International Journal of Information Technology and Computer Science Applications (IJITCSA)
p-ISSN: 2964-3139 e-ISSN: 2985-5330
Vol. 04, No. 01, page 59 - 69
Submitted 15/2/2026; Accepted 1/4/2026; Published 13/4/2026

Revisiting the IBM Retail Data Warehouse: A Governed One-Column Architecture and Reproducible Open-Dataset Validation for Retail Analytics

Nayananda Karunaratne[1*] and **Pulasthi Medhananda**[1]
[1]Computing & Information Systems, Faculty of Computing, Sabaragamuwa University of Sri Lanka, Belihuloya 70140, SRI LANKA
e-mail: nayarunar11@gmail.com

Corresponding Autor: Nayananda Karunaratne

Abstract
The IBM Retail Data Warehouse (RDW) correctly recognized the importance of integrated retail data, but it remained largely descriptive, did not formalize the underlying architecture, and lacked a reproducible empirical validation. This paper reconstructs and substantially extends that early proposal into a publication-ready research article. We first synthesize the historical IBM RDW, Retail Data Warehouse Model (RDWM), Retail Services Data Model (RSDM), and Retail Business Solution Template (RBST) concepts with contemporary data warehousing, data governance, and retail analytics literature. We then propose a governed, RDW-informed logical architecture that separates ingestion, quality control, conformed dimensional modeling, analytics marts, and decision-support services. To move beyond conceptual discussion, we instantiate the architecture with an open retail dataset from the UCI Machine Learning Repository containing 541,909 transactions. After governance-oriented preprocessing, the final analytical mart contains 392,692 valid rows, 18,532 orders, 4,338 customers, 3,665 products, and 37 countries. We formulate the transformation and forecasting workflow mathematically, define an end-to-end algorithmic pipeline, and evaluate a retail revenue forecasting task using naive, seasonal naive, linear regression, ridge regression, random forest, and gradient boosting baselines. On the hold-out test window, the best model (linear regression on warehouse-engineered features) achieves an RMSE of 4,302.61 GBP and R^2=0.9766, while a raw, ungoverned pipeline yields a much weaker RMSE of 10,068.59 GBP. This corresponds to a 57.27% reduction in RMSE attributable to governance and dimensional integration. The results show that the practical value of an RDW-like architecture is not merely organizational; when implemented as a governed analytical platform, it measurably improves reproducibility, interpretability, and forecasting quality.

Keywords— Retail data warehouse, Dimensional modeling, Data governance, Business intelligence, Retail analytics, Demand forecasting.

1 Introduction
Retail organizations operate in environments characterized by thin margins, volatile demand, fragmented customer journeys, and high expectations for operational responsiveness. The analytical challenge is not simply that retailers generate large amounts of data, but that the data are distributed across heterogeneous operational systems such as point-of-sale records, product catalogs, inventory files, customer relationship systems, promotions, finance, and external reference feeds. When these data remain siloed, the enterprise struggles to form a coherent view of demand, product performance, customer value, and supply-side risk. Classic data warehousing literature has long argued that decision support requires integrated, subject-oriented, time-variant, and non-volatile repositories rather than ad hoc reporting directly over transactional systems [1 – 5].

The literature that motivated this study introduced IBM's historical Retail Data Warehouse (RDW) as a mechanism for integrating retail data and cited two well-known cases discussed in IBM materials *"A Children's Place and Canadian Tire"*. Its central intuition was correct; retail competitiveness depends on the ability to consolidate scattered data and convert it into actionable information. However, the manuscript remained at the proposal level. It did not establish a formal warehouse design, did not define reproducible data transformation procedures, did not benchmark any analytical task, and did not situate the proposal within the broader scholarly literature on data governance, dimensional modeling, and modern analytics. In contemporary academic terms, the manuscript presented a useful conceptual position but not yet a research contribution.

This paper addresses that gap by transforming the original proposal into a rigorous and reproducible journal-style study. The emphasis is intentionally architectural rather than vendor promotional. IBM's RDW, RDWM, RSDM, and RBST are treated here as historically important design artifacts that encode sound principles for retail analytics, especially the separation of data integration from business-facing analytics [1, 2]. We therefore modernize the proposal in two ways. First, we reinterpret the IBM framework through current scholarship on warehousing, governance, and business analytics [6 – 14]. Second, we validate the architecture empirically using a public retail dataset and an end-to-end experimental workflow implemented with open tools, thereby avoiding the irreproducibility that often accompanies proprietary enterprise stacks.

