

Business Intelligence for Multidimensional Retail Sales Analysis Using a Data Warehouse

Niko Lugas Prambudi¹, Willy Agung Ramdani², Hendra Marcos^{*3}

^{1,2}Department of Information Systems, Faculty of Computer Science, Universitas Amikom Purwokerto, Indonesia

³Department of Informatics, Faculty of Computer Science, Universitas Amikom Purwokerto, Indonesia

e-mail: nikolugasprambudi2@gmail.com, willyagrm1112@gmail.com, hendra.marcos@amikompurwokerto.ac.id

Article History	Abstract
Received : 06 January 2026 Revised : 22 February 2026 Accepted : 02 March 2026 Published : 28 March 2026 Keywords: <i>Business Intelligence; Data Warehouse; ETL; Retail Analytics; Star Schema</i>	<i>Retail transaction data must be transformed into reliable analytical information to support data-driven decision-making. This study develops a Business Intelligence (BI) framework based on a star-schema data warehouse for multidimensional retail sales analysis across customer, product, and time dimensions. A public Kaggle dataset containing 1,000 synthetic transactions from 2023 and nine attributes was processed through an Extract, Transform, and Load (ETL) pipeline implemented in Pentaho Data Integration and loaded into MySQL. Data validity was assessed using completeness, arithmetic-consistency, and referential-integrity criteria. The pipeline loaded all 1,000 transactions, with no missing required values, no discrepancies between Total Amount and Quantity × Price per Unit, and no unmatched foreign keys. The analysis produced a total sales value of 456,000 monetary units and 2,514 units sold. Electronics generated the highest sales value (156,905), whereas Clothing had the highest volume (894 units). Monthly sales peaked in May (53,150) and were lowest in September (23,620). Female and male customers contributed 51.06% and 48.94%, respectively. The framework demonstrates a traceable approach to multidimensional retail analytics, although generalization is limited by the synthetic dataset.</i>

^{*)}Corresponding author: hendra.marcos@amikompurwokerto.ac.id

1. INTRODUCTION

Retail analytics treats transaction data as a fundamental source of information for understanding demand, customer characteristics, product performance, and sales outcomes. The literature indicates that retail analytics has evolved from descriptive reporting toward more integrated use of data to support operational and strategic decision-making. In this context, decision quality is determined not only by data volume but also by an organization's ability to structure, integrate, and present data in an analytically usable form [1]. Empirical evidence from online retailers further shows that descriptive analytics creates value when analytical dashboards are actively adopted and used in decision processes [12].

Data Warehousing and Business Intelligence (BI) provide an architecture that separates analytical workloads from daily transaction-processing systems. Implementations in retail organizations have demonstrated that data warehouse models facilitate sales analysis across products and time periods, while BI dashboards enable sales trends to be presented in a form that is easier for decision-makers to interpret [2], [3]. Comparable retail implementations also show that multidimensional warehouse structures and dashboard visualization can support faster exploration of sales performance and operational patterns [13], [14].

The reliability of information generated from a data warehouse depends heavily on the Extract, Transform, and Load (ETL) process. ETL extracts source data, performs cleansing and transformation, and loads the processed data into a consistent analytical structure. Research on ETL maturity identifies this process as a central component of data warehouse quality. In sales analytics, dimensional modeling based on a star schema and ETL implementation using Pentaho have also been applied to construct fact and dimension tables that support more efficient data aggregation [4]-[6]. Recent retail data-warehouse work further demonstrates the use of star-schema structures to preserve data integrity and support multidimensional analysis [15].

This study uses the Retail Sales Dataset published on Kaggle. The dataset represents a fictitious retail environment and contains transaction identifiers, dates, customer identifiers, gender, age, product categories, quantities, unit prices, and total transaction amounts. The same dataset has also been used in retail analytics studies to explore sales patterns and customer segmentation, making it suitable as an instructional dataset for designing an analytical pipeline without exposing actual customer data [7], [8].

