UbiSecE: UbiLab's Secure Cloud Environment for Public Health Research in Microsoft Azure

Pedro Augusto Miranda[#]
PhD Candidate
School of Public Health Science
Ubiquitous Health Technology Lab
University of Waterloo, ON, Canada
Pedro.miranda@uwaterloo.ca

Jasleen Kaur[#]
Postdoctoral Fellow
School of Public Health Science
Ubiquitous Health Technology Lab
University of Waterloo, ON, Canada
Jasleen.kaur@uwaterloo.ca

Plinio Morita[*]
Associate Professor
School of Public Health Science
Ubiquitous Health Technology Lab
University of Waterloo, ON, Canada
Plinio.morita@uwaterloo.ca

[#] These authors contributed equally to this work and share the first authorship
[*]Corresponding Author

Abstract— **The use of Personal Health Information (PHI) has become increasingly popular in public health research in recent years. However, many researchers have stored collected PHI in local databases or filesystems with limited centralized storage. This has raised concerns about cybersecurity, the lack of standards, and the absence of a data governance program. To address these issues, a cloud-based infrastructure was developed for public health research over PHI that meets the requirements of Ubilab, a public health research group at the University of Waterloo. UbiSecE, a Secure Cloud-Based Infrastructure for Public Health Research, was designed by adapting Microsoft Azure's cloud infrastructure to meet the needs of Ubilab. Relevant laws, regulations, and standards, such as PIPEDA, GPDR, FIPPA, and PHIPA, that govern the utilization of PHI for public health research were identified. Additionally, the lab's actors, social norms, processes, and collective problems were analyzed to establish the foundation of the data governance program in Azure. Azure's data governance architecture guidelines were followed to provide the primary governance mechanisms for evaluating, guiding, and monitoring UbiSecE resources and processes. To ensure the secure maintenance of PHI, role-based access controls were implemented for all users, and all governance processes were deployed via Azure. Furthermore, NIST 800-53 compliance was integrated for all deployed resources. UbiSecE offers a centralized, private, and secure environment for public health research, which enables different users with different roles to conduct research with PHI.**

Keywords—Public health research, Personal health information (PHI), Cloud-based infrastructure, Data governance, Research data security, Data management

I. INTRODUCTION

The use of personal health information (PHI) in public health research has increased in recent years due to its ability to provide valuable insights into population health and address pressing health issues facing communities [1-3]. However, maintaining collected PHI in local databases or filesystems poses significant cybersecurity risks, as the information is vulnerable to breaches, unauthorized access, or accidental loss [4-6]. Without centralized storage, it becomes difficult to manage and secure the sensitive information. Moreover, the lack of standards and data governance programs exacerbates these concerns, as there is no set framework for securing PHI, limiting the ability of researchers to collaborate and share information. Additionally, the lack of oversight and accountability in the collection, management, and use of PHI can result in ethical violations or unintended consequences [7-8]. Therefore, there is a critical need to implement a data governance framework for the storage and use of PHI in public health research to mitigate cybersecurity risks and ensure privacy, confidentiality, and ethical research practices.

The UbiLab [9] at the University of Waterloo is working on different public health research projects [10]. Previously, the research projects were conducted on local systems. Each student/researcher at UbiLab has access to personalized resources, a development environment, and a database on their own machine, however with limited IT management experience leading to decentralized storage of research data.

The main objective of this research is to establish a private and secure cloud environment for public health research at UbiLab. To achieve this, the paper proposes a framework based on Microsoft Azure's governance architecture [11] and healthcare standards such as Personal Information Protection and Electronics Documents Act (PIPEDA) [12], Healthcare Insurance Portability and Accountability Act (HIPAA) [13], and General Data Protection Regulation (GDPR) [14]. This research focuses on developing the proposed environment design and framework, which could lead to better management of research data and ensure its privacy and security.

II. METHODS

A. Requirements Elicitation

The study commenced with conducting interviews with Ubilab's stakeholders to gain insight into their requirements concerning data governance and cloud computing and storage. Proposed diagrams of architectures and governance structures were presented and discussed to ensure that the stakeholders had a comprehensive understanding of the technical aspects related to security and governance. These discussions aimed to facilitate stakeholder approval of the final architecture and framework.

