HEALER2: a framework for secure data lake towards healthcare digital transformation efforts in low and middle-income countries.

YENG, P.K., DIEKUU, J.-B., ABOMHARA, M., ELHADJ, B., YAKUBU, M.A., OPPONG, I.N., FAUZI, M.A., YANG, B. and EL-GASSAR, R.

2023

HEALER2: A Framework for Secure Data Lake towards healthcare digital transformation efforts in Low and Middle-income Countries

Prosper Kandabongee Yeng
Department of Information Security and Communication Technology
NTNU
Gjøvik, Norway
prosper.yeng@ntnu.no

John-Bosco Diekuu
Department of Information Security and Communication Technology
RGU
Garthdee Road, UK
j.diekuu@rgu.ac.uk

Mohamed Abomhara
Department of Information Security and Communication Technology
NTNU
Gjøvik, Norway
mohamed.abomhara@ntnu.no

Benkhelifa Elhadj
Department of digital transformation and Electronic Communication
Staffordshire university
Staffordshire university, UK
e.benkhelifa@staffs.ac.uk

Mustapha Awinsongya Yakubu
University of Cincinnati
Ohio, USA
yakubuma@mail.uc.edu

Isaac Nyame Oppong
Oracle
Accra, Ghana
nyameoppong84@gmail.com

Adejoke Odebade
Department of digital transformation and Electronic Communication
Staffordshire university
Staffordshire university, UK
adejoke.odebade@research.staffs.ac.uk

Muhammad Ali Fauzi
Department of Information Security and Communication Technology
NTNU
Gjøvik, Norway
muhammad.a.fauzi@ntnu.no

Bian Yang
Department of Information Security and Communication Technology
NTNU
Gjøvik, Norway
bian.yang@ntnu.no

Rania El-Gassar
Department of Business, Marketing and Law
Campus Ringerike (A 312)
University of Southen Norway
Norway
rania.el-gazzar@usn.no

***Abstract*—Low and middle-income countries are vigorously digitizing their operations in the healthcare sector as steps in the digital transformation journey. However, some of the basic principles in information security are being skipped. This has a tendency to introduce fundamental vulnerabilities in the core foundation of their healthcare IT infrastructure. This paper, therefore, assessed e-health strategies in Africa and proposed a data lake framework for healthcare IT infrastructure which is deemed secure, privacy-preserving and economically efficient.**

***Index Terms*—Healthcare, Cyber security, data lake, data warehouse, digital transformation**

I. Introduction

In recent years, the healthcare industry has witnessed a significant s hift t owards d igital t ransformation e fforts worldwide. The infamous COVID-19 clearly exposed the power of digital transformation in healthcare across the globe [1]. Countries or continents that took the lead in healthcare digital transformation exhibited their agility in contact tracing, planning, identifying and vaccinating the appropriate age groups, leading to effective controls. However, jurisdictions that lacked the data did not have the opportunity to leverage automated systems like AI and rule-based to gain the muscles that come with digital transformation.

Low and middle-income countries (LMICs) have been particularly keen on embracing this transformation to improve healthcare delivery, enhance patient outcomes, and address existing challenges [2]. One key aspect of this digital revolution is the adoption of data lakes [3]–[5], which are large repositories of diverse healthcare data from various sources, such as electronic health records, medical imaging, wearable devices, and genomics. Data lakes offer a unique opportunity for LMICs to leverage the power of big data analytics and machine learning algorithms to gain valuable insights and make data-driven decisions in healthcare. However, the successful implementation of a data lake in the healthcare domain of LMICs faces significant challenges, particularly in terms of data security [6]–[9]. Healthcare data is highly sensitive and confidential, containing personal health information (PHI) and personally identifiable information (PII) [10], [10]–[12]. Therefore, ensuring the security and privacy of data within a data lake is of paramount importance to safeguard patient privacy, comply with regulatory requirements, and build trust among stakeholders. The research question addressed in this study is "What data lake framework exists for LMIC through the lens of security, privacy and cost" for health in the digital health transformation era? This paper, therefore, presents a comprehensive and cost-effective framework for securing a data lake in the context of healthcare digital transformation efforts in LMICs. The framework incorporates various security measures, best practices, and technologies while being mindful of the cost burden to address the unique challenges faced by healthcare organizations in these countries. By following this framework, healthcare institutions can establish a secure and robust data lake environment that enables the effective utilization of healthcare data while ensuring compliance with privacy regulations and protecting patient confidentiality. The remainder of this paper is structured as follows. Section 2 provides the approach of the study. Section 3 provides an overview of the current state of healthcare digital transformation efforts in Africa and their security-related gaps. Section 4 presents related works focusing on the security aspects. Section 5 presents the proposed framework for securing a data lake, outlining the key components and security measures involved and providing a discussion of the potential benefits and impact of implementing the framework in LMICs. Finally, Section 6 concludes the paper and provides recommendations for future research.

II. Study approach

In this study, we assessed the various digital health strategies developed by African countries. The security considerations outlined in these digital health strategies were the primary focus of the assessment. In addition to that, related work was found in google scholar, IEEE Explore and PubMed. These data lake frameworks and their security considerations in healthcare were reviewed against security considerations and their cost of implementation. Keywords such as data lake, health and big data were used to identify the relevant articles. A summary of the findings and their related gaps are shown in section III and IV.

