

## Informe Final

### Sistema ETL de Web Scraping con arquitectura de lago de datos

#### Objetivo

Este proyecto implementa un sistema ETL (Extraer, Transformar, Cargar) completo que combina los requisitos de Tarea 1 y Tarea 2. El sistema extrae datos utilizando Scrapy, los transforma a través de canales de validación y los almacena en una arquitectura de lago de datos de tres niveles. Además, se creó un panel Streamlit para visualizar los datos procesados.

#### 1. Decisiones de diseño

Se usaron las siguientes fuentes de la tarea 1 y se añadió una fuente de noticias adicional:

- <https://www.npr.org/>
- <https://www.aljazeera.com>
- <https://quotes.toscrape.com/>
- <http://www.nytimes.com>

Se definieron ítems personalizados con validaciones específicas (por ejemplo, “strip()” y campos obligatorios).

Se desarrolló un “**pipeline**” que limpia, valida y guarda simultáneamente en JSONL y PostgreSQL.

Se usó un archivo “**.jsonl**” en lugar de “**.json**” para evitar problemas de duplicación o estructura.

#### 2. Extracción, limpieza y validación

Se desarrollaron 4 spiders para las fuentes mencionadas.

Se configuró USER\_AGENT, DOWNLOAD\_DELAY, y ROBOTSTXT\_OBEY para scraping responsable y ético.

La ejecución se automatiza cada 2 días mediante un script .bat.

Se empleó XMLFeedSpider para parsear RSS correctamente.

La limpieza se implementó en el archivo pipelines.py, donde se validaron y transformaron los campos requeridos:

- **title:** Eliminación de espacios y caracteres especiales (strip()).
- **url:** Validación de formato URL y unicidad en BD.

- **date:** Transformación al formato estándar ISO YYYY-MM-DD.
- **source:** Asignación automática según el spider.
- **summary:** Limpieza de saltos de línea y truncamiento si es largo.

Inserción en PostgreSQL con control de duplicados usando “ON CONFLICT DO NOTHING”. Además, se implementó control de duplicados tanto en el JSONL como en la base de datos PostgreSQL, utilizando la URL como clave única.

Codificación en UTF-8 para soportar caracteres especiales.

Manejo de errores mediante try/except.

## 2.1.Almacenamiento de datos

Los datos procesados se almacenan en:

