover, the continued innovation in fog computing to optimize the flow of healthcare data across multiple devices and the cloud infrastructure will help solve the performance issues. There are several research opportunities around improving healthcare data streaming algorithms at Fog computing using distributed messaging systems. Within this distributed infrastructure, there are areas around effective governance and management to provide a near real-time data exchange between healthcare data stakeholders to perform timely healthcare-related decisions; therefore, as IoT devices such as patient care monitoring devices continue to be innovated therefore there are plenty of opportunities for improved architecture, efficient algorithms, and application integration mechanisms. Processing healthcare data streams at the Edge, closer to the data source, can reduce network traffic and improve the latency of time-sensitive healthcare application.(Badidi & Moumane, 2019)

There are also significant research opportunities on the handling of the sheer number of medical records produced daily (structured or unstructured) and the paradigm shift of healthcare services from reactive to proactive care which is monitoring. The power of big data and big data analytics continues to be a threat to the security and privacy of healthcare information. There are several policies, acts and laws that govern the security and privacy of healthcare data in different parts of the world. In Canada, the Personal Information and Electronic Documents Act (PIPEDA) describes the personal information that can only be used for the purpose of which it was collected. Individuals also have a right to know what information is collected and consent also should be provided when sharing data. (Canada, 2021) This Act may not apply to all provinces in Canada, Ontario has its own equivalent law which applies specifically to the Personal Health Information Protection Act. Thus, following these privacy guidelines, and the massive amount of healthcare data, the healthcare information system will be a key consumer of big data. These healthcare data are still in silos in multiple locations among different stakeholders, therefore, there are technical challenges in securely sharing those data within the healthcare information system. Therefore, there are research opportunities in helping solve challenges such as normalizing the data, standardizing individual records in terms of architecture design, and data streaming algorithms. Artificial intelligence can also play a big role in providing meaningful insight in reading the consumption of these big data.

Chapter 7 – Conclusion and Future Recommendation

In this research, we tackled the process of securing the privacy of data with an emphasis on healthcare-related data. After we went through several data protection methods implemented in the healthcare information system, we specifically targeted the classification of data in terms of data privacy and introduced a middleware architectural component within the healthcare information system.

There are two main contributions to this research. The use of machine learning algorithms for the purpose of data classification and the introduction of data privacy brokers in the healthcare information system architecture provides transparent means to engage healthcare data owners when their data is accessed.

Therefore, this thesis addressed by providing solutions for the following two major questions:

1. ***Machine learning algorithms can be used to identify crucial and model influencing data elements for the purpose of protecting data privacy.***
2. ***Centralizing data privacy protection processes by introducing the data privacy broker into the healthcare information system can be an effective way to enhance personal privacy protection.***

As technology, healthcare monitoring devices, smartphones, mobile devices and IoTs, continue to grow substantially, further research should continue in keeping the security of data and preserving the privacy of healthcare data. The research needs to close the privacy-preserving gap and the concerns of organizations and users. Several scenarios of the concern include methods in collecting and retaining sensitive personal information; processing personal information in environments, such as the cloud; and information sharing mechanisms, authentication mechanisms, encryption technologies, data masking and other privacy-protecting techniques.

Despite the promising results in the introduction of the data privacy broker in the healthcare information system architecture, there remain several unsolved challenges in safeguarding data privacy in the healthcare domain.

In some cases, the research community might need the healthcare dataset for research purposes or for public healthcare policy decisions on an urgent basis. In this scenario, we need to find out a very agile method of getting the information quickly so that healthcare policy decisions can be made on time. Data privacy brokers can create delays in the information flow if a massive amount of data is needed to be accessed. Even though we can use the data privacy broker for asynchronous data transfers, a hardware appliance-based data privacy broker can solve the performance problem. Intensive research should continue to evolve in healthcare data privacy methodologies by integrating the emerging areas of AI and ML since data management is increasingly converging on business intelligence, operational systems, and advanced analytics.

Industries are currently using Cloud Access Security Broker (CASBs) in their infrastructure when dealing with on-premises and Cloud landscapes to integrate components to make sure that there is a secured transfer of data between applications. This CASB includes several technology features such as monitoring, auditing, authentication methods and network firewalls. There is room for robust architecture research work on CASB to include integration capabilities.