III. Information security gaps in digital health strategies of Africa

Since 2005, the World Health Assembly (WHA) has been encouraging Member States to embrace digital transformation by developing and implementing digital health strategies to contribute towards achieving the Sustainable Development Goals (SDGs) and universal health coverage (UHC) [13]. In 2020, the WHO formulated a global digital health strategy with the vision of "health for all" and SDG-3. Subsequently, the WHO's Regional Committee for Africa resolved to advance eHealth solutions in the African Region. A review of eHealth strategies revealed that out of the 54 African countries, 42 (78.8%) have developed eHealth strategies. These strategies encompass various transformation areas, including telemedicine, electronic health record systems, eLearning, and others. The trajectory of digital transformation implementation involves experimentation, early adoption, development and growth, scale-up, and mainstreaming. Nigeria, for example, is currently in the development and growth stage [14].

One notable advancement in the digital transformation of healthcare is the significant transformation of electronic health records in most African countries through the District Health Information System (DHIS2) [14]. This marks a significant milestone; however, only two countries (South Africa and Tanzania) out of the 42 have explicitly mentioned in their strategies that they have developed dedicated eHealth-specific security strategies [15], [16]. The WHO identified 12 African countries that have implemented security strategies, but these strategies mainly consist of general national regulations or legislations that incorporate some measures for healthcare data protection [13]. Although these strategies consider aspects of digital transformation such as data warehousing or data lakes, they do not adequately address the security dimension. For instance, the Health ICT strategy of Nigeria showed significant progress in its implementation however, there exists a gap in the security incorporation. Whiles the strategy indicated higher progression in the digital implementation, the security and interoperability aspect was yet to be planned for. This might not be in line with the "shift-left" principle of security and privacy [17], [18].

IV. Related studies of data lakes

In addition to an evaluation of diverse eHealth strategies, this section delved into an exploration of pertinent data lakes, specifically focusing on their security and cost implications. Within the realm of digital transformation, data storage and processing form pivotal pillars. When devising digital transformation strategies within healthcare, especially for resource-constrained regions such as Africa, careful deliberation of both cost and security dimensions becomes paramount. This proactive approach ensures the sustainability of transformative initiatives.

In healthcare, a domain of paramount significance, data diversity is pronounced. The spectrum encompasses varied unstructured attributes, including textual medical reports, medical images, genomic data, and vital sign waveforms [19]. Recognizing the absence of a universal Data Lake architecture tailored for healthcare's myriad scenarios, the HEALER framework was conceived. Engineered to aptly ingest, store, and provide efficient access to diverse healthcare data, HEALER furnishes a centralized repository. It uniquely facilitates analysis and querying regardless of data format or type. A practical manifestation of HEALER validated its capability in data ingestion, waveform processing, access provisioning, and natural language report analysis. Unlike traditional systems, HEALER prioritizes the storage of unstructured data – an often overlooked aspect.

Cappiello et al.'s [3] proposition introduces a data lake framework centred on establishing a communal data space within healthcare. This framework embraces data governance, privacy concerns, data sources, ingestion pipelines, storage containers, and access tools. A prime objective was fostering data sharing and collaboration within the healthcare ecosystem to bolster patient care and medical advancements. While addressing big data attributes and data sharing, the framework lacked emphasis on the critical consideration of implementation cost, a factor of substantial significance in healthcare institutions often grappling with financial constraints.

Similarly, the spotlight on cost and security considerations extends to non-healthcare contexts. In the realm of business intelligence architecture, Marilex [5] unearthed the multifaceted role of data lakes through interviews with BI experts. While illuminating typical data lake components, the study was confined to larger enterprises in developed economies. Thus, extrapolation to small businesses and low-income countries is limited. Additionally, the experts' perspectives on privacy and security were conspicuously absent, leaving an information gap crucial for enriching data lake frameworks tailored to LMICs.

Youssef et al.'s contribution [4] spotlighted a data lake framework catering to the utility industry. This framework aspired to centralize data repositories for real-time analytics, capitalizing on a layered structure encompassing data integration, storage, processing, and consumption. Regrettably, the critical dimension of cost was omitted from its considerations, overlooking a vital facet in the development and maintenance of data lakes.

In a broader context, while these frameworks addressed various aspects of data lakes and their implications, the synthesis of cost-efficiency and security in data lake implementation remains a pressing concern that necessitates comprehensive exploration, particularly when tailoring solutions for LMICs.

This paper, therefore, proposed HEALER2: a conceptual framework of a data lake towards addressing the security and cost aspects of data storage in healthcare. This framework can be used in LMICs, thereby filling an important gap in the existing literature.

V. Proposed data lake framework and discussion

A comprehensive framework (HEALER2) was proposed to address various aspects, including cost implications, security, and privacy considerations, in the development of a data lake for LMIC (Low- and Middle-Income Countries).

Regarding security and privacy, we propose the implementation of privacy and data governance frameworks, as depicted in Figure 1. These frameworks can be aligned with the Ministry of Health or the health organization's Data Lake establishment. This involves the establishment of an ethical board for data-driven decisions and the formation of a committee responsible for data governance and privacy management policies and procedures, in accordance with national laws and regulations. Additionally, a quality control committee will be established to define data quality metrics, rules, procedures, data profiling, and data cleansing processes. It is crucial to establish metrics for privacy protection, a compliance framework, monitoring and auditing mechanisms, as well as controls. Reporting and analytics will be followed by training and awareness procedures.

