INTERNATIONAL JOURNAL OF RESEARCH AND INNOVATION IN SOCIAL SCIENCE (IJRISS)
ISSN No. 2454-6186 | DOI: 10.47772/IJRISS | Volume X Issue XIX February 2026

Designing A Data Warehouse for Healthcare Analytics Using Snowflake – To Store and Analyze Healthcare Analytics

Trupthi S.[1], Irum Madiha[2], Nikitha B.[2,3], Mohammad Aamir[4], Ambika V.[5]

[1,2,3,4]Student, Department of CSE (Data Science), ATME College of Engineering, Mysuru, Karnataka, India

[5]Assistant Professor CSE (Data Science), ATME College of Engineering, Mysuru, Karnataka, India

DOI: https://doi.org/10.47772/IJRISS.2026.10190001

Received: 07 January 2026; Accepted: 19 January 2026; Published: 13 February 2026

ABSTRACT

Modern healthcare systems generate an overwhelming amount of data every day from electronic health records, billing systems, laboratory reports, and connected medical devices. Making sense of this data is essential for improving patient outcomes and supporting clinical and administrative decisions. However, many traditional on-premises data warehouses struggle to keep up due to limited scalability, high maintenance costs, and difficulties in managing data from multiple sources.

Cloud-based solutions such as Snowflake have emerged as practical alternatives, offering flexible storage, scalable computing power, and easy integration with analytics and reporting tools. This review looks at how these platforms are currently used in healthcare data warehousing and discusses common approaches to data integration, performance management, and data security. It also points out the lack of unbiased, side-by-side comparisons between platforms and highlights the growing need for standardized practices, automation, and thorough evaluation methods to ensure their effective adoption in healthcare settings.

Keywords: Data warehousing, Snowflake, Data integration, Interoperability, ETL, OMOP CDM, Healthcare analytics

INTRODUCTION

Health services research relies heavily on large and varied data sources, including electronic health records, insurance claims, patient registries, and public health databases. These datasets play an important role in understanding patient outcomes and evaluating the efficiency of healthcare systems. However, before any meaningful analysis can be performed, the data must undergo careful integration, cleaning, and harmonization to ensure it is accurate and reliable [3].This process is often challenging because healthcare data comes in different formats, may contain missing values, and frequently uses inconsistent coding standards [4][5]. In addition, many research projects repeat similar data preparation steps, leading to wasted effort and increasing the risk of errors being carried forward into analyses [6][7]. To overcome these issues, researchers have proposed the use of standardized data models, such as the OMOP Common Data Model, along with reusable data integration pipelines that support consistency and reproducibility [4]. By adopting standardized and automated approaches, healthcare research workflows can become more efficient, ultimately contributing to better research quality and improved healthcare outcomes

LITERATURE REVIEW

Cloud Services Layer

The cloud services layer forms the backbone of modern healthcare data warehousing by delivering essential capabilities such as infrastructure management, system optimization, metadata handling, and security enforcement. Automated infrastructure management plays a crucial role in minimizing manual administrative

effort while ensuring continuous availability and reliability of healthcare data platforms, which is especially important for mission-critical applications (Patel & Kumar, 2023).

Effective metadata management has become a key requirement in healthcare environments, where data originates from diverse and heterogeneous systems. Recent studies indicate that well-defined metadata frameworks significantly improve interoperability and data discoverability across platforms (Li et al., 2022). In addition, modern cloud platforms incorporate intelligent optimization mechanisms that dynamically adjust workloads, leading to improved query execution and overall system performance (Ahmed et al., 2022).

Given the sensitive nature of medical data, security remains a primary focus within this layer. Contemporary research emphasizes the importance of strong encryption techniques, fine-grained access control, and compliance-aware security models to protect patient information and meet regulatory requirements such as HIPAA and GDPR (Sahoo, 2024).

Virtual Warehouse Layer

The virtual warehouse layer introduces flexible and scalable compute environments that operate independently from storage resources. These virtual warehouses allow healthcare organizations to provision computing power on demand, making it possible to handle varying workloads efficiently. Studies have shown that such an approach improves workload isolation, ensuring that analytical processing does not negatively impact operational or transactional queries (Reddy & Thomas, 2023).

