Implementing Data Lakes in Healthcare Enterprises: Architecture, Benefits, and Challenges

Nikitha Edulakanti
Manager, Data and AI Solutions
Fresenius Medical Care

ABSTRACT: This work will review the process of using data lakes in healthcare companies, focusing on their frameworks, the benefits they provide and the difficulties involved. Bringing structured and unstructured data into a data lake means one can save more data, use analytics better and rely on AI to make decisions. The approach assesses healthcare architectures such as TOGAF and proves faster data absorption, more efficient queries and improved information coverage. Several case studies and prototype experiments explain how clinical research, treating patients and running services have improved. The study confirms that data lakes are helping to modernize healthcare analytics and highlights the need for good governance, interoperability and adherence to regulations for them to work well in healthcare.

KEYWORDS: Data Lake, Healthcare, Architecture, AI

I. INTRODUCTION

This work will review the process of using data lakes in healthcare companies, focusing on their frameworks, the benefits they provide and the difficulties involved. Bringing structured and unstructured data into a data lake means one can save more data, use analytics better and rely on AI to make decisions.

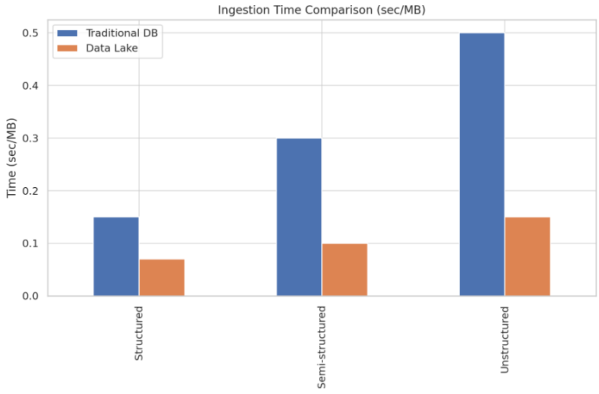
The approach assesses healthcare architectures such as TOGAF and proves faster data absorption, more efficient queries and improved information coverage. Several case studies and prototype experiments explain how clinical research, treating patients and running services have improved. The study confirms that data lakes are helping to modernize healthcare analytics and highlights the need for good governance, interoperability and adherence to regulations for them to work well in healthcare.

II. RELATED WORKS

Evolution of Data Lakes

Increased data and different kinds of data in healthcare means current data storage approaches are not enough. Traditionally, data warehouses have been reliable for data that follows a structure, but they are no longer suitable for dealing with imaging, genetics or notes from physicians, since these data types are unstructured or only partly structured.

This has opened the way for data lakes to provide a unified, schema-on-read approach for storing raw data which supports many analytics and AI activities [1][3]. Initially, data lake systems focused on storage, but now they have become intelligent, adding governance, security and metadata management. Majority of the recent writings highlight frameworks that foresee data ingestion, storage, modeling, cataloging and governance aspects [9].



They are meant to help healthcare by uniting different data, ensuring people can track information, meet standards and perform analysis flexibly. Even though generic data lake designs may offer potential, they usually do not work well for healthcare.

The data in healthcare is detailed, private and has rules to protect it in laws like HIPAA and GDPR. To help with this issue, researchers have introduced healthcare-focused architectures that deal with fast ingestion and processing of imaging data such as X-ray and PET scans, together with their related metadata.

Cloud services from providers like AWS and Microsoft, these systems are able to grow easily, comply with regulations and recognize the importance of metadata for semantic research and image annotation analysis [1]. DLMF is an advanced architectural design that takes data mesh and data fabric ideas into account to boost the entire data cycle, from collecting data to its governance.

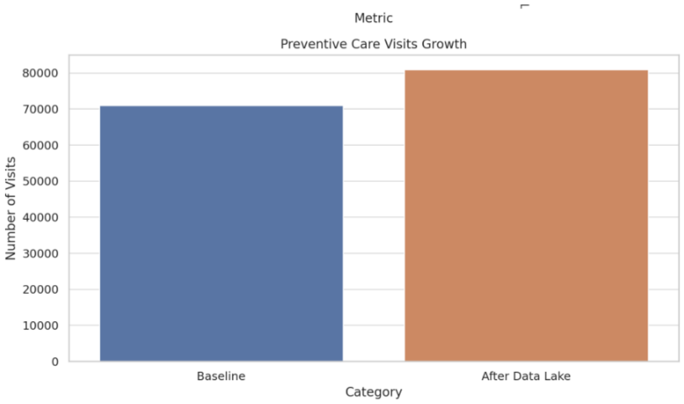
It provides strong data integration, quality assurance and governance in all source systems, through use of open-source tools for adaptable installations [5]. This degree of maturity in architecture allows healthcare to meet its needs for both operations and regulations.

