International Journal of Information Technology and Computer Science Applications (IJITCSA)
p-ISSN: 2964-3139 e-ISSN: 2985-5330
Vol. 04, No. 01, page 35 - 46
Submitted 7/2/2026; Accepted 29/3/2026; Published 3/4/2026

A Reproducible Data Warehouse and OLAP Framework for Retail Analytics: Design, Dimensional Modeling, and Experimental Evaluation in the SwiftMart Case

Victorio Palben Medel
Information System, School of Information Technology, Makati City, PHILIPPINES
e-mail: vipaldel@mapua.edu.ph

Corresponding Autor: Victorio Palben Medel

Abstract
Retail organizations increasingly rely on heterogeneous operational platforms, including point-of-sale systems, customer relationship management applications, cloud data stores, and locally administered databases. Although these platforms are valuable for transaction processing, they often generate fragmented, duplicated, and semantically inconsistent data that constrain enterprise reporting, forecasting, and customer intelligence. This paper substantially extends a conceptual SwiftMart case into a full design-and-evaluation study of a retail data warehouse and Online Analytical Processing (OLAP) framework. The proposed artifact combines a Kimball-style dimensional architecture, a governed extract-transform-load (ETL) pipeline, conformed dimensions, and materialized OLAP summaries for managerial analytics. To ground the case empirically, the framework is evaluated using the open-access UCI Online Retail dataset, which contains 541,909 transaction records from a UK-based online retailer covering 1 December 2010 to 9 December 2011. The experiment transforms raw transactions into a star schema with 524,878 curated fact rows, 19,960 orders, 4,355 customer members, 4,158 product members, and 38 countries. Four representative analytical workloads are benchmarked across three storage designs: a normalized operational data store, a dimensional warehouse, and materialized aggregate tables. The dimensional warehouse reduces mean latency by 42.3% relative to baseline joins, while materialized aggregates reduce latency by approximately 99.9%. A forecasting demonstration on warehouse-generated daily revenue aggregates further shows that a random forest model outperforms a naive benchmark, achieving an RMSE of 23,715.84 versus 34,055.29. The paper contributes an end-to-end reference architecture for retail analytics, together with dimensional design rationale, mathematical formulations, algorithms, empirical results, and implementation guidance relevant to both academic researchers and practitioners.

Keywords— Retail analytics, Data warehouse, OLAP, Dimensional modeling, ETL, Business Intelligence, Forecasting.

1 Introduction

Retail data infrastructures have undergone a profound transition from isolated transaction processing systems to integrated analytical ecosystems. Contemporary retailers do not operate through a single application stack; instead, they depend on a portfolio of operational platforms such as point-of-sale systems, customer relationship management (CRM) systems, e-commerce platforms, cloud logs, marketing tools, and localized databases maintained by individual departments. This heterogeneity is operationally useful, yet analytically problematic. When data are distributed across platforms with different schemas, update rhythms, identifiers, and semantic conventions, organizations struggle to form a consistent view of products, customers, orders, and performance [1–3].

The original SwiftMart manuscript positioned these issues as a conceptual case involving duplicated data, fragmented storage, inconsistent formats, and limited support for integrated analytics . Those observations remain

highly relevant. Retail decision making increasingly requires unified analytical access not only for descriptive reporting but also for forecasting, segmentation, pricing analysis, basket analysis, and operational planning. Without integrated historical data, even basic managerial questions—such as monthly revenue by market, top-selling products by quarter, or top customers by year—require repeated ad hoc joins across systems and significant manual cleansing. The result is a costly analytical environment characterized by low reproducibility, high latency, and weak governance [4–6].

Data warehousing and OLAP continue to provide a robust response to this problem. A warehouse separates analytical workloads from operational systems, stores subject-oriented and time-variant historical data, and supports multidimensional queries through dimensional modeling and precomputed aggregates [3, 7, 8]. In the retail sector, these capabilities are especially important because decision making occurs simultaneously at multiple grains: transaction line, invoice, product, customer, channel, store, region, and calendar hierarchy. Forecasting and customer intelligence additionally depend on reliable historical series and conformed dimensions [9–11]. Even so, many retail organizations still underinvest in analytical integration because of cost concerns, privacy risks, timeliness requirements, and legacy-system inertia [1, 2, 12].

This paper addresses that gap by transforming the SwiftMart case from a purely descriptive proposal into a reproducible design-and-evaluation study. The work does not claim novelty in the fundamental theory of warehousing; rather, its contribution lies in integrating the case requirements, dimensional design choices, ETL governance logic, and benchmark-based empirical validation into a single study that can be inspected, reproduced, and adapted. Because SwiftMart's proprietary data are not publicly available, the empirical validation uses the open-access UCI Online Retail dataset [13]. This strategy preserves the paper's managerial relevance while ensuring methodological transparency.

Three research questions guide the study:

- RQ1: Can a Kimball-style dimensional warehouse resolve the fragmentation and semantic inconsistency implied by the SwiftMart case while remaining analytically expressive for retail decision support?
- RQ2: How much analytical performance improvement can be achieved by moving from normalized operational joins to a star schema and then to materialized aggregate tables for representative retail workloads?
- RQ3: Can warehouse-generated aggregates reliably support downstream predictive analytics, illustrated here through daily revenue forecasting?