This study is deliberately limited to analytical questions that can be directly addressed using the attributes available in the dataset; variables such as promotions, customer loyalty, and profit margin are excluded because they are not represented in the source data. Accordingly, the study addresses three research questions: (RQ1) To what extent does the proposed star-schema and ETL pipeline preserve transaction completeness, arithmetic consistency, and referential integrity? (RQ2) What temporal, product-category, and customer patterns are revealed when the transformed transaction data are analyzed multidimensionally? (RQ3) What decision-support insights can be derived from the integrated dashboard while remaining within the descriptive limits of the synthetic dataset? The study contributes a structured and internally consistent star-schema design, a documented ETL workflow, and an analytical dashboard in which data-driven findings are clearly distinguished from managerial implications.

2. METHOD

2.1 Research Design and Data Source

This study adopts a BI systems engineering approach with descriptive evaluation. The procedure consists of identifying the source-data structure, designing the dimensional model, implementing ETL, validating loaded data, performing analytical aggregations, and producing visualizations. The research workflow is presented in Figure 1. This approach was selected because the primary objective is not to develop a predictive model, but to establish an analytical data structure that can consistently support sales-related queries.

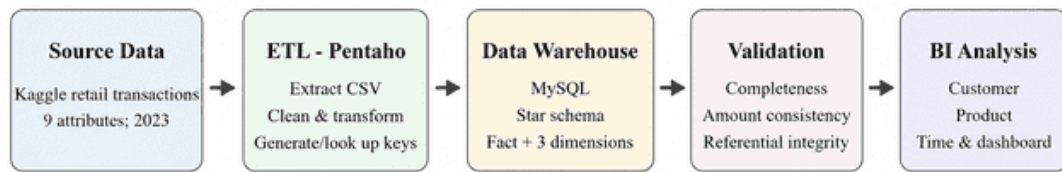


Figure 1. Research and Business Intelligence workflow

The data source is the synthetic/fictitious Retail Sales Dataset available on Kaggle. It contains 1,000 transaction records from 2023 and nine principal attributes. Because the source does not specify a currency, Price per Unit and Total Amount are reported as monetary units rather than being converted into or assumed to represent a particular currency. The variables used in the analysis are summarized in Table 1 [7].

Table 1. Source dataset attributes

Attribute	Role in analysis
Transaction ID	Natural transaction identifier
Date	Transaction date; source for time dimension
Customer ID	Customer identifier
Gender	Customer demographic attribute
Age	Customer demographic attribute
Product Category	Product grouping: Beauty, Clothing, Electronics
Quantity	Number of units in a transaction
Price per Unit	Transaction-level unit price
Total Amount	Transaction value (Quantity $\times$ Price per Unit)

2.2 Dimensional Modeling

The grain of the fact table is defined as one row for each source transaction. Fact_Sales stores dimension keys and three measures: Quantity, Price_per_Unit, and Total_Amount. Dim_Customer contains Customer_ID, Gender, and Age; Dim_Product contains Product_Category; and Dim_Date contains calendar attributes. Price_per_Unit is retained in the fact table rather than Dim_Product because, in the dataset, the attribute is transaction-specific and may vary across transactions. This design follows dimensional-modeling

principles in which numerical measures are centralized in the fact table, while analytical context is represented by dimension tables [5], [6]. The same naming convention—Sales_Key, Date_Key, Customer_Key, and Product_Key—is used consistently in Figure 2, Table 2, and the ETL/database implementation.

