Building a 3-Tier Data Lakehouse for Mobility Analysis in Spain

Project Overview

In this project students, acting as data engineers, to design and implement a 3-tier data lakehouse architecture to ingest, process, and analyze mobility data from Spanish public sources. The primary objective is to create a robust, scalable data infrastructure that supports transport domain experts in deriving insights for urban mobility and planning. The architecture will follow the classic lakehouse tiering model: Bronze (raw/staging layer for ingested data), Silver (core layer for cleaned, transformed, and integrated data), and Gold (mart layer for aggregated, business-ready data products). Data zonification will be applied at three levels: District (census districts), Municipal (municipalities), and GAU (as defined in the Ministry's zoning schemas).
The lakehouse will leverage public domain data to enable answers to two key business questions posed by transport experts. The focus is on big data tools for handling large-scale datasets, with minimal emphasis on domain-specific modeling, the data infrastructure is there to support transport experts that will refine analyses as needed. Deliverables include a code repository with high-quality implementation and documentation for the data infrastructure based on DuckLake (for ACID-compliant storage and versioning) and DuckDB (for efficient SQL analytics), along with a summary article outlining how the infrastructure supports the use cases.

Objectives

- **Data Ingestion and Architecture**: Build a scalable lakehouse as the core deliverable to handle mobility, demographic, and economic data, ensuring it supports downstream analytics by transport experts.

- **Data Processing and Integration**: Implement ELT processes to prepare data across tiers, with simple aggregations and integrations at multiple zonification levels.

- **Use Case Support**: Provide basic analytical examples (e..g., pattern identification and a simple gravity model) to demonstrate how the infrastructure enables transport experts to answer business questions.

- **Documentation**: Produce a code repository and summary research paper to illustrate the infrastructure's value, with substantiated implementation details.

Data Sources

Data must be sourced exclusively from public domain repositories. Focus on scalability for big data volumes (e.g., daily OD matrices can exceed billions of records). From the MITMA data, students must focus on the Estudios Básicos datasets.

1. **Spanish Ministry of Transport, Mobility and Urban Agenda (MITMA) Open Data**:
   - URL: https://www.transportes.gob.es/ministerio/proyectos-singulares/estudios-de-movilidad-con-big-data/opendata-movilidad
   - Key Datasets:
     - **Estudios Básicos**: Daily resident mobility at different zoning/census level, including origin-destination (OD) trip matrices (hourly breakdowns where available), overnight stays, and population counts. Coverage: Full daily data for 2022–2024; for 2025, typical week per month, singular days (e.g., holidays like Semana Santa or bridges). Anomalies noted in data due to incidents (e.g., missing days in October/November 2023 and 2024; antenna registries provided for specific events like Valencia October 2024).
     - Formats: Primarily compressed CSV matrices, with metadata and zoning files (shapefiles or CSVs for districts, municipalities, GAUs).
     - Reference Year: Use 2023 as the baseline for a "typical" year, avoiding anomalies (e.g., data gaps in 2023–2024 due to incidents).

2. **Spanish National Statistics Institute (INE)**:
   - URL: https://www.ine.es/en/ (INEbase portal for downloads).
   - Key Datasets:
     - **Population Demographics**: Official population figures by municipality and census district (e.g., "Continuous Register Statistics" and "Municipal Register Revision"). Includes age, gender, nationality, and density.
     - **Economic Indicators**: GDP, employment rates, income levels, and business activity by municipality/region (e.g., "Spanish Regional Accounts" and "Economically Active Population Survey"). Use NUTS-3 level data for alignment with zones.
     - Formats: CSV; shapefiles for spatial boundaries.
     - Alignment: Census districts match MITMA's district zoning; aggregate to municipal/GAU as needed.

Additional Derived Data:

- **Distances**: Compute Euclidean or network distances between zone centroids using geospatial libraries (e.g., via DuckDB spatial extensions).

Data Lakehouse Architecture

The main focus of the project is to implement a 3-tier lakehouse using DuckLake for storage (on local/object storage for simulation) and DuckDB for processing/querying. This setup enables ACID transactions, schema evolution, and efficient analytics on large datasets, providing a reliable foundation for transport experts.

- **Bronze Layer (Staging/Raw)**:
  - Purpose: Ingest raw data without transformation; store as-is for auditability.
  - Implementation: Use DuckLake tables to partition data by date, zone level (district/municipal/GAU), and source (MITMA/INE). Ingest via Python with DuckLake to handle CSVs.
  - Example Tables: bronze_mitma_od_daily.
- **Silver Layer (Core/Cleaned)**:
  - Purpose: Clean, standardize, and integrate data; apply transformations like null handling, data type casting, anomaly cleaning, and joins across sources.
  - Implementation: Use DuckDB to query Bronze, perform ELT (e.g., normalize units, map zones), and write back to DuckLake. Handle big data scale with partitioning and incremental loads.
  - Example Tables: silver_integrated_od (joined OD matrices with demographics), silver_zone_metrics (population, economy, distances per zone pair).
- **Gold Layer (Mart/Aggregated)**:
  - Purpose: Business-ready views for analytics; aggregates, precomputed models.
  - Implementation: DuckDB for complex queries (e.g., aggregations); store results in DuckLake for consumption.
  - Example Tables: gold_typical_day_patterns (hourly OD aggregates by pattern type), gold_infrastructure_gaps (gravity model outputs with mismatch scores).

