frontiers | Frontiers in Digital Health

TYPE Review
PUBLISHED 16 December 2025
DOI 10.3389/fdgth.2025.1691142

OPEN ACCESS

EDITED BY
Björn Wolfgang Schuller,
Imperial College London, United Kingdom

REVIEWED BY
Roopa Foulger,
OSF HealthCare, United States
L. Raymond Guo,
St Louis VA Medical Center-Jefferson Barracks, United States

*CORRESPONDENCE
Laurent Peyrin-Biroulet
✉ peyrinbiroulet@gmail.com

RECEIVED 22 August 2025
REVISED 17 November 2025
ACCEPTED 19 November 2025
PUBLISHED 16 December 2025

CITATION
Knezevic Ivanovski T, Honap S, Matic R, Markovic S and Peyrin-Biroulet L (2025) Building a healthcare data warehouse: considerations, opportunities, and challenges. Front. Digit. Health 7:1691142. doi: 10.3389/fdgth.2025.1691142

Building a healthcare data warehouse: considerations, opportunities, and challenges

Tamara Knezevic Ivanovski[1,2], Sailish Honap[3,4], Rade Matic[5], Srdjan Markovic[1,2] and Laurent Peyrin-Biroulet[6]*

[1]Department of Gastroenterology and Hepatology, University City Hospital Zvezdara, Belgrade, Serbia, [2]Faculty of Medicine, University of Belgrade, Belgrade, Serbia, [3]Department of Gastroenterology, St George's University Hospitals NHS Foundation Trust, London, United Kingdom, [4]School of Immunology and Microbial Sciences, King's College London, London, United Kingdom, [5]Department for Information Systems and Technologies, Belgrade Academy for Business and Arts Applied Studies, Belgrade, Serbia, [6]Department of Gastroenterology, CHRU Nancy, Inserm NGERE, Lorraine University, Vandoeuvre, France

The increasing digitalization of health systems is accelerating the transition towards a new era of data-driven, evidence-based care. This has profound implications for clinical practice, performance evaluation, policy making and biomedical research. At the heart of this transformation lies a healthcare data warehouse (DW), which functions as a critical infrastructure for aggregating, standardizing, and analyzing diverse clinical and administrative data. When well-designed and implemented, DWs provide clinicians with timely access to comprehensive, longitudinal patient data, enabling more informed decision-making, enhancing care quality, and improving outcomes. For researchers, these repositories offer opportunities for population-level analytics, predictive modeling, and large-scale health service research, enabling insights into disease patterns, healthcare utilization, and system inefficiencies. Centralizing clinical and administrative data in a DW allows for more frequent, nuanced analyses, increasing the precision and responsiveness of care. However, developing an effective DW requires careful consideration of system architecture, data governance, and interoperability. These foundational components support the robust ETL/ELT frameworks that ensure data quality, consistency, and readiness for analysis across diverse and evolving data streams. Beyond supporting individual patient care, DWs act as essential drivers of scalable research, operational efficiency, and evidence-based health policy. Their successful implementation marks a pivotal step toward achieving personalized, high-quality, and cost-effective healthcare in the digital transformation age. This paper reviews the existing literature to outline the process of building and implementing a data warehouse, introducing real-world disease-specific applications. BiotherDW connects theoretical frameworks with practical healthcare applications by demonstrating how traditional data warehouse design can be adapted for national-scale digital health infrastructures.

KEYWORDS

data warehourse, inflammatory bowel disease, ETL/ELT, data integration and interoperability, artificial intelligence in healthcare

1 Introduction

The digitalization of healthcare systems has significant scope for improving healthcare delivery and patient outcomes (1–3), as it moves towards evidence-based, data-led health. Much of this change has been attributed to the widespread adoption of Electronic Health Records (EHRs), which facilitate the entry, processing, storage, and retrieval of digital health data (4). This, in turn, can support vast quantities of data such as clinical practice, healthcare monitoring and evaluation, policymaking, and clinical research (5, 6). Modern healthcare systems produce enormous amounts of data through EHRs. They are built to manage information at the individual patient level, including clinical notes, laboratory results, imaging, prescriptions, procedures, and diagnoses. These contain admission-discharge-transfer records, billing, scheduling, claims processing, human resources, and payroll systems (7, 8). Building on this solid foundation in digital transformation, it is important to evaluate how these systems support evidence-based, data-led healthcare and the challenges encountered when scaling up to enterprise-level data management (9). However, the transition from isolated data silos to integrated, enterprise-level data environments presents new challenges—particularly regarding standardization, interoperability, data modeling, advanced analytics, and the reuse of meaningful data. This highlights the need for centralized repositories that can integrate these diverse data sources (10, 11).

Clinical data warehouses (DWs) are crucial as they are specifically designed for integrating and standardizing information from a variety of sources, including EHR data and administrative databases (12). Through this transformation process, fragmented data becomes centralized and analyzable, helping with healthcare decisions and encouraging teamwork among healthcare professionals (13). The rise of clinical DWs as a mature and widely accepted solution for integrating various healthcare data has provided support for decision-making, enhancing it through research, development, and improved measurement quality (2). This paper aims to integrate traditional architectural frameworks on DW with modern solutions, such as data lakes and lakehouses, analytics based in AI, and data privacy. It also presents several real-world with the goal to help stakeholders be more prepared for the ever-changing landscape of digital health data infrastructure.

DW stores data in a structured form, making it more accessible for analysis. Thus, the DW serves as the backbone of data-driven decision-making across both clinical and organizational domains (14). However, alternative architectures, such as data lakes and data lakehouses, have emerged to address the growing volume of health data (15). To manage data lakes properly, organizations are incorporating data lakehouses, which enhance lake environments with warehouse-level governance and transactional consistency (16). In each of these alternative architectures, there is an inherent trade-off between governance, scalability, and analytical performance parameters (6). A focus on data warehousing provides a grounded and evidence-based perspective on what currently works in healthcare data management and what needs to change to make the next generation of data-driven medicine a reality (17).

2 Methods

This narrative review sought to provide an overview of the existing literature on building health care data warehouses. A comprehensive search was conducted in MEDLINE (via PubMed) of articles published from 2000 to 2025. Keywords related to data warehousing and information technology in health, such as “data warehouse”, “health-care data infrastructure”, “interoperability”, and “electronic health records”, as well as “clinical data” were employed and combined in various ways to find relevant literature. No strict set of pre-defined criteria for article inclusion was utilized for this narrative review—selection was based on the relevance and its originality/contribution towards the field. Additional references were identified through hand searching of reference lists to identify relevant papers.

3 Governance and analytics in healthcare: data architecture patterns

In the context of modern healthcare analytics, there were previously three basic data architecture patterns: DW, data lake, and data lakehouse, supporting business intelligence (BI) and data-driven decision-making (14, 18). Each fulfills a different need an organization has in storing, structuring, and analyzing information, providing a spectrum of uses from standardized reports through exploratory data science projects to AI applications for encoding information (19).

A DW is a tightly governed, centralized repository that brings together detailed data from multiple sources (1). It follows the schema-on-write paradigm, so data is validated and organized before being loaded–usually into star or snowflake schemas to improve report processing speed (20). This approach ensures high data fidelity and supports repeatable, trustworthy analytics so DWs are especially good at making trusted operational reports, regulatory filings and performance metrics for the hospital (19, 20).

Consequently, a data lake is a vast, free-form repository capable of holding raw and partly structured data in its original state (21). With a focus on a schema-on-read pattern, data lakes enable you to take in wide-ranging sets of data types, including structured (like EHR exports), semi-structured (like JSON logs), and unstructured formats (cell notes from patients to their caregivers) (21, 22). This architecture supports real-time data ingestion, large-scale exploration, and data lakehouse pipelines (14, 21). However, it frequently requires supplementary metadata and governance frameworks to ensure data quality and make it discoverable (22). Nevertheless, the lack of angular control and vigorous governance in data lakes often leads to poor data discovery, different quality, and regulatory hazards, making them inappropriate for clinical decision support, where

people must have stabilized tracing information that has been verified (23).

To bridge this gap, the data lakehouse architecture proposes a new pattern where flexibility and scalability coexist with integrity and metadata governance in DW, making possible not only transparent analytics across unstructured and structured health data but also active interrogation (24). It promotes the integration of business intelligence, AI, advanced analytics and other types of data. Lakehouses encourage innovation while maintaining the necessary auditability for application data by allowing concurrent access to raw and refined data sets.

Altogether, these architectures are not distinct, but increasingly interoperable, forming layered data ecosystems that correspond with real-world data capture and demand for different types of analytics, from operational intelligence to predictive modeling (25). Data lakes support exploration, lakehouses enable innovation, and DWs remain the cornerstone of consistent reporting and governance (26). Organizations typically adopt a DW when their main objective is to ensure reliable, standardized business reporting supported by robust governance and regulatory compliance. DW means stable, curated data models that are easy to maintain and integrate with BI tools such as Power BI or Tableau (19, 26). With the potential for better performance on analytical queries, team adoption is easier because it is SQL, and there is greater data accuracy. This is vital where precision and reliability take precedence over flexibility in many industries (2, 26). The choice of DW as the principal data management system for healthcare organizations is prompted by the stringent requirements of domains on data quality, semantic consistency, and regulatory compliance (27). By enforcing schemas and metadata, DW converts diverse clinical and administrative sources into coherent, high-integrity datasets, ensuring temporal consistency that schema-on-read data lakes cannot reliably provide (28, 29). They also offer mature SQL/BI tooling and auditable ETL/ELT and lineage mechanisms, ensuring traceable access to patient records at any point in time, supporting accountability under regulations like the Health Insurance Portability and Accountability Act (HIPAA) and the General Data Protection Regulation (GDPR) (6, 30). Lakes and lakehouses are scalable, but they fail to address governance and interoperability, which remains a challenge in the era of multisource healthcare (31). DWs that are well-governed in practice not only promote analytical adoption and stewardship but also imbue trust, so the advantages of conducting longitudinal, comprehensive analysis or follow-up across time periods are enjoyed by many providers. This is why they remain as one of the primary foundations for high-integrity analytics (29, 32).

