

## **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 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.

Se aplicaron múltiples pipelines para validación, almacenamiento en PostgreSQL y exportación a JSONL.

- **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.

Inserción directa a PostgreSQL, con creación automática de la tabla articles mediante `__init__()`.

Codificación en UTF-8 para soportar caracteres especiales.

Manejo de errores mediante try/except.

## Transformación y Validación

Procesamiento con Pipelines ETL

El sistema ahora está organizado en tres pipelines:

### 1. LandingPipeline

- Guarda los datos crudos en datalake/LANDING\_ZONE/articles\_raw.jsonl
- No realiza validación avanzada
- Imprime en consola:  [Landing] Guardado en JSON

### 2. RefinedPipeline

- Limpia los datos, elimina duplicados y transforma campos
- Inserta en tabla PostgreSQL articles\_refined
- Guarda archivo: datalake/REFINED\_ZONE/articles\_refined.jsonl
- Imprime:  [Refined] Insertado en PostgreSQL

### 3. ConsumptionPipeline

- Crea una vista final para análisis (sin transformación adicional)
- Guarda en datalake/CONSUMPTION\_ZONE/articles\_consumption.jsonl
- Inserta en articles\_consumption
- Imprime:  [Consumption] Insertado en PostgreSQL
- Cada pipeline es modular, activado desde settings.py con prioridad escalonada.

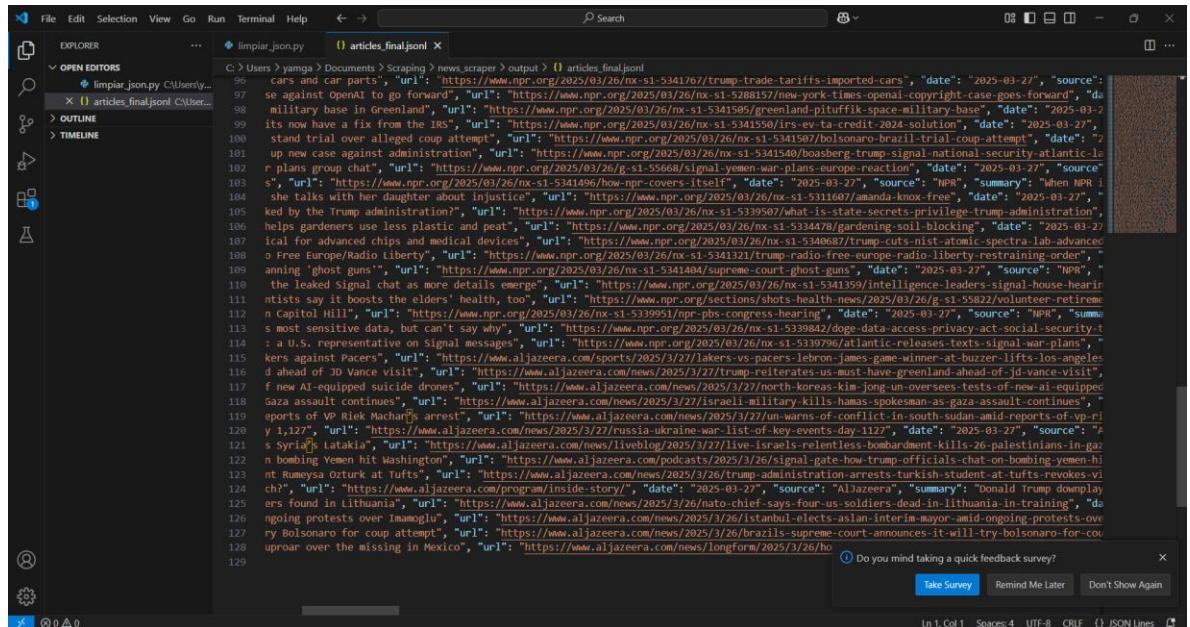
Además:

- Se crea automáticamente la tabla articles si no existe.
- Se genera la carpeta output/ si no está presente.
- Se notifica en consola cada artículo guardado.