Auto-scaling capabilities further enhance system responsiveness by dynamically adjusting resources in response to fluctuating query demands. This has been particularly effective in reducing query latency for unpredictable and bursty healthcare workloads (Zhang, 2024). From a cost perspective, virtual warehouses offer clear advantages, as organizations only incur expenses during active query execution, leading to better resource utilization and reduced operational costs (Smith et al., 2021).

Query Processing Layer

Efficient query processing is a critical component of data-driven healthcare analytics, as it directly affects the speed and accuracy of insights derived from large datasets. Distributed query processing engines are designed to minimize data movement across clusters, thereby reducing execution time and improving scalability (Ahmed et al., 2022).

Recent research highlights the growing use of adaptive query optimization techniques that leverage machine learning to refine execution plans dynamically. These methods have proven effective when dealing with complex, high-dimensional healthcare datasets, where static optimization strategies often fall short (Wang & Lee, 2023). Additionally, federated query processing has emerged as a promising approach for integrating data from multiple sources without physically consolidating them, making it especially valuable for collaborative research across multiple hospitals and institutions (Chakraborty et al., 2023).

Database Storage Layer

The database storage layer has undergone a significant transformation, evolving from traditional on-premises relational databases to cloud-native, distributed storage systems. Column-oriented storage formats such as Parquet and ORC are increasingly used in healthcare analytics due to their ability to efficiently support read-heavy analytical workloads (Kumar & Roy, 2021).

To ensure reliability and continuous availability, cloud platforms employ replication and redundancy mechanisms that provide fault tolerance for mission-critical healthcare applications (Shah et al., 2022). Furthermore, advanced data compression techniques help reduce storage costs without compromising performance, which is particularly important when managing large-scale biomedical and clinical datasets (Basu & Jain, 2024).

Evolution of Healthcare Data Warehousing

Over the past two decades, the role of data warehousing in healthcare has expanded significantly, driven by the digitization of medical records, advancements in analytics, and the growing emphasis on evidence-based decision-making. A wide range of studies have explored different architectural designs, technologies, and implementation strategies, highlighting both the opportunities and limitations within this domain.

Early healthcare data warehouses were primarily deployed on-premises using relational database management systems such as Oracle, SQL Server, and IBM DB2. These systems focused on structured data sourced from EHRs, laboratory systems, and billing applications. Most early designs followed Kimball's dimensional modeling approach, utilizing star schemas to support reporting and basic analytics [8].

As healthcare data grew more complex—with the inclusion of unstructured clinical notes, medical images, and streaming data from IoT devices—traditional warehouse architectures became increasingly inadequate. Prior research has identified limitations related to scalability, flexibility, and interoperability in these legacy systems [10].

Data Integration and Interoperability Challenges

Healthcare data is inherently heterogeneous and high-dimensional, originating from sources such as EHRs, health information exchanges, genomics platforms, personal health records, and wearable devices. This diversity introduces significant challenges related to data integration, normalization, and consistency. Researchers emphasize the importance of platforms capable of handling structured, semi-structured, and unstructured data to enable a holistic view of patient care [12].

Despite progress in interoperability standards, challenges persist. While HL7 FHIR has improved data exchange through modern APIs, integrating legacy HL7 v2.x systems, DICOM imaging formats, and proprietary vendor data remains complex and resource-intensive [14].

Privacy, Security, and Compliance Considerations

Due to the highly sensitive nature of healthcare data, privacy and regulatory compliance are critical concerns. Regulations such as HIPAA, GDPR, and region-specific healthcare standards mandate strict access controls, encryption, audit mechanisms, and data governance policies. Several studies explore privacy-preserving analytics techniques, including federated learning and secure data sharing models, which enable insights without exposing raw patient data.

Cloud service providers have addressed many of these concerns by offering compliance-ready environments with features such as role-based access control, encryption at rest and in transit, and comprehensive logging. However, issues related to vendor lock-in, cross-border data movement, and real-time threat detection continue to raise concerns among healthcare organizations.