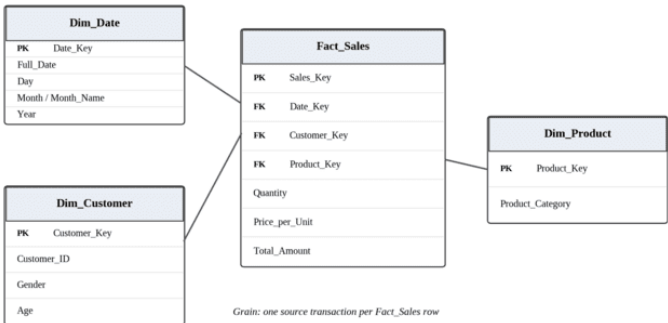


Figure 2. Proposed retail sales star schema
Table 2. Data Warehouse table design

Table	Key attributes / measures	Purpose
Fact_Sales	Sales_Key; Date_Key; Customer_Key; Product_Key; Quantity; Price_per_Unit; Total_Amount	Stores transaction-level measures
Dim_Customer	Customer_Key; Customer_ID; Gender; Age	Customer demographic analysis
Dim_Product	Product_Key; Product_Category	Product-category analysis
Dim_Date	Date_Key; Full_Date; Day; Month; Month Name; Year	Daily and monthly trend analysis

2.3 ETL Implementation and Validation

ETL was implemented using Pentaho Data Integration (Spoon), and the transformed data were loaded into MySQL. During extraction, the CSV file was read as the source dataset. The transformation stage included data-type standardization, date parsing, missing-value checks, duplicate checks for relevant attributes, generation of calendar attributes, and generation or lookup of surrogate keys. Stream Lookup was used during fact-table loading to replace business values from the source data with the corresponding dimension-table keys. Similar ETL and star-schema practices have been reported in sales data warehouse and e-commerce pipeline implementations [5], [6], [9].

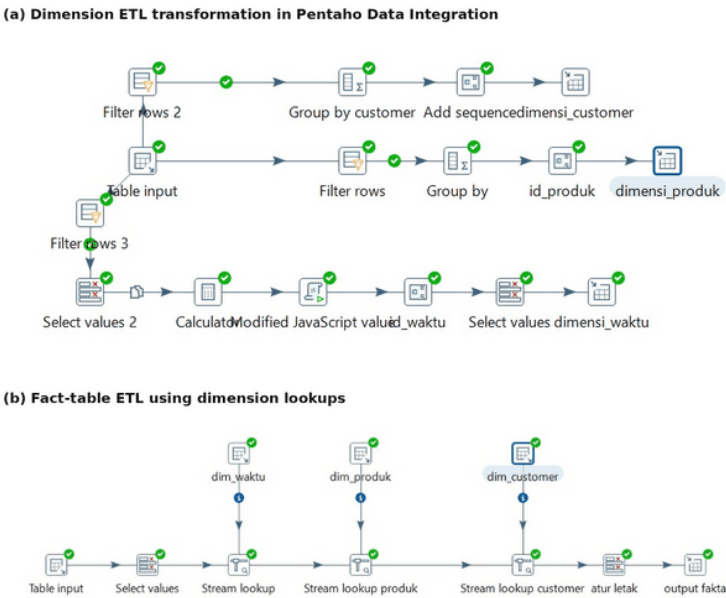


Figure 3. ETL implementation in Pentaho Data Integration

Validation criteria were defined before analysis. Transformed data were considered valid only when (1) completeness reached 100% for all required fields, with no null values in records used for analysis; (2) arithmetic consistency reached 100%, with zero rows for which Total_Amount differed from $\text{Quantity} \times \text{Price_per_Unit}$; and (3) referential integrity reached 100%, with zero orphan Date_Key , Customer_Key , or Product_Key values in Fact_Sales . Any record failing one of these criteria would be corrected or excluded before dashboard aggregation. These criteria provide explicit acceptance thresholds for the ETL output, while the observed validation results are reported in Section 3.1 [4].

2.4 Analytical Measures

The analysis employed descriptive measures that can be derived directly from the fact table, including total sales value, total units sold, monthly sales value, units and sales value by product category, and sales-value contribution by gender. Age was summarized using the observed minimum, first quartile, median, mean, third quartile, and maximum values rather than undefined ordinal labels. Interpretation was restricted to observable patterns; causal explanations involving promotions, seasonality, loyalty, or profit margin were not inferred because these variables are absent from the dataset.

