Journal of Science Research International (JSRI)
ISSN: 2456-6365 (Online)

OPEN ACCESS

CLOUD-BASED DATA WAREHOUSING: CHALLENGES, COMPARATIVE ANALYSIS, AND ML-BASED OPTIMIZATION TECHNIQUES

Nisha Yadav* and Satvika Biswanath Guntha

Department of Information Technology,

Pillai College of Arts, Commerce & Science (Empowered Autonomous), Navi Mumbai, Maharashtra, India 410206

**Corresponding author E-mail: nishayadav@mes.ac.in*

Received: 23 January 2026	Revised: 25 February 2026	Accepted: 17 March 2026	Published: 31 March 2026

DOI: **https://doi.org/10.5281/zenodo.19571008**

Abstract:

Cloud-based data warehousing (CDW) enables organizations to store and analyze large-scale data using scalable, flexible, and cost-effective cloud infrastructure. Modern platforms such as Amazon Redshift, Google BigQuery, and Snowflake offer separated compute and storage, elastic scaling, and usage-based pricing models. Despite these advantages, enterprises often face challenges related to cost predictability, workload performance, elasticity management, and security compliance. This research presents a comparative analysis of major cloud data warehouse platforms based on architecture, performance, scalability, cost management, and security capabilities. Furthermore, a conceptual machine learning (ML)-based optimization framework is proposed to enhance performance and cost efficiency. The framework integrates predictive cost modeling, adaptive resource allocation, and automated query optimization. The study highlights how intelligent optimization techniques can support organizations in improving the efficiency of cloud data warehouse operations and making better platform selection decisions.

Keywords: *Cloud Data Warehousing, Amazon Redshift, Google BigQuery, Snowflake, Elastic Scaling, Machine Learning Optimization, Query Performance.*

1. Introduction

The rapid growth of digital technologies has resulted in a massive increase in data generation across industries. Organizations today collect large volumes of structured, semi-structured, and unstructured data from various sources such as online transactions, IoT devices, and social media platforms. Managing and analyzing this data efficiently has become essential for supporting strategic decision-making. Traditional on-premises data

warehouses have been widely used for enterprise analytics. However, these systems often struggle to meet modern data processing requirements due to limitations such as fixed compute resources, high infrastructure costs, and complex maintenance requirements. As a result, organizations are increasingly adopting cloud-based data warehousing solutions. Cloud data warehouse platforms such as Amazon Redshift, Google BigQuery, and Snowflake provide scalable and flexible environments for data storage and analytics. These platforms offer key advantages including elastic scalability, separation of storage and compute resources, and integration with advanced analytics tools and machine learning services.

Despite these benefits, organizations still encounter several operational challenges when deploying cloud data warehouse systems. These challenges include unpredictable operational costs, performance variations under concurrent workloads, and ensuring security compliance across distributed cloud environments.

Therefore, there is a need to systematically analyze these platforms and explore optimization techniques that improve their operational efficiency. This study aims to evaluate the characteristics of major cloud data warehouse platforms and propose a machine learning-based optimization framework that can enhance performance and cost management.

2. Literature review

Research on cloud-based data warehousing has grown significantly with the rapid expansion of cloud computing and large-scale data analytics. Early developments in distributed data processing architectures laid the foundation for modern cloud-based analytics platforms.

One of the important contributions in this field is the Dremel architecture, introduced by Melnik et al., which enables interactive analysis of massive datasets using a column-oriented storage model and distributed query execution techniques. The architectural principles of Dremel later influenced the development of modern analytics platforms such as Google BigQuery, which supports high-performance processing of large-scale analytical workloads [1].

Another foundational advancement in large-scale data processing is the MapReduce programming model, proposed by Dean and Ghemawat. MapReduce simplified the processing of large datasets by dividing tasks into smaller operations that can be executed in parallel across distributed computing clusters. This approach significantly improved the scalability and efficiency of big data processing systems and influenced the development of several modern distributed computing frameworks [2].

Fan and Bifet conducted research on big data analytics and examined the characteristics of large-scale data processing systems. Their work highlights key challenges associated with managing and analyzing massive datasets, including scalability, computational efficiency, and the ability to process diverse data formats [4]. These insights have contributed to the development of modern cloud-based data warehouse platforms capable of handling complex analytical workloads.