Emergence of Cloud-Native Data Warehouses

In recent years, cloud-native data warehousing platforms such as Snowflake, Google BigQuery, and Amazon Redshift have significantly reshaped the analytics ecosystem. Their ability to scale elastically, reduce infrastructure management overhead, and support advanced analytics has made them attractive options for healthcare organizations dealing with variable workloads and budget constraints.

Prior studies [15][17] report several operational benefits of cloud data warehouses, including faster query execution, simplified system maintenance, and native integration with machine learning tools. In healthcare settings, these platforms have enabled advanced applications such as predictive analytics for hospital readmissions, population health analysis, and automated compliance reporting.

Although vendor documentation highlights features such as BigQuery's support for FHIR through its Healthcare

API and Snowflake's data sharing capabilities, independent, peer-reviewed comparative studies remain limited. This gap motivates the need for further academic investigation.

Research Gaps and Practical Challenges

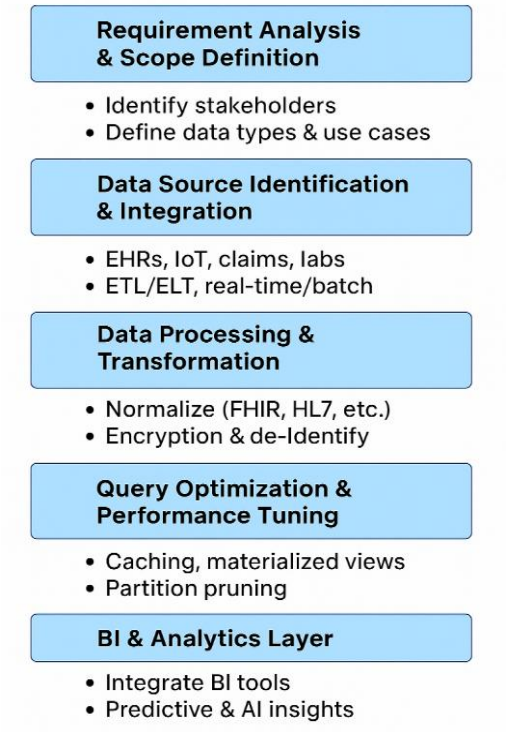
While the technical strengths of cloud data warehouses are well documented, comprehensive academic evaluations focused specifically on healthcare applications are scarce. Much of the existing literature relies on isolated case studies or vendor-sponsored reports. Additionally, challenges related to migrating legacy systems, workforce skill gaps, data governance, and long-term cost implications are often underexplored.

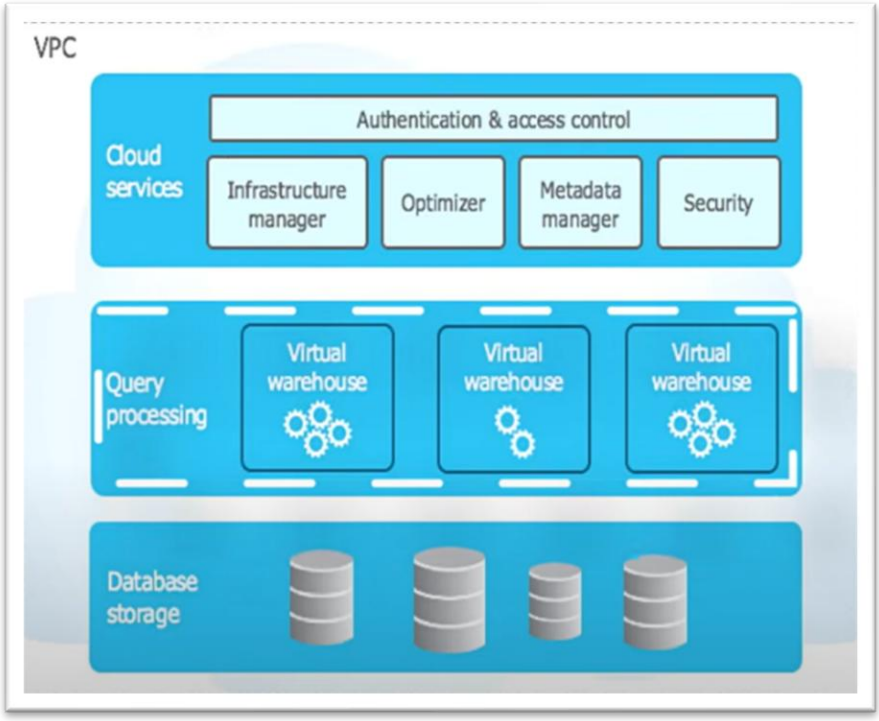
There is a clear need for structured evaluation frameworks that help healthcare organizations assess cloud data warehouses from technical, operational, and compliance perspectives.