3. RESULTS AND DISCUSSION

3.1 Data Warehouse and Data Quality Results

The implementation produced one fact table that preserves transaction-level grain and three dimension tables. The predefined validation criteria were satisfied in all three checks. Completeness validation found no missing required values across the 1,000 transactions (100% complete). Arithmetic-consistency validation identified zero discrepancies between Total_Amount and $\text{Quantity} \times \text{Price_per_Unit}$ (100% consistent). Referential-integrity validation found no orphan Date_Key , Customer_Key , or Product_Key values; all 1,000 fact rows were linked to valid records in the corresponding dimensions. These results indicate that the analytical measures used in the dashboard were derived from a complete, arithmetically consistent, and referentially connected warehouse.

From a design perspective, the star schema simplifies aggregation queries because the fact table is linked to each dimension through surrogate keys. This result is consistent with previous sales data warehouse implementations, which emphasize the analytical simplicity of cross-dimensional queries after ETL and dimensional modeling have been completed [5], [6], [15].

3.2 Overall and Temporal Sales Performance

Aggregating all transactions yielded a total sales value of 456,000 monetary units and 2,514 units sold. Across the time dimension, the highest monthly sales value was recorded in May at 53,150, whereas the lowest occurred in September at 23,620. October reached 46,580 and December 44,690, indicating substantial month-to-month variation in sales value. These differences may serve as preliminary signals for inventory planning and sales monitoring; however, they cannot be directly attributed to promotions or holidays because the dataset does not include promotion, event-calendar, or discount variables.

3.3 Product Category Performance

Two indicators provide complementary perspectives on product performance. By volume, Clothing ranked first with 894 units, followed by Electronics with 849 units and Beauty with 771 units. By sales value, however, Electronics ranked first at 156,905 (34.41%), followed by Clothing at 155,580 (34.12%) and Beauty at 143,515 (31.47%). The difference between the rankings by units and sales value demonstrates that the category with the highest sales volume does not necessarily generate the highest sales value. This interpretation is more appropriate than referring to profit margin, because the dataset does not contain cost-of-goods or profit information.

3.4 Customer Characteristics

Female customers contributed 232,840 monetary units, or 51.06% of total sales value, whereas male customers contributed 223,160, or 48.94%. The 2.12-percentage-point difference indicates a relatively balanced contribution and does not provide sufficient evidence to infer materially different purchasing behavior by gender. Customer ages ranged from 18 to 64 years, with a mean of 41.39 years and a median of 42 years; the first and third quartiles were 29 and 53 years, respectively. These statistics provide an interpretable demographic summary using actual age values. More advanced segmentation using RFM or clustering could be considered if sufficiently detailed customer-frequency and purchase-history data become available [10], [11].

Table 3. Main analytical and demographic results

Indicator	Value	Interpretation
Transactions	1,000	One fact row per source transaction
Total units	2,514	Overall product volume
Total sales value	456,000	Sum of Total_Amount; monetary unit unspecified
Highest month	May: 53,150	Highest monthly sales value
Lowest month	September: 23,620	Lowest monthly sales value
Highest category by units	Clothing: 894 units	Volume leader
Highest category by value	Electronics: 156,905	34.41% of total sales value
Gender contribution	Female 51.06%; Male 48.94%	Relatively balanced
Customer age summary	18–64 years; mean 41.39; median 42	Observed range and central tendency

3.5 Dashboard Interpretation and Managerial Use



Figure 4. Retail sales performance dashboard

The dashboard in Figure 4 integrates monthly sales trends, product-category sales value, gender contribution, customer-age summary statistics, and overall sales indicators. For business users, combining sales value with transaction and volume indicators helps distinguish periods and categories that warrant closer monitoring. The customer-age panel displays actual age statistics (18–64 years, mean 41.39, median 42) rather than undefined ordinal labels. Customer information can provide an initial basis for segmentation; however, marketing decisions should not be based solely on gender or age without additional validation.

