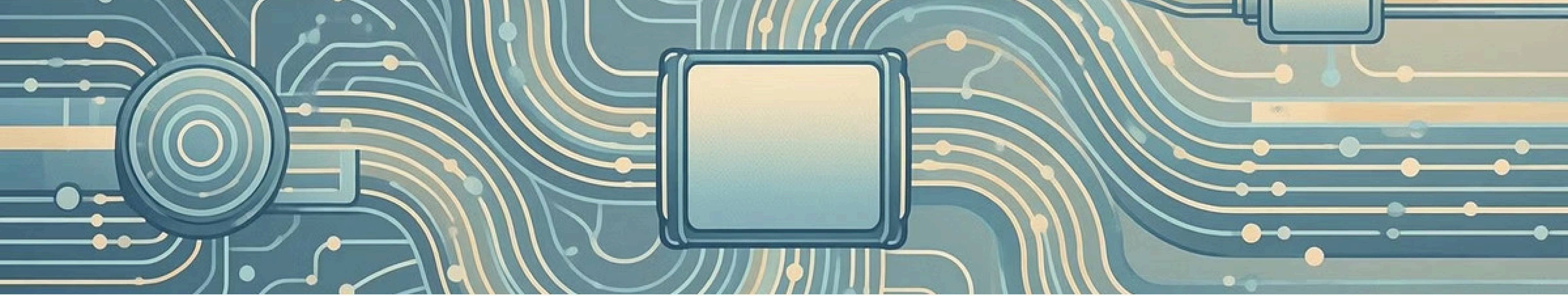
Skincare Product Data Collection & Enrichment Pipeline

A two-stage data pipeline for collecting, validating, and enriching skincare product data from Qudo Beauty's e-commerce platform.



PROJECT OVERVIEW

Building a Reliable Data Foundation

Stage 1: Day 1

Reliable scraping and structuring of skincare product information directly from the source website.

- Product discovery and extraction
- Data structuring and validation
- Initial dataset creation

Stage 2: Day 2

External validation and enrichment using search-based discovery and manufacturer websites.

