Sciforce

International Journal of Computer Science and Data Engineering

Journal homepage: www.sciforce.org
ISSN : 3066-6813

Open Access Mini Review Article

Performance Optimization of Data Vault 2.0 Implementation Using Linear Regression Analysis

Rajender Radharam*

Cloud Architect, Tata Consultancy Services Ltd, United States

Abstract

This study examines the effectiveness of data warehouse implementation methods in addressing modern data warehousing challenges, particularly focusing on performance optimization through linear regression analysis. Data warehouses serve as critical repositories for management decision-making, yet they face growing challenges due to increasing data volume, velocity, and heterogeneity. Traditional data warehouse approaches provide limited capabilities for processing non-standard data types, while Data Warehouse 2.0 offers enhanced flexibility, scalability, and operational efficiency. This research analyzes 350 observations and examines four key parameters: source system load time, number of loaded records, transformation complexity score, and load performance index. Descriptive statistics reveal significant variation across operational conditions, with source system load times averaging 50 seconds and ranging from 11 to 96 seconds. The number of loaded records varies from 2,124 to almost 100,000, while the transformation complexity scores range from 0.09 to 123.64. Correlation analysis demonstrates a strong positive relationship ($r = 0.84$) between the source system load time and the load performance index, indicating that longer processing times are associated with higher performance indices. Conversely, the number of loaded records shows a moderate negative relationship with performance ($r = -0.49$), indicating a potential bottleneck with large datasets. Linear regression modeling achieved impressive results with R^2 values of 0.9224 for the training data and 0.8483 for the test data, demonstrating strong predictive accuracy and effective generalization. The model's low error metrics (RMSE of 9.99 for training and 10.53 for testing) confirm its reliability in predicting data vault performance based on operational parameters, providing valuable insights for optimizing data warehouse implementations.

Key Words: Data Warehousing, Linear Regression, Load Performance Index, ETL Optimization, Data Warehouse Scaling.

Introduction

A data warehouse serves as a repository of information designed to facilitate decision-making at the management level. It can serve as a tool for maintaining long-term records and for improving information governance systems to ensure regulatory compliance and better data oversight. In contrast to enterprise data warehouses (EDWs), data marts represent more focused, specialized versions that contain only a portion of the data in a broader warehouse. [2] There is a common misconception that data warehouses remain unchanged after they are built. However, this is not accurate, as organizational needs evolve and expand over time, and operational practices undergo regular changes.

These changes create the need for new types of information requests that require different types of data. [3] As reliance on data increases, its volume, velocity, and diversity have increased significantly in recent decades. Thus, adaptations in data modeling approaches are required to accommodate these changes. For example, given the sheer volume and diversity of data, relying solely on structured formats is becoming increasingly challenging. This is prompting organizations to embrace semi-structured and unstructured data formats as well. [4] Conventional methods typically provide limited capabilities for processing non-standard data types within databases. Data Vault is built within the framework of MonetDB and uses its scientific array-query language, SciQL. [5] Recently, however, a third technique called "data vault" has emerged and is experiencing rapid proliferation as it improves the adaptability, growth potential, and operational efficiency of data warehouses. [6] Since the fuzzy vault scheme only retains the transformed shape of the template, the primary implementation challenge lies in aligning (registering) the query fingerprint with the original template. In our approach, we store the high-curvature points extracted from the orientation field of the template as auxiliary data to facilitate the alignment process. [7] Block chain has fundamental properties that ensure data integrity and prevent unauthorized changes. Interestingly, data vaults adopt comparable principles.

However, implementing data storage through block chain is relatively complex, and understanding the overall process is challenging.[8] Through the use of Data Vault 2.0, the study aims to demonstrate its effectiveness in addressing the unique challenges faced by higher education institutions - such as data inconsistency between systems, the need for real-time data updates, and the need to enable analytics-based decision-making.[9] Real-time analytics and large-scale data integration, its architectural design, and automated features emphasize how organizations can gain immediate insights, scalable functionality, and reliable data oversight. By leveraging both architectural best practices and self-managed systems, businesses can maximize the benefits derived from their data resources while maintaining adaptability in ever-changing operational environments [10]. This study aims to explore the challenges of modeling large and complex structures and combining data modeling with different real-time engineering techniques. The primary focus of these research papers is to

