

# End-to-End Supply Chain Data Architecture and Analytical Framework

**Subject:** Detailed Technical Documentation and SQL Analytical Findings

---

## 1. Executive Summary

This report details the technical implementation of a data engineering and analytics solution designed to process the DC Supply Chain dataset. The project involves a two-stage architecture:

- **ETL (Extract, Transform, Load) Pipeline:** A Python-based workflow that ingests raw legacy data, enforces data quality standards, performs feature engineering, and migrates the structured data to a PostgreSQL environment.
  - **Analytical Suite:** A comprehensive SQL script comprising advanced queries to derive actionable insights across logistics, finance, customer segmentation, and temporal performance.
- 

## 2. Phase I: Data Engineering Pipeline (Python)

**File Reference:** DCSC.ipynb

The Python workflow serves as the foundational layer, transforming raw CSV inputs into a normalized database schema.

### 2.1 Data Ingestion and Schema Initialization

- **Protocol:** The raw dataset (DCSupplyChainDataset.csv) was ingested using the Pandas library.
- **Encoding Handling:** A latin-1 encoding standard was applied to resolve character encoding conflicts inherent in the source file.
- **Dimensionality:** The initial dataset was inspected for structural integrity, utilizing df.info() and df.head() to assess column data types and null value distribution.

### 2.2 Data Cleaning and Standardization

To ensure compatibility with the downstream PostgreSQL database, rigorous standardization protocols were applied:

- **Nomenclature Normalization:** All column headers were converted to `snake_case` (lowercase with underscores). This prevents syntax errors in SQL queries often caused by spaces or mixed-case headers.

- **Semantic Renaming:** Ambiguous column names were mapped to descriptive identifiers:
  - days\_for\_shipping\_(real) -> actual\_shipping\_days
  - days\_for\_shipment\_(scheduled) -> scheduled\_shipping\_days
  - shipping\_date\_(dateorders) -> shipping\_date
- **Privacy & Optimization:** High-cardinality text fields (product\_description) and Sensitive PII (Personally Identifiable Information) such as customer\_email, customer\_password, and customer\_zipcode were programmatically removed to optimize query performance and ensure data privacy.

## 2.3 Data Integrity & Feature Engineering

- **Date Parsing:** The order\_date and shipping\_date fields were cast to datetime objects to facilitate temporal arithmetic.
- **Logic Validation:** A data quality check was implemented to verify the consistency of shipping records. A calculated duration (shipping\_date - order\_date) was compared against the recorded actual\_shipping\_days to identify discrepancies.
- **Metric Generation:** A new derived metric, delay\_variance, was calculated by subtracting scheduled\_shipping\_days from actual\_shipping\_days. This feature allows for the quantification of logistics efficiency.
- **Text Normalization:** Categorical fields (City, State, Country) underwent string manipulation (stripping whitespace, title casing) to unify inconsistent entries (e.g., merging "India" and "india").

## 2.4 Database Migration

- **Connector:** An interface was established using SQLAlchemy to connect the Python environment to a local PostgreSQL instance (DCSC database).
  - **Load Strategy:** The cleaned dataframe was written to the supply\_chain table using a replace strategy, ensuring the analytical database always reflects the most current state
  - of the pipeline.
- 

# 3. Phase II: Strategic Data Analysis & Findings (SQL)

**File Reference:** DC Supply Chain.sql

The SQL suite performs multi-dimensional analysis using advanced query structures. Below are the specific findings from all queries included in the analysis.

### 3.1 Logistics & Operational Efficiency

#### Query 1: Geographic Bottlenecks

- **Objective:** Identify locations with the highest unpredictability in shipping delays.
- **Findings:** The city of **Thiais, France**, exhibits the highest delay variance (7.60), making it the most unpredictable node. It is followed by **Abbeville, France** (7.09) and **Tinaquillo, Venezuela** (7.08).

#### Query 2: Shipment Mode Reliability

- **Objective:** Compare "Late Delivery" vs. "Shipping on time" rates by mode.
- **Findings:** **First Class** shipping is critically unreliable with a **95.3%** late delivery rate. **Standard Class** performs better but still has a 38.1% late rate. **Same Day** shipping is the most reliable option, though it still only achieves ~50% on-time performance.

### 3.2 Financial Health Assessment

#### Query 3: Market Profitability

- **Objective:** Rank markets by net profit margin.
- **Findings:** **USCA** (US & Canada) leads with an **11.14%** profit margin, followed closely by **Africa** (10.99%) and **LatAm** (10.93%).

#### Query 4: The "Category Star" (Department Revenue)

- **Objective:** Identify top and bottom performing departments by revenue.
- **Findings:** The **Fan Shop** is the top revenue generator (\$17.1M), followed by **Apparel** (\$7.9M). The lowest revenue departments are **Book Shop** (\$12.6k) and **Pet Shop** (\$41.5k).

#### Query 5: The "Loss Leader" Analysis

- **Objective:** Find categories with high sales but low profit.
- **Findings:** The **Strength Training** category is a major loss leader with a margin of only **0.60%**. **As Seen on TV!** products also underperform significantly with a 3.47% margin.

### 3.3 Customer Segmentation & Risk

#### Query 6: Segment Value Analysis

- **Objective:** Determine revenue contribution and Average Order Value (AOV) by segment.
- **Findings:** The **Consumer** segment drives the most revenue (\$19.1M). However, AOV is consistent across all segments: Consumer (\$204), Corporate (\$204), and Home Office (\$202).

### **Query 7: Customer Lifetime Value (CLV)**

- **Objective:** Identify the Top 1% of customers by total spend.
- **Findings:** There are **207** customers in the top 1% tier, with a spending threshold of over **\$6,471**. The top individual customer (ID 791) has spent **\$10,524** to date.

### **Query 8: Fraud Geography**

- **Objective:** Pinpoint cities with the highest percentage of suspected fraud (min. 50 orders).
- **Findings:** **Villemonble, France** has the highest fraud rate at **17.9%**. Other high-risk areas include **Rugby, UK** (16.9%) and **Chaguanas, Trinidad & Tobago** (16.4%).

## **3.4 Temporal Dynamics (Time-Series)**

### **Query 10: MoM and YoY Growth**

- **Objective:** Track monthly sales and profit growth.
- **Findings:** A severe downturn occurred in late 2017. Sales plummeted **-41.6%** Month-over-Month in November 2017 and continued to decline by **-34.2%** in January 2018. Year-over-Year profit in January 2018 was down **-70.5%**.

### **Query 11: Seasonality Patterns**

- **Objective:** Analyze average sales per month to find seasonal peaks.
- **Findings:** **October** is the strongest month for sales, with an average of **\$244.79** per order, significantly outperforming the annual average of ~\$200. This indicates a pre-holiday surge.

---

## **4. Conclusion**

The implemented codebases demonstrate a professional-grade data workflow. The Python pipeline ensures that data entering the database is clean, consistent, and enriched with valuable features. The SQL analysis has successfully identified critical operational bottlenecks in **France**, financial risks in the **Strength Training** category, and a concerning **revenue decline** in the most recent quarter, providing a solid foundation for data-driven strategic adjustments.

---

## **Report Prepared by**

Anshu Raj  
ar.anshu2611@gmail.com