Nevertheless, most healthcare organizations still rely on well-organized DW as the foundation for their analytical infrastructure, using data lakes and lakehouses primarily as auxiliary environments for large-scale learning and data exploration (22, 26). From a strategy perspective, the DW remains the most trusted layer for ensuring accuracy, lineage auditing, and alignment with clinical facts. Newer architectures are gradually adopting these foundational characteristics, reflecting a shift toward a more standardized, "factory assembly line" approach to data management (6). So, while data lakes and lakehouses enable analyses beyond clinical applications, they do not replace the foundational role of DW in healthcare systems. They extend it instead into today's hybrid world, where map-and-reduce clusters, deterministically reworking cloud architectures, historical record management, and analytics meet modern scientific workflows and big-data environments (23, 24).

4 Core architectural models for healthcare data integration

Based on their data integration strategies and governance models, healthcare DWs and integrated data repositories (IDRs) can be divided into four main kinds of architecture (32, 33). The general model of architecture is the traditional and most common, where independent clinical and research stores are built into a centralized staging layer. This handles extraction, transformations, and harmonization. This design provides strong control over data quality and consistency, supporting large-scale medical data mining (30). Based on this, the biobank-specific architecture is tailored for managing blood specimens and gene data through a centralized bio-sample database that connects biological materials to their associated clinical metadata (5). In contrast, the application-layer or user-controlled architecture eliminates the need for a persistent staging layer, instead performing data preprocessing and integration dynamically at the time of query execution, providing flexibility for exploratory and ad-hoc research (34). The federated architecture model enables data from different institutions to be distributed countrywide, with real-time virtual integration through adapters and regular preprocessing rules. This model is particularly suited to multi-institutional collaborations, such as national or international research networks, where secure data sharing is essential (35).

Collectively, these architectural approaches outline the range between centrally and decentrally integrated systems, each presenting different trade-offs in scalability, interoperability, and governance within healthcare data warehousing environments.

5 Data research networks and data registries

Healthcare data ecosystems increasingly combine institutional data architectures, such as DW, data lakes, and lakehouses, with collaborative infrastructures like disease-specific registries and data research networks (DRNs) (14, 36). Each provides distinct advantages and limitations, and together they address the challenges of fragmentation, scalability, and accessibility (22). DRNs are federated infrastructures that connect multiple healthcare organizations to enable large-scale, collaborative research while maintaining local data control and patient privacy. They rely on standard data models such as OMOP or PCORnet to standardize data across institutions, allowing

reproducible and comparable analyses (37). These models provide a *virtual centralization* of data while preserving local control, handy for multi-institutional studies where data-sharing restrictions apply (18). Disease-specific registries, in contrast, are focused databases that systematically collect and manage detailed information on patients with a particular condition, procedure, or treatment within a defined clinical area (e.g., inflammatory bowel disease, oncology, or rare diseases) (7, 15, 38, 39). These registries support clinical research, quality improvement, and epidemiological monitoring by providing curated, high-quality datasets derived from institutional DWs. Together, DRNs and disease-specific registries represent the collaborative layer of modern healthcare data architecture—extending institutional DWs toward population-level and cross-institutional insights (38).

At the organizational level, a DW integrates clinical, administrative, and laboratory data into standardized, high-quality datasets that often serve as the primary source for disease-specific registries (27). These registries, in turn, feed into multi-institutional DRNs that harmonize local data models through frameworks such as OMOP or PCORnet, enabling federated analytics and large-scale clinical studies (40). Meanwhile, data lakes and lakehouses extend this architecture by supporting multimodal data types imaging, genomics, sensor data, allowing registries and DRNs to incorporate richer, unstructured information for advanced analytics and machine learning (24). Together, these layers form a complementary data architecture: institutional DWs provide governance and quality, registries ensure clinical specificity, and DRNs enable collaborative, large-scale research.

DRN, data registries, along with DW, create a complementary ecosystem that integrates diverse data streams into cohesive, high-quality, and actionable insights. This integration facilitates improved clinical decision-making, enhances the efficiency of health systems, and supports informed health policy development on a large scale. However, their true value in interface applications can only be judged by how well they serve their diverse stakeholders, who then turn to consider the perspectives of patients, doctors, administrators, managers, researchers, and politicians.

6 Stakeholders in healthcare data: who benefits and how?

Healthcare data from DWs serves a variety of stakeholders, including patients and service users who depend on information to make informed choices for their health and care (22). Patients and service users rely on these data to make informed choices about their health and care (Figure 1). By providing clinicians with access to each patient's comprehensive medical history, DWs facilitate more personalized and effective care (11, 25).

For clinicians, it provides a complete view of and seamless access to each patient record with relevant patient history, laboratory investigations, and outcome data, leading to improved evidence-based clinical decisions and patient outcomes. However, clinicians' need for up-to-date, real-time data can conflict with other priorities: for example, researchers often require stable, static datasets to conduct rigorous analyses. Balancing these operational needs with research requirements is a key challenge in DW governance.

Health administrators and managers use DWs for operational insights to optimize hospital operations, staffing, and resources. This focus on control can sometimes put administrators at odds with academic stakeholders: administrators may favor strict oversight, whereas researchers advocate for broad data access to drive innovation (41, 42).

Researchers and analysts use robust, integrated datasets from DWs to conduct population-level analyses, derive insights on disease patterns, treatment effectiveness, and patient outcomes, and support innovation through rapid access to standardized data (18, 23, 43). Rapid, standardized data accelerates research and innovation. However, research goals can also conflict with operational constraints due to the requirement for extra curation or approval, potentially delaying analysis (18, 39).

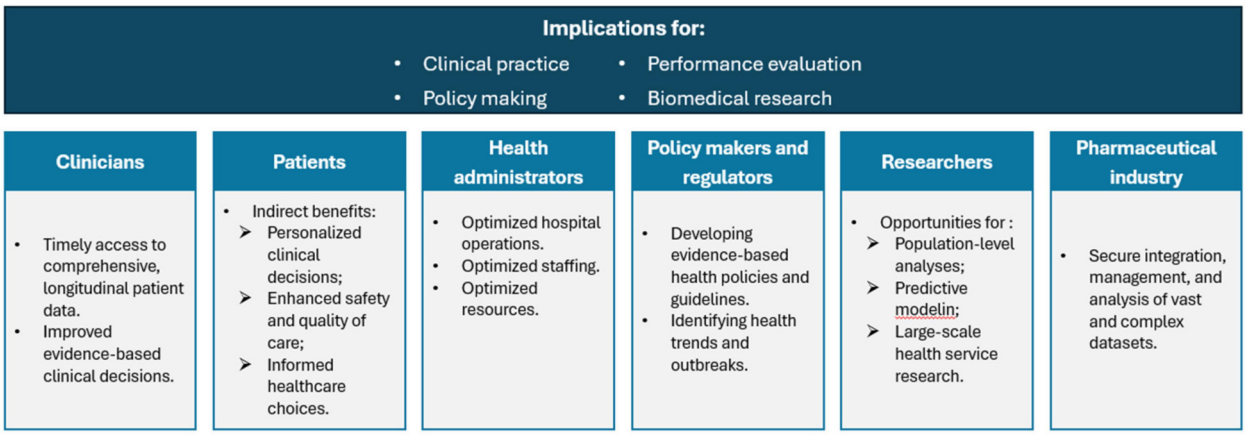

FIGURE 1
Stakeholders in healthcare data: who benefits and how?

Policy makers and regulators rely on DWs for developing evidence-based health policies and guidelines, effectively managing public health by identifying health trends and outbreaks and increasing transparency and accountability through detailed performance assessments. By identifying trends and outbreaks, DWs enable more proactive, evidence-driven policies. Nevertheless, in urgent situations, policy demands can push for rapid data sharing in ways that challenge standard governance or privacy safeguards (5, 18, 38).

Patients benefit indirectly through more personalized clinical decisions, enhanced safety and quality of care due to systematic monitoring, and empowerment to make informed healthcare choices (19). The long-term viability of digital health ecosystems hinges on sustaining patient trust. Ensuring transparency, accountability, and demonstrable value to patients is therefore essential to maintaining the social license for data-driven innovation.

A robustly designed DW represents a critical asset in the pharmaceutical industry, providing a secure and unified platform for integrating, managing, and analyzing large, complex datasets (39). Through its standardized and centralized structure, it facilitates regulatory compliance as an essential requirement in this tightly controlled domain (44). However, the substantial commercial value of such data necessitates transparent oversight and stringent ethical governance.

Further ethical and governance challenges: In addition to tensions specific to stakeholder interests, there are even more important political and moral issues that must be addressed. Questions of data ownership remain unresolved: institutions typically claim ownership of health records, but patients are increasingly asserting their rights over their own data (33). Explicit policies that respect patient rights and promote use of data for the public good should regulate DWs. Moreover, a whole series of questions about the political, moral, and technical aspects new technology has brought to light now need our attention. A DW must be operated with clear policies that safeguard patient rights and ensure data are used for the common good.