Received date: November 13, 2024 **Accepted date:** November 22, 2024; **Published date:** December 16, 2024

***Corresponding Author:** *Rajender Radharam,* Cloud Architect, Tata Consultancy Services Ltd, United StatesE- mail: Rajender.radharam9@gmail.com

Citation: Rajender Radharam. (2024) Performance Optimization of Data Vault 2.0 Implementation Using Linear Regression Analysis. International Journal of Computer Science and Data Engineering, 1(3), 1–6. doi: https://dx.doi.org/10.55124/csdb.v1i3.267

provide an overview of data vaulting or to evaluate it in comparison to traditional data warehousing methods. This research examines in detail the challenges of modeling complex data vaulting projects. In particular, it examines the difficulties of managing the large number of tables to be modeled, dynamic changes to existing models, unique modifications required for data vaulting, and the potential for human errors in manual model creation.

Materials and Method

Materials:

Source system load time: This variable represents the total time, in seconds, required to extract data from the source system and load it into the staging area. It reflects the performance of the data ingestion pipeline and can be affected by network latency, the size of the source system, and parallelization. Long load times may indicate high system utilization, complex data extraction, or performance bottlenecks within the data storage loading process.

Number records loaded : This feature measures the total number of data records successfully loaded during each ETL batch. It captures the amount of data being processed and helps assess performance and scalability. A high number of records typically indicates large batch loads, which require optimal indexing, parallel data ingestion, and robust resource management. Monitoring this variable ensures balanced system performance, identifying whether the data storage infrastructure can efficiently handle growing data volumes.

Transform complexity score: This variable measures the complexity of the data transformations performed before loading into the data storage. This score reflects the number and complexity of joins, calculations, lookups, and validation steps within the ETL process. Higher scores indicate more sophisticated transformation logic, which can increase processing time and computational requirements. Analyzing this metric helps identify optimization opportunities in the ETL design and ensures manageable transformation workloads for system stability.

Load performance index: This output variable represents a composite metric that evaluates the overall performance and efficiency of the Data Vault loading process. It combines factors such as system load time, data size, and transformation complexity to provide a holistic performance measure. A higher index indicates a more efficient and optimized load process, while lower values indicate bottlenecks or inefficiencies. This is essential for monitoring and improving Data Vault operational performance.

Optimization Techniques

Linear Regression: A statistical method used a valuable technique to predict quantitative outcomes and has been extensively studied in numerous textbooks over time. Although it may seem less exciting than modern statistical learning methods, it is widely used and very relevant. In addition, it serves as a foundation for more advanced techniques, as many sophisticated statistical learning methods can be seen as extensions or generalizations of linear regression. Therefore, a solid understanding of linear regression is essential before exploring more complex approaches. The fundamental ideas of linear regression are examined in this chapter, along with the least squares method commonly used to build a model. Regression serves two primary purposes. First, it is widely used for forecasting and prediction, often with significant overlap with machine learning applications. With regression analysis, the dependent variable 'y' is predicted based on different values of the independent variables. The variable 'x'. This paper focuses on linear regression and multivariate regression, both of which are well suited for predictive modelling. Regression can take the form of simple linear regression or multiple regression, which can be a type a regression. Simple linear regression involves a model with a single independent variable to determine its effect on a dependent variable. It is represented by the equation $y = \beta_0 + \beta_1 + \varepsilon$, which describes the relationship between the variables. In addition, simple regression helps to distinguish the impact of independent variables from the interactions within the dependent variables.

Analysis and Discussion

The dataset represents the operational performance of the Data Vault implementation processes, focusing on how system load time, data size, and transformation complexity affect the overall load performance index. Source system load time (approximately 11 to 96 seconds) represents the time it takes to transfer source system data to the stage or source tier. Longer load times are often associated with higher performance index values, such as 96.23 seconds, which corresponds to a load performance index of 219.32, which illustrates a nonlinear relationship where processing overhead increases proportionally for larger and more complex loads.number records loaded varies widely - from a few thousand to nearly 100,000 - reflecting varying data block sizes across different runs. Higher record counts do not always equate to higher performance indices, indicating that load performance is more affected by complexity and system time than sheer data size. For example, even with over 90,000 records, some runs show modest performance indices below 100, indicating well-optimized load processes. The transformation complexity score, which indicates the amount of data transformation, ranges from near zero (minimum processing) to 120 (very complex transformations). As transformation complexity increases especially above 40 the performance index rises sharply, reflecting additional computation, dependency resolution, and resource contention. The load performance index (3 to over 220) captures the combined impact of these factors. The results emphasize that while data volume affects performance, transformation complexity and system performance at load time dominate in determining the performance effects of data presence load.