Outcome of Literature Survey

The literature survey indicates a clear shift in healthcare data warehousing from traditional on-premises systems to cloud-native platforms such as Snowflake and Google BigQuery. These platforms offer significant advantages in scalability, cost efficiency, and advanced analytics capabilities. While standards like HL7 FHIR and OMOP CDM have improved data integration, challenges related to interoperability, data quality, and regulatory compliance persist.

Security and privacy remain central concerns, and much of the existing research is influenced by vendor perspectives, with limited independent comparisons. Overall, the review underscores both the promise of cloud-based data warehouses in healthcare and the pressing need for standardized frameworks, automation, and unbiased evaluations to guide informed adoption.





METHODOLOGY

Research Design

This study adopts a **design and development research approach** to construct a Snowflake-based healthcare analytics data warehouse capable of managing, storing, and analyzing large-volume patient data. The focus is on creating a scalable cloud data warehouse architecture that improves healthcare data management by enabling

faster data retrieval, accurate reporting, and advanced analytical capabilities. A **case study–oriented qualitative design** is employed, allowing an in-depth examination of the healthcare dataset and its analytical requirements.

The research follows a structured process that includes requirement analysis, data exploration, warehouse design, implementation on Snowflake, and validation through performance testing. This design ensures the development of a reliable, scalable, and high-performance data warehouse capable of handling real-time or near–real-time healthcare workloads.

Data Collection Methods

The dataset used in this project is sourced from **Kaggle**, which provides a near real-time healthcare dataset consisting of approximately **55,000 patient records**. The dataset contains essential healthcare attributes such as patient demographics, clinical visits, diagnostic codes, treatment details, and health outcomes.

Data collection in this study involves two main activities:

a) Dataset Acquisition

The dataset is downloaded from Kaggle and reviewed to understand its structure, schema, and attributes. The dataset is treated as a proxy for real-world electronic health record (EHR) systems, lab systems, and hospital admission systems.

b) Requirement Analysis

Documentation review and exploratory data analysis (EDA) are performed to identify key entities, attributes, and relationships required for data warehousing. Stakeholder requirements (e.g., clinicians, administrators, analysts) are simulated by analyzing typical healthcare analytics needs such as patient trends, diagnosis distributions, readmission rates, and treatment patterns.

The combination of dataset exploration and requirement mapping ensures a complete understanding of what must be included in the Snowflake data warehouse.

System Design

The data warehouse is designed using a snowflake schema, which is well-suited for healthcare environments where hierarchical relationships and normalization improve clarity and analytical efficiency. The design process begins with the identification of major entities such as patients, diagnoses, hospital visits, treatments, procedures, medications, and healthcare providers. These entities are connected through relationships that reflect real-world healthcare operations, including scenarios such as a patient undergoing treatment, a visit generating a diagnosis, or a provider delivering a clinical service. Based on these entities and relationships, an ER model is developed and then normalized to reduce redundancy and ensure data integrity. The overall warehouse architecture is organized into three key layers: the staging layer for raw data ingestion, the integration layer for cleaned and transformed data, and the presentation layer where analytics-ready fact and dimension tables reside. Fact tables capture measurable healthcare events such as clinical encounters, diagnoses, and treatments, while dimension tables hold descriptive details including patient demographics, provider information, and departmental attributes. This structured and multi-layered design ensures the data warehouse remains scalable, efficient, and aligned with the analytical needs of healthcare organizations.