Abadi explored the opportunities and limitations of cloud-based data management systems, emphasizing the architectural challenges related to performance, scalability, and resource management in cloud environments [3]. This research highlights the need for efficient system design to support large-scale analytical processing in cloud infrastructures.

Recent studies have also focused on integrating machine learning techniques into database and data warehouse systems to improve operational efficiency. Machine learning models can analyze historical query logs, system performance metrics, and workload patterns to predict resource requirements and optimize query execution strategies. Such intelligent optimization methods can help organizations improve system performance while controlling operational costs. Although previous research provides valuable insights into cloud data warehouse architectures and performance characteristics, many studies focus either on system benchmarking or on optimization techniques independently. Limited research integrates comparative platform analysis with machine learning–based optimization frameworks, which highlights an important research gap addressed in this study.

3. Research gap

Although significant research has been conducted in the field of cloud-based data warehousing, several limitations remain in existing studies. Many prior works primarily focus on evaluating individual cloud platforms or analyzing specific aspects such as system performance, scalability, or cost management independently. However, these factors are rarely examined together within a comprehensive framework that considers both operational efficiency and cost optimization simultaneously. Another important limitation in current research is that most machine learning–based optimization approaches are designed for a single cloud platform environment. In practice, many organizations adopt multi-cloud or hybrid cloud strategies, where different cloud services are used for various workloads. As a result, optimization techniques limited to a specific platform may not be suitable for real-world enterprise environments. Several existing optimization models lack adaptive learning capabilities that can continuously adjust to changing workloads, evolving query patterns, and dynamic resource demands. Without such adaptive mechanisms, it becomes difficult for cloud data warehouse systems to maintain optimal performance and cost efficiency over time.

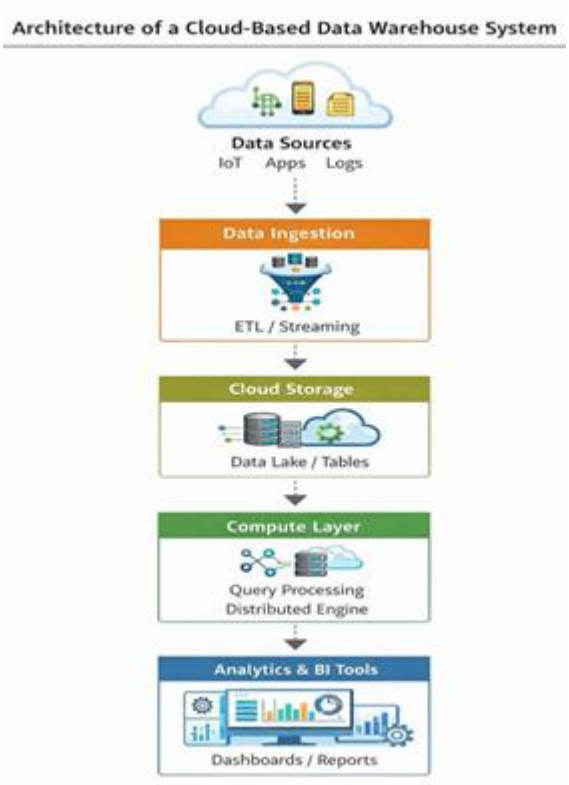
To address these limitations, this research combines a comparative analysis of leading cloud data warehouse platforms with the development of a conceptual machine learning–based optimization framework. The proposed approach aims to support cross-platform optimization, improve resource utilization, and enhance cost efficiency while adapting to changing workload patterns.

4. Objectives

The primary objectives of this research are:

i. To analyze the architectural and operational characteristics of leading cloud data warehouse platforms.
ii. To identify common challenges related to performance, scalability, cost management, and security compliance.
iii. To compare the capabilities of Amazon Redshift, Google BigQuery, and Snowflake.
iv. To propose a machine learning-based framework for optimizing cloud data warehouse performance.
v. To provide insights for organizations planning to adopt cloud-based data warehouse solutions

5. Methodology



This research adopts an analytical research approach based primarily on secondary data sources. Relevant information was gathered from peer-reviewed academic journals, conference publications, cloud service provider documentation, benchmarking reports, and technical whitepapers related to cloud data warehousing technologies. The collected information was examined using a comparative evaluation framework that considered multiple performance and operational factors. These evaluation criteria include system architecture, query processing performance, scalability mechanisms, cost management models, and security features offered by the selected cloud data warehouse platforms.

