Designing a data warehouse for healthcare analytics using snowflake and big query-A Review

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

1,2,3,4 Student, Department of CSE (Data Science), ATME College of Engineering, Mysuru, Karnataka, India
irummadiha2004@gmail.com nikithabgowda04@gmail.com , trupthisrao4@gmail.com
aamirmohammed355@gmail.com

5 Assistant Professor CSE (Data Science), ATME College of Engineering, Mysuru, Karnataka, India ambikav.cd@atme.edu.in

ABSTRACT

Healthcare generates massive datasets from EHRs, claims, labs, and IoT devices, requiring efficient data warehousing for analytics and decision-making. Traditional on-premises systems struggle with scalability and heterogeneous data, while cloud-based platforms like Snowflake and Big Query offer elasticity, cost efficiency, and advanced analytics integration. This review explores current architectures, methodologies, and challenges in healthcare data warehousing, focusing on data integration, privacy, and performance. It also highlights the lack of independent comparative studies and stresses the need for standardization, automation, and rigorous evaluation to support effective adoption in healthcare.

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

I. INTRODUCTION

Health services research depends on large, diverse datasets such as EHRs, insurance claims, registries, and public health databases, which provide insights into patient outcomes and healthcare efficiency. However, meaningful analysis requires extensive data integration, cleansing, and harmonization to ensure accuracy and reliability [3]. Challenges such as heterogeneous formats, missing data, and coding inconsistencies make preprocessing complex [4][5]. Moreover, repetitive and redundant data preparation across projects leads to inefficiencies and risks of error propagation [6][7]. To address this, standardized data models (e.g., OMOP CDM) and reusable integration pipelines have been proposed, enabling reproducibility and reducing duplication [4]. Streamlining these processes through standardization and automation can significantly improve research productivity and healthcare outcomes.

II. LITERATURE REVIEW

II.a Cloud Services Layer

- The cloud services layer provides core functionalities such as infrastructure management, optimization, metadata handling, and security. Automated infrastructure management reduces administrative overhead and ensures high availability of healthcare data systems (Patel & Kumar, 2023).
- Metadata management frameworks have been widely adopted to enable interoperability across heterogeneous healthcare systems (Li et al., 2022).
- Optimizers built into modern platforms enhance query performance by adaptively tuning workloads (Ahmed et al., 2022).

- Security mechanisms, including encryption and fine-grained access control, are emphasized in recent studies to safeguard sensitive patient records and comply with regulations such as HIPAA and GDPR (Sahoo, 2024).

II.b Virtual Warehouse Layer

- Virtual warehouses enable elastic, on-demand compute clusters that scale independently of storage. Recent work demonstrates their effectiveness in improving workload isolation, ensuring analytical tasks do not interfere with transactional queries (Reddy & Thomas, 2023).
- Auto-scaling techniques in virtual warehouses have been shown to reduce latency for unpredictable workloads (Zhang, 2024).
- Moreover, research highlights cost efficiency since resources are consumed only during query execution (Smith et al., 2021).

II.c Query Processing Layer:

- Efficient query processing engines play a critical role in enabling data-driven healthcare analytics. Distributed query optimizers can reduce execution time by minimizing data shuffling across clusters (Ahmed et al., 2022).
- Studies show that adaptive query optimization using machine learning improves execution plans in high-dimensional healthcare datasets (Wang & Lee, 2023).
- Additionally, federated query systems have been proposed to integrate multiple data sources without compromising performance, an approach particularly valuable for multi-hospital research studies (Chakraborty et al., 2023).

II.d Database Storage Layer:

- The storage foundation has evolved from on-premises relational systems to cloud-native, distributed storage architectures. Columnar storage formats such as Parquet and ORC improve analytical efficiency in healthcare by supporting fast, read-intensive workloads (Kumar & Roy, 2021).
- Replication and redundancy strategies enhance reliability and ensure fault tolerance in mission-critical applications (Shah et al., 2022).
- Data compression methods further reduce storage costs while maintaining performance, an increasingly important factor in large-scale biomedical datasets (Basu & Jain, 2024).

The importance of data warehousing in healthcare has grown significantly over the past two decades, driven by the digitalization of healthcare records, advances in analytics, and the need for data-driven decision-making. A substantial body of literature explores various architectures, technologies, and methodologies for designing healthcare data warehouses, highlighting both the opportunities and challenges inherent in the domain.

Traditional healthcare data warehouses were built on-premises using relational database management systems (RDBMS) such as Oracle, SQL Server, or IBM DB2. These systems focused mainly on structured data from electronic health records (EHRs), laboratory systems, and billing applications. Early architectures followed Kimball's dimensional modelling approach, emphasizing star schemas with fact and dimension tables optimized for reporting.[8]

With increasing data complexity, including unstructured clinical notes, imaging, and streaming data from IoT devices, traditional models became less adequate. [10] identified key limitations in scalability, flexibility, and interoperability in traditional healthcare data warehouses.