A. *Data Security, and privacy in data governance*

Security concerns for cloud-system for healthcare include legal and regulatory challenges, communication interference, damages, failures, errors, malicious users, and threats relating to theft. Others include repudiation and attribution, misuse of system resources, legal and regulatory requirements, and misuse of information system resources [20], [21]. According to the European Union Agency for Cybersecurity (ENISA), in providing measures for security in the data lake project, the following measures need to be considered [22]:

- Legal and regulatory compliance with the proposed data governance framework Security and data protection requirements will be identified and this should involve the requisite stakeholders including legal, risk, compliance, and the IT department in the procurement process of the public cloud. The requirement that will be considered includes national legislation for healthcare data protection, legislation for cloud security, cyber security, and data protection. Internal requirements such as information security policies, the legal requirement for specific healthcare data, and national-specific security strategies for healthcare data would be considered. These requirements need to be reconciled with the cloud provider's security controls. Based on the data governance, data will be tagged based on sensitivity levels to ensure that the Cloud Service Provider (CSP) supplies the necessary level of control. Evidence of adherence to recognized standards of the CSP would be examined and ensure that the responsibilities between the health organizations and the CSP are well understood. Service level agreements between CSP and the health organization, on security and privacy requirements are addressed while requiring proof from CSP for ensuring compliance requirements.
- Risk assessment and data protection assessment Conduct risk assessment in accordance with the national guidelines or well-known frameworks such as the methodology provided by the European Union Agency for Cybersecurity (ENISA). Through this, cybersecurity and data protection threats and risks pertaining to cloud services would be identified and evaluated for the entire IT security risks. Data protection impact assessment will be conducted by using recognized tools such as the "ENISA tool for evaluating the risk of personal data processing".
- This initiative centres on the establishment of a comprehensive incident response plan to address security breaches and safeguard data. The primary aim is to

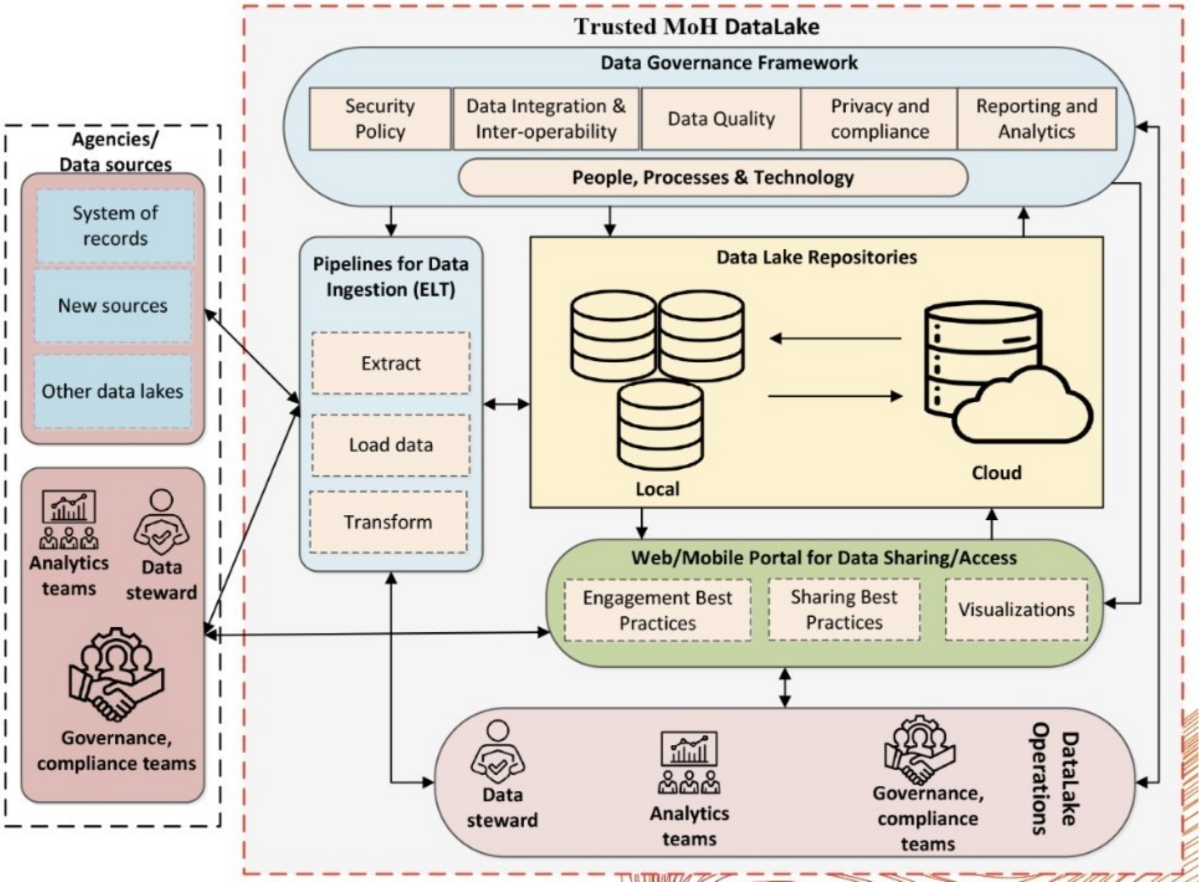

Fig. 1. Data lake framework for LMIC.

outline a robust procedure that delineates the steps to be taken in the event of a security incident occurring within the Cloud Service Provider (CSP). This endeavour ensures that the CSP adheres to established security incident handling protocols, as mandated by the Ghana Cybersecurity Act of 2020. As an integral part of this project, meticulous attention will be directed towards aligning the CSP's internal measures, operational processes, and designated roles with the stringent security requisites set forth by the Ministry of Health (MOH).

A pivotal phase of this endeavor involves the rigorous testing of the security incident processes. This testing will be conducted collaboratively with the CSP, aiming to validate the effectiveness and efficiency of the protocols designed for managing security incidents. By scrutinizing how a security incident is handled, this collaborative assessment ensures that the procedures are finely tuned and capable of swift, adept response.