7 Challenges in healthcare DW

By systematically outlining challenges bridging data integration, quality, security, and regulatory compliance, this review aims to offer a comprehensive perspective for practitioners, researchers, and policymakers. It will not only clarify existing complexities but also establish a foundation for future research and advancement in healthcare DW systems.

Despite their transformative potential, the quality and reliability of the information in healthcare DWs can be severely compromised by issues encountered during their implementation and operation. Inadequate management of these challenges can result in unreliable data insights, which may threaten patient safety and reduce the effectiveness of clinical interventions. Healthcare organizations require scalable, high-performance, and secure infrastructures capable of handling rapidly expanding datasets while maintaining data integrity and accessibility (45).

Healthcare systems operate under stringent regulatory and ethical constraints that shape data governance, privacy, and security. Accordingly, strategies for ensuring patient confidentiality and ethical data use directly influence the integral DW architecture and its operational framework (46).

7.1 Data integration complexity

One of the major challenges in healthcare data warehousing is how to integrate diverse data sources. Healthcare data comes from many sources, including hospitals, clinics, laboratories, and government health agencies, each with its own data formats and standards (47). The variety includes EHRs, lab results, imaging data, billing records, and more, creating inherent complexity when trying to combine this data into a single DW (48). Adding to the difficulty, there are many data types beyond structured information (like coded diagnoses and lab results) and unstructured data (such as clinical notes or x-ray images), which require different processing and storage approaches (49). Managing it in this way ensures that whatever information is received is handled with care so as to be accessible and grasped by people. In addition, the use of different terminologies and standards, such as ICD codes, SNOMED CT, or any local coding scheme, easily makes interoperability difficult to achieve effectively or is responsible for semantic inconsistency (48). Aligning these terminologies demands careful mapping and normalization efforts to support accurate cross-institutional data queries and analytics.

In addition to technical heterogeneity, the biggest obstacle is integrating organizational and infrastructure-related issues (27, 29). This often causes problems when old hospital automation systems need to interface with modern DWs, as well as when clinical information systems try to operate within newer ones (47). Complementary mismatches between legacy platforms and hardware can also pose barriers. These platforms often lack open structures or freely available APIs for information exchange between systems, so data does not flow as freely as users would like (39). Additionally, a lack of communication and collaboration between IT personnel and clinical staff compromises successful requirements gathering and system design. In the worst-case scenario, these issues lead to suboptimal data models that lower the warehouse's utilization factor below expected levels (50). Maintaining accurate data provenance and audit trails presents additional complexity, as healthcare systems must document data origins and transformations rigorously to ensure reliability and compliance with legal standards. Maintaining robust provenance and audit trails adds operational complexity, as healthcare systems must rigorously document data sources and transformation steps to safeguard reliability and meet regulatory requirements (51).

7.2 Data quality and consistency

Organization and completeness of data presentation posed challenges in this paper. Meanwhile, accuracy is always crucial

for successfully rolling out or disseminating data products that truly reach a broad audience. Therefore, both quality and data sources need improvement. For reliable healthcare analysis, high-quality data is essential. Healthcare DWs often aggregate data from multiple institutions, making the preservation of precise, byte-level data integrity a formidable challenge (52). It may later become impossible to determine how much of the original source material was lost due to copy-and-paste operations (another form of substitution error). Various data collection methods, human errors during encoding, and missing, null, "don't know", or "refuse to answer" values threaten to make integrated data sets either incomplete or inaccurate or both. Poor-quality data can carry through the entire analytical process, directly impacting patient outcomes or the validity of decision support systems (53). Nowadays, mitigation strategies increasingly rely on combined verification and data cleaning techniques, AI-assisted anomaly detection, deduplication, and normalization (54). These automated systems keep data clean by continuously reporting anomalies they detect to system administrators, eliminating the need for manual intervention while increasing both trust and efficiency.

The next issue concerning data quality involves bias and imbalances in datasets. EHR and other healthcare data banks can reflect demographic, socioeconomic, or clinical biases embedded in healthcare provision and diagnosis documents (53). Still, many areas underrepresent minorities, which can alter data distribution. Failing to address this can lead to unfairness and errors in AI or ML models trained on such databases (54). Reducing bias requires systems for ongoing algorithm monitoring and model recalibration to uphold fairness. Without these safeguards, models may propagate inequities. Approaches such as bias-aware resampling and transparent evaluation help detect and mitigate underlying data imbalances (55).

In modern healthcare, real-time data integration is increasingly regarded as a crucial tool for rapid intervention. However, the challenge of maintaining data freshness and synchronization continues to be a technological obstacle for healthcare DWs (56). Real-time integration of lab results along with clinical monitoring data disrupts the flow, due to both speed and scale (57). To manage high-velocity data, it is vital to design streaming infrastructures that can handle scalable ingestion and processing. Additionally, it is critically important to keep consistency and cohesion between real-time and batch datasets to ensure accurate clinical decision-making (47). This environment requires a robust streaming architecture capable of handling high data input rates. These challenges necessitate scalable infrastructure and advanced ETL/ELT (Extract, Transform, Load/ Extract, Load, Transform) pipelines specifically designed for the dynamic nature of medical data.

7.3 Security and privacy concerns

Since healthcare data is sensitive, strong data protection is a top priority. Managing patient consent dynamically within health environments is essential for fulfilling legal responsibilities that benefit patients (46). Traditional static consent methods are often insufficient in complex data ecosystems supporting multiple secondary uses. Systems must incorporate transparency and accountability, allowing patients to decide who can access their data and for what purposes (46).

The cornerstones of protecting privacy in healthcare DWs are data encryption and robust access control. There is a growing emphasis on models of attribute-based access control (ABAC) nowadays, which can adaptively define security policies in more detail and include various attributes related to context, rather than just relying on set user roles (30). Protocol-oriented XPOTAM techniques protect valuable information during data sharing, cloud migration, and similar processes (58). The same is true for tokenization with synthetic data. However, these measures not only prevent sensitive data from being disclosed without authorization but also help organizations meet strict regulatory requirements such as the GDPR and HIPAA. Implementing these controls requires technological solutions along with organizational processes that align with the regulatory standards of the respective field (30).

But despite the advances, cloud-based DWs remain vulnerable to cybersecurity threats like data breaches or ransomware (59). In a distributed data architecture, as data moves across multiple systems and networks, it increases the attack surface unless strong protective measures are in place. An ongoing operational challenge is balancing the needs of clinical users and researchers for easy access with strict security controls (46). Effective risk management frameworks and continuous security monitoring are vital for protecting sensitive patient data in this rapidly changing landscape.

7.4 Scalability and performance issues

Healthcare DWs need to load information from a wide variety of sources, such as population health records, images, genetic information and wearables. Big data poses significant challenges for both storage and processing infrastructures, so solutions must be scalable. NoSQL databases outperform traditional SQL relational systems for querying clinical data. This provides a significant advantage in terms of both performance and flexibility (45). To improve operational efficiency, a business needs to grow, and many are moving to cloud native architectures with elastic resource allocation, distributed processing and advanced data storage solutions. These are not only faster and more scalable architectures but also lower the price of entry, which is particularly important if cost is a consideration for project (60).

There is no doubt that limited resources for medical care in emerging countries present a challenge for both healthcare providers and those seeking care. Broken EHR systems, inconsistent patient IDs, and network issues hinder the integration and scalability of healthcare information. This means that during critical times, such as SARS-like outbreaks, we lack real-time capabilities for data processing and disease surveillance

(45). To address these issues, researchers offer various solutions, including data marts designed for targeted analytics, secure ingestion pipelines with code generated and automatically adapted to meet local standards, and a lightweight architecture that requires minimal infrastructure (45).

Optimization strategies for healthcare DWs include employing dimensional modeling during cloud migration to streamline schema design, enhance query efficiency, and increase reporting flexibility (58). Automated ETL/ELT tools minimize manual intervention, enhance process consistency, and shorten deployment cycles (52). In distributed environments, advanced load balancing and provisioning techniques further improve system uptime. Testing with highly successful large-scale deployments demonstrates that the same technical environment can be transformed into something much more responsive than is typical for today's software and hardware combined.

7.5 Resource constraints and technical expertise

Building an effective healthcare DW depends on expertise in data science, medical informatics, and technology management. Reliance on highly specialized personnel can hinder scalability and delay the adaptation of DWs to evolving needs (52). Closing this technical gap requires programs for multi-skills education and capacity-building strategies that enhance both technical capabilities and relevant field knowledge. Complex healthcare DWs often need periodic maintenance to incorporate new data sources, update schemas, and refine procedures (52). However, integrating automated workflows with legacy systems to boost efficiency increases operational demands, requiring precise version control and mutual adaptability management. Additionally, financial constraints are likely to limit investment in advanced infrastructure and software, threatening the sustainability and growth of the warehouse (61).

An integrated, high-capacity data center defines the core of the precinct. This requires finding a way to bridge the cultural and communication gap between disciplines, so that people working in areas such as clinical patient care, technical support services, and administration can understand themselves not only from their own perspective but also from others, ultimately from diverse backgrounds (50). The extensive involvement of firm stakeholders makes it easier to understand expectations, builds more trust in working relationships, and enhances data stewardship. Interdisciplinary project management frameworks help promote shared accountability within stronger, more agile, and responsive healthcare environments (51).

7.6 Regulatory compliance and ethical issues

Healthcare DW must navigate to remain compliant, as data sharing procedures, security mechanisms, and AI governance frameworks constantly require assessment (62). Additionally, certain standards require that activities be pre-certified and auditable, which drives home once more the importance of complete documentation and quality management systems designed specifically for healthcare environments. This is particularly true in the case of healthcare data (46). Ethical considerations are also crucial in integrating AI into healthcare decision-making systems. Eliminating bias, ensuring transparency, and fairness when using AI models are all necessary prerequisites for medical staff to win patients' trust and deliver fair medical care (63). Regarding individual preferences, maintaining patient autonomy involves respecting informed consent and patients' wishes (46). Finding ways to balance innovation with privacy rights remains a complex challenge.