- **output/articles\_final.jsonl:** Archivo acumulativo en formato JSONL (una línea por artículo), sin duplicados.

```

96  cars and car parts", "url": "https://www.npr.org/2025/03/26/nx-si-534176/trump-trade-tariffs-imported-cars", "date": "2025-03-27", "source":
97  se against OpenAI to go forward", "url": "https://www.npr.org/2025/03/26/nx-si-5288157/new-york-times-openal-copyright-case-goes-forward", "da
98  military base in Greenland", "url": "https://www.npr.org/2025/03/26/nx-si-5341585/greenland-pituffik-space-military-base", "date": "2025-03-27",
99  its now have a fix from the IRS", "url": "https://www.npr.org/2025/03/26/nx-si-5341558/irs-ev-ta-credit-2024-solution", "date": "2025-03-27",
100 stand trial over alleged coup attempt", "url": "https://www.npr.org/2025/03/26/nx-si-5341587/bolsonaro-brazil-trial-coup-attempt", "date": "2
101 up new case against administration", "url": "https://www.npr.org/2025/03/26/nx-si-5341540/boasberg-trump-signal-national-security-atlantic-le
102 r plans group chat", "url": "https://www.npr.org/2025/03/26/g-si-55668/signal-yemen-war-plans-europe-reaction", "date": "2025-03-27", "source":
103 s", "url": "https://www.npr.org/2025/03/26/nx-si-5341496/how-npr-covers-itself", "date": "2025-03-27", "source": "NPR", "summary": "When NPR i
104 she talks with her daughter about injustice", "url": "https://www.npr.org/2025/03/26/nx-si-5311607/amanda-knox-free", "date": "2025-03-27", "
105 ked by the Trump administration?", "url": "https://www.npr.org/2025/03/26/nx-si-5339507/what-is-state-secrets-privilege-trump-administration",
106 helps gardeners use less plastic and peat", "url": "https://www.npr.org/2025/03/26/nx-si-5334478/gardening-soil-blocking", "date": "2025-03-27
107 ical for advanced chips and medical devices", "url": "https://www.npr.org/2025/03/26/nx-si-5348887/trump-cuts-nist-atomic-spectra-lab-advancee
108 o Free Europe/Radio Liberty", "url": "https://www.npr.org/2025/03/26/nx-si-5341321/trump-radio-free-europe-radio-liberty-restraining-order",
109 anning 'ghost guns'", "url": "https://www.npr.org/2025/03/26/nx-si-5341404/supreme-court-ghost-guns", "date": "2025-03-27", "source": "NPR",
110 the leaked Signal chat as more details emerge", "url": "https://www.npr.org/2025/03/26/nx-si-5341359/intelligence-leaders-signal-house-hearin
111 ntists say it boosts the elders' health, too", "url": "https://www.npr.org/sections/shots-health-news/2025/03/26/g-si-55822/volunteer-retireme
112 n Capitol Hill", "url": "https://www.npr.org/2025/03/26/nx-si-5339951/npr-pbs-congress-hearing", "date": "2025-03-27", "source": "NPR", "summa
113 s most sensitive data, but can't say why", "url": "https://www.npr.org/2025/03/26/nx-si-5339842/doge-data-access-privacy-act-social-security-t
114 : a U.S. representative on Signal messages", "url": "https://www.npr.org/2025/03/26/nx-si-5339796/atlantic-releases-texts-signal-war-plans",
115 kers against Pacers", "url": "https://www.aljazeera.com/sports/2025/3/27/lakers-vs-pacers-lebron-james-game-winner-at-buzzer-lifts-los-angeles
116 d ahead of JD Vance visit", "url": "https://www.aljazeera.com/news/2025/3/27/trump-reiterates-us-must-have-greenland-ahead-of-jd-vance-visit",
117 f new AI-equipped suicide drones", "url": "https://www.aljazeera.com/news/2025/3/27/north-koreas-kim-jong-un-oversees-tests-of-new-ai-equipped
118 gaza assault continues", "url": "https://www.aljazeera.com/news/2025/3/27/israeli-military-kills-hamas-spokesman-as-gaza-assault-continues",
119 ports of up risk Richard's arrest", "url": "https://www.aljazeera.com/news/2025/3/27/un-warns-of-conflict-in-south-sudan-amid-reports-of-up-ri
120 y 1,127", "url": "https://www.aljazeera.com/news/2025/3/27/russia-ukraine-war-list-of-key-events-day-1127", "date": "2025-03-27", "source": "A
121 s Syria's Latakia", "url": "https://www.aljazeera.com/news/liveblog/2025/3/27/live-israel's-rentless-bombardment-kills-26-palestinians-in-ga
122 nt Rumeysa Ozturk at Tufts", "url": "https://www.aljazeera.com/podcasts/2025/3/26/signal-gate-how-trump-officials-chat-on-bombing-yemen-hi
123 ch?", "url": "https://www.aljazeera.com/program/inside-story/", "date": "2025-03-27", "source": "Aljazeera", "summary": "Donald Trump downplay
124 ers found in Lithuania", "url": "https://www.aljazeera.com/news/2025/3/26/nato-chief-says-four-us-soldiers-dead-in-lithuania-in-training", "da
125 ngoing protests over Imamoglu", "url": "https://www.aljazeera.com/news/2025/3/26/istanbul-elects-aslan-interim-mayor-amid-ongoing-protests-ove
126 ry Bolsonaro for coup attempt", "url": "https://www.aljazeera.com/news/2025/3/26/brazils-supreme-court-announces-it-will-try-bolsonaro-for-cou
127 uproar over the missing in Mexico", "url": "https://www.aljazeera.com/news/longform/2025/3/26/ho
128
129

```

- **Base de datos PostgreSQL:** Conexión usando psycopg2, se creó la base de datos “scraping\_db” y los datos se guardan en la tabla articles que contiene claves únicas basadas en la URL para evitar duplicados.

