Euro Area HICP Nowcast Pipeline

A modular end-to-end data engineering project that demonstrates ingestion, transformation, modelling, and dashboarding for a real-time Euro-area inflation nowcast.

Quick Start

Run everything locally with four steps.

1. Install requirements

```
python3 -m venv .venv
source .venv/bin/activate
pip install --upgrade pip
pip install -r requirements.txt
pip install dbt-duckdb
```

2. Create your `.env`

```
cp .env.example .env
```

Populate it with:

```
FRED_API_KEY=your_fred_api_key_here       # Energy + commodities ingestion
PROJECT_ROOT=/absolute/path/to/this/repo  # e.g. /Users/me/Euro_HICP_MOM2
DB_PATH=/absolute/path/to/this/repo/data/euro_nowcast.duckdb # copy the project root
BRONZE_ROOT=/absolute/path/to/this/repo/data/bronze # copy the project root path be
```

`PROJECT_ROOT` must be the folder where this repo lives.

3. Run the entire pipeline

This single command executes every stage end-to-end:

- API ingestion (Eurostat, ECB, FRED)
- Bronze lakehouse writes
- Silver harmonisation
- Gold Kimball star modelling
- ML feature generation
- XGBoost training
- Next-month nowcast
- Traceability + run logs

```
bash orchestration/run_all.sh
```

4. Launch the dashboard

```
bash orchestration/start_dashboard.sh
```

Open `http://localhost:5000` to see the latest official HICP prints, the model's next-month nowcast, macro indicators, and historical comparisons.

---

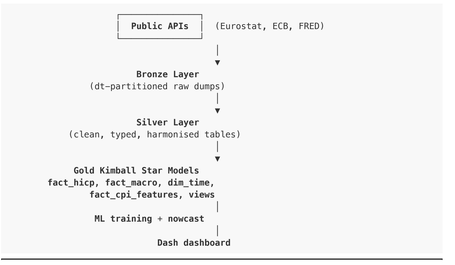
1. Purpose

Produce a real-time Euro-area CPI/HICP nowcast while showcasing:

- Python ingestion scripts
- Bronze → Silver → Gold lakehouse modelling
- dbt-duckdb transformations
- XGBoost-based forecasting
- Local Dash dashboarding
- End-to-end traceability

---

2. Architecture Overview



---

3. Lakehouse Layers

Bronze – Raw, auditable input
- S3-style layout: `data/bronze/<dataset>/dt=YYYY-MM-DD/part-*.csv`
- Direct dumps of API responses
- Schema enforced to `dt, dataset, series_id, geo, value, source, ingest_ts`
- Every ingestion run stores logs plus file paths for traceability

Silver – Clean, consistent tables
- Normalises geos, units, and frequencies
- Ensures indicators share coherent datatypes
- Produces a single consolidated view for modelling
- Acts as the contract for downstream consumers

Gold – Kimball star schema
- `fact_hicp`: Headline CPI values
- `fact_macro`: FX, policy rates, energy indicators
- `fact_cpi_features`: Lagged features + deltas
- `dim_time`: Month-grain dimension
- `gold_nowcast_input`: Model-ready feature table

Why Kimball? Clear fact/dimension separation, consistent join paths, simple to extend with more facts (PPI, logistics) or dims (country-level variants).

---

4. Machine Learning Nowcast

- **Target:** Month-on-month HICP change (stationary), converted back to index levels for scoring.
- **Features:** Lagged CPI, FX, policy rates, energy prices, leakage-safe rolling averages, calendar month & quarter.
- **Training window:** 2000–01 through 2023–06 for base training; everything after becomes a walk-forward evaluation step.
- **Walk-forward loop:** Train on history → predict the next month → log metrics/residuals → repeat.
- **Traceability:** Each run records parameters, MAE/RMSE/R^2, p90 error, feature importance, per-month residuals, and the generated nowcasts inside DuckDB (`ml.*` + `gold.nowcast_output`).

Result: a one-month-ahead HICP index estimate that can be audited back to the exact features and model version used.

---

5. Dashboard

Once the pipeline runs, visit `http://localhost:5000` to explore:

- Latest official HICP vs model nowcast (index + YoY)
- Month-on-month behaviour and historical comparisons
- Macro indicators (FX, energy, policy rate) with optional rolling averages
- Exportable CSVs for CPI and macro panels

This lightweight Dash app is meant to be fast to run locally and easy to redeploy (e.g., containerise for Fargate/AppRunner later).

---

6. Future Improvements

1. **Additional indicators** – Food/agri indices, logistics benchmarks, PPI, expanded FX baskets, commodity futures, retail sentiment.
2. **Multi-model ensemble** – Specialised models per indicator cluster (energy-driven, FX-driven, food-driven) plus PCA/factor inputs and ensemble averaging.
3. **Cloud lakehouse migration** – S3 Bronze/Silver/Gold, Glue/Athena catalogues, Step Functions/Airflow orchestration, hosted dashboard.

---

7. Repository Layout

```
project/
├── ingestion/
├── sql/
│   ├── 00-04_*.sql (bronze setup)
│   └── silver/
├── dbt_project/ (gold models + views)
├── analytics/
│   ├── model_training/
│   └── model_predictions/
├── dashboard/InflationNowcastingDash/
├── orchestration/ (run_all.sh, start_dashboard.sh)
├── data/
├── logs/
└── README.md
```