B. System and Cybersecurity Architecture

The proposed system and cybersecurity architecture framework is built on Microsoft Azure's governance framework, which offers a comprehensive set of best practices and guidelines to ensure the secure and efficient utilization of cloud resources. By using this framework, the proposed architecture offers a robust security mechanism that can safeguard the confidentiality, integrity, and availability of the system's data and services. The architecture framework employs a multi-layered security approach, where security measures are integrated throughout the entire system, including the application, network, data, and user levels.

C. Compliance with Healthcare Standards

The proposed framework has been designed to comply with key healthcare laws and standards, including PIPEDA, HIPAA, and GDPR, to ensure that Protected Health Information (PHI) is stored and managed securely and compliantly. Compliance with these standards ensures that the privacy and confidentiality of individuals are protected, and research is conducted in an ethical and responsible manner. In addition to the North-American requirements, the framework also adheres to the GDPR standard to comply with the research partners across Europe. This ensures that the framework complies with the varying legal requirements across different regions, making it an ideal solution for global research collaboration.

The proposed framework's compliance with these healthcare standards is achieved by implementing a comprehensive set of security and privacy measures, including access controls, data encryption, audit logging, and data backup and recovery. These measures work together to ensure the security and privacy of PHI, while also ensuring that research activities are conducted in a compliant and responsible manner.

D. CyberSecurity Best Practices & Tools

Cloud security best practices, such as those outlined by the National Institute of Standards and Technology (NIST) [15] and Microsoft Azure's governance architecture, have been followed to ensure that the cloud infrastructure is secure and resilient. The use of role-based access controls, encryption, and network segmentation has been implemented. Further, to enhance the security of the cloud infrastructure, different security tools and technologies have been deployed in the proposed framework such as firewalls, intrusion detection and prevention systems, and data loss prevention tools. These security tools helped to mitigate cybersecurity risks and protect the privacy and confidentiality of PHI.

E. Role-Based Access Control

UbiSecE incorporates role-based access controls to facilitate controlled data sharing within the platform. As per the system's design, academics with appropriate permissions can request access to a subset of data that meets their specific requirements. These requirements could include factors such as data types, time frames, demographics, or any other relevant criteria. To enable data sharing with students, an academic/ external can collaborate with the UbiSecE platform to create a controlled environment where the student can conduct their study. This collaboration involves providing the necessary access rights to the student within the established role-based access control framework. The student would be granted access only to the subset of data approved by the academic, ensuring compliance with relevant laws, regulations, and standards governing the use of Personal Health Information (PHI) for public health research. Future plans include removing VPN access and utilizing private links and virtual desktops to avoid data leaks.

F. Data Synchronization

The synchronization between the repository and live systems in UbiSecE is achieved through continuous data integration processes. The cloud-based infrastructure leverages Microsoft Azure's capabilities to establish real time synchronization mechanisms. By utilizing Azure's data integration services, such as Azure Data Factory or Azure Logic Apps, the repository can be seamlessly connected to live systems, ensuring that any updates or changes in the live systems are automatically reflected in the repository. This synchronization process allows researchers in UbiSecE to access the most up-to-date and accurate data for their public health research, promoting data consistency and minimizing the risk of working with outdated information. Although we leverage Azure's capabilities for data synchronization, there is currently no automatic pipeline for data migration. Future work includes automating data migration.

G. Data Vocabularies

When underlying vocabularies or collected data change, the UbiSecE infrastructure can accommodate these modifications through its flexible design. The system supports scalability and adaptability, allowing researchers to update and align the vocabularies and data collection methods as needed. The centralized storage in the cloud-based infrastructure simplifies the process of managing changes, ensuring that all users have access to the most up-to-date information.

H. Data Governance

The union of the system architecture, compliance with healthcare standards, and cloud security best practices are what gives support to data governance in our cloud environment [11], [16]–[23]. The architecture is planned and designed based on the current data agenda and strategy, ensuring that outcomes from PHI research, sharing, and maintenance align with defined objectives. The compliance with healthcare and cybersecurity standards is overseen by UbiLab's governance committee to ensure that all types of research conducted are encompassed and guided by the chosen standards, thus limiting inappropriate PHI utilization and guiding researchers' studies. By following these guidelines, we ensure the proper handling of sensitive healthcare data, maintain the integrity of the research process, and safeguard patient privacy.

III. Results

The developed public health research cloud infrastructure is shown in Figure 1. Our research successfully developed a cloud-architecture tailored for public health research capable of meeting the requirements of stakeholders being cybersecurity, data governance, compliance with healthcare standards (limited to HIPPA, PIPEDA and GDPR for now), whilst providing tailored and customizable environments for students, researchers, and any other stakeholder involved in public health research.