The screenshot shows the pgAdmin 4 interface. The top toolbar includes buttons for connecting, refreshing, saving, and running queries. The query editor contains the following SQL code:

```

1 SELECT * FROM articles LIMIT 20;
2 SELECT source, COUNT(*)
3 FROM articles
4 GROUP BY source
5 ORDER BY COUNT(*) DESC;

```

The results pane displays a table with 5 columns: url, date, source, and summary. It shows 18 rows of data, with the first row being the header. The data is as follows:

	url	date	source	summary
10	https://www.npr.org/2025/03/27/g-s1-56407/exhibit-takes-visitors-inside-the-annex-where-anne-frank-lived	2025-03-27	NPR	For the first time, a re-creation of the annex where Anne Frank and her family hid is available outside of Amsterdam
11	https://www.npr.org/2025/03/26/mx-s1-5341322/greenland-trump-vance-dog-sled-race	2025-03-27	NPR	Second lady Usha Vance has scrapped a plan to attend Greenland's national dog sled race this week. But American
12	https://www.npr.org/2025/03/26/mx-s1-5341780/social-security-administration-identity-requirements	2025-03-27	NPR	Officials said they would now exempt people who apply for Medicare and disability benefits, as well as supplement
13	https://www.npr.org/2025/03/26/g-s1-56392/appeals-court-allen-enemies-act	2025-03-27	NPR	The D.C. Circuit Court of Appeals panel denied the Trump administration's push to restart deportations of alleged c
14	https://www.npr.org/2025/03/26/mx-s1-5341767/trump-trade-tariffs-imported-cars	2025-03-27	NPR	The president's latest action on trade takes effect next week, and will slap a 25% tariff on imported cars and car pe
15	https://www.npr.org/2025/03/26/mx-s1-5288157/new-york-times-openai-copyright-case-goes-forward	2025-03-27	NPR	The legal fight could have far-reaching implications for the media and artificial intelligence industries.
16	https://www.npr.org/2025/03/26/mx-s1-5341505/greenland-pituffik-space-military-base	2025-03-27	NPR	Vice President JD Vance will travel to Greenland this week, including a stop at Pituffik Space Base, the U.S. Defens
17	https://www.npr.org/2025/03/26/mx-s1-5341550/irs-ev-tax-credit-2024-solution	2025-03-27	NPR	Some car owners couldn't claim the EV tax credit for vehicles purchased in 2024 because dealers skipped a key sa
18	https://www.npr.org/2025/03/26/mx-s1-5341507/bolsonaro-brazil-trial-coup-attempt	2025-03-27	NPR	The former far-right populist president, Jair Bolsonaro, will face trial for allegedly attempting to overturn his 2022 r

### 3. Flujo de trabajo

El flujo de trabajo del scraping es el siguiente:

1. Ejecución del spider (Scrapy)
2. Extracción de los artículos
3. Limpieza de los campos con validaciones en pipelines.py
4. Almacenamiento simultáneo en:
  - Archivo JSONL (output/articles\_final.jsonl)
  - Tabla articles en PostgreSQL

### 4. Arquitectura del Datalake

- **Landing\_Zone:** Contiene los datos crudos exportados directamente desde el spider.
- **Refined\_Zone:** Contiene los datos ya limpiados, transformados y validados, extraídos desde PostgreSQL
- **Consumption\_Zone:** Almacena vistas resumidas listas para análisis y visualización.

Con la siguiente estructura:

datalake/

- └─ LANDING\_ZONE/articles\_raw.jsonl
- └─ REFINED\_ZONE/articles\_postgres.csv
- └─ CONSUMPTION\_ZONE/articles\_summary.csv

## 5. Dashboard en Streamlit

Visualiza los datos de la Zona de Consumo conectando a PostgreSQL.

Permite filtrar por fecha usando `st.date_input()`.