An essential aspect embedded within this framework is the formulation of a Service Level Agreement (SLA). This agreement serves as a contractual pact between stakeholders and will encompass various performance indicators. These indicators include quantifying the cloud service's availability and capacity, outlining the promptness of response and reaction times from the CSP's organizational structure, stipulating the protocol for notifying users of planned maintenance or downtime and establishing a mechanism for reporting security incidents. This reporting can occur either by default or upon request, fostering transparency and accountability within the operational landscape.

- Business continuity and disaster recovery: Ensure that the business continuity management plan, provided by the CSP, is effective and it is by the national security strategy, and security policy of the health organization and in line with best practice. For example, planned downtime must be notified
- Termination and secured data deletion: The health organization asset stored on the CSP's premises should be removed, and returned in accordance with the termination of the contractual agreement or if the data retention period is met.
- Auditing logging and monitoring: Define requirements for logging and ensure the CSP abides by the SLA terms, regarding logging. Auditing the cloud should be done at agreed intervals.
- Implement vulnerability and patch management: Assess to ensure the CSP's vulnerability and patch management

practice meets the state of the arts, cybersecurity act, and well-known standards such as the OWASPs.
- Manage assets and classify information: This project will ensure that information is classified to meet data protection and security needs. Assets such as data, cloud services, and related changes will be accounted for. Conduct security testing in collaboration with the CSP.
- Data encryption for data at rest and in transit: Ensure data is encrypted during the whole data life cycle (creation, storing, using, sharing, archiving, deleting, incoming, and outgoing using state-of-the-art methods eg Advanced Encryption Standards (AES), while ensuring the security of encryption keys are protected.
- Authentication and access control: Create access control policies and specify security requirements for user access to data, application interfaces, systems, and network or network components for both private and public cloud services. Access to the cloud services will be secured by strong authentication controls such as multi-factor authentication and other mechanisms based on security.
- Information security awareness, education, and training: Cybersecurity awareness around diverse aspects, such as social engineering attacks and good cyber hygiene would be provided to healthcare professionals to help alleviate common human errors.
- Client and endpoint protection: Identify all endpoint devices such as laptops, mobile devices, and medical devices, connecting to the cloud service and define a security baseline for hardening the endpoints based on the security policies and use tools for facilitating endpoint security offered by the Cloud service provider.
- Network and Database Security: Security controls and measures will be implemented based on assessed risks to Intrusion Protection System, anti-DDoS solutions, WAF, CASB, ATP, and Threat intelligence. Traffic between untrusted and trusted connections of network environments and virtual instances will be monitored and restricted. Based on the identified security risks, various security components of the database such as TLS/SSL encryption for client and inter-node communication, client authentication and authorization will be configured.
- Review isolation between tenants: Review the CSP applies proper segmentation for data, applications (physical and virtual), infrastructure, and network between different tenants to restrict one tenant's access to another tenant's resources.
- Physical and environmental security: Physical security controls to protect data centres and prevent unauthorized physical access would be assessed and implemented. Instances include physical authentication mechanisms or electronic monitoring and alarm systems at private cloud centres.

The inadequate attention given to security and privacy considerations within the digital health strategies of African nations underscores the imperative of this framework. Its purpose is to establish a bedrock for advocating the integration of robust security and privacy protocols during the establishment of health data lakes in Low- and Middle-Income Countries (LMICs).

B. Tools for Security, Privacy, and Data Governance

In pursuit of meticulous metadata management and governance, Apache Atlas emerges as a pivotal selection. This tool offers an encompassing framework designed to oversee metadata pertaining to data residing within Hadoop ecosystems, in conjunction with Apache Spark. The tool boasts an array of functionalities including data lineage tracing, data classification, and data discovery [23]. With metadata management, Apache Atlas propounds a centralized platform that empowers organizations to distinctly define and manage metadata entities—ranging from data sources and databases to tables and columns.

Shifting focus to data lineage tracking, Apache Atlas grants organizations the ability to meticulously trace the journey of data, unearthing its trajectory from its inception to its current state. This traverses numerous data processing stages and data sources [24]. The utility lies in deciphering data origins, comprehending the spectrum of transformations it undergoes, and preserving data provenance.

In terms of data classification and labeling, Apache Atlas furnishes the means to categorize and label data based on variables like sensitivity, confidentiality, or adherence to compliance requisites [25], [26]. This classification enables the imposition of data access policies, ensuring that sensitive information remains shielded from unauthorized access. A vital facet of Apache Atlas is its collaborative and governance features. These capabilities permit multiple users to collaboratively delineate and supervise metadata entities and data lineage. The resultant effect is a synergy that bolsters metadata uniformity and quality across the organization.

Concurrently, Apache NiFi augments this spectrum with an array of security features spanning encryption and authentication, fostering the safeguarding of data privacy and security as it flows through the system. Furthermore, Apache NiFi excels in its support for an extensive array of data formats, encompassing JSON, XML, CSV, Avro, and Parquet. This inherent versatility facilitates seamless interaction with data sourced from diverse origins.

C. Data lake framework development and implementation on the aspect of cost and efficiency

We proposed Apache Spark in this data lake construction. It is open source and provides fast and distributed data processing capabilities and can handle large volumes of data [19], [27]. Spark is designed for in-memory data processing, which allows it to process data much faster than traditional disk-based data processing systems. The number of containers required for Apache Spark depends on the size of the cluster. A typical Spark cluster requires a minimum of one master node and several worker nodes. For instance, if the data lake is started with 10TB, we can estimate the number of worker nodes to be around 10 to 20, depending on the complexity of the processing tasks. So, the total number of Spark containers required could be around 11 to 21. Also, Apache Atlas is proposed because it is open-source software, which means that it is freely available and can be customized to meet specific metadata management and governance requirements.

