Books Pipeline

Enlace repositorio GitHub

https://github.com/Adrimrtz16/books-pipeline

Instalación

```
pip install -r requirements.txt
```

Ejecución

1 Scraper de Goodreads

```
python scrape_goodreads.py
```

Este script realiza scraping en Goodreads para obtener información de libros (título, autores, ISBN_13, puntuación y URL). Guarda el resultado en `landing/goodreads_books.json`.

2 Fetcher y enriquecedor de Google Books

```
python enrich_googlebooks.py
```

Consulta la API de Google Books para enriquecer cada registro extraído de Goodreads (busca ISBN o título, recupera `gb_id`, ISBNs, precios, URL, etc.). Requiere la variable de entorno `GOOGLE_BOOKS_API_KEY` y produce `landing/googlebooks_books.csv`.

3 Merge e integración final

```
python integrate_pipeline.py
```

Combina y normaliza los datos de `landing/goodreads_books.json` y `landing/googlebooks_books.csv`, priorizando campos de Google cuando están disponibles. Escribe la tabla canónica en `standard/dim_book.parquet`, el detalle de fuente en `standard/book_source_detail.parquet` y las métricas de calidad en `docs/quality_metrics.json`.

Outputs

- La tabla canónica normalizada se encuentra en `standard/dim_book.parquet`.
- Las quality metrics se encuentran en `docs/quality_metrics.json`.
- La definición del esquema está en `docs/schema.md`.
- La tabla de detalle de fuentes originales está en `standard/book_source_detail.parquet`.
- Las fuentes de datos en bruto están en `landing/`.

Schema.md

Campo	Tipo	Nullable	Formato	Ejemplo	Reglas
canonical_id	int64	No	numérico	5	Identificador único incremental Debe ser entero positivo

Campo	Tipo	Nullable	Formato	Ejemplo	Reglas
gb_id	object	Sí	string	_1b4nAEACAAJ	ID de Google Books Puede estar vacío
title	object	No	string	Data Science for Business	Texto normalizado (trim espacios) No debe estar vacío
subtitle	object	Sí	string	Using Data Science to Transform Information into Insight	Texto libre opcional Pueden eliminarse espacios extra
author	object	No	array	["Foster Provost","Tom Fawcett"]	Lista separada internamente por comas o pipes Eliminar duplicados Debe tener al menos un autor
publisher	object	Sí	string	John Wiley & Sons	Puede estar vacío Normalizar espacios
published_date	object	Sí	string	2013-11-12	Formato ISO-8601 Admite YYYY, YYYY-MM o YYYY-MM-DD
rating	float64	No	numérico	4.13	Número entre 0 y 5 Convertir string a número
ratings_count	int64	No	numérico	2624	Entero ≥ 0
ISBN_10	object	Sí	string	111866146X	Debe tener 10 caracteres Puede incluir X como dígito de control Validar checksum ISBN-10

Campo	Tipo	Nullable	Formato	Ejemplo	Reglas
ISBN_13	object	Sí	string	9781449361327	Debe tener 13 dígitos numéricos No debe incluir guiones Validar checksum ISBN-13
price_amount	float64	Sí	numérico	27.99	Número decimal válido con punto o coma ≥ 0
price_currency	object	Sí	string	EUR	Moneda válida ISO-4217
book_url_goodreads	object	No	URL	https://www.goodreads.com/book/show/17912916	Debe ser una URL válida
book_url_google_books	object	Sí	URL	https://www.googleapis.com/books/v1/volumes/tXdBAQAAQBAJ	Debe ser una URL válida si existe