## 2.1. Almacenamiento de datos

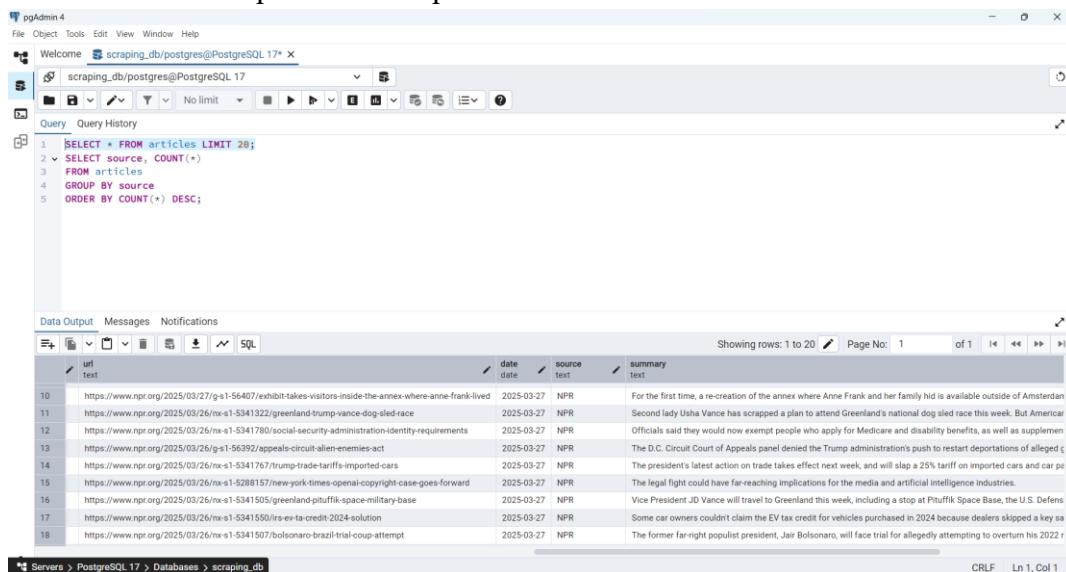
Los datos procesados se almacenan en:

