📄 Enterprise Sales Data Engineering – Project Report

---

1. Introduction

This report presents an end-to-end **Data Engineering project** focused on building an analytics-ready sales data pipeline.
The project simulates a real-world scenario where raw transactional data is ingested, transformed, modeled, and made available for business analytics and visualization.

The objective was to demonstrate practical data engineering skills including ETL development, data modeling, SQL analytics, and dashboard creation.

---

2. Problem Statement

Organizations often receive sales data from multiple source systems in raw and inconsistent formats.
Such data cannot be directly used for analytics or decision-making.

The key challenges addressed in this project were:

- Cleaning and standardizing raw data
- Designing an efficient analytical data model
- Enabling scalable analytical queries
- Delivering meaningful business insights through dashboards

---

3. Data Sources

The project uses three raw CSV datasets:

- **Customers data**: Customer identifiers, names, countries, and signup dates
- **Products data**: Product details, categories, and base prices
- **Transactions data**: Sales transactions including quantity, price, and transaction date

These datasets simulate operational source systems commonly seen in retail or e-commerce environments.

---

4. ETL Pipeline Design

The ETL pipeline was implemented using **Python and Pandas** and consists of the following steps:

1. **Data Ingestion**
   Raw CSV files were loaded into Python for processing.
2. **Data Cleaning & Validation**
   - Removal of duplicates
   - Handling missing values
   - Standardization of date formats
   - Validation of foreign key relationships
3. **Data Transformation**
   - Creation of dimension tables for customers and products
   - Creation of a fact table for transactions
   - Calculation of derived metrics such as total revenue
4. **Data Scaling**
   - Transactional data was programmatically scaled to **50,000 rows** to simulate real-world data volume.
5. **Data Loading**
   - Cleaned and transformed data was loaded into a MySQL data warehouse.
   - Processed datasets were also stored as CSV files for analytics and reporting.

---

5. Data Modeling

A **Star Schema** was designed to support analytical workloads.

Dimension Tables

- **dim_customer** – customer details
- **dim_product** – product and category details

Fact Table

- **fact_transactions** – transactional sales data including quantity, price, and revenue

This schema enables efficient joins and fast aggregation queries.

---

6. Data Warehouse & SQL Analytics

The transformed data was stored in a **MySQL data warehouse**.
Analytical SQL queries were written to answer business questions such as:

- Monthly revenue trends
- Revenue by product category
- Top customers by total spending

These queries validated that the warehouse is analytics-ready.

---

7. Business Intelligence Dashboard

A **Power BI dashboard** was built using the processed datasets to visualize key insights:

- **Monthly Revenue Trend** – to understand sales performance over time
- **Revenue by Product Category** – to identify high-performing categories
- **Top Customers by Revenue** – to highlight valuable customers

The dashboard provides a clear, executive-level view of business performance.

---

8. Tools & Technologies Used

- **Python** – Data ingestion, cleaning, transformation, and scaling
- **Pandas & NumPy** – Data processing
- **MySQL** – Data warehouse
- **SQL** – Analytical queries
- **Power BI** – Data visualization and reporting

---

9. Challenges Faced

- Handling inconsistent raw data formats
- Designing a scalable star-schema model
- Scaling small datasets to realistic volumes
- Managing Power BI connectivity limitations with MySQL

These challenges were resolved using best practices commonly applied in real data engineering projects.

---

10. Conclusion

This project successfully demonstrates an end-to-end **Data Engineering workflow**, from raw data ingestion to business insights.

It showcases practical skills in ETL development, data modeling, SQL analytics, and dashboard creation, reflecting real-world data engineering responsibilities.

---

11. Future Enhancements

- Incremental and automated data loading
- Cloud-based data warehouse integration
- Orchestration using scheduling tools
- Advanced performance optimization