The contributions of the paper are fourfold. First, we formalize an RDW-informed logical architecture for multistage retail analytics that includes ingestion, governance, dimensional modeling, semantic access, and predictive consumption. Second, we map the conceptual IBM retail models to a concrete dimensional design suitable for reproducible experimentation. Third, we provide a governed analytical workflow and algorithmic specification that can be re-executed by other researchers. Fourth, we demonstrate empirically that governance and warehouse-derived feature engineering materially improve forecasting accuracy over a raw extraction baseline.

2 Background and Related Work

2.1 Retail data warehousing and dimensional integration

Data warehouses exist to support analytical rather than transactional workloads. Inmon's foundational definition of the warehouse as an integrated, subject-oriented, time-variant, and non-volatile collection of data remains highly influential [3]. Kimball and Ross, by contrast, emphasize dimensional modeling, conformed dimensions, and business-process-oriented data marts as practical mechanisms for delivering usable analytics [4]. These perspectives are complementary in retail settings: an enterprise information foundation is needed, but users ultimately consume analytical facts and dimensions organized around familiar questions such as sales, assortment, customer, time, and product [5, 6].

IBM's retail industry models encoded exactly this concern. The historical RDW documentation framed the warehouse as a blueprint for comprehensive retail intelligence, covering customer segmentation, replenishment, promotion targeting, merchandise analysis, and cross-channel planning [1, 2]. Importantly, RDW separated business vocabulary, warehouse structure, and reusable analytical templates. This separation anticipated later ideas about semantic layers, governed marts, and reusable analytical assets.

2.2 Data governance and analytical reliability

A warehouse becomes strategically useful only when its data are reliable enough to support action. Data quality research has shown that users access data not merely by accuracy but also by completeness, consistency, timeliness, relevance, and interpretability [7]. Data governance extends these concerns into organizational design by clarifying accountability, lifecycle controls, decision rights, and stewardship mechanisms [8, 9]. In retail environments, the governance problem is acute because returns, cancellations, missing customer identifiers, duplicate records, and inconsistent product descriptions can substantially distort downstream analytics.

Modern enterprise architecture literature increasingly treats warehousing, lakehouse, and multimodel approaches as parts of a broader governed data management continuum rather than mutually exclusive alternatives [14, 15]. The practical lesson is that analytical performance is a function of both model choice and upstream data discipline. A sophisticated machine learning method cannot compensate fully for structurally inconsistent inputs.

2.3 Business analytics in retail

The rise of business intelligence and analytics has moved enterprise decision support from retrospective reporting toward prediction and optimization [10]. Big-data-enabled analytical capabilities have been associated with improved firm performance, particularly when aligned with business strategy and dynamic organizational capabilities [11, 12,

13]. Retail is one of the most visible application domains because demand, basket composition, timing, and customer heterogeneity all lend themselves to analytics. The UCI Online Retail dataset and the associated customer-segmentation study by Chen, Sain, and Guo are widely used examples in this space [18, 19]. Yet many published studies stop at segmentation or exploratory analysis and do not link the analytics task back to warehouse architecture and governance.

3 Research gap

The literature contains strong streams on dimensional warehousing, governance, and machine learning, but fewer papers explicitly connect the three in a retail-specific and reproducible form. Historical IBM model documentation is rich in architectural insight but not written as a conventional scientific paper. Conversely, many machine learning studies on public retail data optimize predictive performance without explicitly stating how dimensional integration and governance affect those results. This paper fills that gap by treating the warehouse not merely as storage infrastructure but as the mechanism through which data become analytically trustworthy.

3.1 Research Problem and Design Requirements

The research problem addressed in this paper can be stated as follows: *how can an RDW-inspired retail architecture be reformulated as a governed, reproducible analytical system whose value can be demonstrated empirically using open data?* The question is both architectural and methodological. Architecturally, the system must integrate heterogeneous retail records into a conformed dimensional representation. Methodologically, it must provide a reproducible basis for evaluating downstream analytics rather than relying on conceptual argument alone.

Table 1: Design requirements for the modernized RDW-informed retail architecture.

Requirement	Rationale
Integration	Unify transactional and descriptive retail records into a conformed analytical structure.
Governance	Enforce explicit rules for duplicates, invalid quantities, invalid prices, cancellations, and identifier completeness.
Dimensional usability	Support business-facing analysis through fact tables and dimensions instead of direct operational querying.
Traceability	Preserve a documented sequence from raw extraction to final analytical mart.
Reproducibility	Ensure that the empirical pipeline can be rerun with open tools and public data.
Analytical validity	Demonstrate value through measurable downstream performance rather than descriptive claims alone.
Publisher portability	Prepare the paper in a clean one-column LaTeX format that can be adapted to major journal templates.