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



```
C:\Users\yanga\Documents\Scraping\news_scrape>output>articles_final.jsonl
[{"url": "https://www.npr.org/2025/03/26/g-s1-5341767/trump-trade-tariffs-imported-cars", "date": "2025-03-27", "source": "NPR", "summary": "The president's latest action on trade takes effect next week, and will slap a 25% tariff on imported cars and car parts. The legal fight could have far-reaching implications for the media and artificial intelligence industries."}, {"url": "https://www.npr.org/2025/03/26/g-s1-5341732/greenland-trump-vance-dog-sled-race", "date": "2025-03-27", "source": "NPR", "summary": "Second lady Jill Biden has scrapped a plan to attend Greenland's national dog sled race this week. But American officials said they would now exempt people who apply for Medicare and disability benefits, as well as supplementals."}, {"url": "https://www.npr.org/2025/03/26/g-s1-5341760/social-security-administration-identity-requirements", "date": "2025-03-27", "source": "NPR", "summary": "The D.C. Circuit Court of Appeals panel denied the Trump administration's push to restart deportations of alleged illegal immigrants. The president's latest action on trade takes effect next week, and will slap a 25% tariff on imported cars and car parts."}, {"url": "https://www.npr.org/2025/03/26/g-s1-563972/appeals-circuit-alien-enemies-act", "date": "2025-03-27", "source": "NPR", "summary": "The president's latest action on trade takes effect next week, and will slap a 25% tariff on imported cars and car parts."}, {"url": "https://www.npr.org/2025/03/26/g-s1-5341767/trump-trade-tariffs-imported-cars", "date": "2025-03-27", "source": "NPR", "summary": "The president's latest action on trade takes effect next week, and will slap a 25% tariff on imported cars and car parts."}, {"url": "https://www.npr.org/2025/03/26/nx-s1-5341505/greenland-ptiflik-space-military-base", "date": "2025-03-27", "source": "NPR", "summary": "Vice President JD Vance will travel to Greenland this week, including a stop at Ptiflik Space Base, the U.S. Defense Department's newest space station."}, {"url": "https://www.npr.org/2025/03/26/nx-s1-5341550/irs-ev-ta-credit-2024-solution", "date": "2025-03-27", "source": "NPR", "summary": "Some car owners couldn't claim the EV tax credit for vehicles purchased in 2024 because dealers skipped a key step in the process. The IRS is trying to fix that."}, {"url": "https://www.npr.org/2025/03/26/nx-s1-5341507/bolsonaro-brazil-trial-coup-attempt", "date": "2025-03-27", "source": "NPR", "summary": "The former far-right populist president, Jair Bolsonaro, will face trial for allegedly attempting to overturn his 2022 re-election victory. The trial is set to begin in September."}, {"url": "https://www.npr.org/2025/03/26/g-s1-5341406/baoberg-trump-national-security-atlantic-leadership", "date": "2025-03-27", "source": "NPR", "summary": "The president's latest action on trade takes effect next week, and will slap a 25% tariff on imported cars and car parts."}, {"url": "https://www.npr.org/2025/03/26/g-s1-5341407/amanda-knox-free", "date": "2025-03-27", "source": "NPR", "summary": "The president's latest action on trade takes effect next week, and will slap a 25% tariff on imported cars and car parts."}, {"url": "https://www.npr.org/2025/03/26/g-s1-5339507/what-is-state-secrets-privilege-trump-administration", "date": "2025-03-27", "source": "NPR", "summary": "The president's latest action on trade takes effect next week, and will slap a 25% tariff on imported cars and car parts."}, {"url": "https://www.npr.org/2025/03/26/g-s1-5341408/gardening-soil-blocking", "date": "2025-03-27", "source": "NPR", "summary": "The president's latest action on trade takes effect next week, and will slap a 25% tariff on imported cars and car parts."}, {"url": "https://www.npr.org/2025/03/26/g-s1-5340687/trump-cuts-nist-atomic-spectra-lab-advanced", "date": "2025-03-27", "source": "NPR", "summary": "The president's latest action on trade takes effect next week, and will slap a 25% tariff on imported cars and car parts."}, {"url": "https://www.npr.org/2025/03/26/g-s1-5341321/trump-radio-free-europe-radio-liberty-restraining-order", "date": "2025-03-27", "source": "NPR", "summary": "The president's latest action on trade takes effect next week, and will slap a 25% tariff on imported cars and car parts."