Healthcare data is heterogeneous and high-dimensional, coming from EHRs, health information exchanges (HIEs), personal health records, genomics, and wearables. This diversity introduces significant integration and normalization challenges. Researchers [12] emphasize the need for platforms capable of ingesting and processing structured, semi-structured, and unstructured data to provide a comprehensive view of patient care.

Interoperability remains a persistent challenge. The adoption of HL7 FHIR (Fast Healthcare Interoperability Resources) as a modern API standard has helped, but integrating legacy HL7 v2.x systems, DICOM imaging data, and proprietary vendor formats continues to pose issues. [14]

Given the sensitive nature of healthcare data, privacy and compliance are central concerns. Regulations such as HIPAA in the United States, GDPR in the EU, and national standards in other regions require robust access controls, audit trails, and encryption. Several studies review privacy-preserving data mining and federated learning methods that allow analytics without compromising patient confidentiality.

Cloud platforms have responded by offering HIPAA-compliant environments and security features such as role-based access control (RBAC), encryption at rest and in transit, and detailed logging. However, concerns remain about vendor lock-in, cross-border data transfer, and real-time breach detection.

In recent years, cloud-native data warehousing platforms like Snowflake, Big Query, and Amazon Redshift have transformed the data analytics landscape. These platforms offer benefits such as elastic scalability, cost-efficiency, and rapid provisioning, which are particularly attractive to healthcare organizations facing unpredictable workloads and budget constraints.

Studies [15][17] outline the operational advantages of cloud data warehouses, including faster query performance, simplified maintenance, and native support for machine learning workloads. In healthcare contexts, cloud DWHs have enabled advanced use cases such as predictive modelling for hospital readmissions, population health monitoring, and automated reporting for regulatory compliance.

Big Query's native support for FHIR through its Healthcare API and Snowflake's data sharing capabilities have received specific attention in whitepapers and case studies by the vendors themselves, but peer-reviewed comparative evaluations are limited. This review seeks to address that gap.

While technical capabilities of cloud data warehouses are well documented, comprehensive academic studies comparing their use in healthcare are scarce. Much of the existing work focuses on individual case studies or vendor-driven success stories. There is also limited discussion on the challenges of migrating legacy healthcare data warehouses to cloud-native platforms.

Furthermore, practical implementation challenges such as workforce readiness, data governance practices, and total cost of ownership are often underreported. There is a pressing need for frameworks that assist healthcare institutions in evaluating cloud DWHs from a strategic, operational, and compliance perspective.

III.OUTCOME OF LITERATURE SURVEY

The literature review shows that healthcare data warehousing has advanced from traditional on-premises systems to cloud-native platforms like Snowflake and Big Query, offering scalability, cost efficiency, and advanced analytics. While standards such as HL7 FHIR and OMOP CDM support integration, challenges in interoperability, data quality, and compliance remain. Security and privacy continue to be central concerns, and most existing studies are vendor-driven with limited independent comparisons. Overall, the review highlights both the potential of cloud-based data warehouses and the need for standardized frameworks, automation, and unbiased evaluations to guide healthcare adoption.

IV. METHEDOLOGY

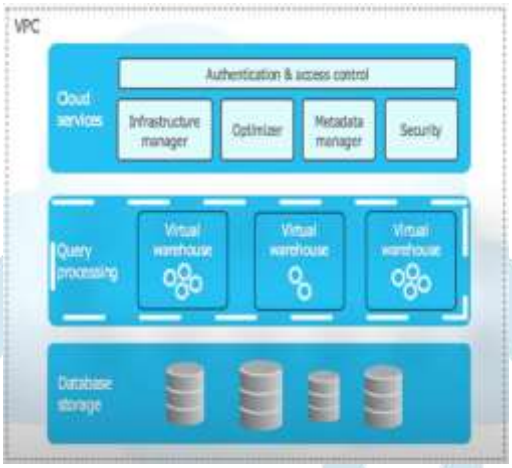

Fig 1 Model Architecture

The methodology for developing a healthcare-focused data warehouse follows a structured, multi-stage approach to ensure scalability, compliance, and analytical efficiency. The process begins with requirement analysis and scope definition, where key stakeholders are identified and the necessary data types, including patient records, clinical data, financial data, and operational data, are outlined. These requirements are further guided by potential use cases such as disease prediction, hospital performance evaluation, and cost optimization. Once the requirements are established, data source identification and integration are carried out by consolidating information from heterogeneous sources such as electronic health records (EHRs), IoT-enabled devices, insurance claims, and laboratory systems. This integration employs ETL and ELT pipelines to support both batch and real-time processing, thereby ensuring efficient and reliable data ingestion.

The next phase involves designing the data warehouse schema, typically using a star or snowflake model comprising fact and dimension tables that are optimized for healthcare analytics. To improve query performance, techniques such as partitioning, clustering, and indexing are implemented. Data processing and transformation follow, where healthcare data is normalized and standardized using frameworks like FHIR, HL7, and ICD-10 to ensure consistency and interoperability. At this stage, strong security measures including encryption, de-identification, and compliance validation are applied to safeguard sensitive information. Security, privacy, and compliance remain central to the methodology, with measures such as HIPAA- and GDPR-compliant encryption, role-based access control, and data masking being adopted. Additionally, robust data governance and auditing frameworks are incorporated to ensure ethical and secure data handling.