To answer these questions, the paper makes four contributions. First, it develops a refined retail data architecture that integrates heterogeneous sources through a governed ETL layer and conformed dimensional model. Second, it formalizes the design through mathematical definitions, dimensional schema descriptions, and algorithms for ETL and benchmark execution. Third, it empirically evaluates the design on a real open dataset using data quality, query latency, and forecasting metrics. Fourth, it packages the study in a reproducible LaTeX manuscript with source files, figures, bibliography, and code artifacts suitable for academic reuse.

The remainder of the paper is organized as follows. Section 2 synthesizes literature on retail data integration, dimensional warehousing, data quality, and predictive analytics. Section 3 defines the SwiftMart problem context and research method. Section 4 presents the proposed architecture, dimensional model, and ETL logic. Section 5 describes the open dataset, preprocessing, benchmark environment, and evaluation protocol. Section 6 reports empirical findings. Section 7 discusses implications, limitations, and governance considerations. Section 8 concludes.

2 Related Work and Research Gap

2.1 Retail data integration challenges

Retail digitalization has created a paradox. On one hand, organizations now capture transactions, customer interactions, loyalty events, web behavior, and inventory events at unprecedented scale. On the other, the accumulation of disparate platforms often intensifies fragmentation, especially when analytical integration is deferred in favor of short-term operational deployment [1, 14]. Aversa et al. describe how retail organizations often possess unevenly developed "big data environments," with investment focused more on data collection than on integrated decision support capability [2]. In such environments, useful information exists, but it is distributed across inaccessible or weakly conformed repositories.

The retail consequences are significant. Fragmented data reduces visibility across channels, inhibit customer-centric decision making, and delay store, assortment, or campaign-level analysis. Wibowo et al. characterize siloed

data environments as difficult to combine analytically, especially when source systems evolve independently and lack harmonized identifiers [15]. In e-commerce and omnichannel settings, the challenge is compounded by the velocity and heterogeneity of semi-structured and external data, which affect the timeliness and interpretability of forecasts and managerial reporting [11, 16]. Akter and Wamba further emphasize that the business value of big data in e-commerce depends less on raw volume than on an organization's ability to integrate, govern, and operationalize it [14].

In the SwiftMart context, the central symptoms are duplicated identifiers, inconsistent formatting, scattered ownership, and difficulty generating integrated sales intelligence. These symptoms align closely with the broader literature, the problem is not the absence of data, but the absence of a governed analytical structure that separates operational heterogeneity from decision-support consistency.

2.2 Data warehousing, dimensional modeling, and OLAP

The classical rationale for data warehousing remains compelling. Chaudhuri and Dayal define the warehouse as a repository optimized for extraction, cleaning, transformation, integration, and multidimensional access for decision support [3]. Kimball and Ross advocate dimensional modeling because it aligns closely with how managers ask questions by slicing facts through dimensions such as time, customer, product, geography, and organizational unit [4]. Inmon, by contrast, emphasizes enterprise integration and subject orientation from a top-down perspective [5]. These schools differ architecturally, but they agree on a central principle decision support requires stable analytical abstractions that are distinct from operational transaction processing.

For retail systems, dimensional modeling is especially attractive because the dominant analytical tasks are aggregative and hierarchical. Revenue, quantity, margin, basket value, promotion response, and customer lifetime value are naturally summarized by calendar period, product category, channel, region, or customer segment. Harinarayan et al. show that precomputing aggregates can dramatically reduce the cost of data-cube style queries by materializing selected views in advance [7]. Golfarelli and Rizzi likewise treat the dimensional model not as a mere storage convenience but as the semantic backbone of analytical communication between technical designers and business users [8].

OLAP operations roll-up, drill-down, slice, dice, and pivot benefit from a warehouse because facts and dimensions are explicitly organized for these navigation patterns. In practice, this improves not only performance but also reproducibility. When the same conformed dimensions are reused across reports and models, analytical outputs become more comparable and less dependent on one-off joins. Sharda et al. emphasize that this alignment between warehouse design and business intelligence workflows remains foundational for trustworthy analytics [6].

2.3 Data quality, privacy, and analytical reliability

A warehouse is only as trustworthy as the data quality controls embedded in its pipelines. Wang and Strong famously argue that data quality must be understood from the consumer perspective rather than reduced to simple correctness [17]. Completeness, consistency, timeliness, and interpretability all shape whether data are fit for managerial use. More recent surveys confirm that data quality monitoring still revolves around a recurring set of core dimensions such as accuracy, completeness, consistency, uniqueness, and timeliness [18]. For retail data, these dimensions manifest concretely as duplicate invoices, null customers, inconsistent product descriptions, returns encoded as negative quantities, or stale snapshots.

Privacy and compliance introduce an additional governance layer. Retail transactions often involve personal and financial information whose analytical use must be bound by minimization, transparency, access control, and lawful processing principles [19]. Aloysius et al. show that customer-facing benefits from big data initiatives are linked to users' perceptions of service quality and trust [12]. Therefore, analytically powerful architecture must also include design provisions for masking, role-based access, and controlled exposure of personally identifiable information.