The findings also demonstrate the role of the data warehouse as an integration layer preceding visualization. Other sales BI studies similarly report that dashboards become more useful when transaction data have first been organized through a warehouse structure and a documented ETL process. Accordingly, the principal value of the implementation lies not merely in the final visualization, but in maintaining consistent definitions of measures and dimensions across all analyses [2], [3], [9], [12]–[15].

The limitations of the study should be explicitly acknowledged. The dataset is synthetic, covers only one year and three product categories, and does not include location, payment method, product cost, profit

margin, discount, promotion, or individual product-identity attributes. The numerical findings therefore should not be generalized as representative of actual retail-industry behavior. Future research may use larger real-world transaction datasets, incorporate location and sales-channel dimensions, and extend the framework toward customer segmentation or predictive analytics once data quality and coverage are sufficient. Research using comparable retail datasets suggests that extensions to predictive modeling and segmentation are feasible after appropriate data exploration and preparation [8].

4. CONCLUSION

This study transformed retail transaction data into a star-schema data warehouse comprising one Fact_Sales table and three dimensions, with ETL implemented using Pentaho Data Integration and MySQL. The resulting structure supports consistent sales analysis across customer, product, and time dimensions. All predefined ETL validation criteria were satisfied: required fields were complete, no arithmetic mismatches were found between Total_Amount and Quantity $\times$ Price_per_Unit, and no orphan dimension keys were detected. From 1,000 transactions, the analysis identified a total sales value of 456,000 monetary units and 2,514 units sold; Electronics recorded the highest sales value at 156,905, Clothing recorded the highest volume at 894 units, the highest monthly sales value occurred in May at 53,150, and sales-value contributions by gender were relatively balanced. The main strength of the approach is a simple, measurable, and traceable analytical structure from source data to dashboard. Its primary limitation is the use of a synthetic dataset with restricted business attributes, preventing causal claims and market-level generalization. Future work should incorporate larger real-world datasets, additional dimensions such as location, sales channel, payment method, promotion, and cost, and more advanced segmentation and predictive techniques after the data-quality pipeline has been validated.