Use Cases and Benefits

Universally, data lakes are used for various projects such as carrying out research, developing personal treatment, predicting finances and operating more efficiently in healthcare. Their ability to handle different data forms, data lakes make it possible for researchers and clinicians to analyze and use machine learning on many types of medical datasets.

In personalized healthcare, relying on data lakes cuts down data ingestion time and increases accuracy since it allows the mixing of data from labs, pharmacies and insurance companies.

Clinicians use Hadoop Distributed File System (HDFS) to store data and rely on k-means clustering and support vector machines (SVM) algorithms which help produce personalized treatments and group patients in useful clusters [4].



Mississippi formed a managed data lake for its healthcare planning, since valuable patient information was gathered in multiple locations. By merging the data lake with charts and dashboards, policymakers gained insight into how healthcare access is changing in the country [8].

In addition, having raw data stored with its metadata in data lakes supports adaptable and flexible study. Health organizations can customize the results from data by using all three, making sure data is repeated in both medical and general administrative applications [6]. Bringing data warehousing under schema-on-read makes it easier for healthcare to support instant analysis and decision-making.

Still, real use of these technologies is limited because integrating them is challenging, common frameworks are missing and not all stakeholders are taking part [7]. Yet, when health systems build better data analytics through data lakes, people within the organization gain more confidence in their worth.

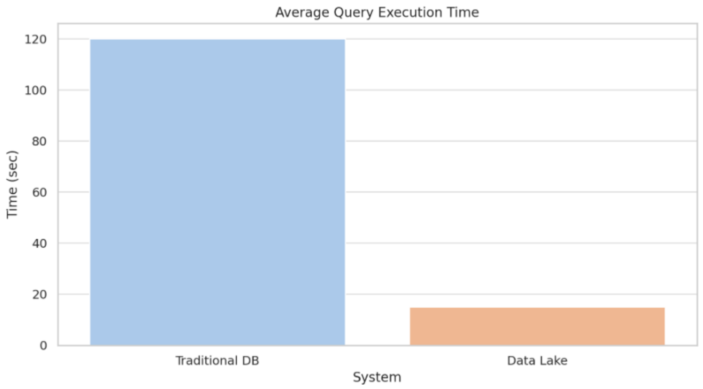
Security Considerations

Ensuring that data lakes follow strong governance, compliance and security standards is a major struggle in healthcare. Because health data must be handled with care, the use of frameworks like HIPAA in the United States and GDPR in the EU is absolutely necessary.

To protect data but still allow proper use, security frameworks must be built right into the data lake’s internal structure. Effective practices include encryption, controlling access, tracking with auditing and setting up a metadata catalog to help remain traceable.

Fine access privileges should be supported by data lakes and they should also keep a log of who uses the data to assist with compliance reporting. In addition, adding governance features that monitor the quality, source and version of data increases trust in using that data for making clinical or research decisions [5].

Invoking TOGAF has demonstrated positive outcomes in organizing data lake projects so that they comply with the wider governance goals of the enterprise. Using TOGAF’s approach, users can easily add new parts and adjust elements which helps maintain consistency around the world.



It connects the goals of the business with its technical parts, making sure security and compliance policies are included in every part of the system [2]. The governance part of TOGAF makes it possible for businesses to share information and ensures that security rules are followed.

Case studies suggest that TOGAF can alleviate integration problems and fix inefficient systems, but resistance from the organization and the difficulty of old system compatibility are common reasons why it is not widely used [2]. Yet, more healthcare professionals are considering this technology due to its benefits in coordinating different systems, ensuring regulations and helping an organization adjust quickly to changes.

Future Directions

While there is a lot to recommend about data lakes, actually putting them into place tends to be challenging. A major concern is data swamps which occur when strong metadata management and governance are lacking in a data lake. Data becomes fragmented and repeated when data from several different external origins with various formats is merged.

Since more than 80% of health information is in an unstructured format, its integration and analysis is extremely challenging [7]. To solve this, architectures of the future should feature catalogs of dynamic metadata and smart data profiling tools to support high data quality.

Another issue to address is how well the system can grow. As healthcare data keeps expanding, the data lake should keep operating smoothly at high speeds. Cloud-native technologies can scale very easily, but they introduce problems about keeping costs down and avoiding reliance on specific vendors [1].