Through this comparative analysis, the study identifies the major strengths and limitations associated with each platform. The insights obtained from the analysis were then used to design a conceptual machine learning–based optimization framework aimed at improving performance efficiency and cost management in cloud data warehouse environments. This framework integrates predictive modeling and intelligent resource allocation strategies to address common operational challenges observed in existing cloud data warehouse systems.

6. Comparative analysis of cloud data warehouse platforms

Table 1: Comparison of Cloud Data Warehouse Platforms

Feature	Amazon Redshift	Google BigQuery	Snowflake
Architecture	Cluster-based MPP	Serverless architecture	Multi-cluster architecture
Storage & Compute	Partially separated	Fully separated	Fully separated
Scaling	Manual scaling	Automatic scaling	Independent scaling
Query Performance	Strong for batch workloads	Efficient for large analytics queries	Good for mixed workloads
Pricing Model	Instance based	Pay per query	Per second compute billing
Cloud Provider	AWS	Google Cloud	Multi-cloud

Amazon Redshift

Amazon Redshift uses a cluster-based massively parallel processing (MPP) architecture. It is well suited for large batch queries and integrates closely with other AWS services. However, it may require manual scaling and performance tuning.

Google BigQuery

Google BigQuery operates as a fully serverless data warehouse platform. It automatically scales computing resources and enables fast analysis of extremely large datasets using distributed processing.

Snowflake

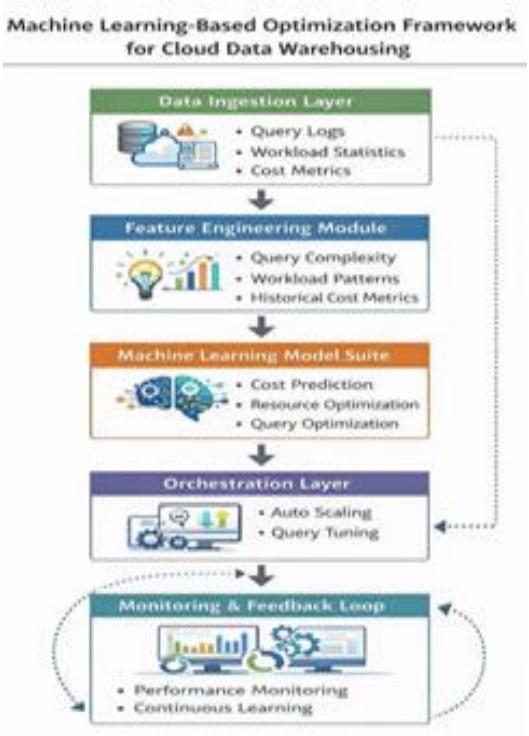
Snowflake separates storage and compute resources and supports multi-cluster architectures. This design allows independent scaling and improved performance for concurrent workloads.

Table 2: Performance and Cost Comparison

Platform	Avg Query Execution Time (sec)	Scaling Time (min)	Estimated Monthly Cost (USD)	Security Features
Amazon Redshift	12	8	4500	Encryption, IAM
Google BigQuery	9	2	3800	IAM, Data Encryption
Snowflake	10	3	4000	Secure Data Sharing, RBAC

7. Proposed ML-based optimization framework

To address the operational challenges associated with cloud data warehouse environments, this study proposes a conceptual machine learning–based optimization framework. The objective of the framework is to improve query performance, resource utilization, and cost efficiency by leveraging intelligent data-driven decision mechanisms.



The framework is composed of several interconnected components, each responsible for a specific stage of the optimization process.

Data ingestion layer

The data ingestion layer is responsible for collecting operational data generated by cloud data warehouse platforms. This includes query execution logs, system performance metrics, workload statistics, and cost-related information obtained from cloud monitoring services. The collected data serves as the foundation for subsequent analysis and model training.

Feature engineering module

The feature engineering module processes the raw data collected from the ingestion layer and transforms it into meaningful analytical features. These features may include query complexity indicators, workload distribution patterns, historical resource utilization, and cost metrics. Proper feature extraction improves the accuracy and effectiveness of the machine learning models used in the framework.