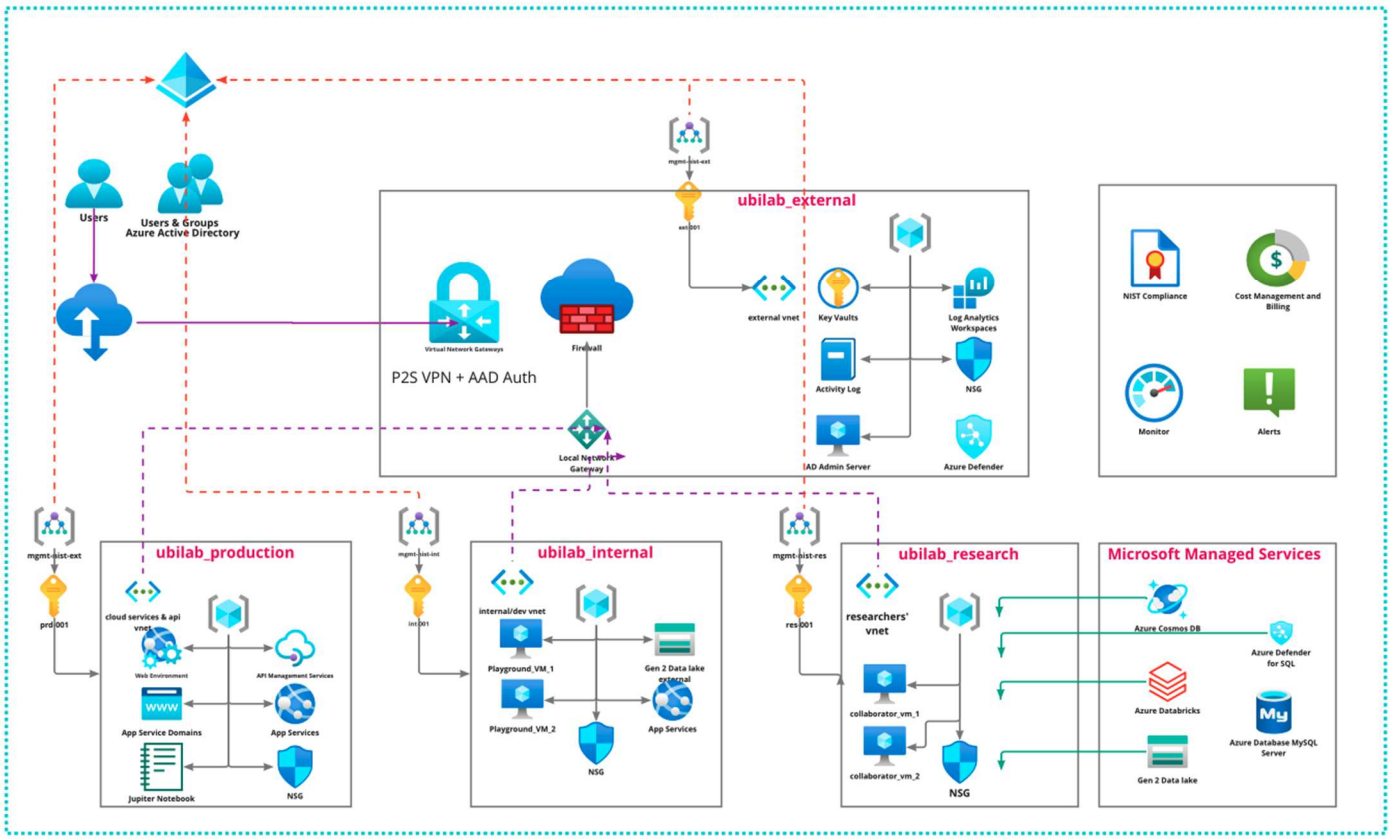

Figure 1: UbiLab's secure cloud architecture for public health research

The developed infrastructure has a root level management group that holds all other management groups and subscriptions. Users access resources through a point-to-site connection to a private VPN unless they have requested otherwise. Personal health information (PHI) is stored separately from all users, and access is restricted and requires authorization from a data custodian. The PHI stored is encrypted at rest and is geo-replicated in the US and Canada only. We use Azure's firewall and subnet to secure and restrict public and internal access to resources and to logically map our cloud network, respectively.

