

# Project Proposal: E-commerce Data Warehouse

Leonardo Ricca 2211129

## Chosen Type of Project

Data Warehouse (DW)

## Description of Dataset

**Dataset:** Brazilian E-Commerce Public Dataset by Olist

**Source:** Kaggle - <https://www.kaggle.com/datasets/olistbr/brazilian-ecommerce>

**Size:** 9 interconnected CSV files (~100MB total)

**Domain:** Real commercial data from Brazilian e-commerce marketplace Olist (2016-2018)

### Dataset Components:

- Orders (99,441 records)
- Order Items (112,650 records)
- Customers (99,441 records)
- Products (32,951 records)
- Sellers (3,095 records)
- Payments (103,886 records)
- Reviews (99,224 records)
- Geolocation (1M+ records)
- Product Category Translation (71 records)

## Brief Description of Intended Work

This project will implement a complete **star schema data warehouse** for e-commerce sales analysis using **PostgreSQL**. The work involves: (1) **ETL process development** in Python to extract, clean, and integrate the 9 source datasets, handling data quality issues and implementing business rules; (2) **Dimensional modeling** with fact table (sales transactions) and 5 dimension tables (time, customer, product, seller, payment); (3) **Data enrichment** through geographic analysis using geolocation data and category translations from Portuguese to English; (4) **OLAP query implementation** using pgAdmin supporting roll-up, drill-down, slice, and dice operations for business intelligence; (5) **Performance optimization** with appropriate indexing and foreign key constraints. The final deliverable will be a fully functional data warehouse supporting complex analytical queries for sales performance, customer behavior, geographic analysis, and product category insights, demonstrating practical application of dimensional modeling principles to real-world e-commerce data.