7.7 Interoperability and standardization

However, challenges arise during the mapping of local health care terminologies to these standards because there are so many linguistic varieties underpinning different places and the practices there vary over time, while documentation consistency is lacking (48). Data quality is affected by these challenges and so the analysis itself is less useful for the development of interoperable information systems (47). For example, with integrating various health IT systems, especially legacy platforms, it takes advanced middleware solutions and APIs to realize real-time data interchange between hospitals, laboratories, and repositories. Both technical and organizational coordination are required when integrating these efforts to cope with the differences between systems, update cycles, and governance silos (46). Middleware architecture ensures applications can interoperate and still specify what values to be taken for functions.

7.8 Managing evolution and digital transformation

Health care organizations are moving preferences from traditional on-site DWs to cloud native and hybrid architectures in order to harness the scalability, cost-effectiveness and improved analytic capabilities of these new kinds of systems (60). Data migration projects mean risk of losing data, system downtime, and performance deterioration. To avoid these risks, careful planning must be combined with rigorous validation procedures, including methods for ensuring continuity assurance (61). Post-migration optimization is necessary to ensure performance levels meet clinical and operational needs. AI-driven ETL/ELT processes and analytics functions are increasingly embedded inside healthcare DWs, supporting real-time analytics, anomaly detection, and predictive modeling (60). Building DWs that last into the future means creating flexible, modular infrastructures capable of integrating new analysis tools as they arrive (58). Automation tools that streamline ETL/ELT as well as data quality monitoring reduce manual labor, accelerate deployment, and provide greater consistency (52).

Automated validation frameworks can improve data integrity, reducing pre-processing required by researchers and allowing for more rapid clinical insights (51). Also, business intelligence platforms have been demonstrated to improve decision-making and operational work within laboratory workflows (57).

The various stakeholders set the strategic goals of a DW. Delivering on these aims requires robust technical support. Further, we will outline practical steps for integrating and standardizing data across systems in healthcare settings.

8 How to construct a modern DW: key points

A modern DW must start with clear business goals and well-defined use cases, such as clinical research, operational reports, or strategic analytics. The key in this phase is identifying main stakeholders and end users and understanding their needs: analysts, clinicians, executives, and administrators all have different requirements, so they help guide the design and operation of your DW (39). Next, is to find all data sources and data integration pipeline links to ensure that incoming data is complete and supports interoperability. Also, consider regulatory requirements early on, such as HIPAA or GDPR, to ensure security and compliance. This builds trust in the DW for decision-making and prepares for future use cases involving new data types that may emerge but will need analysis (30). A modern DW consolidates large volumes of historical data from diverse sources into a centralized, cost-effective repository structured for business intelligence and advanced analytics (64–66). The models for modern DW architecture use cloud computing techniques, such as elastic computing and storage, partitioning data away from the processor itself, providing organizations with flexibility and greater operational efficiency (17, 67). The shift is driven mainly by the dramatic increase in data volumes and the resultant diversity of data sources, collectively known as big data, which necessitates scalable systems that can handle high-velocity and varied data formats (60). It is crucial to recognize the importance of modern DW projects, as they are indispensable for feeding advanced analytics systems that integrate AI/ML. Consisting in predictive insights complementary to traditional reporting (49, 68).

8.1 Architectural frameworks

The architecture of DW has changed from a heavy, unbalanced on-premises design to a flexible and scalable cloud architecture. On-premises DW, based on fixed hardware resources, flexible expansion, and fixed operating costs, inherently has limitations. In contrast, cloud-based DW platforms have elastic performance and cost-efficiency at a massive scale. Furthermore, their storage capability is almost limitless. Using cloud infrastructure, an organization can dynamically allocate system resources, swelling as demand rises and sheltering in times of decline. Thus, handling cost management effectively is possible (67, 69).

Modern DW design rests on a layered approach, with separate components for the ingestion, storage, processing, and analysis of data. Layers of data ingestion handle the extraction and loading from a wide range of heterogeneous sources and often integrate batch as well as streaming processing capabilities. The storage layer needs to absorb low-cost, large-scale retention of both structured and semi-structured data (67). It performs data transformations, aggregations, and query execution. Such layered processing engines often take advantage of vectorized processing and just-in-time compilation for superior performance. A basic concept adopted by modern architecture is to separate storage from computing resources and new resources to be utilized independently at higher efficiency, particularly in cloud environments (24). Additionally, by integrating with particular layers of AI/ML, this analysis procedure can have increased added value. Companies can integrate ML pipelines directly into the DBMS environment to support feature engineering and model production. Lakehouse, an emerging architectural paradigm, demonstrates the adaptability of today's data infrastructure for modern companies. After all, the human factor also counts for a great deal. But these new models will indeed mean great technical complexities as well as whole new ball games in organizational change management issues, for example with necessity for automatic governance systems and strategies spanning all areas (70).

8.2 DW architecture

For each integration and governance need, organizations may take upon themself one of a number of different DW architectures. The Inmon architecture stresses a model of normalized data across the whole enterprise, which ensures integrity and consistency across all areas (64). On the other hand, the Kimball approach emphasizes dimensional modeling to support analytical performance and usability for decision support (66). The Data Vault model further broadens this classical paradigm by creating agile, auditable, and historical structures to support regulatory compliance and incremental evolution–things that are particularly needed in a busy, heterogeneous healthcare data environment. By comparison, emerging architectures such as the medallion architecture and the data lakehouse enforce schema layout and layer data quality functions at scale in scalable big data ecosystems. These approaches make it possible for unified analytics across structured as well as unstructured data sources (24). However, more decentralized paradigms, including data fabric and so forth, emphasize cross-organizational data sharing and independent multi-document governance enabled by metadata, a concern made more and more concrete for multi-institutional healthcare research done today (70). In healthcare, optimal DW design typically has a hybrid aspect, combining features from multiple paradigms in order to achieve interoperability,

compliance and analytical performance across the evolving clinical ecosystems.

8.3 Data modeling techniques

To ensure queries perform well and have a high degree of flexibility in a modern DW, it should make effective data models. Many people still use the dimensional modeling method. Its focus is on both the star schema and snowflake schema. These models consist of fact tables connected to multiple dimension tables, enabling OLAP queries to efficiently perform multidimensional analysis through rapid slicing and dicing of data. The star schema features denormalized dimension tables, enhancing query speed at the cost of data redundancy, while the snowflake schema normalizes dimensions to reduce redundancy but may introduce more complex joins (60, 66). Trade-offs between normalization and denormalization carry with them significant implications for scalability, performance and maintenance. Denormalized schemas make query execution faster, but will complicate ETL/ELT processes and require more storage; normalized schemas facilitate updates and consistency yet introduce the danger of slower query response times (67). Hence schema evolution and flexibility are essential capabilities in modern DWs where one is likely to meet frequent business changes requiring schema design changes or add-ons. Modern table formats enable dynamic schema changes without any outage. The formats have features such as decentralized metadata management, snapshot isolation, compactions happening atomically, and hidden parts, which permit variance in the schema while ensuring that performance will not be compromised and proper compatibility is still guaranteed for older queries (67, 71). These capabilities allow you to process real-time data and compensate for the fact that query optimization in environments with many partitions using standard planning algorithms is too slow (72).

8.4 Data integration and ETL/ELT processes

DW construction is a fundamental process of data integration primarily completed through ETL/ELT tools. Recent developments, such as automated ETL/ELT pipelines optimized for metadata management and AI integration, have transformed this landscape. These modern pipelines are designed to reduce manual workload for data engineers, enhance reliability, adapt dynamically to schema changes, address data quality issues, and ensure timely data availability (73). Nonetheless, even with technological advances as the backdrop ETL/ELT still presents problems relating to crisis management of data quality, coping with changes in data schema and interconnecting different data sources (74).

Maturity models for ETL/ELT processes have been developed to structure improvements from an *ad hoc* to a better-organized mass-movement technology. These models consist of Key Process Areas (KPAs) and Quality Objectives (QOs) that guide organizations in implementing structured, repeatable, and quality-assured ETL processes (74). However, new technologies such as ZeroETL paradigms and Incremental View Maintenance aim to reduce pipeline complexity and performance latency by enabling continuous, high-throughput upstream data processing within DWs. This pattern reduces the reliance on external stream processing engines and increases the support level for complex, high-concurrency workloads. Closer cooperation with cloud platforms and serverless architectures further reduces batch processing cycle times and cost (72).

8.5 Metadata and schema management

Efficient metadata management and schema are essential for coping with the complexity of modern DWs. Decentralized metadata management strategies have emerged to address the centralized bottlenecks characteristic of traditional systems. Apache Iceberg, for its part, makes use of distributed metadata management in conjunction with snapshot isolation and atomic commit support to provide consistent views of data while also allowing concurrent schema updates (71). This capability makes dynamic management of schema practical, therefore promoting greater business agility by reducing downtime and eliminating manual input. Integrating metadata into governance frameworks is vital for securing data lineage, assuring data quality and meeting regulatory commitments. More advanced governance constructs embody AI lifecycle stages within metadata management processes and enable metadata compliance (75). Automated governance processes improve auditability, reduce operational overhead, and align governance activities with privacy policies and compliance standards such as GDPR, Basel III, and CCPA. Schema evolution strategies focus on minimizing impact during schema changes, supporting zero-downtime operations. These include version-control mechanisms that ensure backward compatibility and allow consumers to query data seamlessly under different schema versions (71). Still, a major interoperability challenge arises when systems must integrate across large-scale multi-vendor data ecosystems using a variety of different metadata and schema management tools (75).