2.4 Warehouse-enabled predictive analytics in retail

An integrated warehouse is not an end in itself; its value lies in supporting downstream analytics. Retail forecasting, for example, operates at multiple grains, from aggregate sales to SKU-level demand, and is affected by promotions, calendar structure, and channel interactions [9, 20]. Clean historical aggregates are indispensable for such models. Chen et al. demonstrate how customer-centric analytics can be built from transaction histories through RFM-type features [10]. Papanagnou and Matthews-Amune show that richer structured and external data can improve responses to demand volatility [11]. These works reinforce the proposition that predictive performance depends on the stability and conformance of the underlying data representation.

2.5 Research gap

The literature strongly supports data integration, dimensional warehousing, and predictive analytics in retail. However, many studies focus on either conceptual architecture or downstream models in isolation. Fewer studies present an end-to-end, reproducible evaluation that links:

i. a case-derived problem formulation
ii. dimensional design decisions
iii. ETL quality controls
iv. OLAP performance benchmarking
v. a concrete predictive task using warehouse-generated aggregates.

This paper addresses that gap in the SwiftMart case.

3 SwiftMart problem context and research method

3.1 Case framing

SwiftMart is treated as a pseudonymous retail case representing a common medium-scale analytical maturity problem: multiple operational systems have been adopted over time, but enterprise-level analytical integration remains partial. The original case emphasizes two main pain points: multiple or duplicated data and multiple data sources with non-standardized formats. These are reframed here as design requirements for a warehouse artifact.

The case is analytically realistic for at least three reasons. First, retail departments often optimize their own systems locally, creating fragmented identifiers and inconsistent semantics. Second, reporting requests frequently cross departmental boundaries, requiring joins that operational systems were never designed to support. Third, once predictive analytics are introduced, the lack of historical consistency becomes a bottleneck rather than a mere inconvenience.

3.2 Research strategy

The study follows a design-science logic in which an artifact is built to address an identified organizational problem and then evaluated against explicit performance criteria. The artifact comprises five tightly coupled components which are source abstraction, ETL and staging, dimensional warehouse, OLAP summaries, and analytics services. Evaluation is conducted using an open-access dataset to ensure transparency.

The research process contains six stages:

1. Problem formalization: Translate the SwiftMart case into analytical requirements and research questions.
2. Artifact design: Specify the architecture, dimensional model, ETL rules, and OLAP summaries.
3. Data instantiation: Apply the design to the UCI Online Retail dataset and generate curated warehouse tables.
4. Benchmark execution: Compare representative workloads across operational joins, star schema queries, and materialized aggregates.
5. Predictive demonstration: Evaluate daily revenue forecasting on warehouse-generated aggregates.

Interpretation: Assess managerial implications, limitations, and generalizability.

3.3 Analytical formulations

Let each retail line item i contain quantity q_i, unit price p_i, invoice identifier v_i, product identifier s_i, customer identifier c_i, and timestamp t_i. Line-item revenue is defined as

$$r_i = q_i \times p_i \tag{1}$$

For a dimensional fact table F, monthly country revenue is

$$R_{y,m,g} = \sum_{i \in F(y,m,g)} r_i \tag{2}$$

where $F(y, m, g)$ is the subset of facts associated with year y, month m, and geography g. Data quality is monitored through operational metrics. Completeness of required fields is

$$C = 1 - \frac{M}{N} \tag{3}$$

Where M is the number of missing required values and N is the total number of required field instances. Uniqueness is defined as

$$U = 1 - \frac{D}{N_r} \tag{4}$$

Where D is the number of detected duplicate rows and N_r is the total row count. Benchmark improvement for a workload is measured as

$$I = \frac{T_b - T_x}{T_b} \times 100 \tag{5}$$

Where T_b is baseline mean latency and T_x is latency under the warehouse or aggregate design. For forecasting, mean absolute error (MAE), root mean squared error (RMSE), and mean absolute percentage error (MAPE) are computed as

$$\begin{aligned} \text{MAE} &= \frac{1}{n}\sum_{t=1}^{n}|y_t - \hat{y}_t| \\ \text{RMSE} &= \sqrt{\frac{1}{n}\sum_{t=1}^{n}(y_t - \hat{y}_t)^2} \\ \text{MAPE} &= \frac{100}{n^*}\sum_{t\in\Omega}\left|\frac{y_t - \hat{y}_t}{y_t}\right| \end{aligned} \tag{6}$$

Where Ω excludes zero-demand periods to avoid division by zero.

4 Proposed architecture and dimensional design

4.1 Architecture overview

Figure 1 presents the proposed SwiftMart architecture. The design follows a Kimball-style, bottom-up logic because it balances enterprise consistency with manageable implementation scope [4]. Operational systems remain the sources of record, but analytical workloads are shifted to a warehouse layer populated by governed ETL processes. Departmental marts and materialized summaries are treated as analytical derivatives rather than isolated silos.