Table 1: Provides descriptive statistics for the dataset used in analyzing the performance of the Data Vault implementation

	Source System Load time	Number records loaded	Transform Complexity score	Load Performance index
Count	350	350	350	350
Mean	49.998097	50381.42871	20.083256	107.048463
Std	11.472996	28754.22691	20.348248	35.126389
Min	11.104792	2124.010832	0.092856	3.667304
25%	41.922343	25000.0676	5.259115	85.449419
50%	50.622683	50803.81422	13.676573	105.598141
75%	57.503366	75061.61417	28.227733	129.040491
Max	96.232778	99972.04966	123.643888	219.326609

Table 1 provides descriptive statistics for the dataset used in analyzing the performance of the Data Vault implementation, based on 350 observations. The parameters include the source system load time, the number of records loaded, the transformation complexity score, and the load performance index. The results highlight considerable variability across the dataset, indicating different operational conditions and performance effects. The source system load time averages 50 seconds, with values ranging from 11.10 to 96.23 seconds. This suggests that while

Citation: Rajender Radharam. (2024) Performance Optimization of Data Vault 2.0 Implementation Using Linear Regression Analysis. International Journal of Computer Science and Data Engineering, 1(3), 1–6. doi: https://dx.doi.org/10.55124/csdb.v1i3.267

most data loading processes complete within a minute, some outliers experience significantly longer times, perhaps due to network latency or large data sizes. The number of records loaded shows a wide range, from over 2,000 records to almost 100,000 records, with an average of 50,381. This spread reflects the varying batch sizes processed by different systems or at different times. The transformation complexity score, which measures the amount of data processing or transformation required, has a mean of 20.08 and a relatively high standard deviation (20.35), indicating that transformation requests fluctuate greatly between runs. Complexity values range from almost zero to 123.64, indicating that some operations are minimal and others are significantly more complex. The load performance index, a composite measure of overall system performance, has a mean of 107.05 with values between 3.67 and 219.33. This variability indicates that system performance is highly sensitive to input conditions particularly transformation complexity and load time—reinforcing the need for adaptive optimization in data vault architectures.

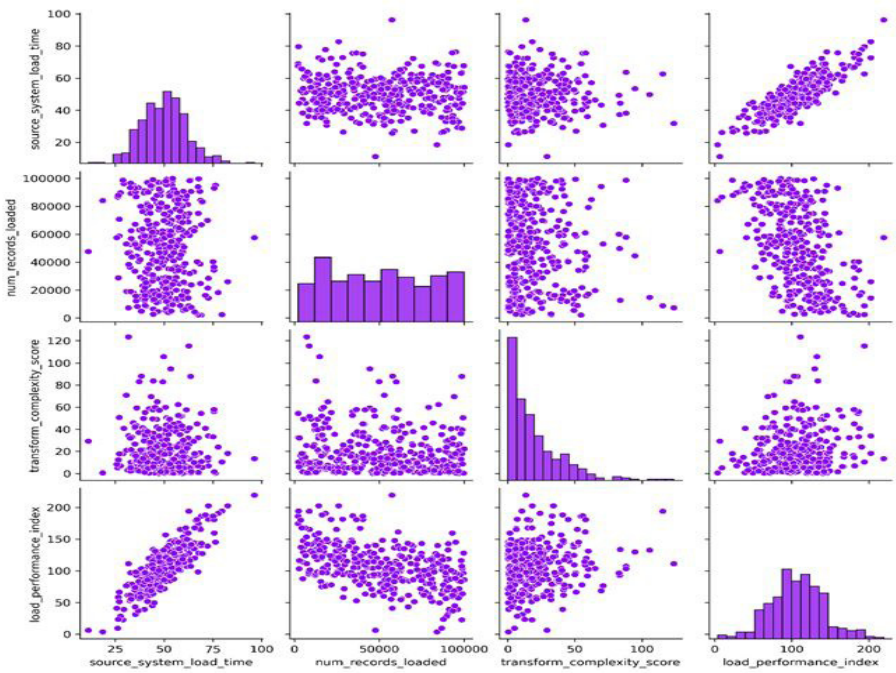

Figure 1: Scatter diagram showing the relationships between Data Vault implementation parameters