8.6 Data governance and quality management

Data governance frameworks ensure the accuracy, consistency, and availability of data assets critical for trustworthy analytics. Automated quality monitoring systems tailored for analytic use cases continuously assess data integrity, completeness, and freshness, thereby supporting reliable decision-making (75). Governance activities must align closely with business objectives and compliance mandates to maintain organizational trust and regulatory adherence (76). The integration of metadata and lineage systems provides comprehensive visibility over data flows and transformations, improving auditability and enabling regulatory reporting (75). Successful governance relies on a

blend of organizational readiness, cultural alignment, and technical capabilities.

8.7 Advanced analytics and AI/ML integration

Modern DWs integrate AI and ML pipelines to generate deeper analytical value, serving as feature engineering centers and operational data foundations that support end-to-end ML processes, including training, deployment, and monitoring (68). Despite the advantages, integrating ML workflows into existing warehouse systems poses challenges, including managing model production environments, data lineage, and operational consistency (77). Real-time architectures incorporating streaming ETL/ELT and incremental view maintenance support low-latency analytics required for predictive modeling and business-critical applications (67). Applications span industries such as healthcare, where predictive analytics drive hospital resource optimization and patient outcome (49). Future trends highlight the growing adoption of cloud-native AI governance frameworks that facilitate automated compliance and continuous improvement (75). However, scalability challenges remain for managing AI workloads at enterprise scale, requiring innovations in autonomous ETL/ELT and data quality monitoring processes (73).

So far, the principles of architecture and methods have been discussed. Now they will be given life in a concrete setting. The following section describes the application of these concepts in actual practice, specifically the development of an IBDDW (inflammatory bowel disease DW).

9 The case study in IBD—BiotherDW

Inflammatory bowel disease (IBD) exemplifies the complexity and chronic nature of modern diseases, requiring longitudinal, cross-disciplinary data integration. It thus serves as an insightful field test for applying DW architecture to real-world medical applications (23). The BiotherDW system, a national data infrastructure, primarily supports biologic therapy in Serbian IBD centers, demonstrating how fundamental DW principles can be effectively applied to deliver real-time clinical benefits as well as long-term research value (25, 65). The Biother team chooses a DW as a foundational layer in a phased modernization plan. Experts first stabilize governance and structure through a DW, then gradually extend toward a lakehouse architecture—integrating curated DW data with raw and semi-structured data in the lake (24, 31). The data architecture involves designing a modern DW using medallion architecture. Metadata-driven ETL/ELT frameworks are responsible for extracting data from source systems, transforming it to meet quality standards, and loading it into the DW. Data modeling emphasizes creating fact and dimension tables optimized for analytical query performance (66, 73). An excerpt of the BiotherDW schema is shown in Figure 2.

9.1 Metadata driven ETL/ELT

The existing BiotherDW solution has been developed based on many years of experience and a deep understanding of the principles of Kimball's and Inmon's methodologies, combining their best practices into a unique, customized medallion architecture (24, 64, 66). The ETL/ELT development is carefully metadata-driven, ensuring high automation, flexibility, and precision in data processing. This approach should guarantee scalability, standardization, and optimal management of the complex BI environment (73, 75). In developing this system, a strategy was used that combines a stable, denormalized model with full metadata control, integration with master data, and separation of views from physical tables with potentially flexible analytical base tables for analytics and AI. Master data, such as patient, is organized and managed at a higher level of aggregation to facilitate easier and consistent data management across the entire BI system. Such an approach ensures more effective data governance, reduces duplication, and promotes integrity by maintaining master data coherently at both detailed and aggregated consolidation levels.

To support real-time queries and audit trails, BiotherDW architecture emphasizes auditability as a key feature. The AuditID framework enables operational monitoring, such as performance diagnostics and pipeline health checks. In modern DW, this framework plays a central role in ensuring full traceability and auditability of data. It serves as a unique identifier, linking each warehouse record to the exact ETL/ELT process execution that created or modified it. This connection allows organizations to trace data lineage, knowing exactly when, how, and by which job each record entered the system. In practice, AuditID functions as a foreign key that connects fact and dimension tables to an audit log table, which contains detailed metadata about ETL/ELT executions. This metadata typically includes the job name, source system, start and end times, number of rows processed, and load status.

9.2 Medallion architecture

Medallion architecture (Figure 3) is a data organization and processing framework primarily used in modern data lakehouse environments (24, 68). The medallion pattern facilitates metadata-driven ETL/ELT processes, modular pipeline development, and gradual data validation, which together ensure data accuracy and seamless support for both traditional BI and AI workloads. The DW is the destination: a stable, curated analytical repository. The medallion architecture is the journey: a pipeline framework that progressively transforms data into trusted repository. Medallion architecture is a multi-layered data design pattern that organizes and progressively refines data as it flows through distinct stages called bronze, silver, and gold layers (24, 31). This structured approach enables scalable data management, improved data quality, and efficient analytics. The bronze layer represents the landing zone for raw, unprocessed

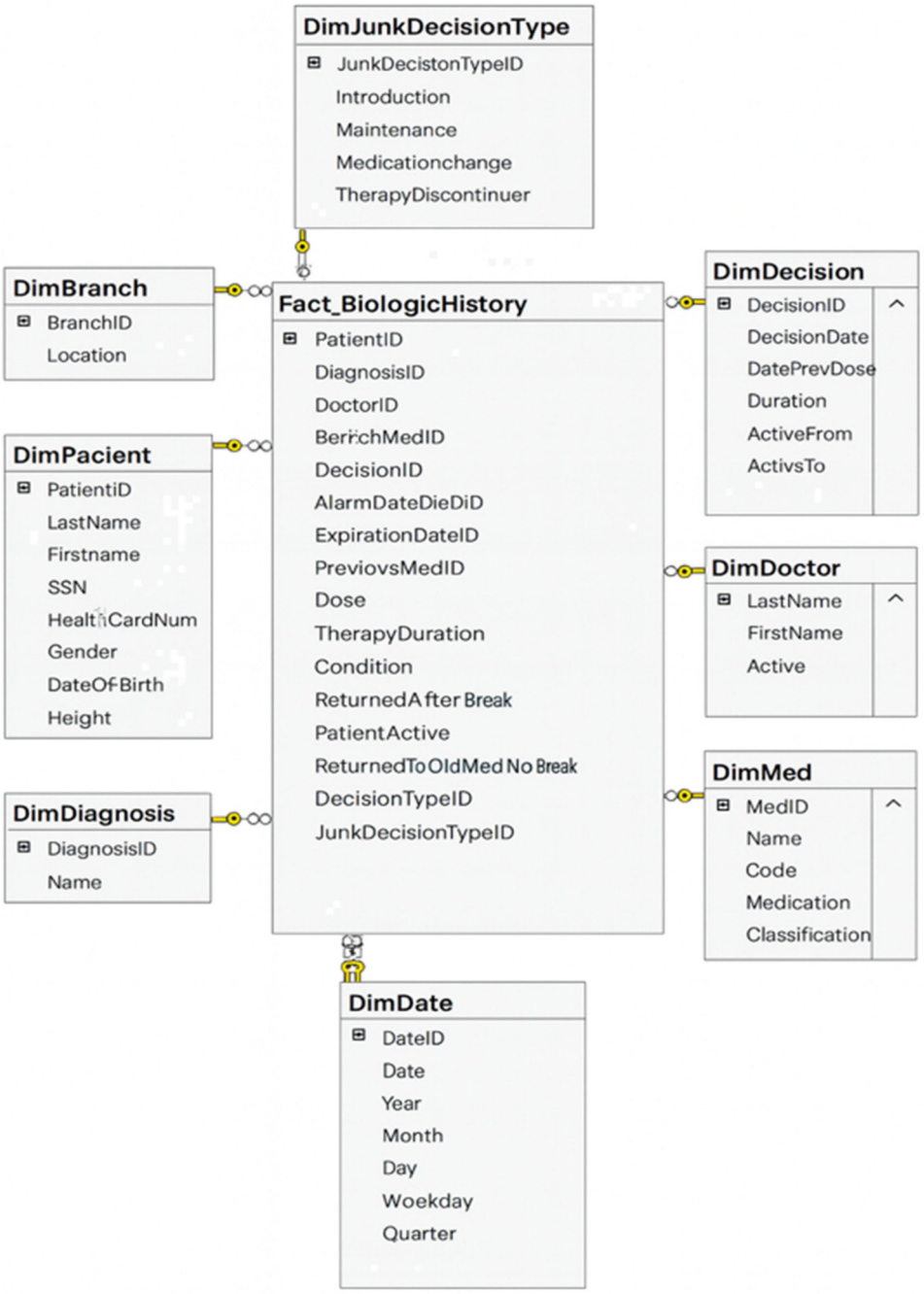

FIGURE 2
Excerpt of the BiotherDW schema.

data as it is ingested from various sources. The silver layer contains cleaned, deduplicated, and conformed data, enriched with metadata and standardized into a consistent schema. The gold layer consists of highly refined, aggregated, and business-ready data arranged for optimal query performance and integration into dashboards, reports, and decision-making workflows. The RPT schema is used as a view layer. This tiered architecture balances flexibility with governance by separating raw data ingestion from curated datasets, thus improving traceability, reliability, and operational efficiency (67).