Ultimately, the challenge is on how to balance the concern in healthcare data privacy and the consumption of healthcare data for the good of society by the healthcare policymakers, healthcare providers, and researchers in academia. In this research thesis, we delivered a couple of state-of-the-art contributions to the healthcare information system in relation to the healthcare data privacy-preserving techniques – identifying and classifying data elements for a dataset based on their impact in exposing the record and architecture that includes the data privacy broker component that provides transparent monitoring and data access awareness solution to all stakeholders including the patient who owns the data.

References

Abbas, A., & Khan, S. U. (2014). A Review on the State-of-the-Art Privacy-Preserving Approaches in the e-Health Clouds. *IEEE Journal of Biomedical and Health Informatics*, *18*(4), 1431–1441. https://doi.org/10.1109/JBHI.2014.2300846

Al Hamid, H. A., Rahman, S. M. M., Hossain, M. S., Almogren, A., & Alamri, A. (2017). A Security Model for Preserving the Privacy of Medical Big Data in a Healthcare Cloud Using a Fog Computing Facility With Pairing-Based Cryptography. *IEEE Access*, *5*, 22313–22328. https://doi.org/10.1109/ACCESS.2017.2757844

Alnemari, A., Romanowski, C. J., & Raj, R. K. (2017). An Adaptive Differential Privacy Algorithm for Range Queries over Healthcare Data. *2017 IEEE International Conference on Healthcare Informatics (ICHI)*, 397–402. https://doi.org/10.1109/ICHI.2017.49

AWS Well-Architected Framework Financial Services Industry Lens. (n.d.). 58.

Azimi, A. (2019, April 8). Is 'Computer Science' and 'Information Systems' the same subject? *Medium*. https://medium.com/@azimidev/is-computer-science-and-information-systems-the-same-subject-4e214292426e

Badidi, E., & Moumane, K. (2019). Enhancing the Processing of Healthcare Data Streams using Fog Computing. *2019 IEEE Symposium on Computers and Communications (ISCC)*, 1113–1118. https://doi.org/10.1109/ISCC47284.2019.8969736

Barril, J. F. H. & Qing Tan. (2017). Integrating privacy in architecture design of student information system for big data analytics. *2017 IEEE 2nd International Conference on Cloud Computing and Big Data Analysis (ICCCBDA)*, 139–144. https://doi.org/10.1109/ICCCBDA.2017.7951899

Bruce, N., Sain, M., & Lee, H. J. (2014). A support middleware solution for e-healthcare system security. *16th International Conference on Advanced Communication Technology*, 44–47. https://doi.org/10.1109/ICACT.2014.6778919

Canada, O. of the P. C. of. (2014, May 15). *Summary of privacy laws in Canada*. https://www.priv.gc.ca/en/privacy-topics/privacy-laws-in-canada/02_05_d_15/

Canada, O. of the P. C. of. (2021, February 11). *The Personal Information Protection and Electronic Documents Act (PIPEDA)*. https://www.priv.gc.ca/en/privacy-topics/privacy-laws-in-canada/the-personal-information-protection-and-electronic-documents-act-pipeda/

CatBoostClassifier. (n.d.). Retrieved December 11, 2021, from https://catboost.ai/docs/concepts/python-reference_catboostclassifier

Confidential Computing Consortium—Open Source Community. (n.d.). Confidential Computing Consortium. Retrieved October 13, 2021, from https://confidentialcomputing.io/

COVID-19 Community Vulnerability Crosswalk—Crosswalk by Census Tract | HealthData.gov. (n.d.). Retrieved December 11, 2021, from https://healthdata.gov/Health/COVID-19-Community-Vulnerability-Crosswalk-Crosswa/x2y5-9muu

Daniels, M., Rose, J., & Farkas, C. (2018). Protecting Patients' Data: An Efficient Method for Health Data Privacy. *Proceedings of the 13th International Conference on Availability, Reliability and Security - ARES 2018*, 1–10. https://doi.org/10.1145/3230833.3230865

Efficient frontier. (2020). In *Wikipedia*. https://en.wikipedia.org/w/index.php?title=Efficient_frontier&oldid=979159350

EHR Connectivity Strategy | eHealth Ontario | It's Working For You. (n.d.). EHealth Ontario. Retrieved December 23, 2021, from https://ehealthontario.on.ca/en/it-professionals/ehr-connectivity-strategy