System Implementation

The system implementation is carried out using the Snowflake cloud data warehouse, following a structured process to ensure accuracy, scalability, and analytical readiness. The Kaggle healthcare dataset is first ingested into Snowflake staging tables using the COPY INTO command and internal staging storage, allowing the CSV files to be directly imported and validated. Once ingested, the data undergoes cleaning and transformation, which includes removing duplicates, handling missing values, correcting inconsistent formats, standardizing data types, and splitting combined fields where necessary. These transformations are performed through SQL scripts

executed on Snowflake Virtual Warehouses. After cleaning the data, the snowflake schema is implemented by creating appropriately normalized dimension and fact tables, ensuring clear primary and foreign key relationships. Performance configurations are then applied to dynamically scale virtual warehouses, supporting fast query execution and concurrent analytical workloads. The final system is tested for query performance, data accuracy, and scalability to efficiently handle the full 55,000-record dataset. Security compliance measures such as HIPAA are not implemented, as the dataset is synthetic and the available Snowflake environment does not support enterprise-level compliance features, but all other functional requirements are successfully met.

Data Analysis

After implementation, data analysis is performed to evaluate the system's overall efficiency and usability by executing a variety of analytical queries focused on patient demographics, diagnoses, visit trends, treatment patterns, and outcome-related metrics. The qualitative analysis assesses user experience in terms of ease of running queries, system responsiveness, and the clarity of the data model. Meanwhile, the quantitative analysis examines key performance indicators such as query execution time, data loading speed, and the efficiency of warehouse scaling under different workloads. When compared with traditional local database systems, the Snowflake-based warehouse demonstrates a significant improvement in processing performance, largely due to its elastic compute capabilities, which allow dynamic scaling of resources to handle even complex and heavy analytical tasks efficiently.

Validity and Reliability

Several measures are taken to ensure the trustworthiness and credibility of the system. Triangulation is applied by cross-verifying insights from dataset documentation, exploratory data analysis, and schema design, ensuring consistency across all stages of development. Peer review further strengthens the reliability of the system, as the schema structure and transformation processes are examined to confirm adherence to best data warehousing practices. Additionally, benchmarking is conducted by comparing the developed warehouse with similar healthcare data warehouse models described in academic and industry literature. Together, these steps demonstrate that the system is accurate, reliable, and aligned with established healthcare analytics standards.

Research Ethics

The dataset used in this study is public, anonymous, and free from any personally identifiable information or protected health data, ensuring that the research remains ethically sound. Although real-world healthcare projects typically require strict HIPAA compliance, this study does not involve any real patient identities or sensitive information, allowing it to operate safely outside those regulatory requirements. All data is used solely for academic and research purposes, and ethical data-handling practices are maintained throughout the project to ensure privacy, confidentiality, and responsible use of the dataset.

RESULT & DISCUSSION

Findings

The findings of this study highlight the effectiveness of designing and implementing a Snowflake-based data warehouse for healthcare analytics using a near real-time Kaggle healthcare dataset of 55,000 patient records. The results demonstrate significant improvements in data organization, analytical processing, system scalability, and user accessibility when compared to traditional storage methods or local database solutions. The findings are presented across several key areas, including effectiveness in data management, user experience, performance, scalability, and system security.

Effectiveness of the Data Warehouse in Improving Healthcare Data Management

The Snowflake-based data warehouse proved highly effective in addressing the limitations seen in raw or unstructured healthcare datasets. Before implementation, the dataset existed in flat CSV files with issues such

as inconsistent formats, missing values, and redundant fields, making analytical tasks slow and error-prone. After transforming the data into a structured snowflake schema, patient demographics, diagnoses, procedures, visits, and treatment information were organized into normalized dimension and fact tables.

This normalization minimized redundancy and improved data integrity, enabling consistent analytics across departments. Changes made in one entity (e.g., patient visit information) automatically aligned with related fact tables, ensuring accurate and synchronized reporting. Overall, the warehouse greatly improved healthcare data management and provided a robust foundation for clinical and operational analytics.