To ensure the security of our cloud infrastructure, the entire tenant is under the enforcement of NIST-800-171 and NIST-800-52 via Azure's policy enforcement services. UbiLab is on its path to full compliance with these standards, but there is still room for improvement in our internal processes and refining our governance program to achieve full compliance.

The developed UbiLab's secure cloud infrastructure is composed of 5 main components or tiers which work together to provide a comprehensive and secure cloud infrastructure for public health research at UbiLab.

UbiLab External - This component holds the entry point for our secure cloud infrastructure. It redirects requests to appropriate resources based on a local network gateway and Azure's firewall. It also includes directory level resources that are only accessible to cloud administrators and are used to manage the cloud.

UbiLab Internal - This component holds internal resources available only for UbiLab's members, such as developers, managers, volunteers, and co-ops students. Access to these resources is based on role-based control access and secure VPN connections.

UbiLab Production - This component holds resources that are accessible without the need for a VPN connection. This component can host websites or public services that we choose to provide to the community.

UbiLab Research - This component holds internal resources that are available only to UbiLab's faculty, PhD, and Post-doc researchers. These resources may include virtual machines, databricks instances, and app services. Access to these resources is authorized via role-based control access, and access to protected health information (PHI) is limited only to data relevant or generated by the researchers' studies.

Microsoft Managed Services - This component is responsible for hosting UbiLab's PHI collected for research and resources that store PHI. Only cloud managers have direct access to its resources, and private links are utilized to provide access for researchers or appropriate members of UbiLab access to 'just enough data' for their tasks.

The proposed secure cloud architecture is currently being used in UbiLab. The results indicate that centralizing the data storage has reduced the overall complexity of data management, where the data management becomes more streamlined and efficient. Additionally, the security of UbiLab's research data has been enhanced due to the reduced risk of potential threats and vulnerabilities. However, granting data access to new collaborators requires management and knowledge of multiple SQL servers, storage accounts, and drivers. Currently, any new

collaborator/ researcher is issued a registered user account with limited access, which is managed by a cloud administrator. The centralization of data has also facilitated the visualization of the lab's entire data storage and content, making it easier to reuse or merge data from different studies. Furthermore, UbiLab's focus on research using public health data from IoT, sensors, and smart devices necessitates NIST compliance, which will simplify future compliances such as HIPAA. Achieving HIPAA compliance would enable UbiLab to extend its data ecosystem to include medical records, leading to more research opportunities and potentially improving ongoing research quality.

IV. Conclusion

To summarize, the private and secure cloud-based data governance framework proposed in this research paper for UbiLab's public health research effectively addresses the challenges faced in the storage and management of PHI in local databases and filesystems. The framework provides a centralized and secure environment for the storage and use of PHI, ensures compliance with healthcare standards, and supports ethical public health research. It is designed based on Microsoft Azure's governance architecture and adheres to key healthcare standards like PIPEDA, HIPAA, and GDPR, which specify legal requirements for safeguarding personal health information.

Using a cloud-based data governance framework offers the benefits of scalability and accessibility of cloud computing and helps ensure that public health research is conducted in an ethical and responsible manner. This research article can be a valuable reference for organizations and public health researchers interested in developing a secure and compliant cloud-based infrastructure for conducting public health research. Furthermore, this secure centralized framework facilitates seamless collaboration among different teams, departments, and public health researchers.

References

[1] Pagliari, C., Detmer, D., & Singleton, P. (2007). Potential of electronic personal health records. British Medical Journal, 335(7615), 330–333. https://doi.org/10.1136/bmj.39279.482963.ad

[2] Peck, E. M., Ayuso, S. E., & El-Etr, O. (2019). Data is personal: Attitudes and perceptions of data visualization in rural Pennsylvania. Conference on Human Factors in Computing Systems - Proceedings, 12. https://doi.org/10.1145/3290605.3300474

[3] Schwartz, P. M., & Solove, D. J. (2011). The PII problem: Privacy and a new concept of personally identifiable information. New York University Law Review, 86(6), 1814–1894. https://heinonline.org/HOL/Page?handle=hein.journals/nylr86&id=1826&div=50&collection=journals

[4] Conaty-Buck, S. (2017). Cybersecurity and Healthcare Records: Tips for Ensuring Patient Safety and Privacy. In American Nurse Today (Vol. 12, Issue 9). https://goo.gl/zG1mjT.

