International Journal of Science and Research (IJSR)
ISSN: 2319-7064
SJIF (2022): 7.942

Unlocking the Potential of Clinical Data Lakes for Future Generations

Mujeebuddin Shaik[1], Uma Priya[2]

Founder and CEO, ClinoSol Research Pvt. ltd
Corresponding Author Email Address: *mujeeb[at]clinosol.com*
Phone Number: 9701158350

Director, ClinoSol Research Pvt. Ltd,
Email Address: *priya[at[clinosol.com*
Phone Number: 9059067107

Abstract: ***Data is the new currency in the life sciences and healthcare industry and the amount of healthcare data is expected to increase dramatically each day. Data consolidation is an important factor for operating systems and analytics. There are different ways to architect a data lake, and each of them has its own merit. Every organization will have a different need for the lake. Once you've established a strong data lake foundation, the next step is to better understand the roots of your data, and how it is processed within your data lake for consumption by analytics and applications. The data lake is beneficial in life science in many ways in reducing costs, lowering risk, increasing competitive advantages, managing comprehensive healthcare, processing the huge volume of data, enhancing research and development, and increasing the speed of data access and query processing. The life sciences industry witnesses a huge block of unstructured data with the data growing each year, the need for robust data management architecture is required to store, access, and run analytics. A data lake pools data from multiple sources and applies analytical models to provide a new approach to information management, reporting, and predictive analytics to help deploy evidence-based care strategies, create advanced analytic insights, and improve patient engagement outcomes, this review provides an overview of the potential of data lakes in their full capacity, it has the potential to improve performance, increase scalability, and make a life science/pharma data-driven industry.***

Keywords: Data lake, data analysis, metadata, healthcare, data warehouse.

1. Background

A huge amount of data is generated by the healthcare industry, and with a large amount of data potentially available for analysis, managing the flow of data efficiently can be a challenge. Rapidly feeding the results of data analysis to multiple devices is critical for a solution.[1,2] An enterprise technology solution in healthcare transformation must solve the four "Vs" (volume, variety, velocity, veracity) of big data: Analytics provide insight to better manage IT resources, optimize staffing schedules, and identify trends in services that can be used for marketing. However, managing big data is a wasted investment if the data cannot be converted into intelligent action A data lake offers a technology infrastructure that accumulates information generated across the health system, including data imported from outside sources and services.[1,3] The data lake may reveal actionable insights about the organization's performance indicators, using an enterprise hybrid cloud framework, a data lake is layered with information-rich data sources, analytic tools, and data science best practices that enable providers to link and correlate information in completely new ways. This paper discusses the limitations of the data warehouse approach and how the data lake concept can overcome these limitations.

What is Data lake- A Conceptual Overview

A clinical data lake is a data storage system that stores medical data in an open format. Data is stored in a lake in a raw, unstructured format that can be queried and accessed by multiple users. This type of data lake allows for easy access to large volumes of healthcare datasets, which can be used for research, analytics, and other applications. A data lake is a highly scalable, cost-effective, and secure way to store data.[2,4] A data lake is a new trend that is gaining a lot of popularity. A data lake makes use of a simple architecture to store data in its raw format. Each data entity in the lake is characterized by a unique identifier and a set of extended metadata. purpose-built schemas can be built by the consumers for query-relevant data. It leads to a smaller set of data that can be analyzed to help answer a consumer's question.[14,5] Since there is no fixed schema defined beforehand, there are questions about the possibility of data becoming incomprehensible. It could cause the lake to turn into a data swamp. So, it is essential to have a metadata repository that records high-level information about data entities (type, time, creator, etc.).[6] A data lake stores data in its actual format and places the responsibility for understanding the data elsewhere. A data lake provides an IT environment that integrates structured, semi-structured, and unstructured data from reliable external and internal sources and ultimately improves the effectiveness and quality of clinical practices.[5,6] In addition, a data lake applies advanced analytics to produce actionable insights that enable timely. interventions to prevent adverse health events and ultimately, elevate overall population wellness.[6] A data lake meets rapidly evolving business and clinical requirements by quickly and efficiently analyzing new combinations of data from multiple sources across the health system. Traditionally, healthcare providers have invested substantial time and effort to extract, transform, and load data from its original format into data warehouses purpose-built for business intelligence. A data lake strategy simplifies storage, management, and analysis of Big Data by consolidating data in real-time, near real-time, or in batch from disparate sources and across multiple protocols.[5]