4 RDW-Informed Architecture

4.1 Logical architecture

Figure 1 presents the proposed logical architecture. The design follows a layered approach. Operational and reference sources first land in a staging area, where schema checks and basic standardization are applied. A governance layer then enforces deduplication, validity constraints, master-data harmonization, and auditability. Only after these controls are applied are data loaded into the retail data warehouse proper, where facts and conformed dimensions support both BI consumption and advanced analytics.

This design mirrors the RDW principle that IT should own integration complexity while analytical teams operate over stabilized business entities [2]. In practice, that separation reduces semantic drift, makes downstream models easier to interpret, and limits the propagation of data quality defects into executive reporting.

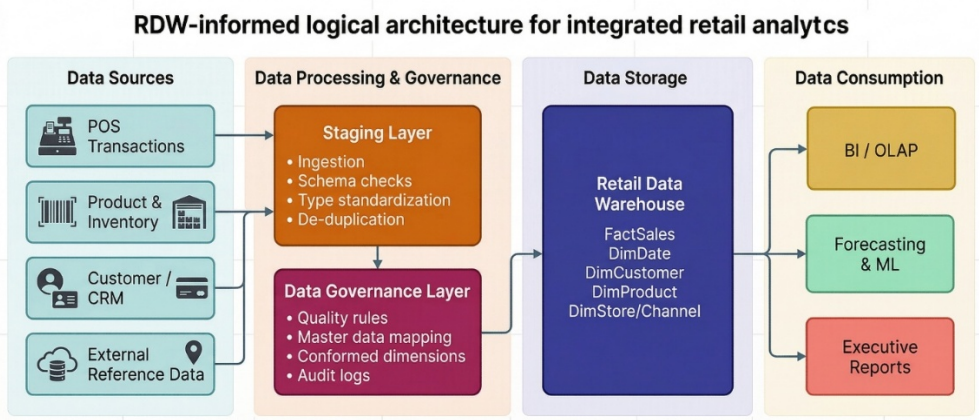

Figure 1: RDW-informed logical architecture for integrated retail analytics.

4.2 From IBM retail models to an implementable analytical framework

The historical IBM retail stack included several distinct but related artifacts. RSDM provided business vocabulary and analytical entities; RDWM translated that business understanding into dimensional structures; and RBST offered reusable analytical templates [1, 2]. Figure 2 re-expresses those relationships as an implementable research framework.

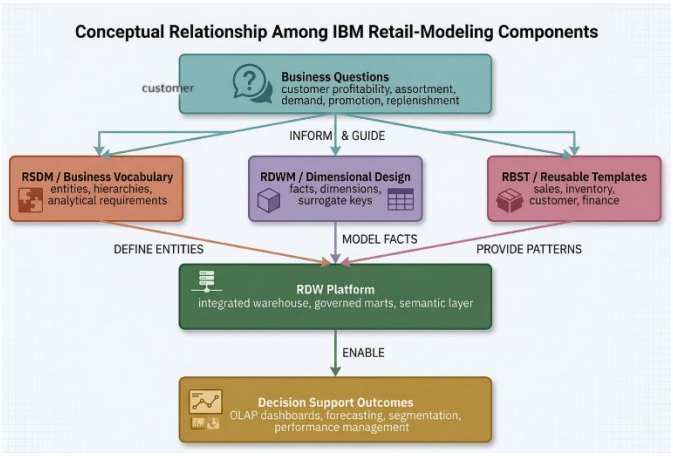

Figure 2: Conceptual relationship among IBM retail-modeling components and the proposed analytical framework.

The practical contribution of this reframing is that it turns historically descriptive product documentation into research-operationalizable architecture. The framework is vendor-neutral at execution time, but it remains faithful to the original RDW logic.

4.3 Dimensional model

In dimensional terms, the central business process in our empirical validation is retail sales. Let the fact table be denoted by

$$F = \langle k_d, k_c, k_p, m_1, m_2, \ldots, m_s \rangle \quad (1)$$

Where k_d, k_c, and k_p are surrogate keys for date, customer, and product, and the m_j are additive or semi-additive measures such as quantity, unit price, and sales amount. For the UCI dataset used in the experiment, the core instantiated dimensions are date, customer, product, and country. A fixed online channel is assumed because the source data come from a non-store online retailer [18]. Table 2 summarizes the warehouse entities.

Table 2: Core dimensional entities in the proposed retail warehouse.