[5] Coventry, L., & Branley, D. (2018). Cybersecurity in healthcare: A narrative review of trends, threats and ways forward. In Maturitas (Vol. 113, pp. 48–52). Elsevier Ireland Ltd. https://doi.org/10.1016/j.maturitas.2018.04.008

[6] Jedrzejczyk, L., Price, B. A., Bandara, A. K., Nuseibeh, B., Hall, W., & Keynes, M. (2009). I Know What You Did Last Summer : risks of location data leakage in mobile and social computing . 1–9. http://computing.open.ac.uk

[7] Micheli, M., Ponti, M., Craglia, M., & Suman, A. B. (2020). Emerging models of data governance in the age of datafication: Https://Doi.Org/10.1177/2053951720948087, 7(2). https://doi.org/10.1177/2053951720948087

[8] Tse, D., Chow, C., Ly, T., Tong, C., & Tam, K. (2018). The Challenges of Big Data Governance in Healthcare. 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), 1632–1636. https://doi.org/10.1109/TrustCom/BigDataSE.2018.00240

[9] "Home | Ubiquitous Health Technology Lab." https://uwaterloo.ca/ubiquitous-health-technology-lab/ (accessed Feb. 11, 2023).

[10] Projects | Ubiquitous Health Technology Lab. (n.d.). Retrieved April 23, 2023, from https://uwaterloo.ca/ubiquitous-health-technology-lab/projects

[11] "Azure Governance - US Partner Community Blog - Microsoft." https://www.microsoft.com/en-us/us-partner-blog/2019/07/24/azure-governance/ (accessed Feb. 11, 2023).

[12] Office of the Privacy Commissioner of Canada, "PIPEDA fair information principles," 2019. https://www.priv.gc.ca/en/privacy-topics/privacy-laws-in-canada/the-personal-information-protection-and-electronic-documents-act-pipeda/p_principle/ (accessed Jan. 26, 2020).

[13] "HIPAA Home | HHS.gov." https://www.hhs.gov/hipaa/index.html (accessed Feb. 11, 2023).

[14] "Art. 4 GDPR - Definitions - GDPR.eu." https://gdpr.eu/article-4-definitions/ (accessed Jan. 18, 2021).

[15] "SP 800-171 Rev. 2, Protecting CUI in Nonfederal Systems and Organizations | CSRC." https://csrc.nist.gov/publications/detail/sp/800-171/rev-2/final (accessed Feb. 11, 2023).

[16] I. Alhassan, D. Sammon, and M. Daly, "Critical Success Factors for Data Governance: A Theory Building Approach," *Information Systems Management*, vol. 36, no. 2, pp. 98–110, Apr. 2019, doi: 10.1080/10580530.2019.1589670.

[17] R. Abraham, J. Schneider, and J. vom Brocke, "Data governance: A conceptual framework, structured review, and research agenda," *International Journal of Information Management*, vol. 49. Elsevier Ltd, pp. 424–438, Dec. 01, 2019. doi: 10.1016/j.ijinfomgt.2019.07.008.

[18] V. Khatri and C. v. Brown, "Designing data governance," *Commun ACM*, vol. 53, no. 1, pp. 148–152, Jan. 2010, doi: 10.1145/1629175.1629210.

[19] M. Al-Ruithe, E. Benkhelifa, and K. Hameed, "A systematic literature review of data governance and cloud data governance," *Pers Ubiquitous Comput*, vol. 23, no. 5–6, pp. 839–859, Nov. 2019, doi: 10.1007/s00779-017-1104-3.

[20] G. Cheng, Y. Li, Z. Gao, and X. Liu, "Cloud data governance maturity model," *Proceedings of the IEEE International Conference on Software Engineering and Service Sciences, ICSESS*, vol. 2017-Novem, pp. 517–520, Apr. 2018, doi: 10.1109/ICSESS.2017.8342968.

[21] D. Tse, C. Chow, T. Ly, C. Tong, and K. Tam, "The Challenges of Big Data Governance in Healthcare," in 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), Aug. 2018, pp. 1632–1636. doi: 10.1109/TrustCom/BigDataSE.2018.00240.

[22] "CIHI's Health Data and Information Governance and Capability Framework," 2020.

[23] "The Seven Simple Principles of Health Care Data Governance." https://www.advisory.com/en/topics/information-security/2018/04/the-seven-simple-principles-of-health-care-data-governance (accessed Aug. 08, 2021.