Incluye:

Título principal del panel

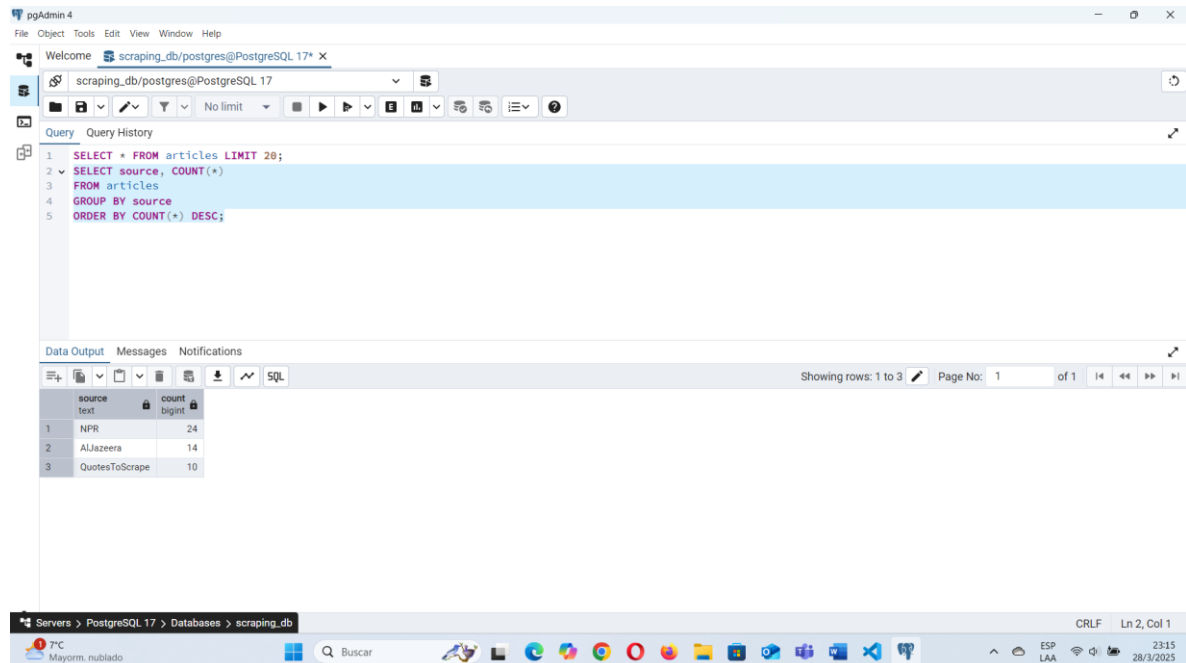
- Total, de artículos scrapeados (con `st.metric()`)
- Tabla de artículos por fuente (`st.dataframe()`)
- Gráfico de barras (`st.bar_chart()`)

Se integra con la API externa **OpenWeatherMap**:

- El usuario ingresa una ciudad
- Se muestra el clima actual (temperatura, descripción, humedad)

## 6. Conclusiones

Se comprobó que los registros en la base de datos y JSONL no se duplican al re- ejecutar los spiders.



Se ejecutó un script de validación (`output/ contar_articulos.py`) para contar artículos por fuente y verificar unicidad de URLs.