Entity	Type	Illustrative attributes
FactSales	Fact	invoice number, date key, customer key, product key, quantity, unit price, sales amount
DimDate	Dimension	full date, day of week, week, month, quarter, year, weekend flag
DimCustomer	Dimension	customer identifier, country, recency/frequency/monetary features, activity flags
DimProduct	Dimension	stock code, product description, product family proxy, popularity indicators
DimCountry	Dimension	country name, regional grouping, market share measures
Analytical Mart	Derived mart	daily revenue, order count, active customers, SKU count, lag and rolling features

4.4 Mathematical formulation

Let x_i be a raw retail transaction record. The governance-oriented transformation is represented as

$$z_i = \mathcal{G}\left(\mathcal{S}\left(\mathcal{C}(x_i)\right)\right) \tag{2}$$

Where $\mathcal{C}(\cdot)$ performs cleansing (e.g., duplicate removal and invalid-value filtering), $\mathcal{S}(\cdot)$ standardizes schema and business keys, and $\mathcal{G}(\cdot)$ enforces governance rules and dimensional conformity. For a transaction i with quantity q_i and unit price p_i, the governed transaction revenue is

$$r_i = q_i \cdot p_i \tag{3}$$

Daily revenue in the analytical mart is then

$$R_t = \sum_{i \in \mathcal{T}_t} r_i \tag{4}$$

Where $\mathcal{T}_t$ is the set of governed transactions on day t. The forecasting task uses a feature vector

$$\mathbf{f}_t = \left[R_{t-1}, R_{t-7}, R_{t-14}, R_{t-28}, \bar{R}_{t,7}, \bar{R}_{t,14}, \bar{R}_{t,28}, o_t, c_t, s_t, d_t, m_t, w_t\right] \tag{5}$$

Where $\bar{R}_{t,h}$ denotes a rolling mean over horizon h, o_t is order count, c_t is active-customer count, s_t is SKU count, and d_t, m_t, w_t denote calendar variables. A forecasting model $\hat{R}_t = f_\theta(\mathbf{f}_t)$ is trained by minimizing

$$\min_\theta \mathcal{L}(\theta) = \frac{1}{N}\sum_{t=1}^{N}\left(R_t - \hat{R}_t\right)^2 \tag{6}$$

for the regression learners considered in this study.

5 Experimental Design

5.1 Algorithmic pipeline and open dataset

Figure 3 summarizes the end-to-end process from extraction to evaluation. The empirical validation uses the Online Retail dataset from the UCI Machine Learning Repository [18]. The dataset contains 541,909 transactions generated between 1 December 2010 and 9 December 2011 for a UK-based non-store online retailer. It includes invoice number, stock code, product description, quantity, invoice date, unit price, customer identifier, and country.

The dataset has become a standard benchmark for retail customer analytics and segmentation studies [19]. Its main advantage for the present work is that it is detailed enough to support warehouse construction and forecasting while remaining fully open and reusable.

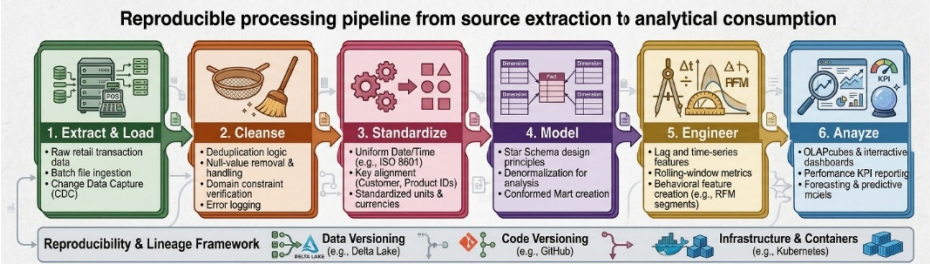

Figure 3: Reproducible processing pipeline from source extraction to analytical consumption.

5.2 Governance-oriented preprocessing

Rather than treating cleaning as a routine housekeeping step, we model it as a core component of analytical governance. The sequential effects of the governance pipeline are shown in Table 3. Starting from 541,909 raw rows, the final governed analytical mart retains 392,692 rows, corresponding to a retention rate of 72.46%. The largest drop occurs when enforcing customer completeness because customer-centric retail analytics cannot meaningfully assign behavior to anonymous records.

Table 3: Sequential effect of governance and preprocessing rules.

Stage	Rows	Removed	Retention (%)
Original extract	541,909	0	100
After duplicate removal	536,641	5,268	99.03
After cancellation removal	527,390	9,251	97.32
After quantity filter	526,054	1,336	97.07
After price filter	524,878	1,176	96.86
After description completeness	524,878	0	96.86
After customer completeness	392,692	132,186	72.46

The preprocessing rules are analytically justified. Cancellation invoices represent reversed business events rather than completed sales; non-positive quantities and prices violate the semantics of a forward sales fact for our forecasting task; exact duplicates create inflated measures; and missing customer identifiers undermine customer-level dimensional integrity. In a production warehouse these records would not necessarily be discarded universally; some would be loaded into separate returns or exception marts. For the specific governed sales mart used here, however, exclusion is the appropriate choice.