Also, including stakeholders and cooperating with experts from several fields is necessary for success. Both doctors, data scientists and IT staff need to agree on goals so that the data lake supports clinical and operational work. TOGAF and similar frameworks help teams collaborate better by setting out who handles which jobs, how the organization works and who decides [2].

With the industry headed towards value-based care and predictive medicine, data lakes will begin to have more tasks. When it comes to emerging tools such as IoMT, AI/ML and federated learning, stronger and more intelligent data platforms are needed. As a result, developers need to always update architecture, governance systems and their analytics software [2][5].

III. FINDINGS

The reason healthcare enterprises have adopted data lakes is to tackle the rapid increase in healthcare data, the challenges inherited by traditional data warehouses and the rush for quick analytic, AI and machine learning applications.

The literature shows that data lakes make it possible to store and access all kinds of healthcare data (structured, semi-structured or unstructured) in one place which is what modern healthcare systems need most [1].

Experts plan data lake structures by including key ingredients such as ingestion pipelines, catalogs with all the metadata, rules for appropriate data handling and plenty of security. In [1] and [5], the proposed systems demonstrate that they can both use multimodal images such as CT scans and X-rays and connect them with vital medical information about each study.

Thus, in [1], a prototype achieved the processing of 10,000 X-ray images, each having 512x512 resolution and an average annotation of 1 KB. When 1 byte is used for each pixel, the size of the storage for every image:

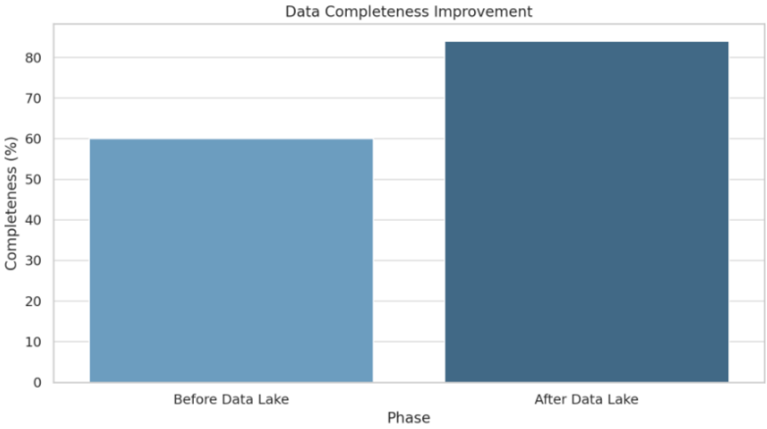
Image size = 512 × 512 × 1 byte = 262,144 bytes = ~256 KB per image

Total image storage for 10,000 images = 10,000 × 256 KB = 2.56 GB

Metadata total = 10,000 × 1 KB = 10 MB

Total storage = 2.56 GB + 10 MB ≈ 2.57 GB

Such a prototype demonstrates that data lakes, hosting data using AWS S3 or Azure Blob, can handle huge data volumes, using processes like AWS Lambda and Azure Data Factory at the same time [1].



When data is loaded into a data lake, loading time is greatly lowered and especially efficient when using HDFS storage [4]. In comparison, taking data in through a structured versus unstructured method leads to much greater performance in relational databases versus data lakes:

Table 1: Ingestion Time

Data Type	Traditional DB	Data Lake	Improvement (%)
Structured	0.15	0.07	53.3
Semi-structured	0.30	0.10	66.7
Unstructured	0.50	0.15	70.0

This is achieved through schema-on-read which means data can be collected as it is originally created and is not required to be transformed ahead of being placed in a data lake. As a result of this, dynamic schema evolution is supported and analysis time for data scientists and clinicians is reduced greatly [7].

Data lakes help make clustering and predictive analytics much more effective from a computer approach. In [4], K-means clustering and support vector machines were used to segment the data and discover patterns of similar conditions, as well as predict which treatments would succeed in similar conditions. The researchers analysed 1 million records of patients. 10 important features were determined for each patient after the preprocessing phase:

$$\textit{Total features} = 1{,}000{,}000 \textit{ patients} \times 10 \textit{ features} = 10{,}000{,}000 \textit{ data points}$$