}, {"url": "https://www.npr.org/2025/03/26/g-s1-5341404/supreme-court-ghost-guns", "date": "2025-03-27", "source": "NPR", "summary": "The president's latest action on trade takes effect next week, and will slap a 25% tariff on imported cars and car parts."}, {"url": "https://www.npr.org/2025/03/26/g-s1-5341397/intelligence-leaders-signal-house-hearings", "date": "2025-03-27", "source": "NPR", "summary": "The president's latest action on trade takes effect next week, and will slap a 25% tariff on imported cars and car parts."}, {"url": "https://www.npr.org/2025/03/26/g-s1-5341399/volunteer-retirement", "date": "2025-03-27", "source": "NPR", "summary": "The president's latest action on trade takes effect next week, and will slap a 25% tariff on imported cars and car parts."}, {"url": "https://www.npr.org/2025/03/26/g-s1-5339951/pbs-congress-hearing", "date": "2025-03-27", "source": "NPR", "summary": "The president's latest action on trade takes effect next week, and will slap a 25% tariff on imported cars and car parts."}, {"url": "https://www.npr.org/2025/03/26/g-s1-5339842/doj-data-access-privacy-act-social-security", "date": "2025-03-27", "source": "NPR", "summary": "The president's latest action on trade takes effect next week, and will slap a 25% tariff on imported cars and car parts."}, {"url": "https://www.npr.org/2025/03/26/g-s1-5339796/atlantic-releases-texts-signal-war-plans", "date": "2025-03-27", "source": "NPR", "summary": "The president's latest action on trade takes effect next week, and will slap a 25% tariff on imported cars and car parts."}, {"url": "https://www.aljazeera.com/sports/2025/3/27/lakers-vs-pacers-lebron-james-game-winner-at-buzzer-lifts-los-angeles", "date": "2025-03-27", "source": "Aljazeera", "summary": "LeBron James scored 36 points and had 10 rebounds in a game-winning performance against the Indiana Pacers."}, {"url": "https://www.aljazeera.com/news/2025/3/27/jd-vance-visits-his-new-ai-equipped", "date": "2025-03-27", "source": "Aljazeera", "summary": "JD Vance, the son of Donald Trump, visited a facility where AI-equipped suicide drones are being developed."}, {"url": "https://www.aljazeera.com/news/2025/3/27/israeli-military-kills-hamas-spokesman-as-gaza-assault-continues", "date": "2025-03-27", "source": "Aljazeera", "summary": "Israel's military killed a Hamas spokesman during a raid in Gaza."}, {"url": "https://www.aljazeera.com/news/2025/3/27/warns-of-conflict-in-sudan-amid-reports-of-vp-ri", "date": "2025-03-27", "source": "Aljazeera", "summary": "U.S. Vice President Kamala Harris warned of conflict in Sudan amid reports of violence."}, {"url": "https://www.aljazeera.com/news/2025/3/27/russia-ukraine-war-list-of-key-events-day-1127", "date": "2025-03-27", "source": "Aljazeera", "summary": "A Russian representative on Signal messages revealed a list of key events in the war between Russia and Ukraine."}, {"url": "https://www.aljazeera.com/news/2025/3/27/live-israeli-relentless-bombardment-kills-26-palestinians-in-gaz", "date": "2025-03-27", "source": "Aljazeera", "summary": "An Israeli airstrike in Gaza killed 26 Palestinians."}, {"url": "https://www.aljazeera.com/news/2025/3/27/trump-administration-arrests-turkish-student-at-tufts-revokes-vi", "date": "2025-03-27", "source": "Aljazeera", "summary": "Donald Trump's administration arrested a Turkish student at Tufts University and revoked their visa."}, {"url": "https://www.aljazeera.com/news/2025/3/27/inside-story-us-soldiers-dead-in-lithuania-in-training", "date": "2025-03-27", "source": "Aljazeera", "summary": "Four U.S. soldiers were found dead in Lithuania while training."}, {"url": "https://www.aljazeera.com/news/2025/3/27/istanbul-elects-asian-interim-major-amid-ongoing-protests-over", "date": "2025-03-27", "source": "Aljazeera", "summary": "Istanbul elected its first Asian mayor despite ongoing protests over the missing president."}, {"url": "https://www.aljazeera.com/news/2025/3/26/brazils-supreme-court-announces-it-will-try-bolsonaro-for-cor", "date": "2025-03-26", "source": "Aljazeera", "summary": "Brazil's Supreme Court announced it will try Jair Bolsonaro for corruption."}, {"url": "https://www.aljazeera.com/news/2025/3/26/longform/2025/3/26", "date": "2025-03-26", "source": "Aljazeera", "summary": "A long-form article by Aljazeera."}], [{"text": "Do you mind taking a quick feedback survey?"}, {"text": "Take Survey"}, {"text": "Remind Me Later"}, {"text": "Don't Show Again"}]
```