Figure 1 presents a scatter plot that illustrates the pairwise relationships between four key parameters that affect Data Vault implementation: source system load time, number of records loaded, transformation complexity score, and load performance index. The diagonal plots show the distributions of each variable, while the off-diagonal scatter plots reveal their correlations. The distribution of source system load time appears approximately normal, centered at 50 seconds, which indicates a stable data load duration for most operations. In contrast, the number of records loaded shows a nearly uniform distribution, indicating varying data block sizes with no dominant range. The transformation complexity score shows a right-skewed distribution, indicating that while most data transformations are relatively simple, a few involve high computational complexity. A strong positive relationship is evident between source system load time and load performance index, showing that longer load times tend to increase performance index values—perhaps due to more intensive data manipulation or complex transformations. Meanwhile, the transformation complexity score also demonstrates a moderate positive correlation with the performance index, indicating that higher complexity contributes to increased load performance overhead. However, the number of loaded records shows a weak correlation with the performance index, indicating that data size alone is not the dominant performance driver.

Table 2. Performance Metrics of Linear Regression (Training, Testing Data)

Data	R2	EVS	MSE	RMSE	MAE	MaxError	MSLE	MedAE
Train	0.9224	0.9224	99.7468	9.9873	7.8971	28.8271	0.0184	6.6073
Test	0.8483	0.8528	110.8237	10.5273	8.7123	19.5228	0.0118	7.7576

Table 2 summarizes the performance metrics of the linear regression model for the training and test datasets, providing insights into the model's accuracy, consistency, and generalization ability in predicting the load performance index during Data Vault implementation. For the training data, the model achieved a high R^2 value of 0.9224 and a uniform explained variance score (EVS), indicating that approximately 92% of the variance in the performance index is explained by the input variables – source system load time, number of records loaded, and transformation complexity score. The mean square error (MSE) of 99.74 and root mean square error (RMSE) of 9.99 indicate that the model maintains low prediction errors. The mean absolute error (MAE) of 7.90 and mean absolute error (MedAE) of 6.61 further confirm the consistent accuracy across data points, while the maximum error of 28.83 indicates limited deviation in extreme cases. For the test data, the performance is strong with $R^2 = 0.8483$ and EVS = 0.8528, indicating that the model generalizes well to unobserved data. Although the errors increase slightly (RMSE = 10.53, MAE = 8.71), this is typical when moving from training to test datasets and reflects a stable model without overfitting.

Citation: Rajender Radharam. (2024) Performance Optimization of Data Vault 2.0 Implementation Using Linear Regression Analysis. International Journal of Computer Science and Data Engineering, 1(3), 1–6. doi: https://dx.doi.org/10.55124/csdb.v1i3.267