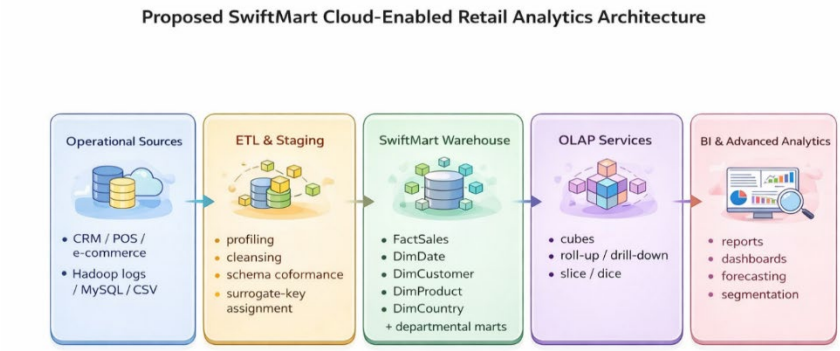

Figure 1: Proposed SwiftMart cloud-enabled retail analytics architecture.

This architecture deliberately separates concerns. Source systems prioritize operational continuity. The staging layer enforces profiling, cleansing, conformance, and surrogate key generation. The warehouse stores line-level facts and descriptive dimensions. OLAP services and materialized summaries accelerate multidimensional analysis. Finally, business intelligence and predictive services consume the warehouse rather than directly interrogating volatile source systems.

4.2 Processing pipeline

Figure 2 summarizes the pipeline from ingestion to analytical consumption. The sequence is intentionally linear at the conceptual level even though, operationally, some stages may be orchestrated incrementally or in parallel.

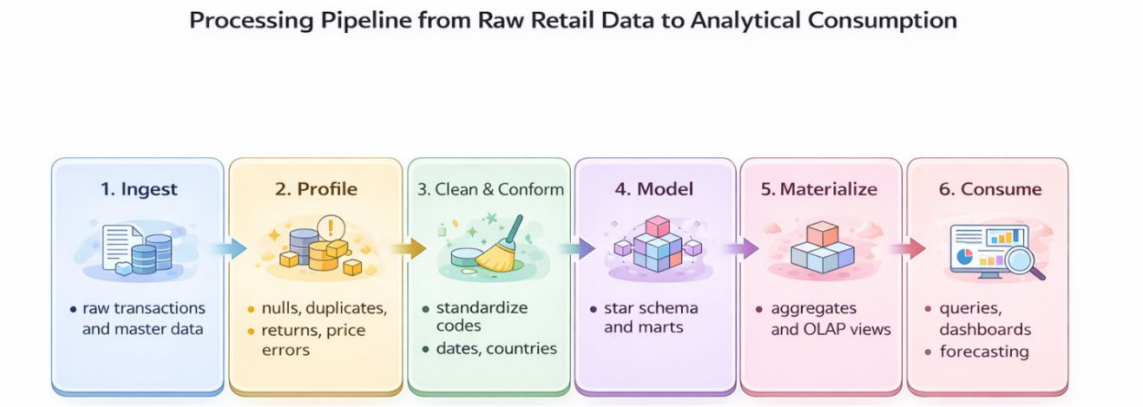

Figure 2: Processing pipeline from raw retail data to analytical consumption.

The pipeline has three design priorities. First, source irregularities must be corrected before analytical exposure. Second, dimensions must be conformed so that the same business concepts are reused consistently across reports. Third, common OLAP workloads should be accelerated through summary tables rather than repeatedly recomputing expensive joins.

4.3 Dimensional model

The warehouse is centered on a FactSales table at invoice-line granularity. This grain was selected because it preserves the full expressive power of the retail transaction history: higher-level views such as invoice-level, daily, monthly, product-level, or country-level analyses can be derived through aggregation without information loss. Table 1 summarizes the core schema.

Table 1: Core dimensional model for the SwiftMart warehouse.

Table	Grain	Key attributes	Main measures / purpose
FactSales	One invoice line for one product on one date for one customer-country combination	FactKey, InvoiceNo, DateKey, CustomerKey, ProductKey, CountryKey	Quantity, unit price, revenue; central transactional fact table
DimDate	One calendar day	Date, year, quarter, month, week, day of week	Supports calendar drill-down and periodic roll-up
DimCustomer	One conformed customer member	Customer ID, country	Customer analysis, retention, top-customer reporting
DimProduct	One conformed product member	Stock code, product description	Product ranking, assortment analytics
DimCountry	One country	Country name	Geographic slicing and country-level KPIs

The evaluation uses country as a separate dimension even though country is also associated with customers, because country-level slicing is a frequent managerial access path and benefits from explicit conformance.

This design is not merely structural; it embodies analytical semantics. The date dimension encodes the reporting calendar. The customer dimension stabilizes customer identity across transactions. The product dimension isolates descriptive product variation from measures. The country's dimension supports geographic aggregation without duplicating textual geography fields inside the fact table. Together, these components create a schema aligned with OLAP navigation.

4.4 Framework illustration

Figure 3 shows the study's integrated evaluation framework. The figure highlights that the artifact is assessed not solely as a schema, but as a connected system linking design, engineering, analytics, and evaluation outcomes.