Other benefits include:

- Scalability: Spark stands out for its exceptional scalability, adept at seamlessly managing extensive data volumes even when faced with rapid data streams. Its capacity to distribute tasks across multiple nodes and clusters empowers it to efficiently tackle the complexities of substantial big data workloads.
- Flexibility: The versatility of Spark is evident in its compatibility with an extensive array of data sources and formats. This adaptability positions Spark as a dynamic data processing solution capable of catering to diverse use cases and data types, enhancing its value across various industries.
- Real-time Data Processing: One of Spark's standout features is its prowess in real-time data processing. This capability empowers organizations to swiftly process and analyze data as it flows in, allowing for immediate insights and responsive decision-making in dynamic environments.
- Advanced Analytics: Spark's capabilities extend beyond conventional data processing, encompassing advanced analytics functionalities like machine learning and graph processing. This comprehensive analytical toolkit equips Spark to not only process data but also uncover intricate patterns, trends, and relationships, making it an invaluable asset for deriving meaningful insights and driving data-informed strategies.

1) Data Ingestion pipelines: Apache NiFi component will be used to design the data ingestion pipelines which will connect to the various agencies' databases as shown in Figure 2. NiFi provides a drag-and-drop user-friendly interface for building data pipelines and provides support for a wide range of data sources and formats [19], [28]. Some of the benefits of Apache NiFi include:

- Scalability: Apache NiFi is designed to scale horizontally, which means it can be easily scaled up or down based on the size of the data lake and the volume of data being processed.
- Ease of use: Apache NiFi provides a user-friendly, drag-and-drop interface for designing and managing data flows, making it easy for users to build and manage data pipelines.
- Real-time data processing: Apache NiFi provides real-time data processing capabilities, allowing it to process and analyze data in real-time.

D. Data Storage

Apache Cassandra will be used as the primary data storage system as illustrated in Figure 3. Cassandra is a highly scalable, open source, and distributed NoSQL database that can handle large volumes of data with high velocity [29], [29]. It provides high availability and fault tolerance and is designed to handle structured, semi-structured, and unstructured data. Cassandra is a schema-free database, allowing for flexible data modelling and enabling the storage of a wide range of data types and structures. It provides real-time read and writes access to data, making it ideal for use cases that require real-time data processing and analysis.

E. Data Visualization and Reporting

Apache Superset will be used for data visualization and reporting. It is s a modern, open-source, business intelligence (BI) web application that allows users to easily create data visualizations, dashboards, and reports [30], [31]. It is built on top of Flask, a popular Python web framework, and leverages the power of SQL databases for data processing. It provides a web-based interface for creating charts, dashboards, and reports. Key benefits include:

- Open-source and community-driven: Apache Superset is free to use and is supported by a large community of contributors who actively develop and maintain the software. Interactive dashboards: Superset provides a variety of interactive charts and graphs that enable users to explore and analyze data in real time.
- User-friendly interface: The intuitive and user-friendly interface of Superset makes it easy for non-technical users to create custom dashboards and visualizations.
- Customizable: Superset can be customized to fit the specific needs of a particular organization or user. It provides a flexible architecture that allows users to extend the functionality of the application.

Machine Learning (ML): We will also be able to pull data from the data lake for analysis leveraging various Machine Learning tools such as Jupiter Notebook, R Studio, Google Colab etc. ML techniques will be useful to explore and enhance effective data analysis and visualization such as Transformers and Generative Adversarial Networks (GAN) for Image processing and ensemble methods such as XGBoost, Random Forest, and Bagging for classification and predictions. These are current ML techniques with high-performance accuracies in predictions.

GIS Portal: The visualization of GIS data will be facilitated through the utilization of a Feature Manipulation Engine (FME), a potent cloud-based GIS tool renowned for its capabilities in data integration, conversion, and transformation [32], [33]. This open-source software enjoys widespread adoption among GIS professionals, data analysts, and data scientists globally. Notably, FME delivers a user-friendly web-based interface enabling the effortless creation, configuration, and management of FME workspaces, virtual machines, and data connections. With its emphasis on security, FME stands as a dependable solution, fortified by a range of security features designed to safeguard data. FME Cloud guarantees secure data access through role-based access control, augmenting security through data encryption for safeguarded data transmission.

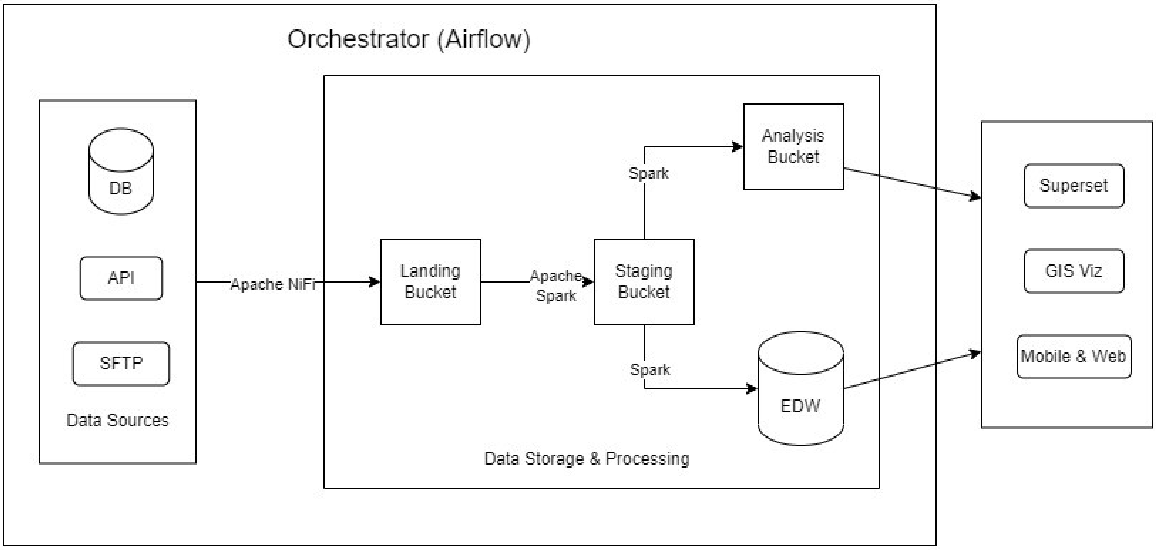

Fig. 2. Data lake injection pipelines.

