

# Reflection on the NordTech ETL Pipeline Project 🙌

## Project Overview ✨

This project involved designing and implementing an end-to-end ETL pipeline for NordTech order data, covering data extraction, exploratory analysis, transformation, sentiment enrichment, and loading into a relational database.

The primary objective was to transform raw, inconsistent transactional data into a reliable, analysis-ready dataset that supports KPI reporting and business insight generation.

The project was approached with a strong emphasis on modularity, reproducibility, and robustness to future data variations.

## Architectural Decisions 🏗️

### Modular ETL Design

The pipeline was intentionally split into separate modules (extract, transform, load, and sentiment) to enforce separation of concerns. This structure improves maintainability, simplifies debugging, and allows individual components to be reused or extended without affecting the entire system.

Such modularity reflects common industry practices in production ETL systems.

Configuration values such as file paths and database locations were centralized in a configuration module to ensure reproducibility and reduce hard-coded dependencies.

# Data Quality Challenges and Solutions

## Handling Real-World Messy Data

The raw dataset contained numerous real-world data quality issues, including:

- Mixed and inconsistent date formats, including Swedish month names
- Textual price values with currency symbols and formatting variations
- Inconsistent categorical labels caused by casing, abbreviations, and synonyms
- Duplicate records
- Invalid logical relationships (e.g., delivery dates preceding order dates)

Addressing these issues required designing robust, defensive cleaning functions that could tolerate malformed input while preserving as much valid information as possible.

## Date Normalization

Dates were parsed using flexible parsing logic to handle multiple formats and languages.

Additional validation was applied to detect and correct reversed order and delivery dates, ensuring temporal consistency across the dataset.

## Quantity (antal) Imputation

The quantity column contained numeric values embedded in text, unit annotations, and Swedish number words.

After normalization, missing quantities were imputed with a value of **1**, based on the business assumption that an order line without an explicit quantity typically represents a single purchased unit.

Using 0 would incorrectly remove revenue contribution, while higher values would inflate sales metrics.

## Customer Ratings (betyg) Handling 🌟

Ratings were validated to ensure they fell within the valid range of 1–5.

Missing values were imputed using the **median** rather than the mean, as ratings are ordinal rather than continuous.

The median preserves the central tendency without introducing unrealistic decimal values or being skewed by extreme ratings.

## Missing Value Strategy 🧩

Missing values were handled based on semantic meaning rather than a single global rule:

- Geographic fields (e.g., region) were kept as null to avoid creating artificial categories.
- Transactional categorical fields (e.g., payment method) used an explicit "unknown" category to preserve row completeness for aggregation.
- Textual placeholders such as "nan" in review text were converted to true missing values to prevent contamination of NLP analysis.

## Sentiment Analysis Integration 🤖

A multilingual BERT model was selected for sentiment analysis due to its contextual understanding and support for Swedish and English text.

Compared to lexicon-based approaches, BERT provides more reliable sentiment classification for short, informal customer reviews.

Sentiment analysis was kept separate from the core ETL pipeline to preserve pipeline purity and allow sentiment enrichment to be recomputed or replaced without reprocessing the entire dataset.

Model outputs were mapped into three interpretable categories: **positive**, **neutral**, and **negative**, making them suitable for KPI reporting.

## Database Integration and Reproducibility

The cleaned dataset was loaded into a local SQLite database, enabling downstream analysis to be performed directly on validated, standardized data rather than raw CSV files.

Validation data was appended rather than overwritten to simulate incremental data ingestion and ensure the pipeline remained robust when exposed to new data.

Reading from SQL in the KPI notebook reinforced the separation between data preparation and analysis layers.

## Business Impact

The cleaned and standardized dataset enables reliable KPI computation, including delivery performance, return rates, customer satisfaction, and sentiment trends.

By combining structured transactional data with sentiment analysis, NordTech gains a more nuanced view of customer experience that goes beyond numerical ratings alone.

These improvements directly support more informed decision-making and reduce the risk of misleading insights caused by poor data quality.

## Limitations and Future Improvements

With additional time, the pipeline could be further improved by:

- Implementing batch sentiment inference for performance optimization

- Adding structured logging instead of print-based debugging
- Writing unit tests for individual transformation functions
- Introducing a command-line interface or workflow orchestrator (e.g., Prefect or Airflow)

These enhancements would move the pipeline closer to production-grade standards.

## Conclusion

This project reinforced the importance of thoughtful data preparation as the foundation of reliable analytics.

Designing a robust ETL pipeline required not only technical implementation but also careful consideration of business assumptions, data semantics, and downstream use cases.

Overall, the project significantly strengthened my ability to design maintainable data pipelines and apply data engineering principles in a realistic, end-to-end workflow.