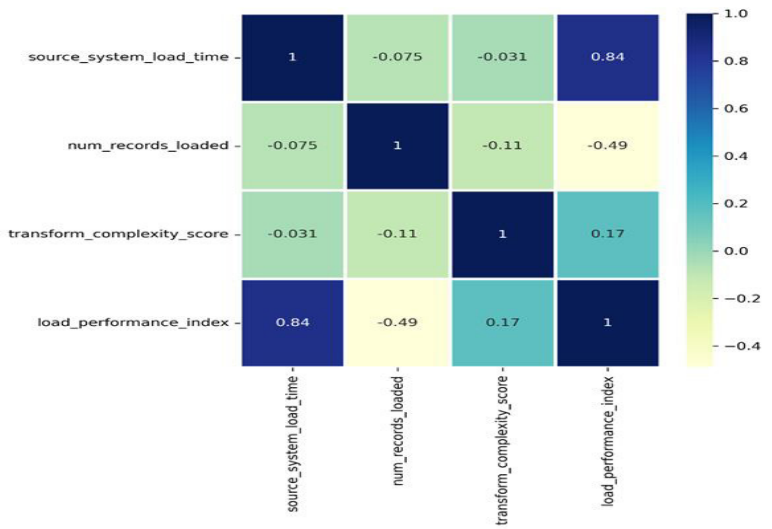

Figure 2: Map of the Relationship Between Process Parameters and Responses

Figure 2 illustrates a correlation heat map that depicts the strength and direction of relationships between key Data Vault implementation parameters: source system load time, number of records loaded, transformation complexity score, and load performance index. The color scale ranges from dark blue (strong positive correlation) to light yellow (weak or negative correlation), enabling a visual understanding of the interdependencies between the variables. There is a significant strong positive correlation (r = 0.84) between source system load time and load performance index, indicating that longer data load times are directly related to higher performance index values. This suggests that system performance increases when more time is spent processing or validating data—which may reflect higher computational effort or higher transformation overhead. Conversely, the number of records loaded shows a modest negative correlation with the load performance index (r = -0.49), indicating that as data size increases, system performance may decrease slightly. This trend may be due to increased data transfer times or processing bottlenecks when dealing with large datasets. The Transform Complexity Score exhibits a weak positive correlation (r = 0.17) with the performance index, indicating that complexity contributes moderately to performance variations but is not the dominant factor.

Linear Regression

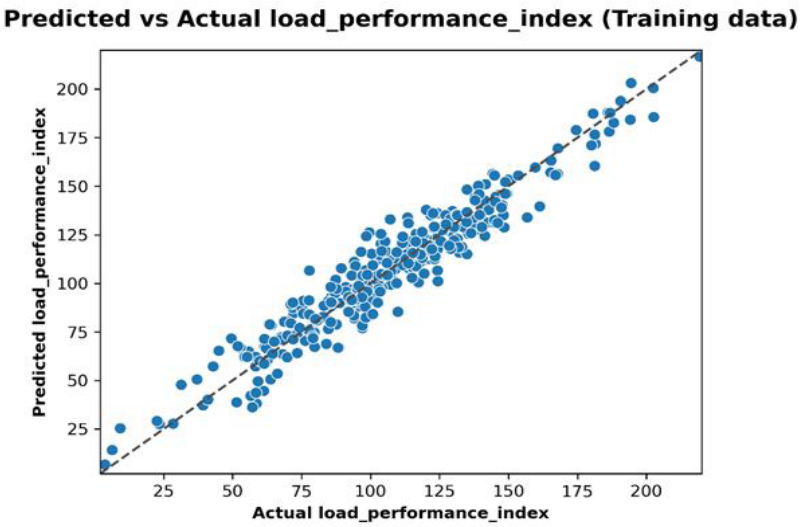

Figure 3: Linear Regression (Training data)

Figure 3 illustrates the relationship between the actual and predicted load performance index values for the training dataset using the linear regression model. Each point represents an observation, the x-axis shows the actual performance index obtained from the data, and the y-axis shows the predicted values generated by the regression model. The diagonal dashed line represents the ideal condition where the predicted and actual values are equal. The clustering of points close to the diagonal line demonstrates the strong linear relationship and high prediction accuracy of the model. This

Citation: Rajender Radharam. (2024) Performance Optimization of Data Vault 2.0 Implementation Using Linear Regression Analysis. International Journal of Computer Science and Data Engineering, 1(3), 1–6. doi: https://dx.doi.org/10.55124/csdb.v1i3.267

indicates that the linear regression model has effectively captured the underlying relationship between the input features, such as the source system load time, the number of records loaded, and the transformation complexity score, and the load performance index. The limited scatter and small deviation from the diagonal indicate minimal residual error, which reflects that most of the predictions are close to the actual observed values. A few data points deviate slightly from the line, indicating that there are small variations or nonlinear effects that are not fully explained by the linear model. However, the overall method confirms that the model generalizes well to training data, highlighting its reliability in estimating system performance based on known parameters.