- Web-based data enrichment
- Manufacturer verification
- Quality enhancement

TECHNOLOGY STACK

Tools & Technologies

Core Stack

Python 3, Requests/httpx, BeautifulSoup, Pandas, Regex, JSON

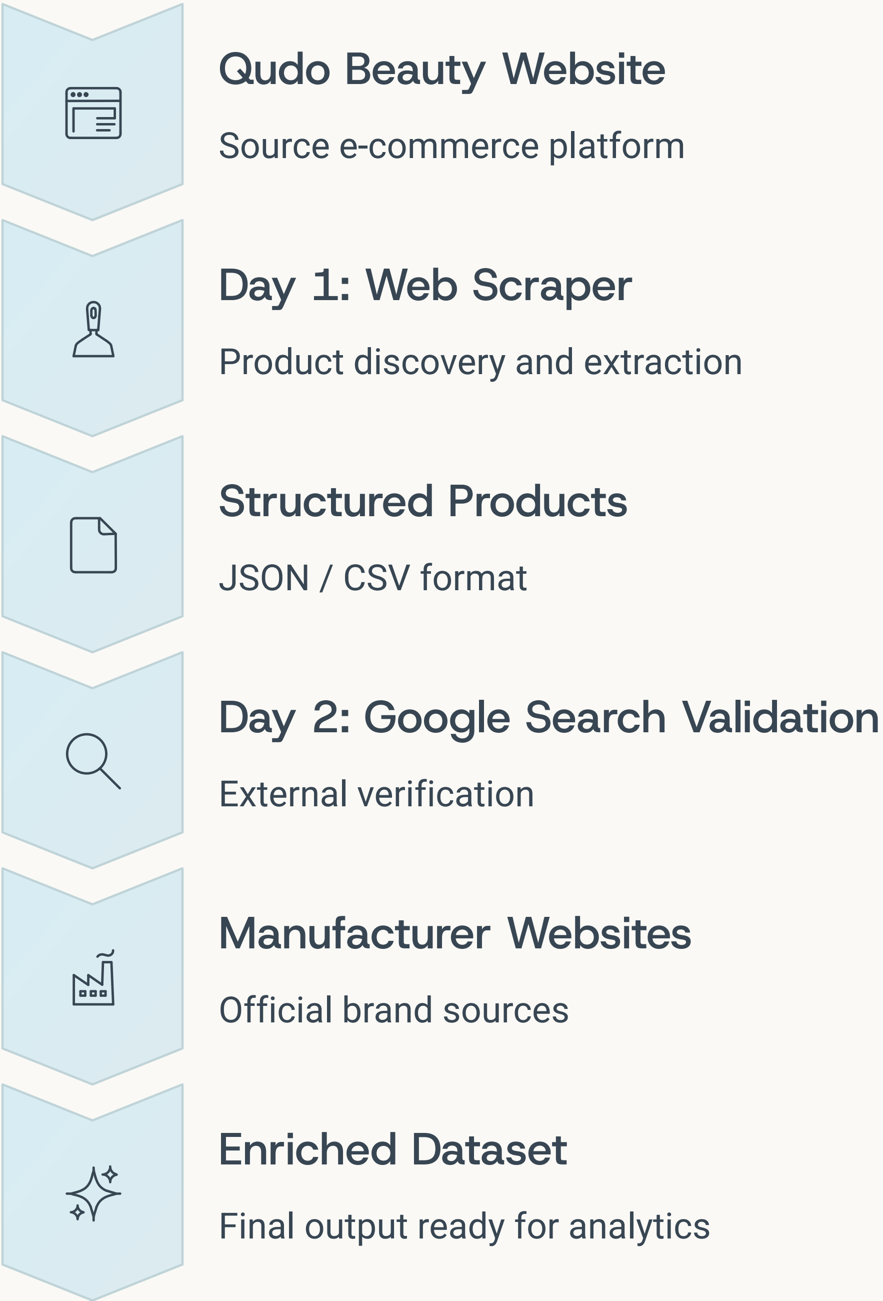
External Services

Google Custom Search API, WooCommerce HTML structure

Data Handling

Structured data processing, export capabilities, machine-readable formats

Pipeline Architecture



Qudo Beauty Website

Source e-commerce platform

Day 1: Web Scraper

Product discovery and extraction

Structured Products

JSON / CSV format

Day 2: Google Search Validation

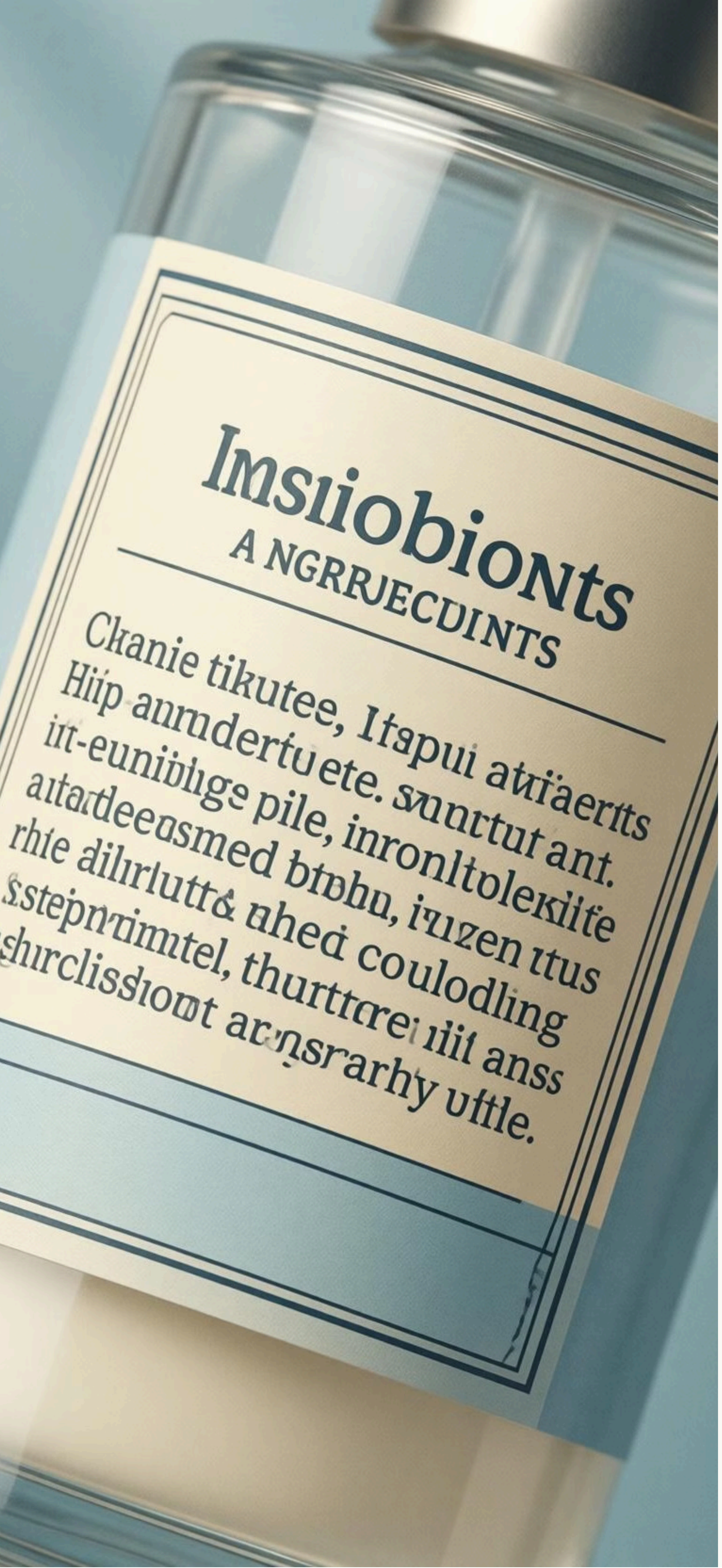
External verification

Manufacturer Websites

Official brand sources

Enriched Dataset

Final output ready for analytics


Imsiiobionts
A NGRRJECUINTS
Ckanie tikutee, Itapui atriaerts
Hiip anmdertuete. sunttut ant.
iit-euniblige pile, inronltolexlite
aitardeeasmed btobu, iuzen itus
rhie dilrlutt& ahed coulodling
sstepnrimtel, thurttrei iiit anss
shirclisslomt ansrarhy uttle.

DATA COLLECTION

Core Product Fields

Day 1 captures essential product information directly from the source website.

Product Identity

Product name, brand, category/type

Ingredients

Complete ingredients list or description

Packaging Details

Size specifications (50ml, 100g, etc.)

Visual Assets

