EcoIndex Project Architecture

Data Layer

**Raw Data
(Excel files)**

**Interim Data
(Parquet)**

**Processed Data
(CSV/Results)**

Core Processing Modules

data_io.py
• save_interim()
• load_interim()

cleaning.py
• normalize_columns()
• cast_types()
• handle_missing()

transform.py
• hellinger_transform()
• log1p_standardize()

dataframe_ops.py
• wrap_columns()
• get_block()
• merge operations

validators.py
• assert_keys()
• data validation

Analysis & Assessment Modules

pca_assessment.py
• PCAContaminationResult
• pca_chemical_assessment()
• compute_pollution_scores()

pca_evaluation.py
• EvalResult class
• permutation_manova()
• evaluate_pca_assessment()

chemical_weights.py
• Variable classifications
• TYPE_WEIGHTS
• Weight calculations

Pipeline & Configuration

config.py
• Path definitions
• File constants
• Project structure

pipeline.py
• make_interim()
• Orchestration
• Workflow control

ingest.py
• Data ingestion
• Initial processing

Data Flow:
1. Raw data ingestion
2. Cleaning & validation
3. Transformation
4. PCA assessment
5. Statistical evaluation
6. Results output

Module Types:

**__init__.py
Public API Interface**

- Data I/O
- Core Processing
- Analysis
- Configuration
- Utilities