Table 1: Data Lake vs Data warehouse

	Data lake	Data Warehouse
Data types	Structured, Unstructured, semistructured, and Raw.	Structured, Clean, and Processed.
Purpose	Undefined, can be used for big data analytics/ predictive analytics and so on.	Specific predefined purpose.
Data Capture	Captures all forms of data for future usage.	Captures only structured data and organizes them in schemas.
Data Source	Native raw data from any source.	Structured data (historic and relational) is typically extracted from transactional systems, operational databases, and applications.
Data Quality	Raw data that may or may not be curated for use.	Centralized curated data is ready for use in analytics.
Data retention	Retains all the data for an unlimited amount of time.	Does not retain data forever, data is purged periodically.
Data processing	Schema-on-read, raw data only transformed when put to use.	Schema-on-write structured and cleansed data.
Agility	Highly agile can be configured and reconfigured when required.	Less agile, pre-configured
Users	Data scientists, and data engineers.	Analysts and business users
Storage costs	Relatively less expensive.	Expensive and time-consuming.
Top use cases	Building data pipelines, stream processing, machine learning, and real-time analysis.	Batch processing, and reporting
Best Suited for	Exploration, discovering patterns, innovation, and flexibility.	Repeatable process, ongoing analyses, and constant operational use such as generating business reports and dashboards.

Benefits of a Clinical Data Lake

A clinical data lake provides many benefits for healthcare organizations. First, it allows for easy access to large volumes of healthcare datasets. This makes it easier for researchers and analysts to access and analyze copious amounts of data quickly and efficiently. Additionally, data lakes are highly secure, so healthcare organizations can be assured that their data is safe and secure. Finally, data lakes are designed to be highly scalable and cost-effective, so organizations can store copious amounts of data without worrying about the cost of buying and maintaining expensive hardware.[11]

Table 2: Benefits of clinical data lake

Reduction in cost	• Reduction in the resources used for gathering/ integrating the data from multiple sources.
	• Elimination of the cost of recruiting experts to manage the data warehouses.
	• Avoids the expenses related to maintaining Data Warehouses.
Lowering risk	• Eliminating manual/ spreadsheet errors in data management.
	• Automating the data access by removing manual data reconciliation.
	•
Increase competitiveness advantage	• Doing resource allocation in a smart and effective manner.
	• Fasten the decision-making process (with more precision and accuracy)
	• Doing resource allocation in a smart manner
	• Create a data pool/knowledge base for best practices
Management of the healthcare system	• Data lakes bring in a combination of technology that can be used to create an efficient health management model.
	• Allowing the providers to enhance their services, make precise decisions, fasten the healthcare processes, and create the best quality health systems.
Excess data processing	• No matter what amount of new data is being generated, it can process them in real-time with ease and accuracy.
	• The raw format data is never lost inside a data lake, so it is always used for more data mining.
Enhanced Research and development	• External and internal data in one place and easy access to all records will enhance the research and development of new drugs, healthcare processes, and equipment.
Increased Speed of Data Access and Query Processing	• Data can be accessed and queries can be performed at a lightning-fast speed.
	• Data Lakes provide better concurrency, remove redundancy, and integrate all the sources.

Clinical **Data Lake Architecture**

Data lakes can be used for a variety of use cases including research and analytics, patient care, disease management, clinical trials, population health management, and more.[3] Data lakes can also be used for predictive analytics, machine learning, and artificial intelligence. The architecture of a clinical data lake is designed to store data in a secure and efficient manner.[2,3] The healthcare industry is the biggest producer of data with nearly 9 petabytes of medical data. With the rise in electronic healthcare records, wearables and digital medical imagery are majorly contributing to this data explosion. Building machine learning models and analytic dashboards on the top of these healthcare organizations can improve the patient experience and drive better health outcomes.[3] Data is stored in a lake in a raw, unstructured format that can be queried and accessed by multiple users, and stored in a repository, which is a secure database that is used to store and manage data. The repository is divided into different zones, which control access to the data and ensure secure storage and access.[3,4] There are five main components of a clinical data lake architecture: the data repository, the data lake, the data lake zones, the data lake services, and the analytics engine. The data repository is the secure database that stores and manages the data. The data lake is the open format in which the data is stored. The data lake zones control access to the data and ensure secure storage and access. The data lake services are the tools and services that are used to process and query the data. Finally, the analytics engine is the software used to analyze the data.[3]