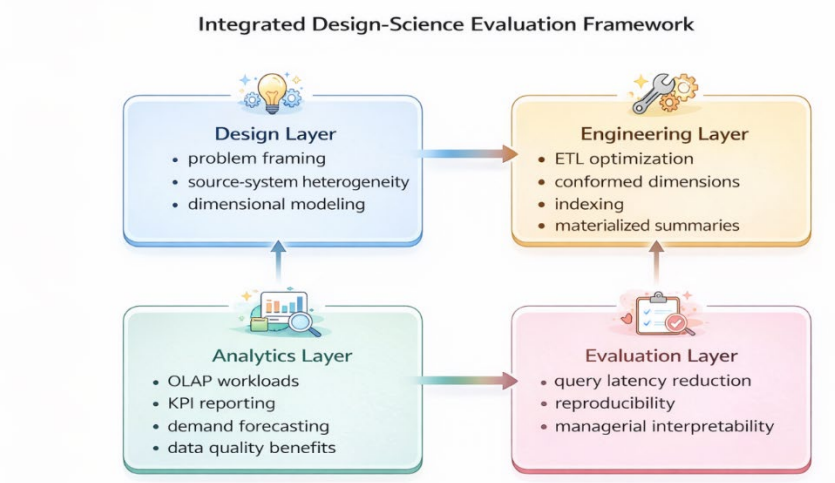

Figure 3: Integrated design-science evaluation framework used in the study.

4.5 ETL logic

Algorithm formalizes the ETL logic used in the study. The procedure intentionally includes both data-quality controls and dimensional loading. This reflects a key principle: in retail warehousing, schema population and data-quality enforcement should not be treated as separate afterthoughts.

Raw transaction table T Curated dimensions and fact table Remove duplicate rows from T Remove cancellation rows where invoice identifier begins with C Remove rows with $q_i \leq 0$ or $p_i \leq 0$ Impute missing product descriptions with a controlled placeholder Assign unknown customer member for missing customer identifiers Parse timestamps and derive calendar attributes Compute line revenue $r_i = q_i \times p_i$ Build conformed dimensions: Date, Customer, Product, Country Generate surrogate keys for each dimension Map transactional rows to dimension keys Load FactSales at invoice-line grain Materialize summary tables for high-frequency OLAP workloads Log row counts, rejected records, and data-quality indicators.

The logic reflects conventional warehouse practice, but the study makes the decisions explicit to aid reproducibility. For example, cancellation rows are excluded from the primary fact table because the benchmark focuses on clean sales analytics rather than reverse-logistics accounting. Missing customer identifiers are not dropped universally; instead, they are mapped to an "unknown" member so that enterprise sales totals remain complete even when customer-level analytics require filtering.

4.6 Governance and privacy controls

Although the experimental dataset is anonymized, the SwiftMart architecture is designed for regulated operational use. Accordingly, the architecture assumes the following controls: role-based access to customer-level dimensions, masking or tokenization of personally identifiable fields before analytical exposure, data-retention rules aligned with business purpose, and audit logging of ETL and reporting jobs [19]. These controls are necessary because analytical integration can amplify privacy risk if left unchecked.

5 Experimental setup

5.1 Dataset

The experiment uses the UCI Online Retail dataset [13]. The official repository describes the dataset as a transactional record of all transactions occurring between 1 December 2010 and 9 December 2011 for a UK-based non-store online retailer, with 541,909 instances and eight variables. The company mainly sells all-occasion gifts, and many customers are wholesalers. These characteristics make the dataset appropriate for warehouse-oriented retail evaluation because it contains line-item transactions, customer identifiers, products, timestamps, prices, and country information.

5.2 Preprocessing and warehouse instantiation

The preprocessing rules are derived from the warehouse grain and intended use. Duplicate rows, cancellations, non-positive quantities, and non-positive prices were removed from the fact-table loading path. Missing descriptions were imputed with a controlled label. Missing customer identifiers were mapped to an unknown member when enterprise totals were needed, while customer-level analyses excluded the unknown member where appropriate.

The resulting warehouse instance contains 524,878 fact rows, 19,960 distinct orders, 4,355 customer members, 4,158 product members, 305 date members, and 38 countries. Total curated revenue amounts to £10,642,110.80 across 5,572,420 units.

Two points are important here. First, the volume of missing customer identifiers is substantial, which illustrates why a warehouse should distinguish enterprise sales completeness from customer-analytics completeness. Second, the dataset's retained line-item volume remains large enough to meaningfully benchmark analytical workloads.

5.3 Benchmark environment

The prototype was implemented in Python 3.13 with pandas 2.2.3, scikit-learn 1.8.0, matplotlib 3.10.8, and SQLite 3.46.1. SQLite was selected not because it represents a production-scale distributed warehouse, but because it provides a transparent, easily reproducible single-node benchmark suitable for demonstrating the structural effect of design choices. Three storage/query designs were compared:

1. **ODS join baseline**: normalized operational-style tables requiring explicit joins at query time.
2. **Star schema warehouse**: conformed dimensions and line-level fact table with indexes.
3. **Materialized aggregates**: precomputed summary tables for high-frequency OLAP workloads.

5.4 Representative workloads

Four workloads were chosen because they reflect standard retail management questions:

1. Monthly revenue by country.
2. Top products by quarter.
3. Average basket value by month.
4. Top customers by year.

These workloads span different analytical shapes: country-period aggregation, product hierarchy ranking, order-level composition, and customer-level consolidation. Mean latency across repeated executions is reported in milliseconds.