In healthcare, medallion architecture offers a clear lineage and trust framework for clinical data pipelines. BiotherDW consolidates data from various clinical and systemic operational systems, integrating EHRs (diagnoses; prescriptions), laboratory platforms (fecal calprotectin; CRP), radiology archives (MRE; ultrasound), histopathology, and national registries. The BiotherDW system primarily sources data from two databases: the Biother transactional database (Biother) and the Serbian patient database (IZIS, Integrated Healthcare Information System). Biother is a software solution designed to manage

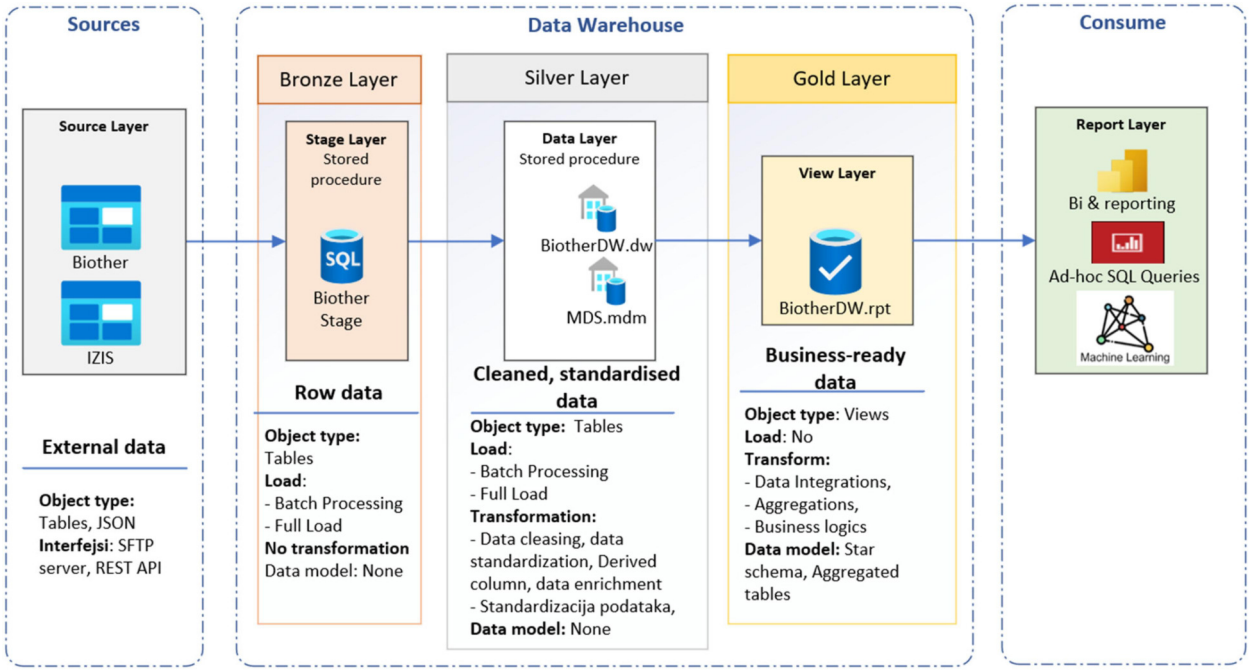

FIGURE 3
Medallion architecture of BiotherDW.

patients undergoing biological therapy, providing numerous benefits. These include digital form submission and tracking for each patient, access to their health history, scheduling colonoscopy appointments, daily therapy planning and monitoring, and automatic generation of monthly medical decisions and reports for commission approval. It also helps forecast medication needs more accurately each month. IZIS is the central electronic health system that manages comprehensive medical and health data of patients, healthcare professionals, institutions, medical interventions, electronic referrals and prescriptions, as well as appointment scheduling and diagnostics. It offers unified patient data management across the entire healthcare system, enhancing efficiency, quality, and access to services through digitalization and interoperability among public, private, and military health institutions.

The current core BiotherDW includes machine learning and AI components, though their use remains limited and mainly applied to specific analytical tasks rather than large-scale operations. However, it has been designed from the beginning to support AI. Moving forward, developing federated learning capabilities that allow collaborative model training among distributed BiotherDW nodes without transferring raw data. Our current infrastructure features modules like dimensionality reduction and interactive binning for feature engineering, using Python-based preprocessing.

10 Conclusion

Healthcare organizations need a data repository to integrate fragmented clinical, administrative, and operational data into a single, consistent source of truth that supports reliable reporting and decision-making. It enables data quality, traceability, and regulatory compliance, providing a trusted foundation for analytics, research, and performance monitoring across the healthcare system. Most healthcare organizations continue to use a well-structured DW as the core of their analytical infrastructure, while data lakes and lakehouses serve mainly as complementary environments for large-scale learning and exploration. The DW remains the most trusted layer for ensuring data accuracy, lineage auditing, and clinical consistency, even as newer architectures evolve toward greater standardization. Thus, rather than replacing the DW, lakes and lakehouses extend its role within modern hybrid ecosystems that combine historical data management with cloud-based analytics and big-data workflows.

Stakeholders in healthcare data include patients, clinicians, researchers, administrators, policymakers, and technology providers, each relying on data to inform decisions, improve outcomes, and optimize operations. They benefit through enhanced care quality, evidence-based research, operational efficiency, and policy planning, enabled by secure, well-governed data sharing and advanced analytics within healthcare data infrastructures.

The complexity of healthcare DWs arises from multiple interrelated challenges, including data integration across heterogeneous sources, ensuring data quality amid bias and imbalance, safeguarding security and privacy, and maintaining scalable performance in resource-constrained environments. Technical skill shortages, regulatory constraints, interoperability barriers, and digital transformation pressures further complicate

effective DW deployment. The success of healthcare DWs increasingly depends on interdisciplinary collaboration, continual workforce development, and adaptive governance frameworks. The construction of a modern DW is a multifaceted endeavor requiring the integration of scalable architectures, flexible data modeling techniques, robust ETL/ELT pipelines, and comprehensive governance frameworks. Security considerations, real-time processing capabilities, and AI/ML integration are critical to meeting contemporary business demands and regulatory landscapes. Emerging paradigms such as data lakehouses offer promising directions for next-generation data platforms, addressing the limitations of traditional centralized models and enhancing agility and scalability. Despite advances, notable challenges persist, including the need for greater automation in ETL/ELT and data quality management, scalability constraints in AI-driven governance frameworks, and overcoming cultural and organizational obstacles associated with new data platform adoption. Practitioners and researchers are encouraged to prioritize cloud-native architectures with adaptable data models, invest in metadata-driven governance frameworks, and leverage real-time data processing engines to maximize business agility and analytical effectiveness. By addressing these key components thoughtfully, organizations can build resilient, high-performance modern DWs that not only meet current analytical needs but also evolve with future technological and business transformations.

In diseases like IBD, a DW architecture is exemplified by the BiotherDW implementation, as reported. By combining ETL/ELT modularity and medallion architecture with the flexibility of a semantic schema that can adapt to various needs, along with regulatory-grade auditability and built-in privacy tools like synthetic data and federated learning, this system has become a modern blueprint for clinical platforms based on health data.

Author contributions

TK: Conceptualization, Writing – original draft. SH: Writing – review & editing, Methodology. RM: Writing – review & editing. SM: Conceptualization, Writing – review & editing. LP-B: Conceptualization, Supervision, Writing – review & editing.

Funding

The author(s) declare that no financial support was received for the research and/or publication of this article.

Acknowledgments

We are grateful to Srdjan Markovic for her help regarding illustrations.

Conflict of interest

SH served as a speaker, a consultant, an advisory board member, or received travel grants from Pfizer, Janssen, AbbVie, Takeda, Ferring, Lilly, Alfasigma, Banook Group, and Pharmacosmos. LP-B reports consulting fees from AbbVie, Abivax, Adacyte, Alimentiv, Amgen, Applied Molecular Transport, Arena, Banook, Biogen, Bristol Myers Squibb, Celltrion, Connect Biopharm, Cytoki Pharma, Enthera, Ferring, Fresenius Kabi, Galapagos, Genentech, Gilead, Gossamer Bio, GlaxoSmithKline, IAC Image Analysis, Index Pharmaceuticals, Inotrem, Janssen, Lilly, Medac, Mopac, Morphic, Merck Sharp Dohme, Nordic Pharma, Novartis, Oncodesign Precision Medicine, ONO Pharma, OSE Immunotherapeutics, Pandion Therapeutics, Par' Immune, Pfizer, Prometheus, Protagonist, Roche, Samsung, Sandoz, Sanofi, Satisfay, Takeda, Telavant, Theravance, Thermo Fischer, Tigenix, Tillots, Viatris, Vectivbio, Ventyx, and Ysopia; reports grants from Celltrion, Fresenius Kabi, Medac, Merck Sharp Dohme, and Takeda; lecture fees from AbbVie, Amgen, Arena, Biogen, Celltrion, Ferring, Galapagos, Genentech, Gilead, Janssen, Lilly, Medac, Merck Sharp Dohme, Nordic Pharma, Pfizer, Sandoz, Takeda, Tillots, and Viatris; and reports travel support from AbbVie, Amgen, Celltrion, Connect Biopharm, Ferring, Galapagos, Genentech, Gilead, Gossamer Bio, Janssen, Lilly, Medac, Morphic, Merck Sharp Dohme, Pfizer, Sandoz, Takeda, Thermo Fischer, and Tillots.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declare that Generative AI was used in the creation of this manuscript. AI-assisted technology (ChatGPT-4o) was used to enhance the readability and language of the manuscript with careful human oversight, verification, and editing by the authors to ensure accuracy, originality, and integrity of the content.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

