Project Documentation: Sales ETL Pipeline

1. Project Overview

This project implements an ETL pipeline to extract, validate, transform, and model sales data to ensure it is clean, consistent, and ready for analytical and reporting use cases.

2. ETL Process

2.1 Extract Phase

- Sales data was provided in **flattened JSON format**.
- The dataset was loaded into a **Pandas DataFrame**.
- **UTF-8 encoding** was applied to avoid character encoding issues.

2.2 Schema Validation

- Each record was validated against a predefined schema.
- Records were checked for unmapped or unexpected fields.
- **No schema violations were detected.**

Result:
The dataset fully conforms to the expected structure.

2.3 Data Exploration & Profiling

Exploratory Data Analysis (EDA) included:

- Dataset shape and data type inspection
- Null value analysis
- Duplicate detection

Key Findings:

- OrderDate was not stored as a proper datetime type.
- Duplicate records were present.
- Name, Education, and Occupation contain ~63% null values.

2.4 Distribution Analysis – Net Price

- Distribution was analyzed using boxplots and histograms.
- NetPrice showed a **right-skewed distribution** with high-value transactions.

- These values were retained as valid, representing premium products aligned with business expectations.

2.5 Data Quality & Redundancy Checks

- Column names and string values were checked for whitespace issues (none found).
- Redundant and low-quality attributes were identified:
    - Color duplicated Subcategory values and lacked semantic meaning.
    - CustomerKey and CustomerCode represented the same entity (1:1 relationship).

Decisions:

- Color and CustomerCode were removed.
- CustomerKey was retained due to better performance, join efficiency, and stability.

3. Transformation Phase

- Column names were standardized for consistency.
- OrderDate was converted to a proper datetime format.
- Duplicate handling was reviewed carefully:
    - Due to the absence of a unique order identifier or timestamp, duplicates were **retained** to avoid removing valid transactions.
- Columns with high null ratios and low analytical value were dropped:
    - Education
    - Occupation

4. Data Modeling & Loading Phase

The refined dataset was modeled using a **dimensional approach** to support analytical queries.

Product Dimension (dim_product)

- Attributes: ProductKey, ProductName, Brand, Subcategory, Category

Customer Dimension (dim_customer)

- Attributes: CustomerKey, Name
- Non-essential attributes were excluded for performance and relevance.

Geography Dimension (dim_geography)

- Attributes: City, State, CountryRegion, Continent
- A surrogate key (GeographyKey) was generated.

Date Dimension (dim_date)

- Derived from OrderDate
- Includes DateKey, Year, Month, MonthName, and Quarter

Sales Fact Table (fact_sales)

- Foreign keys: ProductKey, CustomerKey, GeographyKey, DateKey
- Measures: Quantity, NetPrice, SalesAmount

All tables were exported as CSV files and validated for row consistency.