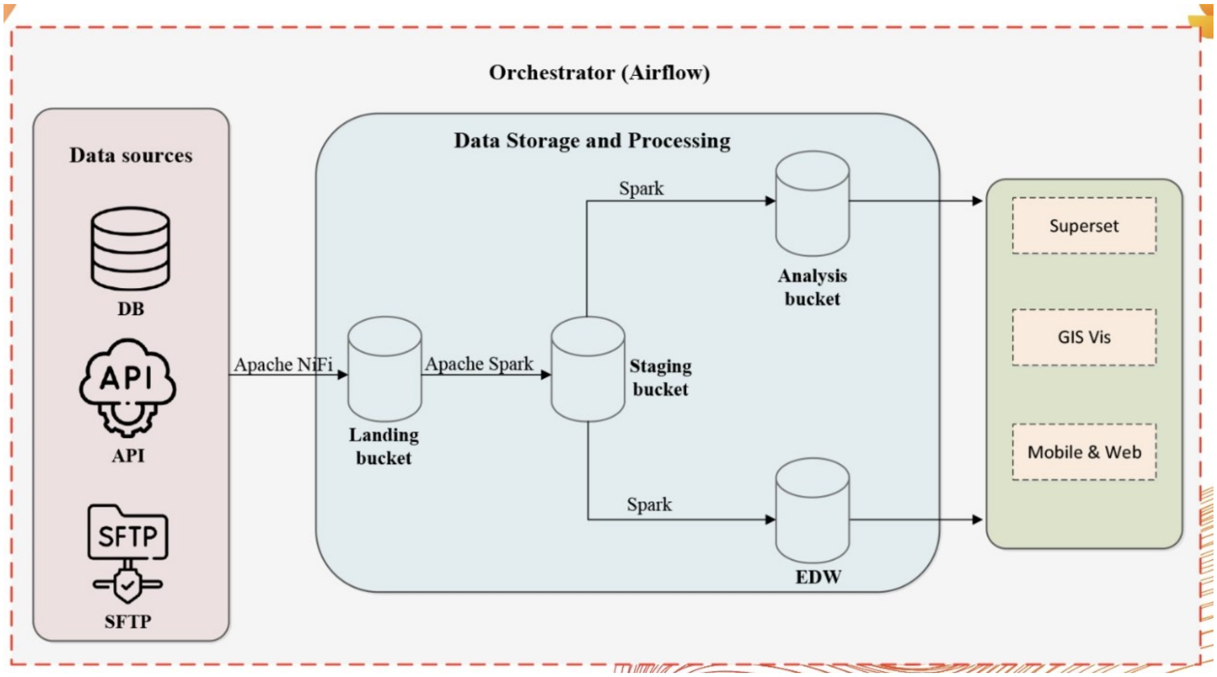

Fig. 3. Data lake storage

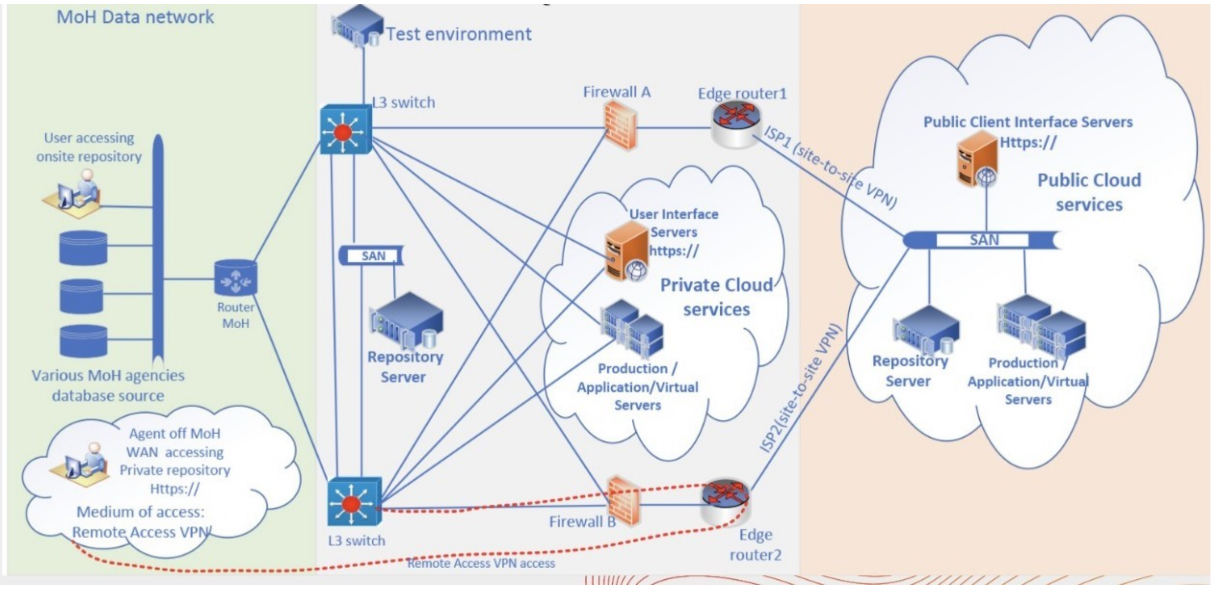
MoH Data network
Test environment
L3 switch
Firewall A
Edge router1
ISP1 (site-to-site VPN)
Public Client Interface Servers
Https://
Public Cloud services
SAN
User accessing onsite repository
User Interface Servers https://
Private Cloud services
SAN
Repository Server
Router MoH
Various MoH agencies database source
Production / Application/Virtual Servers
ISP2(site-to-site VPN)
Repository Server
Production / Application/Virtual Servers
Agent off MoH WAN accessing Private repository Https://
Medium of access: Remote Access VPN
L3 switch
Firewall B
Edge router2
Remote Access VPN access