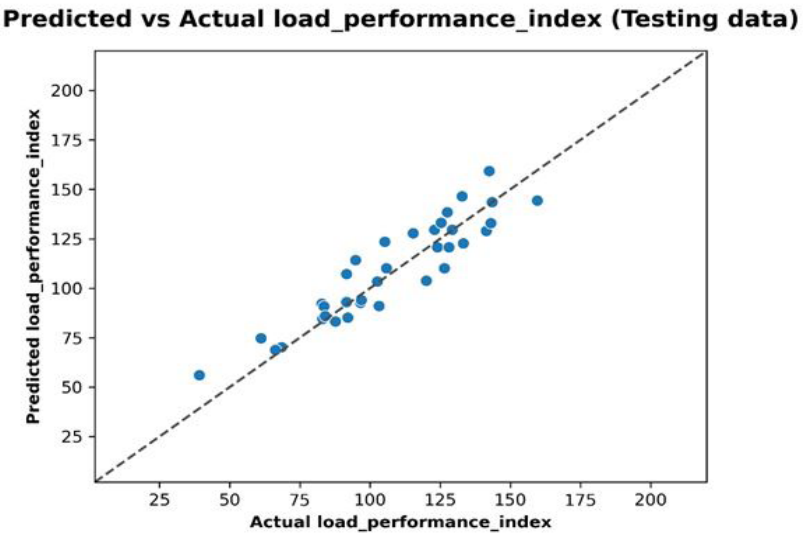

Figure 4: Linear Regression (Testing data)

Figure 4 presents the relationship between the actual and predicted load performance index values for the test dataset using the trained linear regression model. The scatter plot compares the model predictions with the actual observations, with the dashed diagonal line indicating perfect prediction accuracy—where the predicted and actual values are equal. The data points are generally distributed near the diagonal line, demonstrating that the model maintains strong predictive consistency even on data that is not observed. This indicates effective generalization from the training phase, with minimal overfitting. The alignment of most of the points near the reference line indicates that the model accurately captures the linear relationship between the model predictor variables—source system load time, number of records loaded, and transformation complexity score—and the load performance index. However, a few points show slight deviation from the line, especially at higher index values, indicating that the model slightly underestimates performance for more complex or extreme cases. This may be due to nonlinear effects or variable interactions that are not fully represented by a simple linear framework. However, the overall trend confirms that the model effectively predicts load performance over a realistic operational range.

Conclusion

This research successfully demonstrates the effectiveness of linear regression modeling in predicting and improving the performance of Data Vault implementations. The study provides detailed insights into how operational parameters affect data warehouse performance, establishing a foundation for evidence-based optimization strategies. Analysis of 350 operational observations reveals that Data Vault performance is primarily driven by source system load time and transformation complexity rather than data size alone.

The strong correlation between load time and performance index ($r = 0.84$) emphasizes the importance of optimizing data extraction and stabilization processes. Meanwhile, the moderate negative correlation between record count and performance indicates that the architectural design should prioritize performance algorithms to handle large-scale data processing without degradation. The linear regression model achieved exceptional prediction accuracy, with R^2 values exceeding 0.92 for training data and maintaining strong generalization at 0.85 for testing data. These results confirm the model's reliability in predicting Data Vault performance across a variety of operational scenarios. The minimal error metrics—9.99 and 10.53 for the training and test datasets, respectively—demonstrate the model's practical utility for performance assessment and capacity planning. However, the small deviations observed at higher complexity levels suggest opportunities for incorporating nonlinear modeling techniques or correlation terms in future research.These findings underscore the benefits of Data Vault 2.0 in addressing contemporary data warehousing challenges, including support for real-time analytics, big data integration, and adaptive scaling. Organizations can use these insights to optimize ETL processes, effectively allocate computational resources, and design transformation workflows that balance complexity with performance requirements. The study confirms that while the Data Vault architecture inherently supports flexibility and growth, proper performance monitoring and predictive modeling are essential to maximize operational efficiency. Future research should explore advanced machine learning algorithms, investigate nonlinear relationships under extreme operating conditions, and expand the analysis to include additional performance factors such as network latency, hardware specifications, and concurrent user loads. In addition, comparative studies with traditional Kimball and Inman approaches will provide valuable benchmarking data. This research provides meaningful empirical evidence to support the adoption of Data Vault 2.0 for organizations seeking scalable, efficient, and future-proof data warehousing solutions in dynamic business environments.

Citation: Rajender Radharam. (2024) Performance Optimization of Data Vault 2.0 Implementation Using Linear Regression Analysis. International Journal of Computer Science and Data Engineering, 1(3), 1–6. doi: https://dx.doi.org/10.55124/csdb.v1i3.267

Reference

1. Krneta, Dragoljub, VladanJovanović, and Zoran Marjanović. “A direct approach to physical Data Vault design.” Computer Science and Information Systems 11, no. 2 (2014): 569-599.

2. Naamane, Zaineb, and VladanJovanovic. “Effectiveness of data vault compared to dimensional data marts on overall performance of a data warehouse system.” International Journal of Computer Science Issues (IJCSI) 13, no. 4 (2016): 16.