- **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.



	url	date	source	summary
10	https://www.npr.org/2025/03/26/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/g-s1-5341322/greenland-trump-vance-dog-sled-race	2025-03-27	NPR	Second lady Jill Biden has scrapped a plan to attend Greenland's national dog sled race this week. But American officials said they would now exempt people who apply for Medicare and disability benefits, as well as supplementals.
12	https://www.npr.org/2025/03/26/g-s1-5341760/social-security-administration-identity-requirements	2025-03-27	NPR	The D.C. Circuit Court of Appeals panel denied the Trump administration's push to restart deportations of alleged illegal immigrants.
13	https://www.npr.org/2025/03/26/g-s1-563972/appeals-circuit-alien-enemies-act	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 parts.
14	https://www.npr.org/2025/03/26/g-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 parts.
15	https://www.npr.org/2025/03/26/nx-s1-5328157/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/nx-s1-5341505/greenland-ptiflik-space-military-base	2025-03-27	NPR	Vice President JD Vance will travel to Greenland this week, including a stop at Ptiflik Space Base, the U.S. Defense Department's newest space station.
17	https://www.npr.org/2025/03/26/nx-s1-5341550/irs-ev-ta-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 step in the process.
18	https://www.npr.org/2025/03/26/nx-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 re-election victory.

