pipeline_design

Christina Myalla

2026-02-13

Health Facility ETL Pipeline Design

Overview

This repository implements a standardized multi-country ETL pipeline for health facility datasets.

The pipeline transforms heterogeneous national facility registries into a harmonized geospatial dataset.

Pipeline Architecture

Raw Data → Extraction → Transformation → Validation → Standardization → Loading

---

Directory Structure

config/ Pipeline configuration (country metadata, schema, paths) data/ raw/ Original datasets (unchanged) interim/ Intermediate processing outputs processed/ Final standardized datasets

scripts/ extraction/ Data reading functions transformation/ Cleaning and harmonization functions loading/ Export functions utils/ Helper utilities

pipelines/ Country pipeline runners

docs/ Documentation

Configuration-Driven Design

All country-specific logic is stored in YAML configuration files.

This allows:

- reusable transformation scripts
- minimal code duplication
- scalable multi-country processing

Processing Steps

1. Extraction

Reads raw data from CSV, Excel, or other formats.

2. Standardization

Maps country-specific column names to global schema.

3. Data Cleaning

- numeric coordinate conversion
- column harmonization
- type enforcement

4. Spatial Validation

Facility coordinates tested against country boundary polygons.

5. Variable Selection

Only relevant standardized variables retained.

6. Loading

Country datasets saved to:

data/processed/country_standardized/

Multi-Country Scalability

Pipeline designed to process any number of countries via configuration.

Reproducibility

Pipeline execution is deterministic and configuration-driven.

Future Extensions

- global facility master dataset
- facility deduplication across countries
- temporal versioning
- data quality dashboards