5.5 Forecasting demonstration

To illustrate warehouse-enabled predictive analytics, the curated fact table was aggregated to a continuous daily revenue series. Missing calendar days were explicitly represented as zero-revenue periods. Lag features at 1, 7, 14, and 28 days, rolling means at 7, 14, and 28 days, and calendar features (day of week, month, ISO week) were derived from the warehouse aggregate. The first 80% of the chronologically ordered observations were used for training and the final 20% for testing. Three models were compared: a naive last-value baseline, linear regression, and random forest.

6 Results

6.1 OLAP workload latency

Figure 4 visualizes the chart that materialization is particularly valuable when the query combines heavy grouping and sorting, as in the top-product workload. The dimensional warehouse also substantially reduces complexity even without precomputation, because conformed keys and indexed facts simplify aggregation paths.

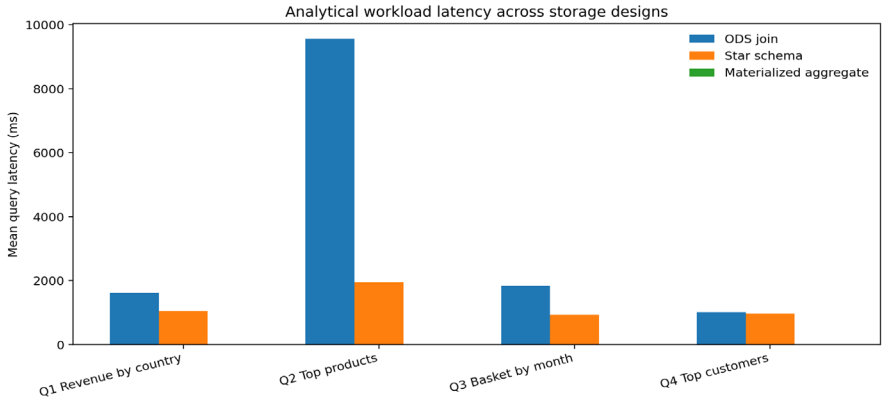

Figure 4: Analytical workload latency across operational joins, dimensional warehouse queries, and materialized aggregates.

These findings support RQ1 and RQ2. The warehouse is not merely a semantic cleanup layer; it materially changes the cost profile of analytical access. The difference is especially relevant in retail, where many managerial dashboards repeatedly issue similar group-by queries across calendar and product hierarchies.

6.2 Managerial descriptive insights from the warehouse

The curated warehouse also supports interpretable business insights. Revenue is highly concentrated in the United Kingdom, followed by the Netherlands, EIRE, Germany, and France. Among products, DOTCOM POSTAGE, REGENCY CAKESTAND 3 TIER, and PAPER CRAFT, LITTLE BIRDIE dominate revenue. The concentration pattern suggests that cross-border analysis and assortment optimization would benefit from explicit country-product cubes rather than undifferentiated global aggregates.

Average basket value is strongest in early 2011 and varies meaningfully across months, indicating the importance of calendar-aware analysis. Such signals are easy to miss when transaction data remain trapped in operational form without consistent temporal summarization.

6.3 Forecasting performance

Figure 5 compares actual revenue with the random forest and naive predictions over the test horizon. Although the random forest does not perfectly track peaks, it follows level changes more closely than the naive baseline. This result supports RQ3: once data are integrated and summarized coherently, even relatively straightforward models can produce materially stronger predictive signals than a persistence benchmark.

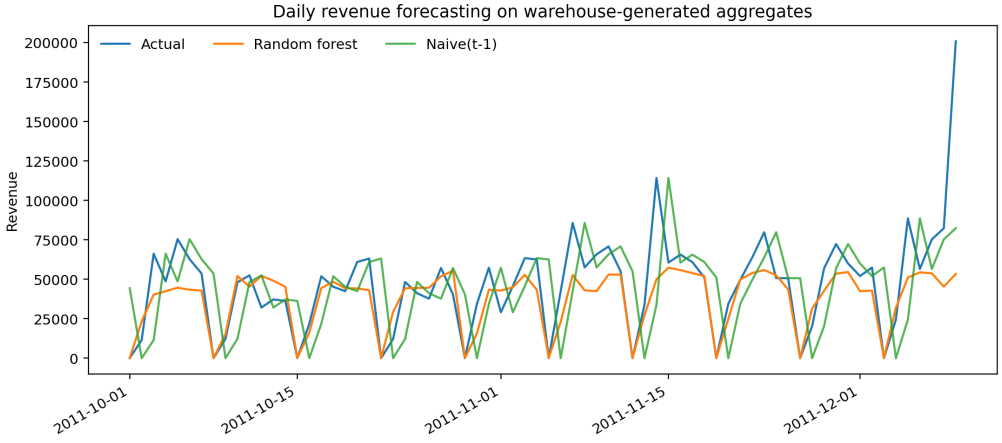

Figure 5: Daily revenue forecasting on warehouse-generated aggregates.

The forecasting exercise should not be interpreted as a claim that random forests are universally optimal for retail demand. Rather, it demonstrates that the warehouse provides a stable and semantically coherent feature space from

which forecasting can proceed reproducibly. This is a crucial architectural point. Many organizations debate model choice before stabilizing the data foundation on which those models depend.