EHR_Connectivity_Strategy_Summary-en.pdf. (n.d.). Retrieved July 16, 2020, from https://www.ehealthontario.on.ca/images/uploads/pages/documents/EHR_Connectivity_Strategy_Summary-en.pdf

Elmisery, A. M., Rho, S., & Botvich, D. (2016). A Fog Based Middleware for Automated Compliance With OECD Privacy Principles in Internet of Healthcare Things. *IEEE Access*, *4*, 8418–8441. https://doi.org/10.1109/ACCESS.2016.2631546

Esposito, C., De Santis, A., Tortora, G., Chang, H., & Choo, K.-K. R. (2018). Blockchain: A Panacea for Healthcare Cloud-Based Data Security and Privacy? *IEEE Cloud Computing*, *5*(1), 31–37. https://doi.org/10.1109/MCC.2018.011791712

Eze, B., Kuziemsky, C., & Peyton, L. (2018). Operationalizing privacy compliance for cloud-hosted sharing of healthcare data: A case study. *Proceedings of the International Workshop on Software Engineering in Healthcare Systems - SEHS '18*, 18–25. https://doi.org/10.1145/3194696.3194701

Fadheel, W., Salih, R., & Lilien, L. (2018). PHeDHA: Protecting Healthcare Data in Health Information Exchanges with Active Data Bundles. *2018 17th IEEE International Conference On Trust, Security And Privacy In Computing And Communications/ 12th IEEE International Conference On Big Data Science And Engineering (TrustCom/BigDataSE)*, 1187–1195. https://doi.org/10.1109/TrustCom/BigDataSE.2018.00164

Frej, M. B. H., Dichter, J., & Gupta, N. (2019). Comparison of Privacy-Preserving Models Based on a Third-Party Auditor in Cloud Computing. *2019 IEEE Cloud Summit*, 86–91. https://doi.org/10.1109/CloudSummit47114.2019.00020

Fritchman, K., Saminathan, K., Dowsley, R., Hughes, T., De Cock, M., Nascimento, A., & Teredesai, A. (2018). Privacy-Preserving Scoring of Tree Ensembles: A Novel Framework for AI in Healthcare. *2018 IEEE International Conference on Big Data (Big Data)*, 2413–2422. https://doi.org/10.1109/BigData.2018.8622627

Ghafour, S. A., Ghodous, P., & Bonnet, C. (2015). Privacy Preserving Data Integration across Autonomous Cloud Services. *2015 IEEE 8th International Conference on Cloud Computing*, 1099–1102. https://doi.org/10.1109/CLOUD.2015.160

Goldstein, M., & Segall, I. (2015). Automatic and Continuous Software Architecture Validation. *2015 IEEE/ACM 37th IEEE International Conference on Software Engineering*, 59–68. https://doi.org/10.1109/ICSE.2015.135

Harari, Y. N. (2020, March 20). Yuval Noah Harari: The world after coronavirus | Free to read. *Financial Times*. https://www.ft.com/content/19d90308-6858-11ea-a3c9-1fe6fedcca75

He, Z., Cai, Z., Sun, Y., Li, Y., & Cheng, X. (2017). Customized privacy preserving for inherent data and latent data. *Personal and Ubiquitous Computing*, *21*(1), 43–54. https://doi.org/10.1007/s00779-016-0972-2

Hossein, K. M., Esmaeili, M. E., Dargahi, T., & khonsari, A. (2019). Blockchain-Based Privacy-Preserving Healthcare Architecture. *2019 IEEE Canadian Conference of Electrical and Computer Engineering (CCECE)*, 1–4. https://doi.org/10.1109/CCECE.2019.8861857

IEEE Standard for System, Software, and Hardware Verification and Validation. (n.d.). IEEE. https://doi.org/10.1109/IEEESTD.2017.8055462

Janjic, V., Bowles, J. K. F., Vermeulen, A. F., Silvina, A., Belk, M., Fidas, C., Pitsillides, A., Kumar, M., Rossbory, M., Vinov, M., Given-Wilson, T., Legay, A., Blackledge, E., Arredouani, R., Stylianou, G., & Huang, W. (2019). The SERUMS tool-chain:

Ensuring Security and Privacy of Medical Data in Smart Patient-Centric Healthcare Systems. *2019 IEEE International Conference on Big Data (Big Data)*, 2726–2735. https://doi.org/10.1109/BigData47090.2019.9005600