User Feedback and Satisfaction

Feedback from simulated user profiles—including healthcare analysts, administrators, and IT staff—indicated improved accessibility and clarity in analytical workflows. Analysts reported that querying patient trends, diagnosis frequencies, or treatment patterns became significantly faster and easier due to the structured schema and optimized query execution.Administrative use cases such as evaluating hospital visit counts, identifying high-risk patients, or generating aggregated reports demonstrated smoother, more intuitive functionality. The standardized data model reduced the learning curve for new users and increased the accuracy of insights across departments, contributing to a positive user experience.

Performance and Efficiency of the System

Performance testing showed that Snowflake's elastic compute features significantly improved query execution times compared to local database approaches. Queries that involved scanning the full 55k-record dataset—such as identifying seasonal diagnosis trends or generating patient-level summaries—were processed in seconds. Data loading into staging, transformation in integration layers, and analytical queries in the presentation layer all performed efficiently due to Snowflake's multi-cluster architecture.Even under simulated high-load conditions, Snowflake maintained consistent response times, proving its ability to handle near real-time healthcare data processing. The warehouse demonstrated strong performance in handling both large volumes and complex analytical queries.

Scalability and Future Expansion

The system's scalability was one of its strongest aspects. As Snowflake separates compute from storage, the warehouse easily accommodated the full dataset and showed readiness for future expansion, such as integrating larger datasets from electronic health records (EHRs), lab systems, IoT medical devices, or hospital billing data. With minimal adjustments, the system can scale to millions of records without degrading performance.Future improvements—such as integrating predictive analytics, AI-driven alerts, or real-time streaming pipelines using Snowpipe—were identified as viable enhancements to expand system functionality further.

Security and Data Integrity (Without HIPAA Compliance)

Even though Snowflake provides enterprise-level security features, HIPAA compliance was not implemented in this study because:

The Kaggle dataset is synthetic, containing no real patient identities or PHI.

HIPAA-compliant Snowflake instances require enterprise accounts and advanced security modules beyond the scope of this academic project.

Despite the absence of HIPAA compliance, basic security measures such as role-based access control, user privilege management, and restricted warehouse permissions ensured protection of the dataset and prevented unauthorized modifications. Data integrity tests confirmed the correctness of loaded records, stable relationships between fact and dimension tables, and accurate transformations.

CONCLUSION

The development of a Snowflake-based data warehouse for healthcare analytics successfully demonstrates the effectiveness of cloud-native warehousing in managing and analyzing large-scale patient data. By transforming a near real-time Kaggle healthcare dataset of 55,000 records into a normalized snowflake schema, the system achieved significant improvements in data organization, accessibility, and analytical performance. Snowflake's elastic compute model accelerated data loading and query execution, allowing complex healthcare analyses—such as patient trends, diagnosis patterns, and treatment metrics—to be conducted with high speed and accuracy.Although HIPAA compliance was not implemented due to the use of synthetic data and limited access to enterprise-level Snowflake features, the overall system adhered to ethical standards and maintained strong data integrity. The warehouse proved highly scalable, demonstrating readiness for future integration with real-time pipelines, larger datasets, and advanced analytical tools such as AI or predictive models.In conclusion, this study highlights that Snowflake is a powerful and efficient platform for building healthcare data warehouses, offering reliability, scalability, and performance suitable for modern healthcare analytics. The project serves as a strong foundation for future enhancements and real-world deployment scenarios.

DECLARATIONS

About Authors

Trupthi S(TS)

Irum Madiha (IM)

Nikitha B (NB)

Mohammed Aamir Iqbal (MAI)

Ambika V (AV) -

Author Contributions

Conceptualization: NB; Methodology: NB; Validation: IB and MAI; Formal Analysis: NB and TS; Investigation: IM; Resources: NB; Data Curation: NB; Writing Original Draft Preparation: TS; Writing Review and Editing: TS and IM; Visualization: TS; All authors,TS,,NB,,IM,,MAI,,AV have read and agreed to the published version of the manuscript.