Product image URL, product page URL

DATA ENRICHMENT

Enhanced Fields Added on Day 2

- **Manufacturer Website**

  Official brand/manufacturer domain

- **Meta Description**

  Verified product description

- **Confirmed Brand**

  External brand confirmation

- **SKU / Barcode**

  Product identifiers (UPC, EAN)

- **Country of Origin**

  Manufacturing origin

- **External Ingredients**

  Verified ingredient text

DAY 1 PROCESS

Scraping & Structuring Logic

01

Product Discovery

Starts from /shop/ page, iterates through paginated listings using WooCommerce-specific selectors with safety controls (MAX_PAGES, MAX_PRODUCTS).

02

Product Page Parsing

Multiple fallback strategies handle inconsistent layouts: header tags, metadata, gallery images, breadcrumbs, attribute tables, and regex patterns.

03

Ethical Scraping

Custom browser-like User-Agent, request timeouts, controlled delays between requests, and graceful error handling ensure reliability.

☆ DAY 2 PROCESS

Data Enrichment & Validation

Day 2 addresses common scraping limitations: missing brands, incomplete ingredients, unverified descriptions, and lack of manufacturer metadata.

Google Search Discovery

Query constructed using product name + brand. Google Custom Search API retrieves top results and evaluates candidate manufacturer domains.