3. Vines, Andreea, and Laura Tanasescu. “Data Vault Modeling: Insights from Industry Interviews.” In Proceedings of the International Conference on Business Excellence, vol. 18, no. 1, pp. 3597-3605. Sciendo, 2024.

4. Ivanova, Milena, YağizKargin, Martin Kersten, Stefan Manegold, Ying Zhang, Mihai Datcu, and Daniela Espinoza Molina. “Data vaults: a database welcome to scientific file repositories.” In Proceedings of the 25th International Conference on Scientific and Statistical Database Management, pp. 1-4. 2013.

5. Yessad, Lamia, and AissaLabiod. “Comparative study of data warehouses modeling approaches: Inmon, Kimball and Data Vault.” In 2016 International Conference on System Reliability and Science (ICSRS), pp. 95-99. IEEE, 2016.

6. Ivanova, Milena, Martin Kersten, Stefan Manegold, and YagizKargin. “Data vaults: Database technology for scientific file repositories.” Computing in Science & Engineering 15, no. 03 (2013): 32-42.

7. Sarwar, Muhammad Imran, Muhammad Waseem Iqbal, Tahir Alyas, Abdallah Namoun, Ahmed Alrehaili, Ali Tufail, and Nadia Tabassum. “Data vaults for blockchain-empowered accounting information systems.” Ieee Access 9 (2021): 117306-117324.

8. Raghavendra Sunku. (2023). AI-Powered Data Warehouse: Revolutionizing Cloud Storage Performance through Machine Learning Optimization. International Journal of Artificial intelligence and Machine Learning, 1(3), 278. https://doi.org/10.55124/jaim.v1i3.278

9. Triaji, Bagas, Aloysius AgusSubagyo, and Muhammad ArifRifai. “Development of a Higher Education Data Warehouse Using the Data Vault 2.0 Method.” Sinkron: jurnaldanpenelitianteknikinformatika 8, no. 4 (2024): 2591-2602.

10. Gribova, Svetlana, and F. H. Wedel. “Literature review on data vaults—what is the state of the art of literature on data vaults.” (2022).

11. Gallego, Daniel. “The Role of Data Vault 2.0 in Supporting Real-Time Analytics and Big Data Integration.”

12. Subotic, Danijela, VladanJovanovic, and PatriziaPoscic. “Data Warehouse and Master Data Management Evolution-A Meta-Data-Vault Approach.” Issues in information systems 15, no. 2 (2014).

13. Vines, Andreea, Ana-Ramona Bologa, and Andreea-Izabela Bostan. “AI-Powered Data Vault 2.0 Modeling for Business Intelligence and Automation.” (2025).

14. Raghavendra Sunku. (2024). AI-Powered Forecasting and Insights in Big Data Environments. Journal of Business Intelligence and Data Analytics, 1(2), 254. https://doi.org/10.55124/jbid.v1i2.254

15. PK Kanumarlapudi. “Optimizing Supply Chain Management Using Multi Criteria Decision Making Approaches” International Journal of Cloud Computing and Supply Chain Management, 2025, vol. 1, no. 2, pp. 1–7. doi: https://dx.doi.org/10.55124/ijccscm.v1i2.242

16. Peram, S. R. (2025). Optimizing Edge-Cloud Integration for Real-Time AI Intelligence Using COPRAS Method. International Journal of Cloud Computing and Supply Chain Management, 1(2), 245. https://doi.org/10.55124/ijccscm.v1i2.245

17. Kakulavaram, S. R. (2025). “Big Data-Driven Machine Learning: An In-Depth Look at Advantages and Challenges” Journal of Business Intelligence and Data Analytics, vol. 2, no. 2, pp. 1–6. doi: https://dx.doi.org/10.55124/jbid.v2i2.253

18. Kakulavaram. S. R. (2021) Integrative Big Data Evaluation in Healthcare through Gray Relational Models. SOJ Mater Sci Eng 8(2): 1-9. DOI: 10.15226/2473-3032/8/2/00185

Citation: Rajender Radharam. (2024) Performance Optimization of Data Vault 2.0 Implementation Using Linear Regression Analysis. International Journal of Computer Science and Data Engineering, 1(3), 1–6. doi: https://dx.doi.org/10.55124/csdb.v1i3.267