Fig. 4. Network architecture.

The tool also accommodates various authentication protocols, including OAuth and LDAP, assuring secure data access.

FME's versatility extends to seamless data connections and transformations across divergent file formats, databases, and web services. With support for over 400 data formats encompassing structured, semi-structured, and unstructured data, FME empowers comprehensive data manipulation and transformation functions, including data cleansing, validation, amalgamation, and filtering.

As illustrated in Figure 4, the implementation will orchestrate the integration of data sources hailing from the ministry of health agencies' network. This journey encompasses data loading and transformation phases, executed under the purview of established governance's security and privacy controls. Subsequently, data within the private cloud enclave will harmonize with its public cloud counterpart, all the while adhering to prevailing national regulations governing healthcare data storage in cloud environments.

F. Limitations of Open Source Tools

Opting for open-source tools to construct data lakes presents a more economical alternative compared to the financial burden associated with proprietary tools entailing high licensing fees [34]. The cost-effectiveness of open-source tools renders them a natural preference for cultivating sustainable data lakes, especially for economies with modest resources. However, prudence is advised as concerns arise over the security and privacy of healthcare data. Notably, the potential vulnerabilities stemming from deprecated or unsupported open-source tools could undermine the integrity of the data lake ecosystem. Vigilant management is therefore imperative, necessitating a proactive approach to rectify such susceptibilities and enhance system security.

VI. Conclusion

Numerous nations are collectively striving to achieve the United Nations' Sustainable Development Goal 3, which focuses on ensuring healthy lives and promoting well-being for all, alongside the objective of universal health coverage by the year 2030. To realize these ambitions, countries have been strongly encouraged to embrace digital transformation within their healthcare systems. This strategic approach is regarded as a remedy for various challenges. However, the process of digital transformation in healthcare generates a significant and diverse array of data. This data encompasses complex structures, semi-structured formats, and unstructured information.

Consequently, concerns have emerged regarding the sustainable and secure storage of this data, especially tailored to the needs of low and middle-income countries. Addressing these concerns requires a comprehensive evaluation of existing digital health strategies in regions such as Africa. This evaluation primarily focuses on privacy and security considerations. Additionally, an assessment of data lakes, both within the healthcare sector and beyond, has been conducted.

This analysis has pinpointed certain deficiencies, particularly in the realms of security and the cost associated with implementation and maintenance. As a response to these shortcomings, a novel framework named HEALER2 has been proposed. HEALER2 stands out as a solution that effectively balances security, privacy, and cost benefits. However, it's worth noting that while HEALER2 enhances security and cost-efficiency, vigilance is imperative in adopting open-source systems. Such systems can potentially introduce vulnerabilities due to factors like deprecation or inadequate maintenance.

In consideration of the aforementioned aspects, future endeavours will involve the practical implementation and thorough evaluation of the HEALER2 framework. This evaluation will gauge its efficacy across dimensions such as security, cost-effectiveness, and overall performance.

References

[1] R. Gabryelczyk, "Has covid-19 accelerated digital transformation? initial lessons learned for public administrations," *Information Systems Management*, vol. 37, no. 4, pp. 303–309, 2020.

[2] G. E. Iyawa, S. Hamunyela, A. Peters, S. Akinsola, I. Shaanika, B. Akinmoyeje, and S. Mutelo, "Digital transformation and global health in africa," *Handbook of Global Health*, pp. 1–32, 2020.

[3] C. Cappiello, M. Gribaudo, P. Plebani, M. Salnitri, and L. Tanca, "Enabling real-world medicine with data lake federation: A research perspective," in *Heterogeneous Data Management, Polystores, and Analytics for Healthcare: VLDB Workshops, Poly 2022 and DMAH 2022, Virtual Event, September 9, 2022, Revised Selected Papers*. Springer, 2023, pp. 39–56.

[4] H. Y. Youssef, M. Ashfaque, and J. V. Karunamurthy, "Dewa r&d data lake: Big data platform for advanced energy data analytics," in *2023 International Conference on IT Innovation and Knowledge Discovery (ITIKD)*. IEEE, 2023, pp. 1–6.

[5] M. R. Llave, "Data lakes in business intelligence: reporting from the trenches," *Procedia computer science*, vol. 138, pp. 516–524, 2018.

[6] P. K. Yeng, B. Yang, and M. S. Pedersen, "Assessing cyber-security compliance level in paperless hospitals: An ethnographic approach," in *2022 9th International Conference on Internet of Things: Systems, Management and Security (IOTSMS)*. IEEE, 2022, pp. 1–8.

[7] P. K. Yeng, M. A. Fauzi, B. Yang, and S. Y. Yayilgan, "Analysing digital evidence towards enhancing healthcare security practice: The kid model," in *2022 1st International Conference on AI in Cybersecurity (ICAIC)*. IEEE, 2022, pp. 1–9.