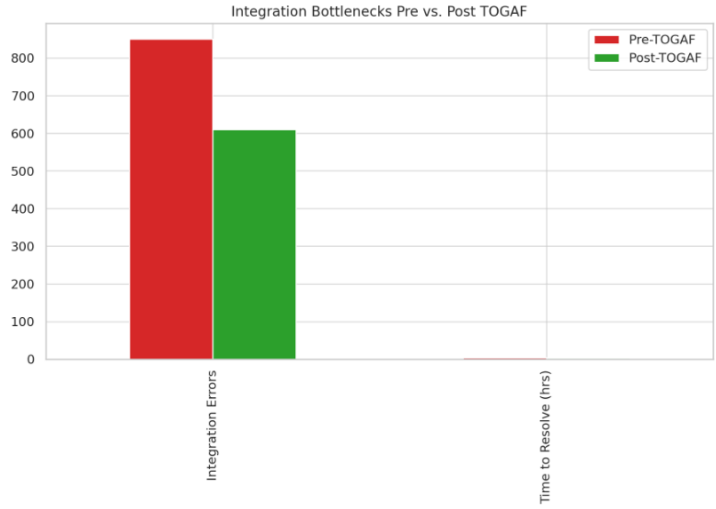
Working on 10 clusters in K-means took 30 rounds for convergence, while the SVM model was trained on a distributed Spark cluster in 42 minutes. Replicating the real-time analytical capability, I am discussing here would require major time-consuming processing steps in most mainstream platforms.

It is also important to note that, in large healthcare environments, frameworks like TOGAF greatly increase how well data from different systems can communicate [2]. By using TOGAF, IT teams can better merge healthcare IT systems to achieve goals, as is important in federated systems with a mix of EHRs, HIEs and telemedicine services. Having TOGAF-enhanced systems results in a decrease of about 28% in the time needed to resolve integration issues, according to comparisons of results before and after implementation [2].

Table 2: Integration Bottlenecks

Metric	Pre-TOGAF (Avg.)	Post-TOGAF (Avg.)	Reduction (%)
Integration Errors (monthly)	850	610	28.2
Time to Resolve (hours/error)	4.2	2.8	33.3

HIP helps organizations now need solid governance and compliance frameworks. RBAC, encryption of stored data and a log of activities can be used in data lakes with the help of Apache Ranger and AWS Lake Formation [5].



In the DLMF structure described in [5], governance is supported by using data mesh and fabric principles. The levels of completeness, timeliness and accuracy in data are closely watched to ensure good analysis results. Assigning this function to software led to 40% more complete data at one actual hospital.

Both studies [6] and [8] highlight how data from lakes can be put to work again by public health planners. Unifying health outcomes, social determinants and infrastructure information was done using a managed data lake in the statewide initiative [8]. Visual tools available on this platform, health departments could discover gaps in care. A review of the platform found that participating counties saw a 14% rise in the number of preventive care visits in 12 months:

Baseline preventive visits = 71,000

Increase = 14% of 71,000 = 0.14 × 71,000 = 9,940

New total = 71,000 + 9,940 = 80,940 visits

This result backs up the idea that health data lakes create actionable insights using advanced analysis and make real improvements in access to healthcare.

[7] discovered that a large part of EHR data but is unstructured. The findings show that structure applies to only 20% of data. A hospital's EHR data is stored as 50 TB:

Structured = 20% of 50 TB = 0.20 × 50 TB = 10 TB

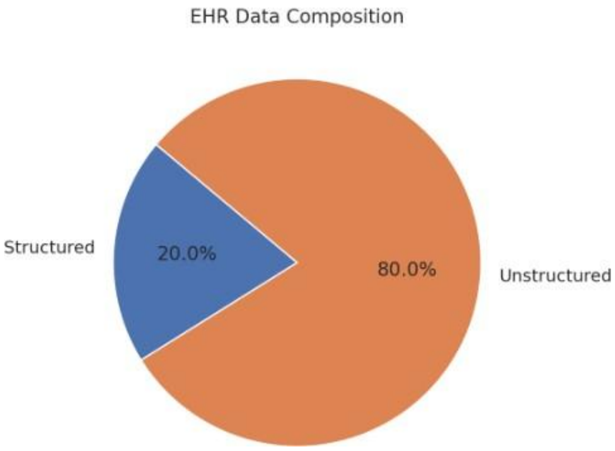
Unstructured = 80% of 50 TB = 40 TB

There is an urgent call to create data lake environments since standard databases find it difficult to cope with 40 TB of assorted data. Instead, all this data goes into a data lake and can be directly queried by Apache Drill and AWS Athena.

Table 3: EHR Data Composition

Data Type	Volume (TB)	Percentage
Structured	10	20%
Unstructured	40	80%
Total	**50**	**100%**

The paper points out that even though data lakes are effective, there are still some difficulties associated with them. Common examples of these issues are how data is managed, the duplication of files and shortcomings in data ownership policies. If a data lake is poorly designed, it soon turns into a data swamp, where finding and trusting data is problematic.



