SQLite ETL Pipeline Simulation - Project Report

Student Name: Y. Harshitha
Project Duration: 2 Weeks
Technology: SQLite Database Management
Date: September 2025

Introduction

This project implements a comprehensive **ETL (Extract, Transform, Load) pipeline** with SQLite. The pipeline simulates real-world retail data workflows for processing customer, product, and sales data, demonstrating advanced SQL and data engineering best practices including validation, transformation, analytics, and business reporting.

Abstract

Over 80 raw data records across customers, products, and sales transactions were processed. The pipeline achieved a 92% data quality rate for customer data and 96% for products, showcasing automated validation and cleansing. Key deliverables include development of audit logging, business rules enforcement, and advanced analytics—like customer segmentation and sales trend analysis.

Tools Used

- **SQLite 3.x:** Light, serverless database engine
- **DB Browser for SQLite:** Graphical database interface
- **SQL Technologies:** DDL, DML, triggers, joins, subqueries, window functions, aggregations, indexing
- **Other:** SQL script version control, database modeling, automated validation/testing, documentation

Steps Involved in Building the Project

1. **Database Architecture:** Multi-layer schema with staging, production, audit, and analytics tables; 12+ tables, 13 indexes, 6 triggers for validation and automation.
2. **Data Extraction & Staging:** Imported sample data (25 customers, 25 products, 32 sales), preserved data for audit integrity.
3. **Data Transformation & Validation:** Applied email/price/quantity checks, duplicate removal, referential integrity, and logged invalid records for review.
4. **Data Loading:** Loaded validated data into production tables; added computed fields (profit margins, totals, lifetime value) with robust error handling and rollback.
5. **Analytics & Reporting:** Generated sales trends (7 months, revenue $949–$11,096), customer segmentation (VIP, regular, new), product/category profitability, and sales rep analytics.
6. **Quality Assurance:** Automated validation, performance testing, audit trails, and detailed documentation ensure data and process integrity.

Conclusion

The project demonstrates advanced data engineering skills with SQLite—real-time validation, transformation, analytics, indexing, and professional documentation. 82 raw records were transformed into 70 clean, validated, and auditable ones, generating valuable business intelligence and ensuring data quality for reporting. The pipeline is robust, automated, and suitable for portfolio or early-career data engineering roles.

GitHub: https://github.com/Harshithayalipi
Contact: yalipiharshitha@gmail.com

LinkedIn: https://www.linkedin.com/in/yalipi-harshitha-711735254