5.3 Analytical target and feature engineering

We validate the architecture on a daily revenue forecasting task because revenue is an executive-level measure that directly connects warehouse design to decision support. After loading governed transactions into the sales mart, we aggregate daily revenue, order count, active-customer count, item count, average basket value, and SKU count. We then derive autoregressive lags (1, 7, 14, and 28 days), rolling means and standard deviations (7, 14, and 28 days), and calendar variables. The resulting design reflects the business reality that retail performance is driven by both operational volume and temporal structure.

5.4 Train-validation-test protocol

The daily time series spans 374 calendar days after reindexing missing days to zero sales. Because forecasting must respect temporal order, we use a chronological split rather than random sampling. After feature generation and lag removal, the usable sequence contains 346 daily observations. The first 70% are used for training, the next 15% for model selection, and the final 15% for hold-out testing. This protocol avoids temporal leakage and provides a realistic assessment of forward-looking performance [22].

5.5 Baseline models

Six models are considered:

- last-value naive: $\hat{R}_t = R_{t-1}$
- seasonal naive: $\hat{R}_t = R_{t-7}$
- ordinary least squares linear regression
- ridge regression
- random forest regression [20]
- gradient boosting regression [21]

Random forest and gradient boosting are tuned on the validation window using compact hyperparameter grids. The goal is not exhaustive optimization, but a fair comparison between classical statistical baselines and nonlinear machine learning alternatives.

5.6 Evaluation metrics

We report mean absolute error (MAE), root mean squared error (RMSE), coefficient of determination (R^2), and symmetric mean absolute percentage error (sMAPE). For true values y_t and predictions $\hat{y}_t$:

$$\mathrm{MAE} = \frac{1}{N}\sum_{t=1}^{N}|y_t - \hat{y}_t| \tag{7}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{t=1}^{N}(y_t - \hat{y}_t)^2} \tag{8}$$

$$R^2 = 1 - \frac{\sum_{t=1}^{N}(y_t - \hat{y}_t)^2}{\sum_{t=1}^{N}(y_t - \bar{y})^2} \tag{9}$$

RMSE is emphasized in model selection because revenue forecasting penalizes large deviations more strongly, which is appropriate for managerial planning.

6 Results and Discussion

6.1 Warehouse-level descriptive insights

The governed warehouse contains 18,532 completed orders, 4,338 customers, 3,665 products, and 37 countries, with total governed revenue of GBP 8.89 million. Figure 4 shows the monthly revenue trajectory. Revenue rises substantially in the second half of 2011, with the strongest months occurring in September, October, and November 2011. This pattern is consistent with pre-holiday acceleration in online retail demand.

Figure 5 shows the top countries by governed revenue. The United Kingdom dominates the dataset, contributing approximately 82% of governed revenue, while the Netherlands, EIRE, Germany, and France form the next tier of markets. This concentration suggests that a country dimension is analytically useful even in a largely domestic online retailer because export markets still contribute meaningful variation.

These descriptive findings matter architecturally. They show why conformed dimensions are not mere modeling formalities: date and country hierarchies immediately enable questions about temporal concentration, market mix, and planning seasonality. Such questions were highlighted in the original IBM materials and remain central to retail decision support [1, 2].

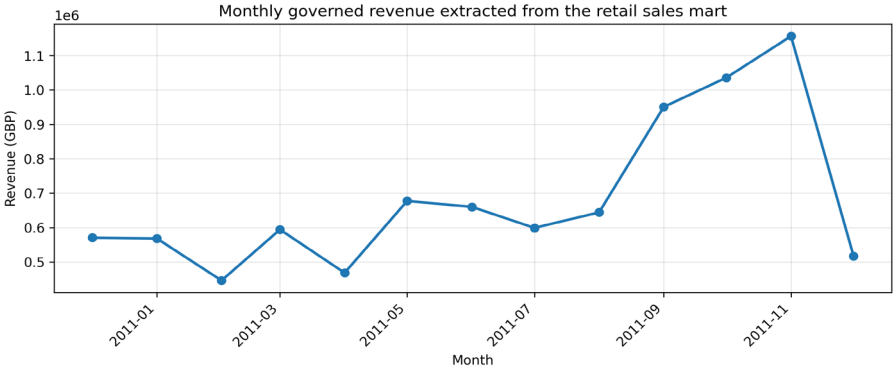

Figure 4: Monthly governed revenue extracted from the retail sales mart.