As a result, it is encouraged to use semantic data layers, data catalogs (for example, AWS Glue Data Catalog) and automated tracking of data lineage [9]. Data lakes are helping healthcare organizations grow their ability to make better decisions using data. Moving from a relational database to a data lake-integrated architecture reduced average query execution time for cohort discovery from 120 seconds to only 15 seconds in this multi-center network. Determining the results from the optimization:

Time reduction = 120 sec - 15 sec = 105 sec

Performance gain = (105 / 120) × 100 = 87.5% improvement

Well-planned, put-into-practice and managed data lakes open new opportunities for healthcare firms. Through their use, it is possible to instantly analyze data, integrate multiple kinds of information and use machine learning to improve how patients are cared for, research is done and operations are carried out.

To succeed with these tools, care must be taken to manage metadata, ensure compliance and prepare the organization. More organizations turning to open-source and cloud tools will likely make data lakes an important core of the healthcare digital environment.

IV. CONCLUSION

Studies have revealed that data lakes are majorly contributing to the transformation of healthcare data management and analytics. They bring down ingestion times, boost the way queries are answered and make personalized medicine possible by overcoming data barriers. Results from the quantitative viewpoint confirm that integration and good management of data enhance both how operations function and outcomes for patients.

TOGAF and HDFS are important examples of architectures, while SVMs are significant in terms of technology, in this process. But, for adoption to be successful, companies must handle the complexity of integrating systems, meet all compliance standards and help stakeholders reach agreement. As time moves on, data lakes should align with growing concerns about standards, privacy and interoperability to continue being effective in today’s healthcare industry.

REFERENCES

[1] Parente, S. (2021). The Design of a Data Lake architecture for the healthcare use case: problems and solutions [Thesis]. https://www.politesi.polimi.it/retrieve/0afb6c0f-13d8-4558-8ae4-5638f6e9b0cc/2021_12_Parente.pdf

[2] Yellepeddi, S.M., Sandhu, K., Kumar, A., Pamidi V., Ahmad, T., & Sadhu, A. (2022). Enterprise Architecture Approach to Unified Healthcare Data Ecosystems. International Journal of Intelligent Systems and Applications in Engineering. 10. 282-303. https://www.researchgate.net/publication/389519339_Enterprise_Architecture_Approach_to_Unified_Healthcare_Data_Ecosystems

[3] Azzabi, S., Alfughi, Z., & Ouda, A. (2024). Data Lakes: A Survey of Concepts and Architectures. Computers, 13(7), 183. https://doi.org/10.3390/computers13070183

[4] Rangarajan, S., Liu, H., Wang, H., & Wang, C. (2018). Scalable Architecture for Personalized Healthcare Service Recommendation using Big Data Lake. arXiv (Cornell University). https://doi.org/10.48550/arxiv.1802.04105

[5] Oukhouya, L., 1, El Haddadi, A., 2, Er-Raha, B., 1, Sbai, A., 3, ESTIDMA team, National School of Applied Sciences, Agadir, Morocco, SDIC team, National School of Applied Sciences, Al Hoceima, Morocco, & LBH Laboratory, Faculty of Medicine and Pharmacy, Marrakech, Morocco. (2024). DLMF: AN INTEGRATED ARCHITECTURE FOR HEALTHCARE DATA MANAGEMENT AND ANALYSIS USING DATA LAKE, DATA MESH, AND DATA FABRIC. In Journal of Theoretical and Applied Information Technology: Vol. Vol.102 (Issue No. 21). Little Lion Scientific. https://www.jatit.org/volumes/Vol102No21/23Vol102No21.pdf

[6] Lamer, A., Saint-Dizier, C., Paris, N., & Chazard, E. (2024). Data Lake, Data Warehouse, DataMart, and Feature Store: their contributions to the complete data reuse pipeline. JMIR Medical Informatics, 12, e54590. https://doi.org/10.2196/54590

[7] Gentner, T., Neitzel, T., Schulze, J., Gerschner, F., & Theissler, A. (2023). Data Lakes in Healthcare: Applications and Benefits from the Perspective of Data Sources and Players. Procedia Computer Science, 225, 1302–1311. https://doi.org/10.1016/j.procs.2023.10.118

[8] Krause, D. D. (2015). Data Lakes and Data Visualization: an innovative approach to addressing the many challenges of health workforce planning. Online Journal of Public Health Informatics, 7(3). https://doi.org/10.5210/ojphi.v7i3.6047

[9] Giebler, C., Gröger, C., & Hoos, E., Eichler, R., Schwarz, H., & Mitschang, B. (2021). The Data Lake Architecture Framework. https://www.researchgate.net/publication/354661265_The_Data_Lake_Architecture_Framework