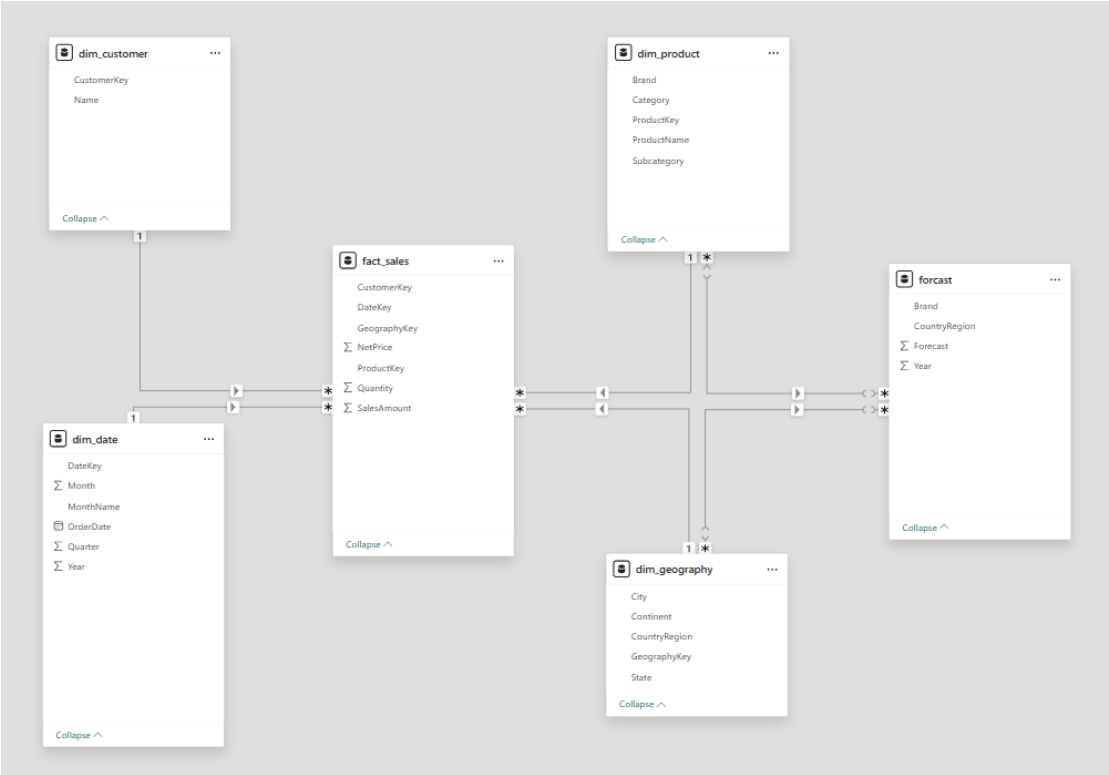
5. Forecast Data Integration

- A clean forecast dataset covering **2008–2009** was integrated.
- Forecast values are provided by:
    - CountryRegion
    - Brand
    - Year
- No data quality issues were detected.

Use Cases:

- Forecast vs actual comparison
- Trend analysis
- Regional and brand-level performance evaluation

6. Data Model



Overall Performance (Global)

Total Sales: $42.64M

Year-over-Year decline:

- 2008: $23.1M
- 2009: $19.55M
- YoY Growth: –15.38%

The decline is structural, not seasonal, and visible across most markets and categories.

United States

Largest market, but declining

Total Sales: $34.33M

YoY Growth: –18.72%

Top categories:

- Home Appliances (largest contributor)
- Computers
- Cameras & Camcorders

Key Insight:

- Heavy reliance on Home Appliances makes the US market highly sensitive to economic slowdowns.
- 2009 sales underperformed 2008 in most months.

Germany

Small market with the sharpest decline

Total Sales: $6.24M

YoY Growth: – 49.56%

Actual sales in 2009 were significantly below forecast.

Key Insight:

- Germany is a strategic weakness.
- Requires pricing, cost optimization, or reduced investment.

China

Emerging market with strong potential

Total Sales: $2.07M

2009 Sales ≈ $2.06M

Extremely high YoY growth (inflated due to very low 2008 base).

Key Insight:

- Strong demand in Home Appliances and Computers.
- Market is in early growth phase, not saturation.

Product Performance

Best-selling products:

- Washers & Dryers
- Refrigerators

Sales are highly concentrated in the top 10 products.

Risk:

- Overdependence on a limited product set increases revenue volatility.

Forecast vs Actual (2009)

Actual sales missed forecasts in all regions.

Largest gaps observed in:

- United States
- Germany

Key Insight:

- Forecasting models were overly optimistic and did not fully account for market contraction.

Monthly Sales Trends

2008: Strong peaks in Q2 and Q3

2009: Lower peaks and weaker momentum, especially in Q4

Seasonality Insight:

- Best-performing periods remain Q2–Q3
- End-of-year performance weakened in 2009

Recommendations

1. **Reduce exposure to Germany** or re-evaluate go-to-market strategy
2. **Increase investment in China** (marketing, distribution, product expansion)
3. **Diversify product portfolio** to reduce concentration risk
4. **Rebuild forecasting models** with conservative economic assumptions
5. **Focus campaigns around Q2–Q3**, where demand is strongest