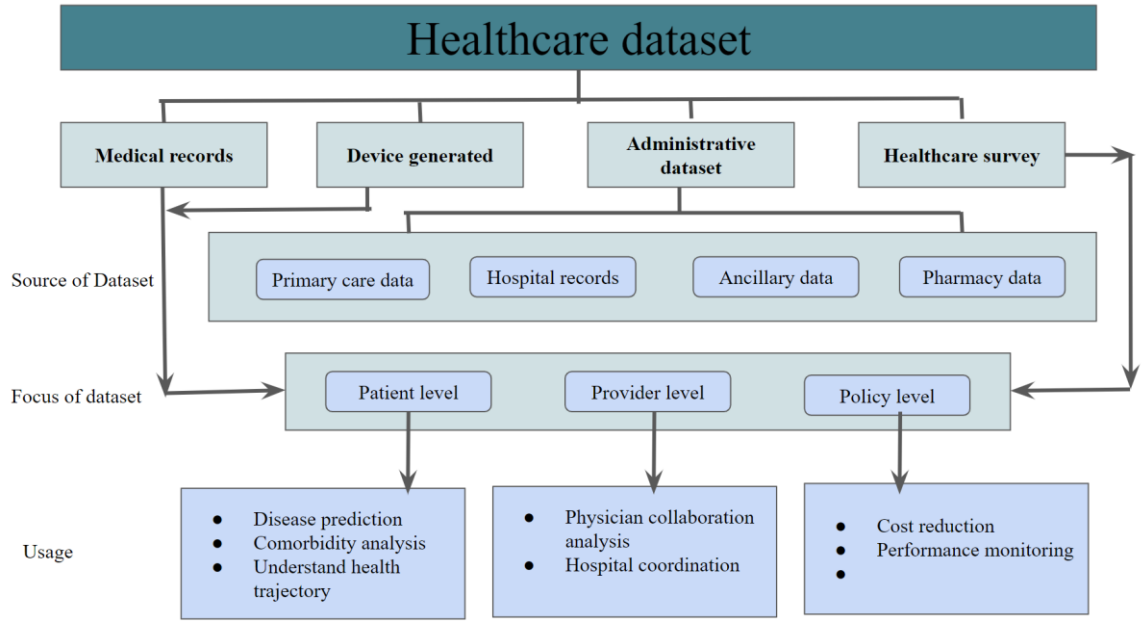

Figure 2: Types of healthcare data and their prospective use

Data lakes and evidence-based care

Evidence-based care/ practice (EBP) is the combination of patient values, clinical expertise, and the best research evidence in the decision-making process for patient care. The patients encounter their own personal preferences and unique concerns, values, and expectations.[2,6] The best research evidence is found in clinically relevant research that has been conducted using sound methodology. The data lake constitutes a cost-effective platform for EBP. Evidence-based implies that some types of evidence are sufficient to guide general analysis but fail to provide the desired level of clinical rigor which can be required for a broader approach to multivariate analysis, capturing the data from several areas outside of the EHR.[6,7] Resources of evidence include electronic health records/ electronic medical record systems (reflecting internal-based records), clinical or trial-published research (datasets, publications/ programs), government assets (publications, surveys, collective libraries), genomic research, and of course personal health records (patient-reported data, family medical history, exercise or diet regime, data from smart devices).[3] The body of evidence is broad, deep, and expansive. Therefore, it would be a struggle for any enterprise data warehouse to incorporate such vast amounts of information with tremendously variable content, so the data lake provides an optimal platform. Professor Dhavendra Kumar, a Consultant in Clinical Genetics at the University Hospital of Wales, Cardiff University, United Kingdom has proposed that the practice of evidence-based medicine (EBP) would include analysis of genomic or genetic profiles and its success would depend upon the strength of translational research.[8] The most successful application of this kind of data is- it has been in the characterization of human cancers, which includes the ability to predict clinical outcomes, while utilizing the data lake approach, we can clearly see how it would support the multivariate analysis of combining genomic research data sets with personal health records (family history) and electronic health records (patient treatment) to develop evidence-based care models. Modeling tools (SPSS, SAS, R, Python, or Scala) then analyze statistical relevance between the genome markers, patient information, treatment, and its outcomes.[9]