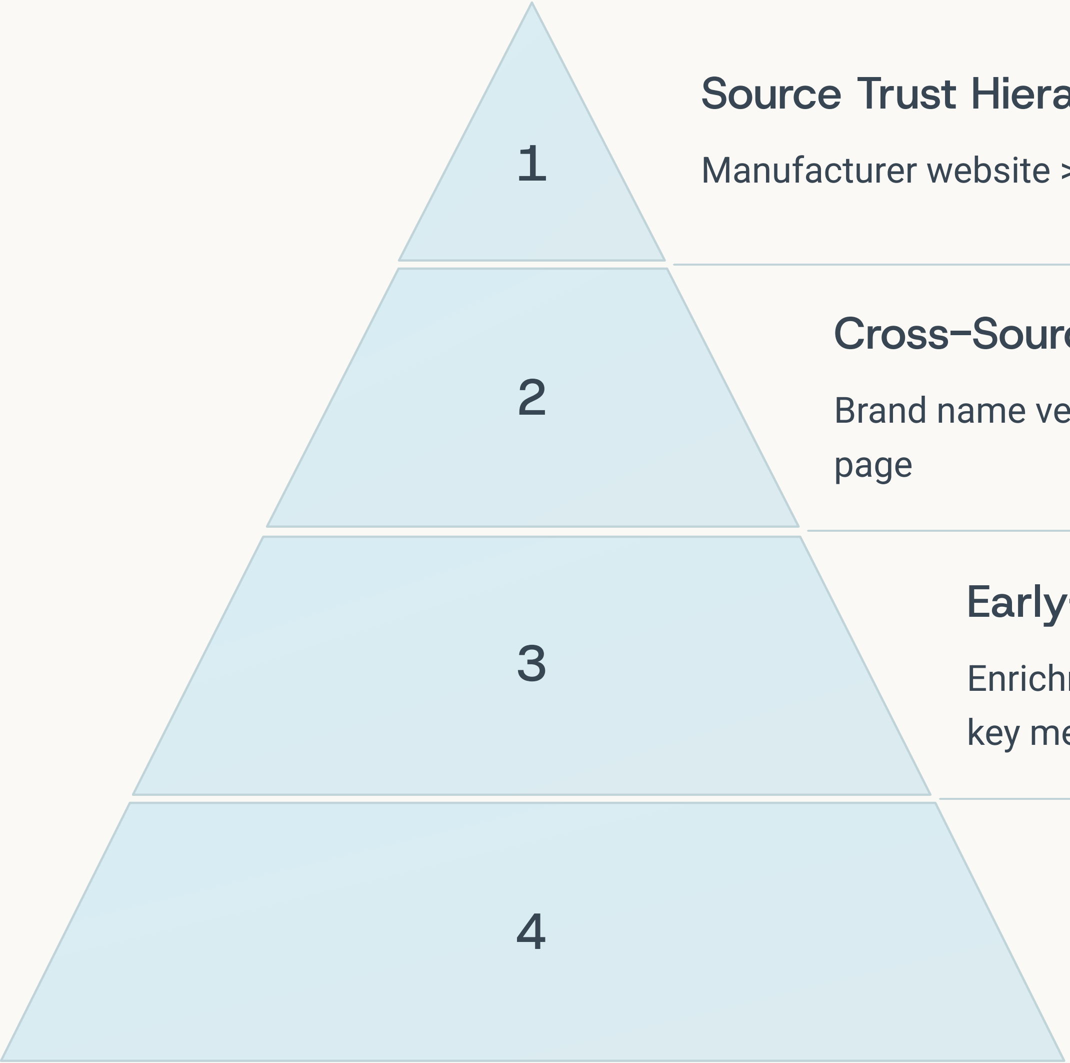
Manufacturer Validation

Brand presence checked in page content, domain credibility assessed. Stops early once confidence threshold is met.

HTML & Regex Extraction

From verified pages: meta descriptions, ingredient text blocks, SKU/barcode patterns, and country of origin references.

Multi-Layer Reliability & Trust



Source Trust Hierarchy

Manufacturer website > e-commerce site > inferred data

Cross-Source Confirmation

Brand name verified across scraped data, Google results, and manufacturer page

Early-Stop Confidence

Enrichment halts once brand confirmed, official domain identified, and key metadata extracted

Fallback Transparency

Inferred fields only used when explicit data unavailable—no silent assumptions

Final Outputs & Key Takeaways

Dataset Organization

- products.json / products.csv (Day 1)
- day2_enriched_products.json
- day2_enriched_products.csv
- One row per product, flat schema
- Analysis-ready with absolute URLs

Limitations & Assumptions

Assumptions: Only shop-listed products valid, ingredient text in descriptions acceptable, brand inference as fallback.

Limitations: Google API rate limits, variable manufacturer metadata, SKU/barcode availability varies.

The final output is a clean, enriched, machine-readable dataset suitable for analytics, product catalogs, recommendation systems, or downstream AI pipelines.