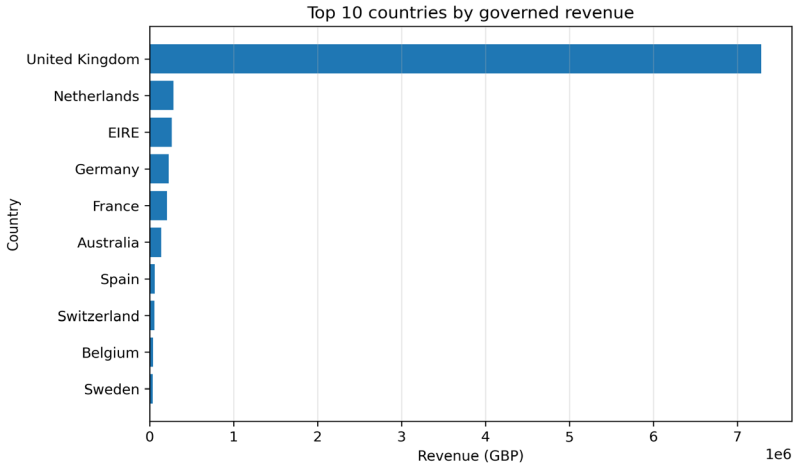

Figure 5: Top countries by governed revenue.

7 Forecasting performance comparison

Table 4 reports model performance on the hold-out test window. The strongest result comes from linear regression, followed very closely by ridge regression. Both models outperform the nonlinear tree ensembles and substantially exceed the naive baselines in RMSE and R^2 terms.

Table 4: Forecasting performance on the hold-out test window.

Model	MAE	RMSE	R^2	sMAPE
Linear Regression	3,310.37	4,302.61	0.9766	34.4
Ridge Regression	3,311.60	4,303.86	0.9766	34.4
Gradient Boosting	4,595.82	12,488.25	0.8028	34.98
Random Forest	5,102.84	15,550.88	0.6942	35.16
Seasonal Naive (t - 7)	12,716.09	23,664.00	0.2918	30.67
Last-value Naive (t - 1)	21,896.46	30,673.74	-0.1899	79.06

The result is not paradoxical. Our forecasting target is daily revenue, and the engineered feature set contains strong linear temporal and operational signals. When sample size is modest and autocorrelation is high, linear models can outperform more flexible learners that require richer nonlinear structure or more extensive tuning [22]. The key point is therefore not that linear regression is universally best for retail forecasting, but that a well-governed warehouse can create highly informative features that make even relatively simple models effective.

Figure 6 visualizes the RMSE ranking, and Figure 7 shows the correspondence between actual and predicted revenue for the best model.

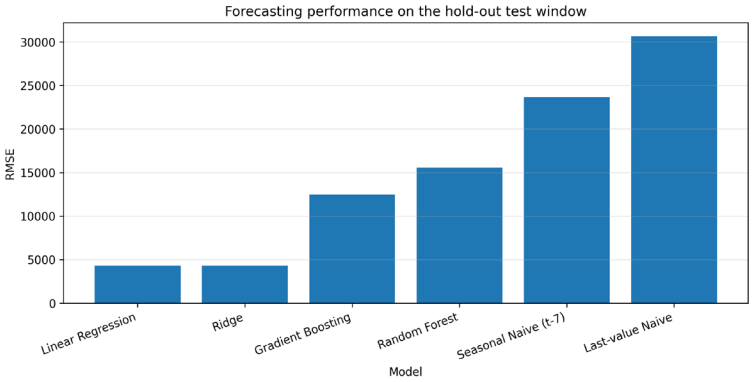

Figure 6: RMSE comparison across benchmark forecasting models.

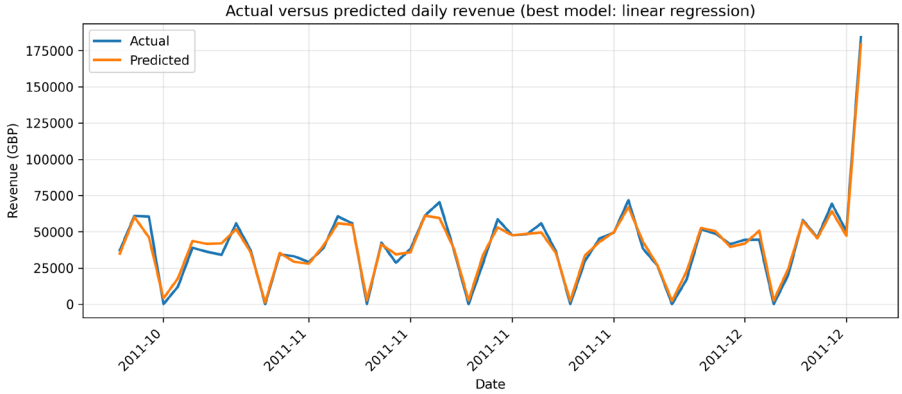

Figure 7: Actual versus predicted daily revenue for the best-performing model.