Planning a data structure is an important factor before you land it into a data lake. Having a plan helps you use security, partitioning, and processing in the most effective manner. Clinical data lake zones are used to control access to the data and ensure secure storage and access.[10] The zones are organized into four levels: public, private, secure, and restricted. The public zone is open to anyone and stores publicly available data. The private zone is used to store data that is for internal use only and requires authentication to access. The secure zone is used to store sensitive data and requires additional authentication to access. Finally, the restricted zone is used to store extremely sensitive data and requires special permission to access.Clinical data lakes can store a variety of data types. Examples of clinical data include patient records, clinical trials, medical images, lab results, genomics data, and more. This data can be used for a variety of applications, such as research, analytics, and decision-making.[9]A data warehouse for healthcare is a data storage system that stores healthcare data in an organized format. Data warehouses are used to store substantial amounts of healthcare data and can be used for research, analytics, and decision-making. Data warehouses are highly secure, so healthcare organizations can be assured that their data is safe and secure.[11] Healthcare analytics is the process of analyzing healthcare data to gain insights and make decisions. Analytics can be used to drive patient care, improve operational efficiency, and identify trends in healthcare data. Healthcare analytics can also be used to improve patient outcomes and reduce costs.

Strategies for Implementing a Clinical Data Lake

Implementing a clinical data lake can be a complex process. Organizations should start by assessing their data needs and identifying the data sources they need to store.[10] They should also develop a data governance strategy to ensure

data security and compliance. Additionally, organizations should consider the cost of implementing a data lake and the resources they need to maintain it. Finally, organizations should develop an analytics strategy to ensure they are getting the most out of their data.[10] A successful data governance program applies standards and processes to create high-quality data and ensure it is appropriately used across the organization. Data governance has originally focused on structured data in traditional data warehouses and relational databases, but things have changed. If your organization consists of a data lake environment and wants to get accurate analytics results from it, you also should engage in proper data lake governance as part of the overall governance initiative. But data lakes consist of various challenges across all the disciplines of enterprise data management including data governance.[9,10]

Benefits of data lake governance

Effective data governance enables organizations to improve data quality and maximize the use of data for business decision-making, which can lead to operational improvements, stronger business strategies, and better financial performance that applies to governing data lakes as it does with other types of systems.[11] Some of the specific benefits that data lake governance provides include the following: *Increased access to relevant data for advanced analytics*- in a well-governed data lake, it can be easier for the analytics team, data scientists, and other members to find the data they need for predictive analysis, machine learning, and other data science application, *Less time is spent preparing data for analytics uses*- in a data lake the data is commonly left in its raw form until it's needed for specific applications, the data preparation process can be shortened in a governed environment (upfront data cleansing reduces the need to fix data errors and other issues later on), *Lower IT and data management costs*- by preventing a data lake from sprawling out of control, the data processing or storage resources it requires may reduce. Overall data management needs can also be decreased by improving data quality, cleanliness, and consistency, *Improved security and regulatory compliance on sensitive data*-customer analytics to aid in sales and marketing is a common use case for data lakes. As a result, they usually include sensitive information about customers, and strong governance of the data lake helps ensure that such data is properly secured and doesn't get misused.[11]

Healthcare use cases for a data lake

Traditional enterprise data warehouse approaches were built with the intention to attract a large number of self-service business users with an interest in day to day clinical, operational and financial reporting that draws from structured and pre processed data. Data lakes are designed as highly agile, configurable alternatives for answering complex questions and leveraging all available data sources. The two basic use cases for a data lake in healthcare are- predicting healthcare costs and evidence-based practice.

Analysis of healthcare data comes from reports developed against the enterprise data warehouse, which had extracted the data from electronic health records that are commonly found in electronic medical records systems. The records are cast against the dimensions such as place, time, diagnosis, or treatment category where observations are the measured performances that have a rate of change based on past measurements. With the advent of data lake strategy, providers and healthcare plans are attempting to enrich their data and predict the possible patterns of risk or greater cost using an expanded data set. These new predictive measures can show a high degree of relevance with added values. Advanced users include the new models into actionable change with factors which may not always be found in an electronic health record, for example personal health records (health information record related to the care of a patient that is maintained by the patient). The purpose of adding a patient health record to analysis is to provide a complete and accurate summary of a person's medical history.