Kifer, D., & Machanavajjhala, A. (2011). No free lunch in data privacy. *Proceedings of the 2011 International Conference on Management of Data - SIGMOD ’11*, 193. https://doi.org/10.1145/1989323.1989345

Kissi Mireku, K., Zhang, F., & Komlan, G. (2017). Patient knowledge and data privacy in healthcare records system. *2017 2nd International Conference on Communication Systems, Computing and IT Applications (CSCITA)*, 154–159. https://doi.org/10.1109/CSCITA.2017.8066543

Kundalwal, M. K., Singh, A., & Chatterjee, K. (2018). A Privacy Framework in Cloud Computing for Healthcare Data. *2018 International Conference on Advances in Computing, Communication Control and Networking (ICACCCN)*, 58–63. https://doi.org/10.1109/ICACCCN.2018.8748480

Kupwade Patil, H., & Seshadri, R. (2014). Big Data Security and Privacy Issues in Healthcare. *2014 IEEE International Congress on Big Data*, 762–765. https://doi.org/10.1109/BigData.Congress.2014.112

L-diversity. (2020). In *Wikipedia*. https://en.wikipedia.org/w/index.php?title=L-diversity&oldid=969789348

Li, N., Li, T., Venkatasubramanian, S., & Labs, T. (n.d.). *t-Closeness: Privacy Beyond k-Anonymity and -Diversity*. 10.

Li, Z., & Pino, E. J. (2019). D&D: A Distributed and Disposable Approach to Privacy Preserving Data Analytics in User-Centric Healthcare. *2019 IEEE 12th Conference on Service-Oriented Computing and Applications (SOCA)*, 176–183. https://doi.org/10.1109/SOCA.2019.00033

Lu, R., Lin, X., & Shen, X. (2013). SPOC: A Secure and Privacy-Preserving Opportunistic Computing Framework for Mobile-Healthcare Emergency. *IEEE Transactions on Parallel and Distributed Systems*, *24*(3), 614–624. https://doi.org/10.1109/TPDS.2012.146

Lundberg, S. M., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems*, *30*. https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html

Machanavajjhala, A., Gehrke, J., Kifer, D., & Venkitasubramaniam, M. (2006). L-diversity: Privacy beyond k-anonymity. *22nd International Conference on Data Engineering (ICDE'06)*, 24–24. https://doi.org/10.1109/ICDE.2006.1

Mivule, K. (n.d.). *Targeted Data Swapping and K-Means Clustering for Healthcare Data Privacy and Usability*. 5.

Navuluri, K., Mukkamala, R., & Ahmad, A. (2016). Privacy-Aware Big Data Warehouse Architecture. *2016 IEEE International Congress on Big Data (BigData Congress)*, 341–344. https://doi.org/10.1109/BigDataCongress.2016.53

Nishi, Y. (2015). Design principles in test suite architecture. *2015 IEEE Eighth International Conference on Software Testing, Verification and Validation Workshops (ICSTW)*, 1–4. https://doi.org/10.1109/ICSTW.2015.7107426

O'Connell, J. (n.d.). *Jennifer O'Connell: Covid-19 is normalising mass surveillance of citizens*. The Irish Times. Retrieved July 17, 2020, from https://www.irishtimes.com/opinion/jennifer-o-connell-covid-19-is-normalising-mass-surveillance-of-citizens-1.4213829

Pfluger, A., Golubski, W., & Queins, S. (2011). Model Driven Validation of System Architectures. *2011 IEEE 13th International Symposium on High-Assurance Systems Engineering*, 25–28. https://doi.org/10.1109/HASE.2011.46

Pflüger, A., Golubski, W., Queins, S., & Cramergasse, V. (n.d.). *System Architecture Validation with UML*. 6.