### 3. Flujo de trabajo

El flujo de trabajo del scraping es el siguiente:

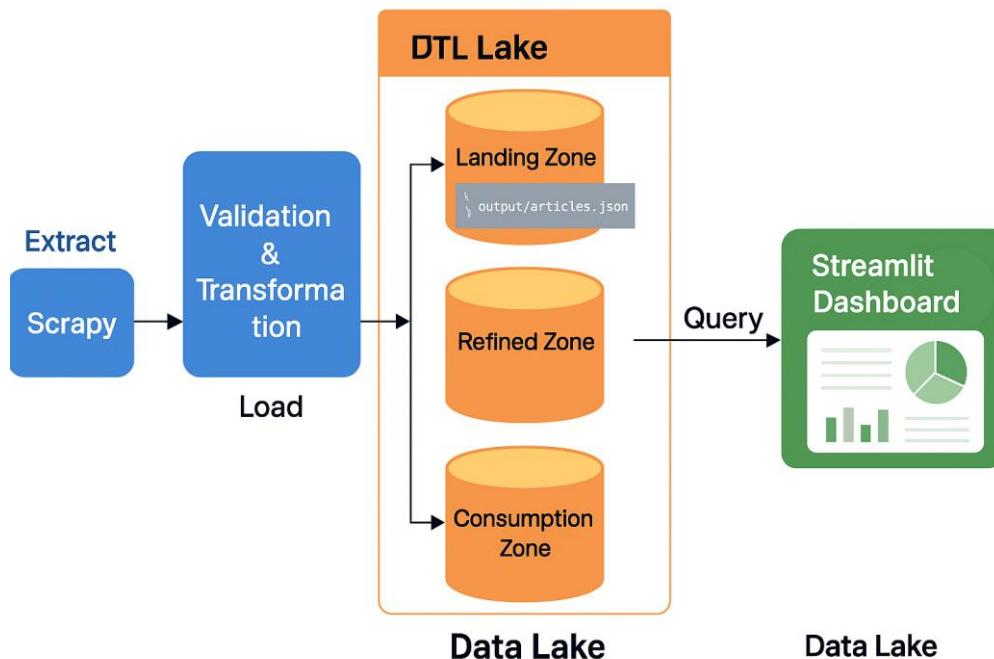
1. Ejecución del spider (Scrapy)
2. Extracción de los artículos
3. Limpieza de los campos con validaciones en pipelines.py
4. Almacenamiento simultáneo en:
  - Archivo JSONL (output/articles\_final.jsonl)
  - Tabla articles en PostgreSQL

### 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

Archivo: dashboard/dashboard.py

Usa dotenv para variables de conexión (.env)

Conecta con PostgreSQL y muestra:

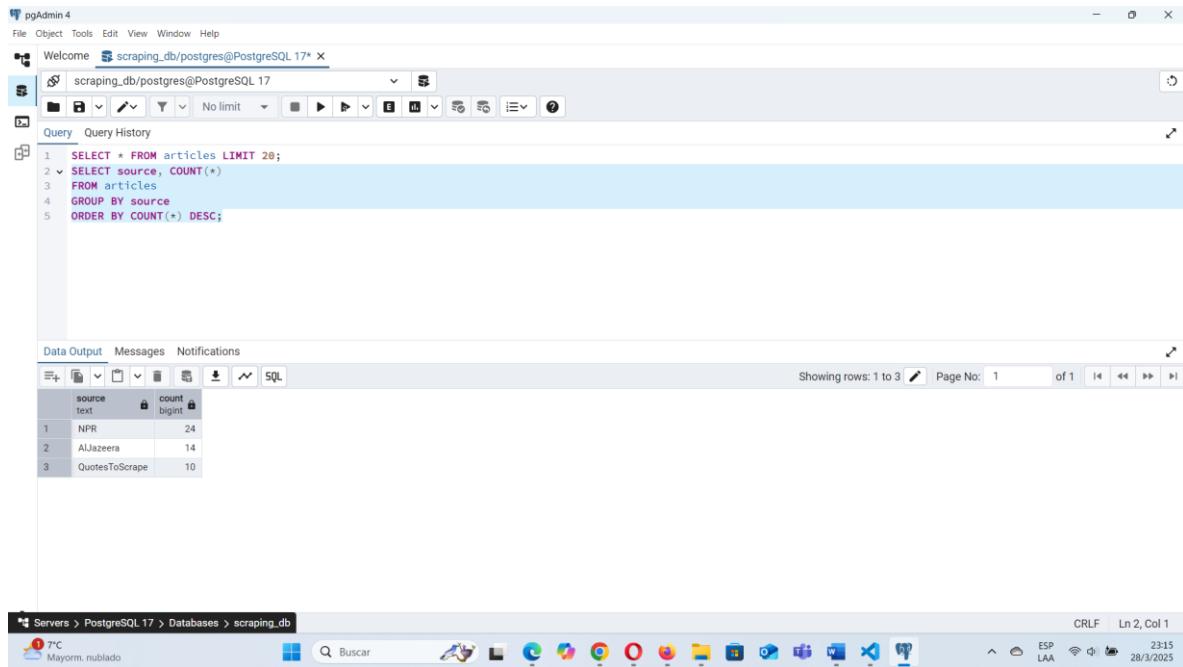
- Total, de artículos scrapeados (st.metric)
- Tabla con st.dataframe()
- Gráfica de fuentes por cantidad (st.bar\_chart())