6.4 Interpretation of empirical findings

Three broader findings emerge. First, data-quality interventions are not peripheral; they are integral to the success of the warehouse. Had cancellations, invalid prices, and duplicates remained in the fact table, both OLAP summaries and forecasting signals would have been distorted. Second, the dimensional warehouse already provides meaningful acceleration over operational joins, but materialized aggregates are essential for truly interactive analytical performance. Third, the warehouse creates a shared analytical substrate on which descriptive and predictive workflows can coexist.

7 Discussion

7.1 Why SwiftMart architecture is analytically justified

The proposed SwiftMart architecture is justified not only because it conforms to established warehousing theory, but because it addresses the exact pathologies described in the case. Duplicated data are mitigated through centralized conformance and surrogate-key mapping. Format inconsistency is handled in the ETL layer. Historical analysis becomes feasible because the fact table is time-aware by design. Departmental access needs are served through derived marts and summaries rather than through persistent silo re-creation.

In practical terms, this means SwiftMart can answer both stable and emergent questions from one analytical foundation. Stable questions include routine management dashboards, monthly country performance, product rankings, or customer revenue concentration. Emergent questions include campaign impact analysis, forecast model experimentation, or new cube materialization for a fast-changing business problem. In other words, architecture improves not only current reporting but also future adaptability.

7.2 Relation to Kimball and Inmon perspectives

The study adopts Kimball-style dimensional implementation, but the results also reinforce a broader point bottom-up marts are most effective when they are built with enterprise conformance in mind. A common criticism of mart-oriented approaches is that they can degenerate into local optimizations. The SwiftMart design avoids that failure mode by using shared dimensions and a unified fact grain. In this sense, the study occupies a pragmatic middle ground between Kimball's usability emphasis and Inmon's integration emphasis [4, 5].

7.3 Implications for data quality management

The data-quality outcomes are especially important for reviewers and practitioners. The experimental dataset, though public and widely used, still contains duplicates, cancellations, and missing identifiers. This underlines a fundamental lesson such as open access does not imply analytical readiness. In production retail environments, the problem is usually worse because additional sources introduce inconsistent codes, delayed refreshes, and governance exceptions. Therefore, a publishable retail analytics architecture should explicitly specify quality controls rather than simply assuming clean upstream data.

The operational metrics used duplicates removed, invalid rows rejected, missing values mapped, and curated row counts logged are intentionally simple, but they provide a minimum viable quality ledger that supports auditability. This aligns with the long-standing view that data quality must be measured in relation to analytical use rather than treated as an informal preprocessing step [17, 18].

7.4 Privacy, ethics, and trust

The SwiftMart case also raises privacy and trust concerns. Personalized retail analytics can improve service quality, but it can also intensify customer concerns if data reuse becomes opaque. Architecture therefore assumes privacy by design such as minimum necessary exposure of customer-level data, separation between personally identifiable information and analytics views, and role-based access controls [19]. These governance provisions were not executed on the open dataset because they are already anonymized, but they remain necessary for any real deployment.

7.5 Limitations

Several limitations should be acknowledged. First, the empirical evaluation uses one open dataset from an online retailer rather than a multi-source proprietary enterprise corpus. This constrains direct realism. However, the dataset is large, transactional, and representative enough to support meaningful warehouse instantiation. Second, the

benchmark is implemented on a single-node SQLite prototype. The absolute latency values would change under cloud-native engines, but the relative structural effect of schema design and materialization is still informative. Third, the open dataset does not include internal staff or commission fields; hence, the original SwiftMart bonus-calculation scenario is discussed architecturally rather than validated with proprietary HR data. Fourth, the forecasting exercise is illustrative rather than exhaustive and does not compare advanced sequence models.

7.6 Future work

Future research can extend the study in at least four directions. First, additional dimensions such as channel, promotion, salesperson, or supplier can be added where organizational data permit. Second, slowly changing dimensions and late-arriving fact handling can be modeled explicitly. Third, warehouse benchmarking can be repeated on columnar or cloud-native analytical engines to assess scalability under larger workloads. Fourth, privacy-preserving analytics mechanisms, including masking, differential access, or privacy-aware cube computation, can be evaluated directly [21].

8 Conclusion

This paper transformed a conceptual SwiftMart manuscript into a rigorous, experimentally grounded study of retail data warehousing and OLAP. The original case correctly identified the central analytical pain points of fragmented retail data such as duplication, heterogeneous formats, siloed ownership, and limited support for integrated reporting and advanced analytics. The present study extends that foundation by specifying a reproducible warehouse artifact, implementing a dimensional model, formalizing the ETL logic, and evaluating the design empirically on a real open-access retail dataset.

The findings are clear. A conformed dimensional warehouse materially improves analytical access over operational joins, and materialized summary tables make core OLAP workloads effectively interactive. The warehouse also supports predictive analytics by producing stable daily aggregates from which forecasting features can be derived. More broadly, the study shows that the path from fragmented operational data to reliable retail intelligence is not achieved through modeling alone; it requires the coordinated design of data quality controls, dimensional semantics, performance-aware summaries, and governance principles.

