ETL Process Report - UCI Online Retail Dataset

Overview

This report documents the ETL (Extract, Transform, Load) process applied to the UCI Online Retail dataset, containing transaction records for a UK-based online retailer between December 2010 and December 2011.

Environment

- IDE: Visual Studio Code (VSCode)
- Language: Python (Jupyter Notebook .ipynb)
- Database: SQLite (retail_dw.db)

Dataset

- Source: UCI Machine Learning Repository - Online Retail
- Size: 541,909 rows
- Columns: InvoiceNo, StockCode, Description, Quantity, InvoiceDate, UnitPrice, CustomerID, Country

---

ETL Steps

1. Extraction

- Loaded online_retail.csv into a Pandas DataFrame.
- Converted InvoiceDate to datetime.
- Dropped rows with missing CustomerID or Description.

Logs:

2025-08-12 17:30:38,009 - INFO - Extraction: 541909 rows loaded from online_retail.csv

2025-08-12 17:30:38,286 - INFO - Transformation: After removing missing values -> 406829 rows

---

2. Transformation

- Removed Outliers: Dropped rows with Quantity < 0 or UnitPrice <= 0.
- New Column: TotalSales = Quantity * UnitPrice.
- Filtered to Last Year:
  Latest invoice date: 2011-12-09
  Cutoff date: 2010-12-09
  Rows retained: 384,529
- Customer Summary: Aggregated by CustomerID (total purchases, country).
- Time Dimension: Extracted date, quarter, month, and year.

Logs:

2025-08-12 17:30:38,339 - INFO - Transformation: After removing outliers -> 397884 rows

2025-08-12 17:30:38,474 - INFO - Transformation: Latest invoice date is 2011-12-09 12:50:00, cutoff date is 2010-12-09 12:50:00

2025-08-12 17:30:38,475 - INFO - Transformation: After last-year filter -> 384529 rows

2025-08-12 17:30:38,580 - INFO - Transformation: Customer summary has 4277 rows

2025-08-12 17:30:38,622 - INFO - Transformation: Time dimension has 16630 rows

---

3. Loading

- Created SQLite database retail_dw.db.
- Inserted data into:
  SalesFact - 384,529 rows
  CustomerDim - 4,277 rows
  TimeDim - 16,630 rows

Logs:

2025-08-12 17:30:41,845 - INFO - Loading: Inserted 384529 rows into SalesFact

2025-08-12 17:30:41,846 - INFO - Loading: Inserted 4277 rows into CustomerDim

2025-08-12 17:30:41,847 - INFO - Loading: Inserted 16630 rows into TimeDim

2025-08-12 17:30:41,848 - INFO - ETL completed successfully

Special Consideration - Exam Requirement Adaptation

The exam required filtering to the "last year" assuming August 12, 2025 as the current date. This dataset only covers 2010-2011, so applying the filter literally would result in zero rows.

To preserve analytical value:

- We used the dataset's latest invoice date (2011-12-09) as the "current date".
- Applied a one-year filter from that date (2010-12-09), keeping 384,529 rows.

Screenshots of Loaded Data

(Screenshots are referenced in the markdown version)

Deliverables

- Script: etl_retail.ipynb / etl_retail.py
- Database: retail_dw.db
- Report: This PDF document
- Screenshots: Samples of each table

Conclusion

The ETL pipeline:

1. Extracted, cleaned, and transformed the Online Retail dataset.
2. Created one fact table and two dimension tables for analytics.
3. Adapted "last year" logic to fit the dataset while following the task's intent.