Ecommerce Data Warehouse

End-to-end sales analytics pipeline

By: Dickens

Agenda

1. Project Overview & Architecture

High-level goals, scope and target outcomes.

2. Star Schema & Data Model

Dimension and fact design, grain, and key relationships.

3. Challenges Faced

Data alignment, joins, staging gaps and resolution approach.

4. Results & Reporting Output

Data quality, aggregates and sample visual output.

5. Business Insights

Top categories, seasonality and long-tail observations.

6. Key Learnings

Operational and technical takeaways for future pipelines.

Architecture & Star Schema

Layered architecture: RAW → STAGING → STAR_SCHEMA → REPORTING. The star schema centralises **fact_sales** with conformed dimensions (product, customer, date, store, channel) to support performant aggregation and flexible analysis.

Challenges Faced

Timestamp vs Date Mismatch

Joins failed when timestamp fields had different timezones or formats; required deterministic cast/normalisation.

Dimension Rebuilds

Empty staging runs caused dimension truncation; implemented guardrails and idempotent upserts.

Fact Table Collapsing

INNER JOIN logic removed rows when dimension lookups failed—resolved with LEFT JOINs plus integrity checks.

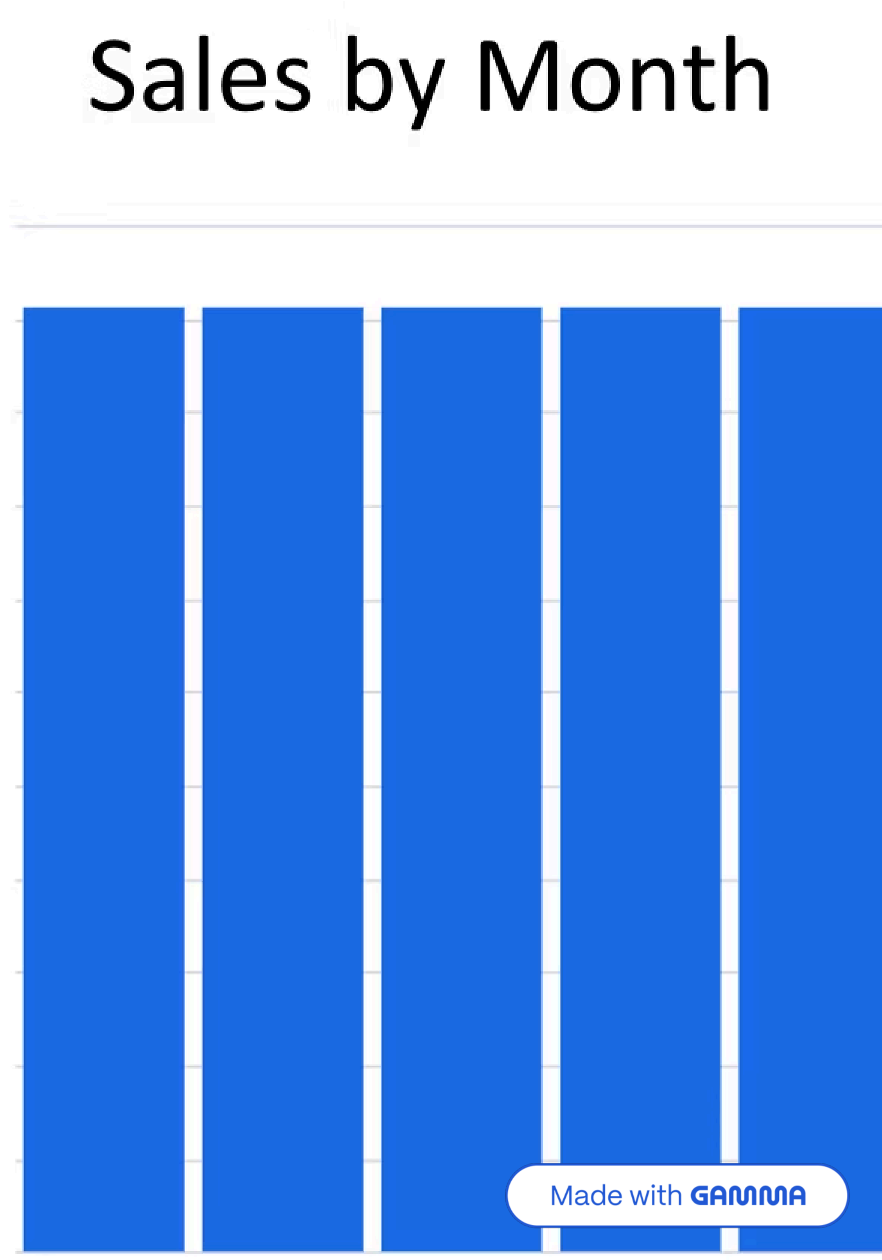
Source Misalignment

Schema drift after RAW upload required schema validation and automated alerts.

Resolution: implemented layer-by-layer validation, schema contracts, and automated dbt tests to detect and prevent regressions.

Results — Visual Summary

Cleaned and reconciled data now feeds reporting layers with clear lineage. Visual above shows sample dashboard output: product mix, revenue by channel and monthly trend.



Reporting Snapshot

Examples of outputs produced: daily sales trend, category performance, product-level top-N lists and a cube for cross-dimensional slicing. Dashboards updated via scheduled ELT and materialised aggregates for low-latency queries.

Technical Results

Fact & Aggregate Counts

fact_sales populated with 300 rows; sales_by_month generated 6 monthly aggregates; sales_cube_product_month produced 272 product–month combinations.

Validation

End-to-end pipeline fully validated with dbt tests, row counts and checksum comparisons across layers.

Performance

Materialised aggregates deliver sub-second responses for common queries; incremental models reduced compute costs.

Business Insights

- Revenue concentration: Electronics dominate — headphones and smartwatches lead sales and margin contribution.
- Stable monthly distribution: No extreme seasonality detected in the analysed window; consistent demand across months.
- Long-tail observed: A wide tail of low-volume SKUs contributes a non-trivial portion of catalogue revenue.
- Actionable: Prioritise top-performing SKUs for inventory and promotions; rationalise long-tail SKUs where carrying cost outweighs value.

Key Learnings

1

Dimension Completeness

Ensure dimensions are fully populated before fact ingestion; use referential integrity checks to avoid orphans.

2

Dependency Management

Explicit dbt dependencies and well-defined run order prevent downstream failures during partial loads.

3

Join Logic Matters

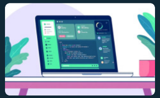
Prefer LEFT JOINs with validation guards for optional dims; INNER JOINs are only safe when referential integrity is guaranteed.

4

Layered Validation

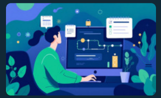
Automated checks across RAW → STAGING → STAR_SCHEMA → REPORTING ensure data quality and enable fast troubleshooting.

Next Steps & Recommendations



Automate Tests

Expand dbt tests to cover schema drift, row-level checks and business rules; integrate alerts into ops channel.



Scheduling & Monitoring

Implement orchestration with retry policies and SLA monitoring to reduce production incidents.

Business Partnership

Work with product and merchandising teams to action insights: focus on high-margin electronics and rationalise long-tail inventory.

Contact: Dickens — ready to support implementation and handover to analytics operations.