For practitioners, the study offers a concrete blueprint for moving from siloed reporting toward integrated analytics. For researchers, it provides a transparent benchmark-oriented case that links warehousing theory with empirical evaluation. For SwiftMart-like organizations, the main message is straightforward: historical retail data becomes strategically valuable only when they are modeled, governed, and exposed through an architecture built for analysis rather than transaction processing.

8.1 Representative benchmark workloads

The benchmark executed four representative SQL workload families:

i. monthly revenue by country
ii. top products by quarter
iii. average basket value by month
iv. top customers by year.

Each workload was executed against the normalized baseline, the dimensional warehouse, and the relevant materialized summary table. Repeated execution was used to estimate mean latency. The workload selection deliberately balances common managerial interest with differences in query shape.

BIBLIOGRAPHY

[1]. E. Aktas and Y. Meng, “An exploration of big data practices in retail sector,” Logistics, vol. 1, no. 2, pp. 12, 2017. doi: https://doi.org/10.3390/logistics1020012.

[2]. J. Aversa, T. Hernandez, and S. Doherty, “Incorporating big data within retail organizations: A case study approach,” Journal of Retailing and Consumer Services, vol. 60, pp. 102447, 2021. doi: https://doi.org/10.1016/j.jretconser.2021.102447.

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

[4]. R. Kimball and M. Ross, “The data warehouse toolkit: The definitive guide to dimensional modeling,” 3rd ed., Wiley, 2013.

[5]. W. H. Inmon, "Building the data warehouse," 4th ed., Wiley, 2005.
[6]. R. Sharda, D. Delen, and E. Turban, "Business intelligence, analytics, and data science: A managerial perspective," 4th ed., Pearson, 2018.
[7]. V. Harinarayan, A. Rajaraman, and J. D. Ullman, "Implementing data cubes efficiently," SIGMOD Record, vol. 25, no. 2, pp. 205–216, 1996. doi: https://doi.org/10.1145/235968.233333.
[8]. M. Golfarelli and S. Rizzi, "Data warehouse design: Modern principles and methodologies," McGraw-Hill, 2009.
[9]. R. Fildes, S. Ma, and S. Kolassa, "Retail forecasting: Research and practice," International Journal of Forecasting, vol. 38, no. 4, pp. 1283–1318, 2022. doi: https://doi.org/10.1016/j.ijforecast.2019.06.004.
[10]. 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, no. 3, pp. 197–208, 2012. doi: https://doi.org/10.1057/dbm.2012.17.
[11]. C. I. Papanagnou and O. Matthews-Amune, "Coping with demand volatility in retail pharmacies with the aid of big data exploration," Computers & Operations Research, vol. 98, pp. 343–354, 2018. doi: https://doi.org/10.1016/j.cor.2017.08.009.
[12]. J. A. Aloysius, H. Höhle, S. Goodarzi, and V. Venkatesh, "Big data initiatives in retail environments: Linking service process perceptions to shopping outcomes," Annals of Operations Research, vol. 270, no. 1-2, pp. 25–51, 2018. doi: https://doi.org/10.1007/s10479-016-2276-3.
[13]. D. Chen, "Online retail [dataset]," UCI Machine Learning Repository, 2015. doi: https://doi.org/10.24432/C5BW33.
[14]. S. Akter and S. F. Wamba, "Big data analytics in E-commerce: a systematic review and agenda for future research," Electronic Markets, vol. 26, no. 2, pp. 173–194, 2016. doi: https://doi.org/10.1007/s12525-016-0219-0.
[15]. M. Wibowo, S. Sulaiman, and S. M. Shamsuddin, "Machine learning in data lake for combining data silos," in International conference on data mining and big data, pp. 294–306, 2017. doi: https://doi.org/10.1007/978-3-319-61845-6_30.
[16]. A. Gandomi and M. Haider, "Beyond the hype: Big data concepts, methods, and analytics," International Journal of Information Management, vol. 35, no. 2, pp. 137–144, 2015. doi: https://doi.org/10.1016/j.ijinfomgt.2014.10.007.
[17]. 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: https://doi.org/10.1080/07421222.1996.11518099.
[18]. L. Ehrlinger and W. Wöß, "A survey of data quality measurement and monitoring tools," Frontiers in Big Data, vol. 5, pp. 850611, 2022. doi: https://doi.org/10.3389/fdata.2022.850611.
[19]. P. Voigt and A. Bussche, "The EU general data protection regulation (GDPR): A practical guide," Springer, 2017. doi: https://doi.org/10.1007/978-3-319-57959-7.
[20]. S. Beheshti-Kashi, H. R. Karimi, and K. D. Thoben, "A survey on retail sales forecasting and prediction in fashion markets," Systems Science & Control Engineering, vol. 3, no. 1, pp. 154–161, 2015. doi: https://doi.org/10.1080/21642583.2014.999389.
[21]. A. Cuzzocrea, L. Bellatreche, and I. Y. Song, "Data warehousing and OLAP over big data: Current challenges and future research directions," in Proceedings of the 16th international workshop on data warehousing and OLAP, pp. 67–70, 2013. doi: https://doi.org/10.1145/2513190.2517828.