Tools and Tech Stack:

- Storage: DuckLake (with Neon-Postgres and S3 for the cloud for the final

version, or DuckDB and local file storage for quick local iterations).

- Processing: DuckDB (SQL-based analytics).
- Orchestration: Use Airflow for ELT pipelines.
- Visualization: Integrate with tools like Matplotlib in notebooks for article figures.

Methodology

To demonstrate the lakehouse's utility, implement basic data-driven methods in the Gold layer to support transport experts in answering the business questions. Keep domain specifics minimal—focus on data a data infrastructure that supports domain expert users.

Business Question 1: Typical Day in Mobility for a Reference Year (2023)

- **Approach**: Aggregate MITMA OD matrices in the Gold layer to characterize average hourly origin-destination flows and demand.
  - Sub-Question: Identify basic mobility pattern types (e.g., weekdays, weekends, holidays) via simple clustering.
  - Example: Apply k-means clustering on normalized hourly flow data (per zone level) using DuckDB and Python libraries. Use features like total trips and peak hours; determine clusters (e.g., 3–5) via basic metrics.
  - Output: Pre-compute average hourly OD matrices and demand totals per pattern, aggregated across zones (district municipal GAU).

Business Question 2: Where is Transport Infrastructure Most Lacking?

- **Approach**: Prepare data for a simple gravity model to estimate potential demand, then compare to actual demand from MITMA.
  - **Gravity Model Example**: Use a basic formula like

$$T_{ij} = k \cdot \frac{P_i \cdot E_j}{d_{ij}^2}$$

    where T_{ij} is estimated trips between zones i and j, P_i is population at i, E_j is economic activity at j, and d_{ij} is distance. Fit simple parameters (e.g., k) using the data.

  - **Assessment**: Compute a mismatch ratio between actual and estimated trips. Aggregate to rank zones by service level (weighted by population/economy), identifying best- and worst-served areas.

Project Staging and Assessment

The project will be developed following an agile methodology, structured in three main sprints. At the end of each sprint, teams will present their results and be assessed.

- **Sprint 1: Schema Design and Prototyping**
  - **Task**: Propose the basic data schemas for the Bronze, Silver, and Gold tiers.
  - **Task**: Develop and test basic Proof of Concept SQL queries for transformations (Bronze -> Silver, Silver -> Gold) using a provided subset of the data spanning a week.
  - **Assessment**: Review of schema design (normalization, partitioning strategy) and query logic.
- **Sprint 2: Lakehouse Implementation and Ingestion**
  - **Task**: Build the core DuckLake lakehouse structure and implement ingestion scripts for all data sources (MITMA, INE).
  - **Tech Note**: For rapid iteration, you can use Ducklake with a local file system with DuckDB.
  - **Assessment**: Successful ingestion, data validation in Bronze, and creation of initial Silver tables.
- **Sprint 3: Orchestration and Final Analytics**
  - **Task**: Create an Airflow pipeline with DAGs to orchestrate the full ELT process from ingestion to the Gold layer.
  - **Task**: Finalize all Gold layer tables and analytical queries supporting the two business questions.
  - **Assessment**: A fully functional, automated pipeline and review of final analytical outputs.

Deliverables

1. **Code Repository**: Containing high-quality code for the lakehouse implementation, ELT pipelines, with appropriate documentation (e.g., README, inline comments, architecture diagrams, schema definitions,

and sample queries). Include scalability considerations (e.g., handling 2022–2025 data volumes) and demonstration of how it supports transport experts.

2. **Summary Article** (10–15 pages, Research Paper Style):
   - Introduction to the infrastructure and its role in mobility analysis.
   - Methodology overview (focus on data processes).
   - Results: Basic visuals (e.g., OD heatmaps, pattern summaries, gap rankings via tables).
   - Discussion: How the lakehouse enables expert insights.
   - Conclusions and Limitations (e.g., data quality issues).

Evaluation Criteria

- **Sprint Assessments (20%)**: Incremental evaluation of results from Sprints 1, 2, and 3.
- **Final Technical Implementation (40%)**: Overall lakehouse functionality, ELT efficiency, scalability, and pipeline robustness.
- **Final Data Integration and Processing (20%)**: Quality of the final data tiers, handling of zonification, and data integrity.
- **Use Case Support (10%)**: Quality and clarity of the basic analytical examples demonstrating utility.
- **Documentation (10%)**: Clarity, completeness, and focus on infrastructure in the repository and article.