Evidence-based practice is the integration of patient values, clinical expertise, and the best research evidence into the decision-making process for patient care. Clinical expertise refers to a clinician's education, experience and clinical skills. The patients experience their own personal preferences and unique concerns, values and expectations. The best research evidence can be found in clinically relevant research that has been conducted using sound methodology.[11] A data lake represents a cost-effective platform for evidence based practice. Evidence- based implies that some types of evidence are sufficient to guide general analysis but fail to offer the desired level of clinical outcomes required for a broader approach to multivariate analysis, drawing on data from several areas outside of the electronic health records. Resources of evidence include electronic health record/ electronic medical record systems (reflecting internal-based records), clinical published research (publication, datasets, and sometimes programs), government assets (such as, surveys, publications, collective libraries), genomic research, and of course personal health records including- patient reported data, family medical history, exercise or diet regimen, data from smart devices. The body of evidence is deep, broad, and expansive. Therefore, it would be a struggle for any enterprise data warehouse to incorporate vast amounts of information with such tremendously variable content, so big data and more specifically, the data lake provide an optimal platform.[10, 11]

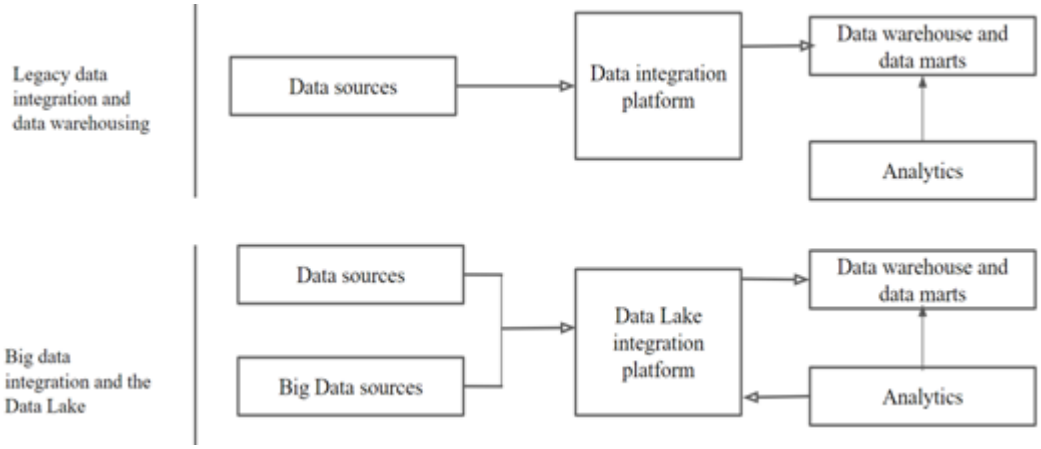

Figure 1: Transitioning to the Data Lake

2. Conclusion

As more than 80% of data generated across the world is unstructured, organizations have acknowledged the requirement of big data architecture for boosting growth, looking at the current scenario adoption of data lakes will surely rise as organizations will start reaping from data lake implementations. Therefore organizations should develop a data governance strategy, consider the cost of implementing a data lake, and develop an analytics strategy to ensure they are getting the most out of their data. By following these strategies, organizations can unlock the potential of clinical data lakes for future generations. The value of a data lake can be enhanced significantly by including strong data governance combined with data quality, metadata management, and data security processes in the design, loading, and maintenance of the environment, active participation by experienced professionals in all of these areas is also crucial, otherwise, your data lake might become more of a data.

References

[1] Hai, R., Geisler, S., Quix, C.: Constance: An intelligent data lake system. In: Proceedings of the 2016 International Conference on Management of Data. *SIGMOD '16, New York, NY, USA, ACM* (2016) 2097– 2100
[2] Maini E, Venkateswarulu B. Data Lake-An Optimum Solution for Storage and Analytics of Big Data in Cardiovascular Disease Prediction System. *IJCEM* (2018) 2230-7893.
[3] Inmon, B. Data Lake Architecture: Designing the Data Lake and Avoiding the Garbage Dump. *Technics Publications* (2016).
[4] Wang, H., Zhang, Z., Taleb, T. Special issue on security and privacy of IoT. *World Wide Web* (2017) 1–6.
[5] Mathew, P.S., Pillai, A.S.: Big data challenges and solutions in healthcare: A survey. In: Innovations in BioInspired Computing and Applications. *Springer (*2016) 543–55.
[6] Zaino J, Wageman, JV.: Data lakes take healthcare analytics to the next level. *Healthtech* (2019).
[7] Shashank A. Why data lake is the future of healthcare. Health IT answers (2017).
[8] LaPlante A. Architecting data lakes. O'Reilly Media; 2016.
[9] Roski J, Bo-Linn GW, Andrews TA. Creating value in health care through big data: opportunities and policy implications. Health affairs. 2014 Jul 1;33(7):1115-22.
[10] Krause DD. Data Lakes and Data Visualization: An Innovative Approach to Address the Challenges of Access to Health Care in Mississippi. *Online J Public Health Inform*. 2015;7(3):e225
[11] Data lakes: Driving the digital transformation journey. ebook.