Se integra con la API externa OpenWeatherMap:

- El usuario ingresa una ciudad
- Se muestra el clima actual con descripción y temperatura

## 6. Conclusiones

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



The screenshot shows the pgAdmin 4 interface. In the top-left corner, it says "pgAdmin 4". The menu bar includes "File", "Object", "Tools", "Edit", "View", "Window", and "Help". Below the menu is a toolbar with various icons. The main window has a title bar "Welcome scraping\_db/postgres@PostgreSQL 17\*". Underneath is a toolbar with buttons for file operations like "New", "Open", "Save", etc. A dropdown menu shows "scraping\_db/postgres@PostgreSQL 17". Below the toolbar is a "Query History" section with a dropdown menu set to "Query". The main area contains a code editor with the following SQL script:

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

Below the code editor is a "Data Output" tab. It displays a table with three rows:

source	count
NPR	24
AlJazeera	14
QuotesToScrape	10

At the bottom of the pgAdmin window, there's a status bar showing "Showing rows: 1 to 3 | Page No: 1 of 1".

Below the pgAdmin window, the Windows taskbar is visible. It shows the system tray with icons for battery, signal, and date/time (23:15, 28/3/2025). The taskbar also has icons for the Start button, Search, Task View, File Explorer, and other system applications.

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_scaper\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_scaper\output>
```

Y ambos arrojan los mismos resultados.

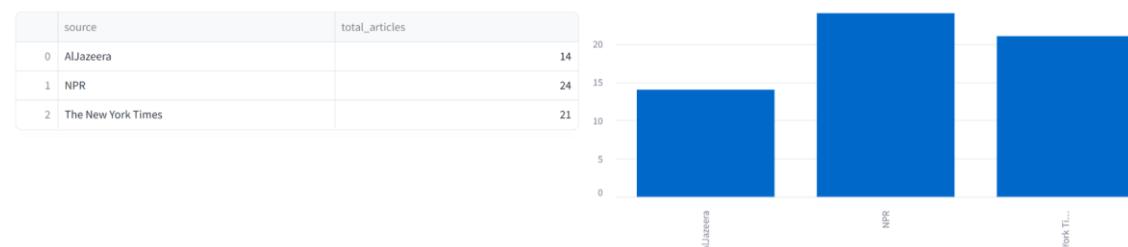
Se extrajo la base de datos desde el PostgreSQL en formato csv en (output/articles\_postgres.csv)

ID	title	url	date	source	summary
1	"The world as we have created it is a process of our thinking. It cannot be changed without changing our	https://quotes.toscrape.com/page/1/	27/3/2025	QuotesToScrape	Quote by Albert Eins
3	"This life is what you make it. No matter what, you're going to mess up sometimes, it's a universal truth.	https://quotes.toscrape.com/page/2/	27/3/2025	QuotesToScrape	Quote by Marilyn Mo
4	"I love you without knowing how, or where, or from where. I love you simply, without problems or pride."	https://quotes.toscrape.com/page/3/	27/3/2025	QuotesToScrape	Quote by Pablo Ner
5	"The more that you read, the more things you will know. The more that you learn, the more places you'll	https://quotes.toscrape.com/page/4/	27/3/2025	QuotesToScrape	Quote by Dr. Seuss
6	"A reader lives a thousand lives before he dies, said Jojen. The man who never reads lives only one."	https://quotes.toscrape.com/page/5/	27/3/2025	QuotesToScrape	Quote by George R.R. J
7	"There is nothing I would not do for those who are really my friends. I have no notion of hating people by	https://quotes.toscrape.com/page/6/	27/3/2025	QuotesToScrape	Quote by Jane Auste
8	"That's the problem with drinking, I thought, as I poured myself a drink. If something bad happens you'd	https://quotes.toscrape.com/page/7/	27/3/2025	QuotesToScrape	Quote by Charles Bu
9	"If I had a flower for every time I thought of you... I could walk through my garden forever."	https://quotes.toscrape.com/page/8/	27/3/2025	QuotesToScrape	Quote by Alfred Ein
10	"Anyone who has never made a mistake has never tried anything new."	https://quotes.toscrape.com/page/9/	27/3/2025	QuotesToScrape	Quote by Albert Eins
11	"The truth." Dumbledore sighed. "It is a beautiful and terrible thing, and should therefore be treated with	https://quotes.toscrape.com/page/10/	27/3/2025	QuotesToScrape	Quote by J.K. Rowlin
12	"Climate change and overfishing threaten Vietnam's tradition of making fish sauce	https://www.npr.org/2025/03/27/11-5342247/vietnam-fish-sauce-climate-change-overfishing	27/3/2025	NPR	Climate change and
13	Exhibit takes visitors inside the annex where Anne Frank lived	https://www.npr.org/2025/03/27/11-564077/exhibit-takes-visitors-inside-the-annex-where-anne-fra	27/3/2025	NPR	For the first time, a
14	As U.S. foreign aid grants get slashed, Greenland gets money for a dog race	https://www.npr.org/2025/03/26/11-5341322/greenland-trump-vance-dog-sled-race	27/3/2025	NPR	Second lady Jillia Vi
15	Social Security officials partially walk back plans for in-person verification	https://www.npr.org/2025/03/26/11-5341780/social-security-administration-identity-requirement	27/3/2025	NPR	Officials said they w
16	Appeals court sides with judge who blocked deportations under wartime authority	https://www.npr.org/2025/03/26/11-56392/appeals-circuit-alien-enemies-act	27/3/2025	NPR	The D.C. Circuit Cou
17	Trump announces new 25% tariff on imported cars and car parts	https://www.npr.org/2025/03/26/11-531767/trump-trade-tariffs-imported-cars	27/3/2025	NPR	The president's late
18	Judge allows 'New York Times' copyright case against OpenAI to go forward	https://www.npr.org/2025/03/26/11-5288157/new-york-times-openai-copyright-case-goes-forward	27/3/2025	NPR	The legal fight could
19	What to know about Pituffik, the only U.S. military base in Greenland	https://www.npr.org/2025/03/26/11-5341505/greenland-pituffik-space-military-base	27/3/2025	NPR	Vice President JD Va
20	EV buyers who missed out on their tax credits now have a fix from the IRS	https://www.npr.org/2025/03/26/11-5341550/irs-ev-a-credit-2024-solution	27/3/2025	NPR	Some car owners ca
21	Brazil's Supreme Court says Bolsonaro must stand trial over alleged coup attempt	https://www.npr.org/2025/03/26/11-5341507/bolsonaro-brazil-trial-coup-attempt	27/3/2025	NPR	The former far-right
22	Federal judge who drew Trump's anger picks up new case against administration	https://www.npr.org/2025/03/26/11-5341540/baasberg-trump-signal-national-security-atlantic-la	27/3/2025	NPR	James Baasberg, ch
23	How Europeans are reacting to the Yemen war plans group chat	https://www.npr.org/2025/03/26/11-5568/ignal-yemen-war-plans-europe-reaction	27/3/2025	NPR	In Europe, there's no
24	Why Amanda Knox returns to Italy — and how she talks with her daughter about injustice	https://www.npr.org/2025/03/26/11-5341496/how-npr-covers-itself	27/3/2025	NPR	When NPR is in the r
25	Why Amanda Knox returns to Italy — and how she talks with her daughter about injustice	https://www.npr.org/2025/03/26/11-5311607/amanda-knox-free	27/3/2025	NPR	Amanda Knox spent
26	What is the 'state secrets privilege' invoked by the Trump administration?	https://www.npr.org/2025/03/26/11-5338507/what-is-state-secrets-privilege-trump-administrativ	27/3/2025	NPR	The state secrets pri

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



### Estadísticas por fuente de noticias



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.

## Verificación de funcionamiento:

- Mensajes  Guardado: Título visibles en consola
- Confirmación visual en archivo articles\_final.jsonl
- Consulta directa a PostgreSQL para ver los artículos

## 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.

La arquitectura del lago de datos fue correctamente implementada.

Se aplicaron buenas prácticas de scraping y almacenamiento.

La integración de un dashboard Streamlit con API externa enriquece el análisis de los datos.