Machine learning model suite

This component consists of multiple machine learning models designed to support different optimization tasks. These models may include predictive algorithms for estimating operational costs, models for recommending optimal resource allocation strategies, and query optimization models that improve execution efficiency. By analyzing historical workload data, the models can identify patterns that help optimize system performance.

Orchestration layer

The orchestration layer acts as the decision-making component of the framework. It applies the recommendations generated by the machine learning models to automatically adjust system configurations. This may include scaling compute resources, optimizing query execution plans, or scheduling workloads more efficiently to improve performance and reduce operational costs.

Monitoring and feedback loop

The monitoring and feedback mechanism continuously evaluates system performance after optimization actions are applied. Updated operational data is collected and fed back into the machine learning models for retraining and refinement. This adaptive learning process allows the framework to respond dynamically to changes in workload patterns and evolving system conditions.

Conclusion

This study presented a comparative evaluation of three major cloud data warehouse platforms: Amazon Redshift, Google BigQuery, and Snowflake. The analysis examined key aspects such as system architecture, query performance, scalability mechanisms, cost models, and security capabilities. The findings indicate that each platform offers specific advantages depending on the analytical workload requirements, scalability demands, and cost considerations of an organization. For instance, serverless platforms provide strong support for large-scale analytical queries, while multi-cluster architectures enable better performance for mixed workloads and concurrent users. In addition to the comparative analysis, this research proposed a conceptual **machine learning–based optimization framework** aimed at improving the operational efficiency of cloud data warehouse systems. The framework integrates workload monitoring, feature engineering, predictive modeling, and automated orchestration to support intelligent resource allocation and query optimization. By leveraging machine learning techniques, the framework can help organizations enhance performance, reduce operational costs, and improve the overall efficiency of cloud-based data warehousing environments. Future research can extend this work by implementing and validating the proposed framework in real-world cloud environments. Experimental evaluation using real workload data could further assess its effectiveness in optimizing resource utilization, improving query performance, and achieving cost-efficient data warehouse operations.

References

1. Melnik, S., Gubarev, A., Long, J. J., Romer, G., Shivakumar, S., Tolton, M., & Vassilakis, T. (2010). Dremel: Interactive analysis of web-scale datasets. *Proceedings of the VLDB Endowment, 3*(1–2), 330–339.
2. Dean, J., & Ghemawat, S. (2008). MapReduce: Simplified data processing on large clusters. *Communications of the ACM, 51*(1), 107–113.
3. Abadi, D. J. (2009). Data management in the cloud: Limitations and opportunities. *IEEE Data Engineering Bulletin, 32*(1), 3–12.
4. Fan, W., & Bifet, A. (2013). Mining big data: Current status and forecast to the future. *ACM SIGKDD Explorations, 14*(2), 1–5.
5. Stonebraker, M., Abadi, D., DeWitt, D., Madden, S., Paulson, E., Pavlo, A., & Rasin, A. (2005). C-Store: A column-oriented DBMS. *Proceedings of the VLDB Conference*.
6. Armbrust, M., Fox, A., Griffith, R., Joseph, A. D., Katz, R., Konwinski, A., Lee, G., Patterson, D., Rabkin, A., Stoica, I., & Zaharia, M. (2010). A view of cloud computing. *Communications of the ACM, 53*(4), 50–58.

7. Zaharia, M., Chowdhury, M., Franklin, M., Shenker, S., & Stoica, I. (2010). Spark: Cluster computing with working sets. *Proceedings of the USENIX Conference on Hot Topics in Cloud Computing*.
8. Chaudhuri, S., & Narasayya, V. (2007). Self-tuning database systems: A decade of progress. *Proceedings of the VLDB Conference*.
9. Amazon Web Services. (2024). *Amazon Redshift documentation*. https://docs.aws.amazon.com/redshift
10. Google Cloud. (2024). *BigQuery documentation*. https://cloud.google.com/bigquery/docs
11. Snowflake Inc. (2024). *Snowflake data cloud documentation*. https://docs.snowflake.com
12. Elgamal, T., Sandur, S., & Das, S. (2021). Cost-aware query optimization for cloud data warehouses. *IEEE International Conference on Cloud Computing*.
13. Kraska, T., Beutel, A., Chi, E., Dean, J., & Polyzotis, N. (2018). The case for learned index structures. *Proceedings of the ACM SIGMOD International Conference on Management of Data*.