Wizards of Oz

AN END-TO-END FOOD DATA ENGINEERING PIPELINE FOR NUTRITIONAL ANALYSIS

Ingestion

JSON
ingest USDA Dataset

Parquet
ingest OpenFoodFacts Dataset

Staging

clean_usda_data
extract FoundationFoods
normalise/flat

clean_openfoodfacts_data
chunked streaming
extract nutritions

enrich_and_merge
merge both Datasets together
harmonise rows

Production

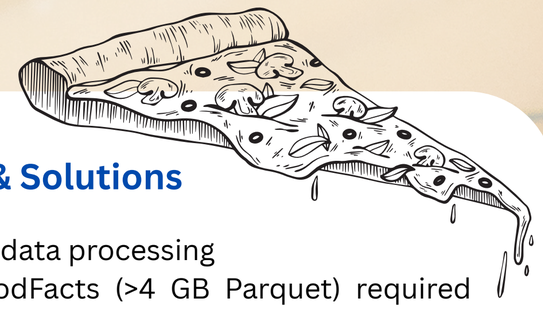
Data Warehouse
create star schema

Foodgraph
produce graph schema

We designed an end-to-end ETL pipeline orchestrated with Apache Airflow.
Two heterogeneous food datasets are ingested, cleaned, normalized, merged and exposed through both a relational data warehouse and an optional graph database.

Context & Objective

Public food datasets are large, heterogeneous and difficult to compare. The goal of this project is to build a reproducible data engineering pipeline that transforms raw nutrition data into analytics-ready datasets, enabling reliable comparisons between homemade and industrial food products.

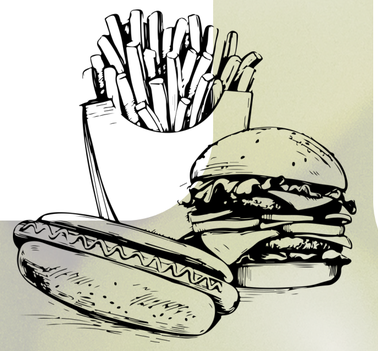


Challenges & Solutions

- Large-scale data processing
  - OpenFoodFacts (>4 GB Parquet) required chunked reading to prevent memory issues and ensure stable execution.
- Data quality issues
  - Incorrect parsing of the USDA JSON caused extensive NaN values, requiring careful debugging and validation.
- Heterogeneous schemas
  - Inconsistent nutrient naming across sources was resolved through flexible matching and normalization logic.
- Dataset imbalance
  - Due to the significantly larger size of OpenFoodFacts, some aggregated comparisons may be biased toward this source. This imbalance should be considered when interpreting comparative analytics results.

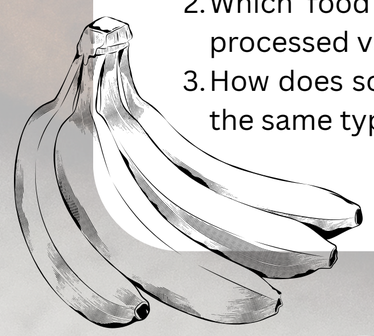
Key Design Choices

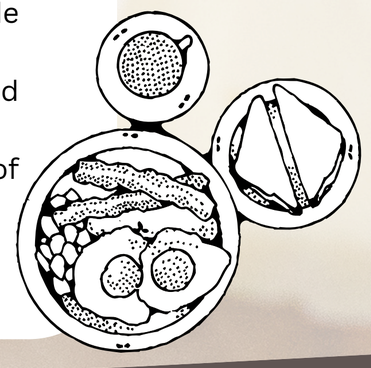
- Simple, idempotent ingestion using immutable landing files
- Deterministic data cleaning and unit normalization (per 100g)
- Chunked processing for large OpenFoodFacts Parquet files
- Star schema in Postgres for analytical workloads
- Optional Neo4j graph for exploratory relationship analysis



Questions that can be answered based on the data

1. How does the total nutrient content (protein, magnesium, vitamins) of a homemade dish compare to that of a premade product?
2. Which food categories show the largest fat content differences between raw and processed versions?
3. How does sodium (salt) content differ between homemade and commercial dishes of the same type?





INSA INSTITUT NATIONAL DES SCIENCES APPLIQUÉES LYON

Ahmed MANSOUR ahmed.mansour@insa-lyon.fr
Robert MICHEL robert-franz.michel@insa-lyon.fr
Stefan SEVERIN stefan.severin@insa-lyon.fr