To optimize system performance, query optimization strategies are employed through caching, materialized views, clustering keys, and partition pruning, all of which reduce query execution time for large datasets. The business intelligence and analytics layer builds upon this foundation by integrating visualization and reporting tools such as Looker, Tableau, and Power BI. This enables not only traditional descriptive analytics but also predictive and prescriptive modelling, while advanced features such as AI-driven insights and natural language processing (NLP) expand analytical capabilities to unstructured data. Finally, monitoring and maintenance ensure the long-term reliability and sustainability of the warehouse. Automated logging, continuous monitoring, and backup strategies strengthen system resilience, while cost optimization is achieved through storage tiering and iterative performance tuning.

Overall, this comprehensive methodology establishes a secure, efficient, and scalable framework for healthcare data warehousing, capable of supporting advanced analytics while adhering to strict privacy, security, and regulatory standards.

V. CHALLENGES AND GAP

Despite the promising capabilities of cloud-based data warehouses such as Snowflake and Big Query for healthcare analytics, several challenges and gaps remain unaddressed in current implementations and literature. First, the complexity of data integration per-sists as a major obstacle. Healthcare data sources vary widely in structure, format, and semantics, including electronic health records, laboratory systems, insurance claims, and patient generated data. Harmonizing these heterogeneous datasets into a coherent data ware-house requires substantial effort in data map-ping, transformation, and validation, often resulting in time-consuming and resource intensive processes.[5]

Second, maintaining data quality presents on-going difficulties. Issues such as missing values, duplication, inconsistent coding, and erroneous entries necessitate comprehensive cleansing strategies. While ETL pipelines and automated workflows can mitigate some errors, fully ensuring data accuracy and completeness remains challenging in practice [2]
Third, privacy, security, and regulatory compliance are critical concerns in handling sensitive patient information. Cloud platforms must implement robust safeguards including encryption, role-based access controls, and audit trails to comply with regulations such as HIPAA and GDPR. Furthermore, cross juris-dictional data transfer rules and patient con-sent complexities add layers of difficulty to secure data management [5]

Performance optimization and cost management also represent significant issues. Cloud data warehouses offer scalability, but inefficient schema design, improper indexing, or large uncurated datasets can lead to query latency and inflated operational costs. Developing healthcare-specific optimization strategies is essential yet underexplored [7]
Interoperability challenges remain despite standardization efforts through frameworks such as HL7 FHIR and OMOP CDM. Mapping legacy and proprietary data sources to these standards is labore intensive and prone to se-mantic discrepancies, limiting seamless data exchange and unified analytics [8].

Moreover, there is a notable gap in comparative, peer-reviewed studies that rigorously evaluate Snowflake and Big Query in healthcare contexts. Existing literature primarily consists of vendor-led case studies, with limited independent analyses to guide healthcare organizations in platform selection. Lastly, skill shortages and organizational change management represent nontechnical barriers. The transition to cloud-native data warehouses demands expertise in cloud computing, data engineering, and healthcare informatics, which many institutions currently lack. Adequate training and strategic change management plans are necessary to realize the benefits of these technologies fully [4]. Addressing these challenges and gaps is crucial for maximizing the potential of cloud-based data warehouses in advancing healthcare analytics and patient care.

Future research in healthcare data warehousing should focus on developing automated and intelligent data integration methods to better handle diverse and complex datasets. Enhancing data quality through real-time monitoring and advanced cleansing techniques will im-prove the reliability of analytics. Privacy preserving approaches like federated learning and differential privacy are essential to enable secure, compliant data sharing across institutions. Additionally, optimizing performance and cost management tailored to healthcare workloads will support scalable solutions. More independent comparative studies of cloud platforms are needed to guide healthcare organizations in selecting the most suitable technologies. Workforce training and effective change management will be critical for successful cloud adoption. Finally, integrating emerging data types such as genomics and wearable devices will expand the scope of healthcare analytics and personalized medicine.

VI. CONCLUSION

Cloud-based data warehouses like Snowflake and Big Query offer powerful solutions for managing and analysing vast healthcare datasets, enabling improved patient care and operational efficiency. However, challenges related to data integration, quality, privacy, and cost optimization must be addressed to fully realize their potential. Standardization efforts, advanced automation, and privacy-preserving technologies are critical future steps. Furthermore, building skilled teams and conducting rigorous platform evaluations will support effective adoption in healthcare settings. By overcoming these obstacles, healthcare organizations can leverage modern data warehousing technologies to advance analytics, enhance decision-making, and ultimately improve health outcomes.

ACKNOWLEDGEMENT

The authors would like to express their sincere gratitude to Prof. Ambika .V, Assistant Professor of Computer Science and Engineering (Data Science), ATME College of Engineering, Mysuru, for her constant support, valuable guidance, and encouragement throughout the preparation of this review paper. Her insightful suggestions and constructive feedback greatly contributed to the improvement and completion of this work. Finally, the Authors also appreciate the support of ATME College of Engineering, Mysore

REFERENCE

[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.