Proyecto Pipeline de Libros

Autor: Antonio Valverde Soto

Documentación Técnica Completa

01 Scraping

Bloque 1 — Scraping de Goodreads

Este bloque implementa la fase de **extracción de datos (Extract)** mediante web scraping en la plataforma Goodreads, obteniendo una muestra de libros a partir de una búsqueda.

Objetivo del bloque

Obtener:

- título

- autor principal

- rating

- número de valoraciones

- URL del libro

- ISBN10 / ISBN13

Guardado en:

```

landing/goodreads_books.json

```

Pasos del Bloque

1. Realizar búsqueda pública

Ejemplo:

```

https://www.goodreads.com/search?q=data+science

```

2. Extraer tabla de resultados

Selectores:

- `table.tableList tr`

- `a.bookTitle span`

- `a.authorName span`

- `span.minirating`

3. Acceder a la ficha del libro

Selectores:

- `#bookDataBox .clearFloats`

- `.infoBoxRowTitle`

- `.infoBoxRowItem`

4. Scraping ético

- Pausas

- User-Agent realista

- Límite de páginas

5. Salida del bloque

Archivo JSON con un libro por registro.

02 Enrichment

Bloque 2 — Enriquecimiento con Google Books API

Objetivo

Ampliar los datos de Goodreads con metadata completa de Google Books:

- título/subtítulo

- autores

- editorial

- fecha publicación

- idioma

- categorías

- ISBN normalizados

- precio y moneda

Salida:

```

landing/googlebooks_books.csv

```

Funcionamiento

1. Cargar goodreads_books.json

Se lee como fuente principal.

2. Construcción de la consulta

Orden de prioridad:

1. `isbn:ISBN13`

2. `isbn:ISBN10`

3. `intitle:TITULO+inauthor:AUTOR`

3. Petición a Google Books API

Endpoint:

```

https://www.googleapis.com/books/v1/volumes

```

4. Extracción de datos

Desde `volumeInfo`, `saleInfo`, `industryIdentifiers`.

5. Normalización parcial

- Autores → "A | B | C"

- Categorías → "X | Y"

6. CSV final

Separador `;`, UTF-8.

03 Integration

Bloque 3 — Integración, Normalización y Modelo Canónico

Objetivo

Unir Goodreads + Google Books, limpiar, deduplicar y generar datasets en Parquet listos para análisis.

Salidas:

```

standard/dim_book.parquet

standard/book_source_detail.parquet

docs/quality_metrics.json

docs/schema.md

```

Pasos del bloque

1. Cargar datos de landing/

Incluye:

- source

- row_id

2. Normalizar campos

- fechas → ISO

- idioma → BCP-47

- moneda → ISO-4217

- autores/categorías → listas

3. book_id_candidato

Regla:

- usar ISBN13 si existe

- si no, `titulo+autor+editorial` normalizado

4. Deduplicación

Reglas:

- título más largo

- primer autor no nulo

- unión de listas

- precio más reciente

- idioma no nulo

- editorial no nula

5. Modelo canónico

Campos:

- book_id

- titulo

- autor_principal

- autores

- editorial

- anio_publicacion

- fecha_publicacion

- idioma

- isbn10 / isbn13

- categorias

- precio / moneda

- ts_ultima_actualizacion

6. Trazabilidad

`book_source_detail.parquet` contiene todos los valores originales.

7. Calidad

`quality_metrics.json` calcula nulos, duplicados, etc.

8. Esquema

`schema.md` describe el modelo.

04 Quality

Métricas de Calidad del Pipeline

El pipeline genera:

```

docs/quality_metrics.json

```

¿Qué contiene?

1. Conteos

- registros_goodreads

- registros_googlebooks

- libros_finales_dim

2. Porcentaje de nulos

- titulo

- isbn13

- precio

3. Duplicados por book_id_candidato

4. Validación de rango

- precio >= 0

5. Distribución por fuente

- goodreads

- googlebooks

Utilidad

Permite validar:

- integridad

- consistencia

- calidad del scraping

- efectividad del enriquecimiento