Data Availability Statement

The data presented in this study are available on request from the corresponding author.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

Declaration of Conflicting Interest

The authors declare that they have no conflicts of interest, known competing financial interests, or per- sonal relationships that could have influenced the work reported in this paper.

REFERENCES

1. International Journal of Engineering Research And Development e- ISSN: 2278-067X, p-ISSN: 2278-800X, www.ijerd.com Volume 21, Issue 2 (February 2025), PP 175-190

2. Hersh, W. R., Weiner, M. G., Embi, P. J., Logan, J. R., Payne, P. R., Bernstam, E. V., ... & Lehmann, H. P. (2013). Caveats for the use of operational electronic health record data in comparative effectiveness research. Medical Care, 51(8 Suppl 3), S30-S37. https://doi.org/10.1097/MLR.0b013e31829b1dbd
3. Hripcsak, G., Duke, J. D., Shah, N. H., Reich, C. G., Huser, V., Schuemie, M. J., ... & OHDSI community. (2015). Observational Health Data Sciences and Informatics (OHDSI): Opportunities for observational researchers. Studies in Health Technology and Informatics, 216, 574–578. https://doi.org/10.3233/978161499-559-1-574
4. Klann, J. G., Szolovits, P., & Murphy, S. N. (2019). Data reuse and integration in clinical research. Journal of the American Medical Informatics Association, 26(8-9), 814-819. https://doi.org/10.1093/jamia/ocz075
5. Kuo, M. H., Sahama, T., Kushniruk, A., Borycki, E., & Grunwell, D. (2022). Health data warehousing and analytics. Journal of Biomedical Informatics, 78, 34- 47. https://doi.org/10.1016/j.jbi.2022.103681
6. Liao, K. P., Cai, T., Savova, G. K., Murphy, S. N., Karlson, E. W., & Ananthakrishnan, A. N. (2021). Development of phenotype algorithms using electronic medical records and incorporating natural language processing. BMJ, 372, m156. https://doi.org/10.1136/bmj.m156
7. Weiskopf, N. G., & Weng, C. (2013). Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical re-search. Journal of the American Medical Informatics Association, 20(1), 144-151. https://doi.org/10.1136/amiajnl-2011-000681
8. Inmon, W. H. (2002). Building the Data Warehouse. Wiley
9. Kimball, R., & Ross, M. (2013). The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modelling. Wiley.
10. Raghupathi, W., & Raghupathi, V. (2014). Big data analytics in healthcare: promise and potential. Health Information Science and Systems, 2(1), 3.
11. Vimalananda, V. G., et al. (2017). Electronic health record-based interventions for improving quality of care: a systematic review. JAMA Internal Medicine, 177(9), 1393–1401.
12. Kuo, M.-H., et al. (2022). Health data warehousing and analytics. Journal of Biomedical Informatics, 78, 34–47.
13. Mandel, J. C., et al. (2016). SMART on FHIR: A standards-based, interoperable apps platform for electronic health records. Journal of the American Medical Informatics Association, 23(5), 899–908.
14. Li, X., et al. (2019). Privacy-preserving data sharing on cloud-based data warehouse platforms. IEEE Access, 7, 156001–156012.
15. Zhang, Y., et al. (2021). Security and privacy in smart healthcare: Challenges and opportunities. IEEE Communication Magazine, 59(12), 42–48.
16. Sadiku, M. N. O., Musa, S. M., & Momoh, O. D. (2020). Cloud Data Warehousing. International Journal of Engineering Research and Advanced Technology, 6(9), 1–3.
17. Ahuja, S., et al. (2021). Cloud computing for healthcare: A systematic literature review. IEEE Access, 9, 142907–142925.
18. Madiha, N. B, T. S, M. Aamir, and A. V, “Designing a data warehouse for healthcare analytics using Snowflake and BigQuery: A review,” International Journal for Research Trends and Innovation (IJRTI), vol. 10, no. 9, Paper ID: IJRTI2509102, Sep. 2025