References

1. Sebaa A, Chikh F, Nouicer A, Tari A. Medical big data warehouse: architecture and system design, a case study: improving healthcare resources distribution. *J Med Syst*. (2018) 42(4):59. doi: 10.1007/s10916-018-0894-9

2. Karami M, Rahimi A, Shahmirzadi AH. Clinical data warehouse: an effective tool to create intelligence in disease management. *Health Care Manag*. (2017) 36(4):380–4. doi: 10.1097/HCM.0000000000000113

3. Sukumar SR, Natarajan R, Ferrell RK. Quality of big data in health care. *Int J Health Care Qual Assur*. (2015) 28(6):621–34. doi: 10.1108/IJHCQA-07-2014-0080

4. Ozaydin B, Zengul F, Oner N, Feldman SS. Healthcare research and analytics data infrastructure solution: a data warehouse for health services research. *J Med Internet Res*. (2020) 22(6):e18579. doi: 10.2196/18579

5. Holmes JH, Elliott TE, Brown JS, Raebel MA, Davidson A, Nelson AF, et al. Clinical research data warehouse governance for distributed research networks in the USA: a systematic review of the literature. *J Am Med Inform Assoc*. (2014) 21(4):730–6. doi: 10.1136/amiajnl-2013-002370

6. Pavlenko E, Strech D, Langhof H. Implementation of data access and use procedures in clinical data warehouses. A systematic review of literature and publicly available policies. *BMC Med Inform Decis Mak*. (2020) 20(1):157. doi: 10.1186/s12911-020-01177-z

7. Gavrilov G, Vlahu-Gjorgievska E, Trajkovik V. Healthcare data warehouse system supporting cross-border interoperability. *Health Informatics J*. (2020) 26(2):1321–32. doi: 10.1177/1460458219876793

8. Polton D. Les données de santé. *Méd Sci*. (2018) 34(5):449–55. doi: 10.1051/medsci/20183405018

9. Kunjan K, Toscos T, Turkcan A, Doebbeling B. A multidimensional data warehouse for community health centers. *AMIA Annu Symp Proc*. (2015) 2015:1976–8.

10. Berndt DJ, Hevner AR, Studnicki J. The catch data warehouse: support for community health care decision-making. *Decis Support Syst*. (2003) 35(3):367–84. doi: 10.1016/S0167-9236(02)00114-8

11. Persell SD, Kaiser D, Dolan NC, Andrews B, Levi S, Khandekar J, et al. Changes in performance after implementation of a multifaceted electronic-health-record-based quality improvement system. *Med Care*. (2011) 49(2):117–25. doi: 10.1097/MLR.0b013e318202913d

12. Wang Z, Craven C, Syed M, Greer M, Seker E, Syed S, et al. Clinical data warehousing: a scoping review. *J Soc Clin Data Manag*. (2024) 4(1):8, p. 1–19. doi: 10.47912/jscdm.320

13. Visweswaran S, McLay B, Cappella N, Morris M, Milnes JT, Reis SE, et al. An atomic approach to the design and implementation of a research data warehouse. *J Am Med Inform Assoc*. (2022) 29(4):601–8. doi: 10.1093/jamia/ocab204

14. Campion TR, Craven CK, Dorr DA, Knosp BM. Understanding enterprise data warehouses to support clinical and translational research. *J Am Med Inform Assoc*. (2020) 27(9):1352–8. doi: 10.1093/jamia/ocaa089

15. Oueslati W, Akaichi J. A survey on data warehouse evolution. *Int J Database Manag Syst*. (2010) 2(4):11–24. doi: 10.5121/ijdms.2010.2402

16. Alexander CA, Wang L. Data Warehousing: A Literature Review on Effective Implementation Approaches. Data Warehous. (2023).

17. Kahn MG, Mui JY, Ames MJ, Yamsani AK, Pozdeyev N, Rafaels N, et al. Migrating a research data warehouse to a public cloud: challenges and opportunities. *J Am Med Inform Assoc*. (2022) 29(4):592–600. doi: 10.1093/jamia/ocab278

18. Curtis LH, Weiner MG, Boudreau DM, Cooper WO, Daniel GW, Nair VP, et al. Design considerations, architecture, and use of the Mini-sentinel distributed data system. *Pharmacoepidemiol Drug Saf*. (2012) 21(S1):23–31. doi: 10.1002/pds.2336

19. Goers R, Coman Schmid D, Jäggi VF, Paioni P, Okoniewski MJ, Parker A, et al. SwissPKcdw—a clinical data warehouse for the optimization of pediatric dosing regimens. *CPT Pharmacomet Syst Pharmacol*. (2021) 10(12):1578–87. doi: 10.1002/psp4.12723

20. Garcelon N, Neuraz A, Benoit V, Salomon R, Kracker S, Suarez F, et al. Finding patients using similarity measures in a rare diseases-oriented clinical data warehouse: Dr. Warehouse and the needle in the needle stack. *J Biomed Inform*. (2017) 73:51–61. doi: 10.1016/j.jbi.2017.07.016

21. Martins TGDS, Rangel FDS. Data warehouse and medical research. *Einstein São Paulo*. (2022) 20:eED6324. doi: 10.31744/einstein_journal/2022ED6324

22. Bauer C, Ganslandt T, Baum B, Christoph J, Engel I, Löbe M, et al. The integrated data repository toolkit (IDRT): accelerating translational research infrastructures. *J Clin Bioinforma*. (2015) 5(1):S6. doi: 10.1186/2043-9113-5-S1-S6

23. Murphy SN, Visweswaran S, Becich MJ, Campion TR, Knosp BM, Melton-Meaux GB, et al. Research data warehouse best practices: catalyzing national data sharing through informatics innovation. *J Am Med Inform Assoc*. (2022) 29(4):581–4. doi: 10.1093/jamia/ocac024

24. Armbrust M, Ghodsi A, Xin R, Zaharia M. Lakehouse: A New Generation of Open Platforms that Unify Data Warehousing and Advanced Analytics. (2021).

25. Chelico JD, Wilcox AB, Vawdrey DK, Kuperman GJ. Designing a clinical data warehouse architecture to support quality improvement initiatives. *AMIA Annu Symp Proc AMIA Symp*. (2016) 2016:381–90.

26. Crowson AN, Harvey M, Stout S. Data warehouse strategies and the modern anatomic pathology laboratory: quality management, patient safety, and pathology productivity issues and opportunities. *Semin Diagn Pathol*. (2019) 36(5):294–302. doi: 10.1053/j.semdp.2019.05.001

27. Karakachoff M, Goronflot T, Coudol S, Toublant D, Bazoge A, Constant Dit Beaufils P, et al. Implementing a biomedical data warehouse from blueprint to bedside in a regional French university hospital setting: unveiling processes, overcoming challenges, and extracting clinical insight. *JMIR Med Inform*. (2024) 12:e50194. doi: 10.2196/50194

28. Lamer A, Popoff B, Delange B, Doutreligne M, Chazard E, Marcilly R, et al. Barriers encountered with clinical data warehouses: recommendations from a focus group. *Comput Methods Programs Biomed*. (2024) 256:108404. doi: 10.1016/j.cmpb.2024.108404

29. Doutreligne M, Degremont A, Jachiet PA, Lamer A, Tannier X. Good practices for clinical data warehouse implementation: a case study in France. *PLoS Digit Health*. (2023) 2(7):e0000298. doi: 10.1371/journal.pdig.0000369

30. Thantilage RD, Le-Khac NA, Kechadi MT. Healthcare data security and privacy in data warehouse architectures. *Inform Med Unlocked*. (2023) 39:101270. doi: 10.1016/j.imu.2023.101270

31. Harby AA, Zulkernine F. Data lakehouse: a survey and experimental study. *Inf Syst*. (2025) 127:102460. doi: 10.1016/j.is.2024.102460

32. Lyu S, Craig S, O'Reilly G, Taniar D. The development and use of data warehousing in clinical settings: a scoping review. *Front Digit Health*. (2025) 7:1599514. doi: 10.3389/fdgth.2025.1599514

33. Gagalova KK, Leon Elizalde MA, Portales-Casamar E, Görges M. What you need to know before implementing a clinical research data warehouse: comparative review of integrated data repositories in health care institutions. *JMIR Form Res*. (2020) 4(8):e17687. doi: 10.2196/17687

34. Grossberg LB, Papamichael K, Feuerstein JD, Siegel CA, Ullman TA, Cheifetz AS. A survey study of Gastroenterologists' attitudes and barriers toward therapeutic drug monitoring of anti-TNF therapy in inflammatory bowel disease. *Inflamm Bowel Dis*. (2018) 24(1):191–7. doi: 10.1093/ibd/izx023

35. Doherty NF. The role of socio-technical principles in leveraging meaningful benefits from IT investments. *Appl Ergon*. (2014) 45(2):181–7. doi: 10.1016/j.apergo.2012.11.012

36. Martins TGDS, Costa ALFDA, Martins TGDS. Big data use in medical research. *Einstein São Paulo*. (2018) 16(3):eED4087. doi: 10.1590/S1679-45082018ED4087

37. Shahid A, Nguyen TAN, Kechadi MT. Big data warehouse for healthcare-sensitive data applications. *Sensors*. (2021) 21(7):2353. doi: 10.3390/s21072353

38. Arnold CG, Sonn B, Meyers FJ, Vest A, Puls R, Zirkler E, et al. Accessing and utilizing clinical and genomic data from an electronic health record data warehouse. *Transl Med Commun*. (2023) 8(1):7. doi: 10.1186/s41231-023-00140-0

39. Walters KM, Jojic A, Pfaff ER, Rape M, Spencer DC, Shaheen NJ, et al. Supporting research, protecting data: one institution's approach to clinical data warehouse governance. *J Am Med Inform Assoc*. (2022) 29(4):707–12. doi: 10.1093/jamia/ocab259