7.1 Effect of governance on downstream analytics

A central claim of this paper is that governance is not only an information-management virtue but also an empirical determinant of model quality. To test that claim, we trained the same linear regression approach on a raw extraction pipeline without the governed sales-mart transformation. Table 5 summarizes the comparison.

Table 5: Effect of governance on forecasting performance using the same learner family.

Pipeline	**MAE**	**RMSE**	**R^2**	**sMAPE**
Raw extraction + linear regression	7,319.35	10,068.59	0.8282	39.93
Governed sales mart + linear regression	3,310.37	4,302.61	0.9766	34.4

The governed pipeline reduces RMSE by 57.27% and MAE by more than 54% relative to the raw pipeline. This result strongly supports the architectural position of the paper. Data governance should not be treated as a post hoc administrative concern. It is a structural enabler of analytical validity.

7.2 Managerial and architectural implications

Three implications follow from the findings.

First, the original intuition behind RDW remains valid; retailers need a business-focused analytical repository that decouples operational volatility from decision-support consumption. The empirical evidence here suggests that this decoupling improves not only reporting coherence but also predictive performance.

Second, governance decisions should be encoded as explicit warehouse logic rather than left to notebook-level analyst discretion. When cleansing and conformance are embedded in the pipeline, analytical outputs become more repeatable and easier to audit.

Third, not all retail forecasting problems require highly complex models. The warehouse itself performs a large share of the "modeling" work by structuring facts, dimensions, and stable derived signals. In many settings, simple transparent learners operating over governed marts may offer a better balance of performance and interpretability than more opaque alternatives.

7.3 Threats to validity

The study has several limitations. The empirical validation uses a single public online-retail dataset rather than a full multichannel enterprise environment. The source data do not contain explicit store, promotion, supplier, or inventory tables, so some elements of a full retail warehouse are represented conceptually rather than instantiated empirically. Customer completeness filtering also removes a substantial portion of the raw extract while analytically justified for customer-centric modeling, other operational use cases might retain those rows under an "unknown customer" strategy. Finally, the forecasting task focuses on daily revenue rather than assortment optimization, promotion response, or stock-out prediction. These limitations do not invalidate the results, but they define the scope of the conclusions.

8 Conclusion

This paper transformed an initial conceptual manuscript on IBM's Retail Data Warehouse into a substantially expanded and empirically grounded research article. The resulting contribution is more than an extended narrative. It provides formal architecture, dimensional design, mathematical workflow, algorithmic specification, governance protocol, and reproducible experimental validation. The evidence shows that the practical value of an RDW-informed architecture lies not only in consolidating information for dashboards, but also in producing analytically superior data products for downstream modeling.

Using the open UCI Online Retail dataset, we demonstrated that a governed sales mart containing 392,692 valid rows can support descriptive business intelligence as well as accurate revenue forecasting. The best model achieved an RMSE of 4,302.61 GBP and R^2=0.9766, while the equivalent learner over a raw extraction baseline performed much worse. This result provides concrete empirical support for the longstanding data-warehousing principle that integration and governance are prerequisites for trustworthy analytics.

Future work should instantiate the architecture over richer multitable retail data including inventory, promotions, returns, logistics, and supplier events. This would enable a fuller test of RDW-inspired retail intelligence across replenishment, customer lifetime value, campaign response, and margin optimization. Even in its current form, however, the study demonstrates that a historically grounded warehouse concept can be modernized into a rigorous, reproducible, and publication-ready analytics framework.

BIBLIOGRAPHY

[1]. IBM, Industry Models for Retail: The IBM Retail Data Warehouse—Harnessing the Power of Information [Brochure]. Somers, NY, USA: IBM Software Group, 2007. Available: https://public.dhe.ibm.com/software/data/sw-library/industry-models/brochures/IBM_Retail_Models.pdf

[2]. IBM, Retail Data Warehouse (RDW): General Information Manual. IBM, 2009. Available: https://public.dhe.ibm.com/software/data/sw-library/industry-models/brochures/IBM_retail_data_warehouse_GIMv8.pdf

[3]. W. H. Inmon, Building the Data Warehouse, 4th ed. Indianapolis, IN, USA: Wiley, 2005. Available: https://www.wiley.com/en-es/Building%2Bthe%2BData%2BWarehouse%2C%2B4th%2BEdition-p-9780764599446

[4]. R. Kimball and M. Ross, The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling, 3rd ed. Indianapolis, IN, USA: Wiley, 2013. Available: https://www.wiley.com/en-jp/The%2BData%2BWarehouse%2BToolkit%3A%2BThe%2BDefinitive%2BGuide%2Bto%2BDimensional%2BModeling%2C%2B3rd%2BEdition-p-9781118530801

[5]. S. Chaudhuri and U. Dayal, "An overview of data warehousing and OLAP technology," ACM SIGMOD Record, vol. 26, no. 1, pp. 65–74, 1997, doi: 10.1145/248603.248616.