BIBLIOGRAPHY

- [1] R. P. Roodekerk, N. DeHoratius, and A. Musalem, "The past, present, and future of retail analytics: Insights from a survey of academic research and interviews with practitioners," *Production and Operations Management*, vol. 31, no. 10, pp. 3727-3748, 2022, doi: 10.1111/poms.13811.
- [2] I. P. S. Handika and P. P. Santika, "Perancangan datawarehouse dan teknologi business intelligence untuk analisa penjualan pada perusahaan retail PT. ABC," *RABIT: Jurnal Teknologi dan Sistem Informasi Univrab*, vol. 5, no. 2, pp. 76-85, 2020, doi: 10.36341/rabit.v5i2.1309.
- [3] S. Anardani, M. N. L. Azis, and M. Y. Asyhari, "The Implementation of Business Intelligence to Analyze Sales Trends in the Indofishing Online Store Using Power BI," *Brilliance: Research of Artificial Intelligence*, vol. 3, no. 2, pp. 300-305, 2023, doi: 10.47709/brilliance.v3i2.3232.
- [4] M. Khan, I. Ali, S. Khurram, S. Naseer, S. Ahmad, A. T. Soliman, A. A. Gardezi, M. Shafiq, and J.-G. Choi, "ETL Maturity Model for Data Warehouse Systems: A CMMI Compliant Framework," *Computers, Materials & Continua*, vol. 74, no. 2, pp. 3849-3863, 2023, doi: 10.32604/cmc.2023.027387.
- [5] D. Tan, J. Wiratama, and S. F. Wijaya, "Sales Analysis on Garment Industry with Datawarehouse and ETL Implementation on Star Schema," *The Indonesian Journal of Computer Science*, vol. 13, no. 1, 2024, doi: 10.33022/ijcs.v13i1.3770.
- [6] A. Yang, J. Wiratama, and S. F. Wijaya, "Empowering Data Transformation: Transforming Raw Data into A Strategic Planning for E-Commerce Success," *Journal of Information Systems and Informatics*, vol. 6, no. 1, pp. 339-348, 2024, doi: 10.51519/journalisi.v6i1.665.
- [7] M. Talib, "Retail Sales Dataset," Kaggle. [Online]. Available: <https://www.kaggle.com/datasets/mohammadtalib786/retail-sales-dataset>. [Accessed: Aug. 18, 2026].
- [8] J. Tang, "Unlocking Retail Insights: Predictive Modeling and Customer Segmentation Through Data Analytics," *Journal of Theoretical and Applied Electronic Commerce Research*, vol. 20, no. 2, art. 59, 2025, doi: 10.3390/jtaer20020059.
- [9] P. Hutabalian et al., "Implementasi Pipeline ETL/ELT dan Model Dimensional untuk Analisis Penjualan Shopee Menggunakan PostgreSQL, Docker, dan Apache Superset," *Jurnal Informatika dan Teknik Elektro Terapan*, vol. 13, no. 3S1, pp. 1321-1332, 2025, doi: 10.23960/jitet.v13i3S1.8093.
- [10] A. Wibowo and A. R. Handoko, "Segmentasi Pelanggan Ritel Produk Farmasi Obat Menggunakan Metode Data Mining Klasterisasi Dengan Analisis Recency Frequency Monetary (RFM)

-
- Termodifikasi,” *Jurnal Teknologi Informasi dan Ilmu Komputer*, vol. 7, no. 3, pp. 573-580, 2020, doi: 10.25126/jtiik.2020702925.
- [11] N. S. Niko, A. Rahman, D. M. U. Atmaja, and A. Basri, “Klasterisasi Stok Produk Retail Untuk Menentukan Pergerakan Kebutuhan Konsumen Dengan Algoritma K-Means,” *Bulletin of Information Technology (BIT)*, vol. 4, no. 3, pp. 306-312, 2023, doi: 10.47065/bit.v4i3.736.
- [12] R. Berman and A. Israeli, “The Value of Descriptive Analytics: Evidence from Online Retailers,” *Marketing Science*, vol. 41, no. 6, pp. 1074–1096, 2022, doi: 10.1287/mksc.2022.1352.
- [13] W. G. S. Parwita, N. L. W. S. R. Ginantra, I. M. D. P. Asana, and I. G. L. A. S. Dharma, “Retail Data Visualization in Business Intelligence System for Ayu Nadi Group,” in *Proceedings of the 4th International Conference on Vocational Education and Technology (IconVET 2021)*, EAI, 2022, doi: 10.4108/eai.27-11-2021.2315530.
- [14] A. R. I. Sipayung, N. E. Zendrato, T. A. Marbun, J. N. Telaumbanua, and O. Sihombing, “Penerapan Teknologi Business Intelligence dalam Meningkatkan Strategi Penjualan dengan Metode OLAP pada Café Le Kahve,” *Jurnal Teknik Informasi dan Komputer (Tekinkom)*, vol. 7, no. 2, pp. 628–636, 2024, doi: 10.37600/tekinkom.v7i2.1529.
- [15] A. Wijaya and A. Novaldi, “Perancangan Data Warehouse Menggunakan Model Star Schema untuk Analisis Penjualan Retail Berbasis PostgreSQL,” *Journal of Informatics and Business*, vol. 3, no. 3, pp. 305–314, 2025, doi: 10.47233/jibs.v3i3.3759.