40. Fleurence RL, Curtis LH, Califf RM, Platt R, Selby JV, Brown JS. Launching PCORnet, a national patient-centered clinical research network. *J Am Med Inform Assoc*. (2014) 21(4):578–82. doi: 10.1136/amiajnl-2014-002747

41. Kent S, Burn E, Dawoud D, Jonsson P, Østby JT, Hughes N, et al. Common problems, common data model solutions: evidence generation for health technology assessment. *Pharmacoeconomics*. (2021) 39(3):275–85. doi: 10.1007/s40273-020-00981-9

42. Liman L, May B, Fette G, Krebs J, Puppe F. Using a clinical data warehouse to calculate and present key metrics for the radiology department: implementation and performance evaluation. *JMIR Med Inform*. (2023) 11:e41808. doi: 10.2196/41808

43. Velentgas P, Bohn RL, Brown JS, Chan KA, Gladowski P, Holick CN, et al. A distributed research network model for post-marketing safety studies: the meningococcal vaccine study. *Pharmacoepidemiol Drug Saf*. (2008) 17(12):1226–34. doi: 10.1002/pds.1675

44. Blandi L, Amorosi A, Leoni O, Clemens T, Brand H, Odone A. The potential of digital health records for public health research, policy, and practice: the case of the lombardy region data warehouse. *Acta Biomed Atenei Parm*. (2023) 94(S3):e2023121. doi: 10.23750/abm.v94iS3.14407

45. Soumma SB, Shahriar F, Mahi UN, Abrar MH, Fahad MAR, Hoque ASMdL. Design and Implementation of a Scalable Clinical Data Warehouse for Resource-Constrained Healthcare Systems. arXiv. (2025). Available online at: https://arxiv.org/abs/2502.16674 (Accessed October 15, 2025).

46. Singh S, Dulai PS, Vande Casteele N, Battat R, Fumery M, Boland BS, et al. Systematic review with meta-analysis: association between vedolizumab trough concentration and clinical outcomes in patients with inflammatory bowel diseases. *Aliment Pharmacol Ther*. (2019) 50(8):848–57. doi: 10.1111/apt.15484

47. Dhayne H, Haque R, Kilany R, Taher Y. In search of big medical data integration solutions—a comprehensive survey. *IEEE Access*. (2019) 7:91265–90. doi: 10.1109/ACCESS.2019.2927491

48. Hechtel N, Apfel-Starke J, Köhler S, Fradziak M, Schönfeld N, Steinmeyer J, et al. Harmonisation of German health care data using the OMOP common data model—a practice report. *Stud Health Technol Inform*. (2023) 305:287–90. doi: 10.3233/SHTI230485

49. Jaiswal VK. Designing a predictive analytics data warehouse for modern hospital management. *Int J Sci Res Comput Sci Eng Inf Technol*. (2025) 11(1):3309–18. doi: 10.32628/CSEIT251112337

50. Björneld O, Carlsson M, Löwe W. Case study—feature engineering inspired by domain experts on real world medical data. *Intell Based Med*. (2023) 8:100110. doi: 10.1016/j.ibmed.2023.100110

51. Marteau BL, Hornback A, Zhong Y, Lowson C, Woloff J, Smith BM, et al. Improving a large healthcare system research data warehouse using OHDSI's data quality dashboard. In: *2024 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI)*. Houston, TX: IEEE (2024). p. 1–8. doi: 10.1109/BHI62660.2024.10913573

52. Ravipati T, Andrew NE, Srikanth V, Beare R. Challenges in public healthcare research data warehouse integration and operationalisation. *Int J Popul Data Sci*. (2022) 7(3):1859. doi: 10.23889/ijpds.v7i3.1859

53. Rahman AU, Saqia B, Alsenani YS, Ullah I. Data quality, bias, and strategic challenges in reinforcement learning for healthcare: a survey. *Int J Data Inform Intell Comput*. (2024) 3(3):24–42. doi: 10.59461/ijdiic.v3i3.128

54. Chen H, Hailey D, Wang N, Yu P. A review of data quality assessment methods for public health information systems. *Int J Environ Res Public Health*. (2014) 11(5):5170–207. doi: 10.3390/ijerph110505170

55. Hooshafza S, Mc Quaid L, Stephens G, Flynn R, O'Connor L. Development of a framework to assess the quality of data sources in healthcare settings. *J Am Med Inform Assoc*. (2022) 29(5):944–52. doi: 10.1093/jamia/ocac017

56. Seethala SC. Transforming healthcare data warehouses with AI : future-proofing through advanced ETL and cloud integration. *Int J Sci Res Comput Sci Eng Inf Technol*. (2023):749–52. doi: 10.32628/CSEIT23902180

57. Mansoor I, Dar FJ. Utilizing data analytics and business intelligence tools in laboratory workflow. *EJIFCC*. (2024) 35(1):34–43.

58. Inukonda J. Leveraging dimensional modeling for optimized healthcare data warehouse cloud migration: data masking and tokenization. *Int J Sci Res IJSR*. (2024) 13(10):437–41. doi: 10.21275/SR241004233606

59. Onyebuchi A, Matthew UO, Kazaure JS, Okafor NU, Okey OD, Okochi PI, et al. Business demand for a cloud enterprise data warehouse in electronic healthcare computing: issues and developments in E-healthcare cloud computing. *Int J Cloud Appl Comput*. (2022) 12(1):1–22. doi: 10.4018/IJCAC.297098

60. Master's of Science in Information Technology, Washington University of Science and Technology, Virginia, USA, Uddin MKS, Hossan KMR. A review of implementing ai-powered data warehouse solutions to optimize big data management and utilization. *Acad J Bus Adm Innov Sustain*. (2024) 4(3):66–78. doi: 10.69593/ajbais.v4i3.92

61. Lakum S. Healthcare data migration: a technical framework for digital transformation success. *Int J Sci Res Comput Sci Eng Inf Technol*. (2024) 10(6):1362–9. doi: 10.32628/CSEIT241061177

62. Ranjbar A, Mork E, Ravn J, Brøgger H, Myrseth P, Østrem HP, et al. Managing risk and quality of AI in healthcare: are hospitals ready for implementation? *Risk Manag Healthc Policy*. (2024) 17:877–82. doi: 10.2147/RMHP.S452337

63. Familoni BT. Ethical frameworks for AI in healthcare entrepreneurship: a theoretical examination of challenges and approaches. *Int J Front Biol Pharm Res*. (2024) 5(1):057–65. doi: 10.53294/ijfbpr.2024.5.1.0032

64. Inmon WH. *Building the Data Warehouse*. Boston: QED Technical Pub. Group (1993). p. 298.

65. Evans RS, Lloyd JF, Pierce LA. Clinical Use of an Enterprise Data Warehouse.

66. Kimball R, Ross M. The Data Warehouse Toolkit.

67. Zhang F, Wu M, Xu C, Bao Y, Qiao J, Zhou Y, et al. Streaming view: an efficient data processing engine for modern real-time data warehouse of alibaba cloud. *Proc VLDB Endow*. (2025) 18(12):5153–65. doi: 10.14778/3750601.3750634

68. Aileni AR. AI/ML optimized lakehouse architecture: a comprehensive framework for modern data science. *World J Adv Eng Technol Sci*. (2025) 15(2):2099–104. doi: 10.30574/wjaets.2025.15.2.0754

69. Khan B, Khan W, Jan S, Chughtai MI. An overview of ETL techniques, tools, processes and evaluations in data warehousing. *J Big Data*. (2024) 6(1):1–20. doi: 10.32604/jbd.2023.046223

70. Blohm I, Wortmann F, Legner C, Köbler F. Data products, data mesh, and data fabric: new paradigm(s) for data and analytics? *Bus Inf Syst Eng*. (2024) 66(5):643–52. doi: 10.1007/s12599-024-00876-5

71. Lingala AR. Comparison of table formats for data warehouse. *Int J Sci Res Eng Manag*. (2024) 08(12):1–9. doi: 10.55041/IJSREM15425

72. Zhang F, Yin C, Fan H, Fang F, Chen Y, Wang X, et al. AnalyticDB-PG: a cloud-native high-performance data warehouse in alibaba cloud. *Proc VLDB Endow*. (2025) 18(12):5139–52. doi: 10.14778/3750601.3750633

73. Chanda D. Automated ETL pipelines for modern data warehousing: architectures, challenges, and emerging solutions. *Eastasouth J Inf Syst Comput Sci*. (2024) 1(03):209–12. doi: 10.58812/esiscs.v1i03.523

74. Khan M, Ali I, Khurram S, Naseer S, Ahmad S, Soliman AT, et al. ETL Maturity model for data warehouse systems: a CMMI compliant framework. *Comput Mater Contin*. (2023) 74(2):3849–63. doi: 10.32604/cmc.2023.027387

75. Dibouliya A. Unified data governance framework for AI-enabled data warehouses in banking. *Eur Mod Stud J*. (2025) 9(4):67–76. doi: 10.59573/emsj.9(4).2025.7

76. Dheeraj Kumar Bansal. Enterprise data engineering: architecting modern data warehouses for business success. *Int J Sci Res Comput Sci Eng Inf Technol*. (2025) 11(1):3266–77. doi: 10.32628/CSEIT251112348

77. Li H, Wang X, Feng Y, Qi Y, Tian J. Integration methods and advantages of machine learning with cloud data warehouses. *Int J Comput Sci Inf Technol*. (2024) 2(1):348–58. doi: 10.62051/ijcsit.v2n1.36