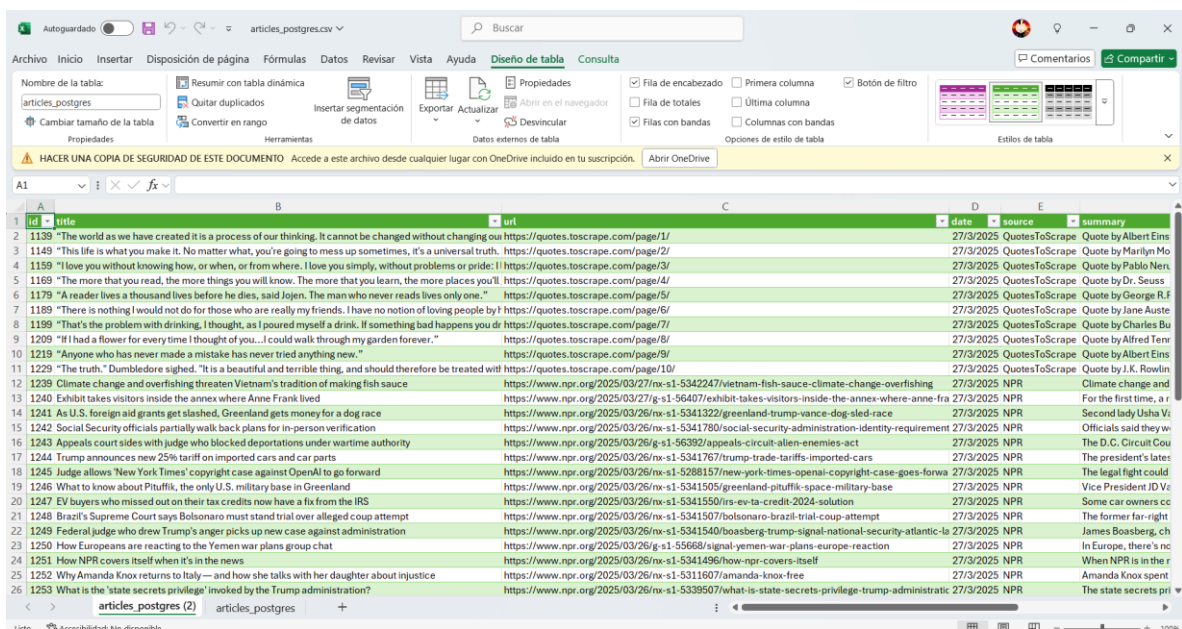
```
C:\Users\yamga\Documents\Scraping\news_scraper\output>python contar_articulos.py
Total de URLs únicas: 48

URLs únicas por fuente:
• QuotesToScrape: 10
• NPR: 24
• AlJazeera: 14

C:\Users\yamga\Documents\Scraping\news_scraper\output>
```

Y ambos arrojan los mismos resultados.

Se extrajo la base de datos desde el PostgreSQL en formato csv en (output/articles\_postgres.csv)



Se ejecuto el dashboard con los datos que agarran los spiders.



Y se añadió una API para conocer los datos del tiempo según la ciudad.

#### Consulta el clima actual

Ciudad:

la paz

```
{
  "Ciudad" : "La Paz"
  "Temperatura (°C)" : 8.99
  "Clima" : "muy nuboso"
  "Humedad (%)" : 87
}
```

### 6.1.Scripts auxiliares incluidos

- **contar\_articulos.py:** Script para contar artículos totales y por fuente.
- **limpiar\_postgres.py:** Script opcional para borrar todos los datos de la tabla articles.
- **limpiar\_json.py:** Script para limpiar el archivo JSONL si se requiere.
- **exportar\_postgres\_csv.py:** Script usa csv.writer con el parámetro quotechar="'" y quoting=csv.QUOTE\_ALL para que cada campo quede entre comillas y exporte desde PostgreSQL respetando la estructura de las columnas.
- **exportar\_resumen.py:** Hace la conexión a la base de datos, y realiza la consulta para exportar el summary.

### automatización

ejecutar\_scrapers.bat para ejecutar spiders cada 2 días periódicamente.

### 6.2.Problemas Resueltos

- **Duplicación de artículos:** Se resolvió con una restricción 'UNIQUE' en el campo "url".
- **Inconsistencias en codificación:** Se forzó 'utf-8' y manejo de errores con try/except.
- **Error exportación desde PostgreSQL:** Se codifico un script que exporta correctamente los datos almacenados respetando los delimitadores de columnas.

## 7. Conclusiones

El sistema es funcional, automatizado, modular y escalable. La estructura de almacenamiento en (JSON + PostgreSQL) garantiza que sea más fácil para el análisis. La implementación de control de duplicados asegura la calidad del dataset evitando errores de duplicados o datos sobrescritos.

Se aplicó correctamente el patrón ETL con tres zonas del lago de datos.

El uso de Scrapy y Streamlit permitió integrar extracción, transformación y visualización de forma eficiente.

La integración con una API externa enriquece la experiencia del usuario y permite combinar fuentes estructuradas con servicios en tiempo real.