Puppala, M., He, T., Yu, X., Chen, S., Ogunti, R., & Wong, S. T. C. (2016). Data security and privacy management in healthcare applications and clinical data warehouse environment. *2016 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)*, 5–8. https://doi.org/10.1109/BHI.2016.7455821

Rao, S., Suma, S. N., & Sunitha, M. (2015). Security Solutions for Big Data Analytics in Healthcare. *2015 Second International Conference on Advances in Computing and Communication Engineering*, 510–514. https://doi.org/10.1109/ICACCE.2015.83

Rghioui, A., Laarje, A., Elouaai, F., & Bouhorma, M. (2015). Protecting E-healthcare Data Privacy for Internet of Things Based Wireless Body Area Network. *Research Journal of Applied Sciences, Engineering and Technology*, *9*(10), 876–885. https://doi.org/10.19026/rjaset.9.2638

Robson, B., & Srinidhi Boray. (2016). *Studies Using a Universal Exchange Language Solution for Application of a BlockChain Approach in Healthcare.* https://doi.org/10.13140/RG.2.1.4199.3209

Salahuddin, M. A., Al-Fuqaha, A., Guizani, M., Shuaib, K., & Sallabi, F. (2017). Softwarization of Internet of Things Infrastructure for Secure and Smart Healthcare. *Computer*, *50*(7), 74–79. https://doi.org/10.1109/MC.2017.195

Samarati, P., & Sweeney, L. (n.d.). *Protecting Privacy when Disclosing Information: K-Anonymity and Its Enforcement through Generalization and Suppression*. 19.

Sharif, M. I., Li, J. P., Ullah, S., Ul Haq, A., & Alam, G. (2019). An Efficient Access Privacy Protocol for Healthcare Patient Information System. *2019 16th International Computer Conference on Wavelet Active Media Technology and Information Processing*, 461–465. https://doi.org/10.1109/ICCWAMTIP47768.2019.9067602

Sharma, S., Chen, K., & Sheth, A. (2018). Toward Practical Privacy-Preserving Analytics for IoT and Cloud-Based Healthcare Systems. *IEEE Internet Computing*, *22*(2), 42–51. https://doi.org/10.1109/MIC.2018.112102519

Singh, N., Jangra, A., Elamvazuthi, I., & Kashyap, K. (2017). Healthcare data privacy measures to cure & care cloud uncertainties. *2017 4th International Conference on Signal Processing, Computing and Control (ISPCC)*, 402–407. https://doi.org/10.1109/ISPCC.2017.8269712

Su, X., Hyysalo, J., Rautiainen, M., Riekki, J., Sauvola, J., Maarala, A. I., Hirvonsalo, H., Li, P., & Honko, H. (2016). Privacy as a Service: Protecting the Individual in Healthcare Data Processing. *Computer*, *49*(11), 49–59. https://doi.org/10.1109/MC.2016.337

Tao, Y., Wang, X., Xu, X., & Yu, J. (2016). A Hybrid Transaction Processing and Data Analysis Framework: A Use Case Study for Multi-Source Healthcare Data Management. *2016 6th International Conference on Digital Home (ICDH)*, 165–169. https://doi.org/10.1109/ICDH.2016.043

Tekinerdogan, B., Clements, P., Muccini, H., Chaudron, M., Polini, A., & Woods, E. (2011). Architecture-Based Testing and System Validation—Workshop Summary. *2011 Ninth Working IEEE/IFIP Conference on Software Architecture*, 341–341. https://doi.org/10.1109/WICSA.2011.53

The AWS Security Reference Architecture—AWS Prescriptive Guidance. (n.d.). Retrieved December 23, 2021, from https://docs.aws.amazon.com/prescriptive-guidance/latest/security-reference-architecture/architecture.html

Top_10_Strategic_Tec_726890_ndx.pdf. (n.d.).

What is a CASB (Cloud Access Security Broker)? Definition from WhatIs.com. (n.d.). SearchCloudSecurity. Retrieved July 31, 2020, from https://searchcloudsecurity.techtarget.com/definition/cloud-access-security-brokers-CABs

White, S. K. (2018, August 10). *What is TOGAF? An enterprise architecture methodology for business*. CIO. https://www.cio.com/article/3251707/what-is-togaf-an-enterprise-architecture-methodology-for-business.html

Xiaohui Bai. (2008). Study of C4ISR Architecture Simulation Validation with UML and Object-Based Petri Nets. *2008 IEEE 8th International Conference on Computer and Information Technology Workshops*, 571–576. https://doi.org/10.1109/CIT.2008.Workshops.61

Yuxuan, Y., Xianyu, Z., & Yuanjie, J. (2020). Sociological Aspects of Big Data Privacy. *Proceedings of the 2020 12th International Conference on Machine Learning and Computing*, 230–235. https://doi.org/10.1145/3383972.3384075