[8] P. K. Yeng, M. A. Fauzi, L. Sun, and B. Yang, "Assessing the legal aspects of information security requirements for health care in 3 countries: Scoping review and framework development," *JMIR Human Factors*, vol. 9, no. 2, p. e30050, 2022.

[9] P. K. Yeng, M. A. Fauzi, B. Yang, and P. Nimbe, "Investigation into phishing risk behaviour among healthcare staff," *Information*, vol. 13, no. 8, p. 392, 2022.

[10] P. Yeng, B. Yang, and E. Snekkenes, "Observational measures for effective profiling of healthcare staffs' security practices," in *2019 IEEE 43rd Annual Computer Software and Applications Conference (COMPSAC)*, vol. 2, 2019, pp. 397–404.

[11] H. Moghaddasi and M. M. Ghaemi, "A comparative study of three standards of data security in health systems," *Journal of Health and Biomedical Informatics*, vol. 2, no. 3, pp. 184–194, 2015.

[12] P. Yeng, S. D. Wolthusen, and B. Yang, "Comparative analysis of threat modeling methods for cloud computing towards healthcare security practice," 2020.

[13] WHO, "Framework for implementing the global strategy on digital health in the who african region," pp. 1–10, 2021. [Online]. Available: https://www.afro.who.int/sites/default/files/2021-07/AFR-RC71-10\%20Framework\%20for\%20implementing\%20the\%20Global\%20strategy\%20on\%20digital\%20health\%20in\%20the\%20WHO\%20African\%20Region.pdf

[14] M. of Health, "National health ict strategic framework," pp. 1–63, 2020. [Online]. Available: https://www.health.gov.ng/doc/HealthICTStrategicFramework.pdf

[15] MOH, "National digital health strategy for south africa 2019-2024," pp. 1–36, 2020. [Online]. Available: https://drive.google.com/file/d/1BZd1srARhDyaeajzpVThlF2bNE1J6f5J/view

[16] M. of Helth, "The national digital health strategy 2019 – 2024," pp. 1–47, 2020. [Online]. Available: https://drive.google.com/file/d/1UvC3Ybcbnl3qrg-B1YKBJLQ8gPVXpkBr/view

[17] M. Pitchford, "The 'shift left'principle," 2021.

[18] G. Stoneburner, C. Hayden, and A. Feringa, "Engineering principles for information technology security (a baseline for achieving security)," Booz-Allen and Hamilton Inc Mclean VA, Tech. Rep., 2001.

[19] C. Manco, T. Dolci, F. Azzalini, E. Barbierato, M. Gribaudo, and L. Tanca, "Healer: A data lake architecture for healthcare," 2023.

[20] R. Aiswarya, R. Divya, D. Sangeetha, and V. Vaidehi, "Harnessing healthcare data security in cloud," in *2013 International Conference on Recent Trends in Information Technology (ICRTIT)*, 2013, pp. 482–488.

[21] M.-H. Kuo *et al.*, "Opportunities and challenges of cloud computing to improve health care services," *Journal of medical Internet research*, vol. 13, no. 3, p. e1867, 2011.

[22] ENISA, "Cloud security for healthcare services," 2023. [Online]. Available: https://www.enisa.europa.eu/publications/cloud-security-for-healthcare-services

[23] B. HU, W. WANG, and H. L. Chi, "Open source initiatives for big data governance and security: A survey," *ZTE Communications*, vol. 16, no. 2, pp. 55–66, 2019.

[24] J. Freche, M. d. Heijer, and B. Wormuth, "Data lineage," *The Digital Journey of Banking and Insurance, Volume III: Data Storage, Data Processing and Data Analysis*, pp. 5–19, 2021.

[25] N. MIGOTTO, "A metadata model for healthcare: the health big data case study," 2022.

[26] A. Agrahari and D. Rao, "A review paper on big data: technologies, tools and trends," *Int Res J Eng Technol*, vol. 4, no. 10, p. 10, 2017.

[27] A. Antolínez García, "Introduction to apache spark for large-scale data analytics," in *Hands-on Guide to Apache Spark 3: Build Scalable Computing Engines for Batch and Stream Data Processing*. Springer, 2023, pp. 3–21.

[28] H. Silva and H. Thathsarani, "Live data ingestion & attacks detection analysis system," 2023.

[29] A. Rafay, "Performance analysis/measurements with cassandra and hbase," 2023.

[30] S. Greca, I. Shehi, and J. Nuhi, "Analyzing climate changes impacts using big data hadoop," *Proceedings of RTA-CSIT*, vol. 2023, 2023.

[31] G. Galliano, "The importance of data visualization tools in modern enterprises. cost-effective solutions and empowering of an open source project." Ph.D. dissertation, Politecnico di Torino, 2023.

[32] B. Williamson, "Governing through infrastructural control: Artificial intelligence and cloud computing in the data-intensive state," *The SAGE Handbook of Digital Society. Thousand Oaks, CA: SAGE*, 2023.

[33] M. Uggla, P. Olsson, B. Abdi, B. Axelsson, M. Calvert, U. Christensen, D. Gardevärn, G. Hirsch, E. Jeansson, Z. Kadric *et al.*, "Future swedish 3d city models—specifications, test data, and evaluation," *ISPRS International Journal of Geo-Information*, vol. 12, no. 2, p. 47, 2023.

[34] P. P. Khine and Z. S. Wang, "Data lake: a new ideology in big data era," in *ITM web of conferences*, vol. 17. EDP Sciences, 2018, p. 03025.