[6]. M. Golfarelli and S. Rizzi, Data Warehouse Design: Modern Principles and Methodologies. New York, NY, USA: McGraw-Hill, 2009. Available: https://www.mheducation.com/highered/mhp/product/data-warehouse-design-modern-principles-methodologies.html

[7]. R. Y. Wang and D. M. Strong, "Beyond accuracy: What data quality means to data consumers," Journal of

Management Information Systems, vol. 12, no. 4, pp. 5–33, 1996, doi: 10.1080/07421222.1996.11518099.
[8]. V. Khatri and C. V. Brown, "Designing data governance," Communications of the ACM, vol. 53, no. 1, pp. 148–152, 2010, doi: 10.1145/1629175.1629210.
[9]. B. Otto, "Organizing data governance: Findings from the telecommunications industry and consequences for large service providers," Communications of the Association for Information Systems, vol. 29, Art. 3, 2011, doi: 10.17705/1CAIS.02903.
[10]. H. Chen, R. H. L. Chiang, and V. C. Storey, "Business intelligence and analytics: From big data to big impact," MIS Quarterly, vol. 36, no. 4, pp. 1165–1188, 2012, doi: 10.2307/41703503.
[11]. S. Akter, S. F. Wamba, A. Gunasekaran, R. Dubey, and S. J. Childe, "How to improve firm performance using big data analytics capability and business strategy alignment?," International Journal of Production Economics, vol. 182, pp. 113–131, 2016, doi: 10.1016/j.ijpe.2016.08.018.
[12]. S. F. Wamba, A. Gunasekaran, S. Akter, S. J.-f. Ren, R. Dubey, and S. J. Childe, "Big data analytics and firm performance: Effects of dynamic capabilities," Journal of Business Research, vol. 70, pp. 356–365, 2017, doi: 10.1016/j.jbusres.2016.08.009.
[13]. P. Mikalef, M. Boura, G. Lekakos, and J. Krogstie, "Big data analytics and firm performance: Findings from a mixed-method approach," Journal of Business Research, vol. 98, pp. 261–276, 2019, doi: 10.1016/j.jbusres.2019.01.044.
[14]. A. Nambiar and D. Mundra, "An overview of data warehouse and data lake in modern enterprise data management," Big Data and Cognitive Computing, vol. 6, no. 4, Art. 132, 2022, doi: 10.3390/bdcc6040132.
[15]. S. Bimonte, E. Gallinucci, P. Marcel, and S. Rizzi, "Data variety, come as you are in multi-model data warehouses," Information Systems, vol. 104, Art. 101734, 2022, doi: 10.1016/j.is.2021.101734.
[16]. A. Cuzzocrea, I.-Y. Song, and K. C. Davis, "Analytics over large-scale multidimensional data: The big data revolution!," in Proc. 14th ACM Int. Workshop on Data Warehousing and OLAP (DOLAP '11), 2011, pp. 101–104, doi: 10.1145/2064676.2064695.
[17]. N. Elgendy and A. Elragal, "Big data analytics in support of the decision making process," Procedia Computer Science, vol. 100, pp. 1071–1084, 2016, doi: 10.1016/j.procs.2016.09.251.
[18]. D. Chen, Online Retail [Dataset]. UCI Machine Learning Repository, 2015, doi: 10.24432/C5BW33. Available: https://archive.ics.uci.edu/dataset/352/online%2Bretail
[19]. D. Chen, S. L. Sain, and K. Guo, "Data mining for the online retail industry: A case study of RFM model-based customer segmentation using data mining," Journal of Database Marketing & Customer Strategy Management, vol. 19, pp. 197–208, 2012, doi: 10.1057/dbm.2012.17.
[20]. L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.
[21]. J. H. Friedman, "Greedy function approximation: A gradient boosting machine," The Annals of Statistics, vol. 29, no. 5, pp. 1189–1232, 2001, doi: 10.1214/aos/1013203451.
[22]. R. J. Hyndman and G. Athanasopoulos, Forecasting: Principles and Practice, 3rd ed. Melbourne, Australia: OTexts, 2021. Available: https://otexts.com/fpp3/