

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.

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

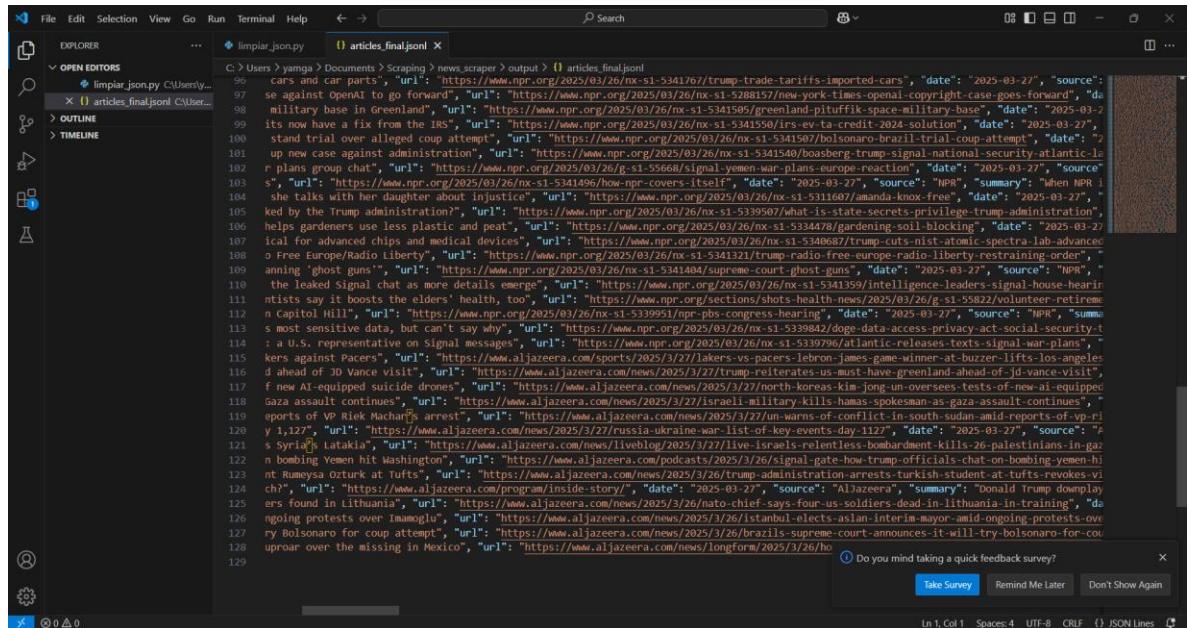
Codificación en UTF-8 para soportar caracteres especiales.

Manejo de errores mediante try/except.

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.



```

File Edit Selection View Go Run Terminal Help ⌘ Search
C:\> Users > yampa > Documents > Scraping > news_scraping > output > articles_final.jsonl
OPEN EDITORS
limpiar_json.py C:\Users\...
articles_final.jsonl C:\Users\...
OUTLINE
TIMELINE

1 cars and car parts", "url": "https://www.npr.org/2025/03/26/nx-s1-5341767/trump-trade-tariffs-imported-cars", "date": "2025-03-27", "source": "NPR", "summary": "Trump trade tariffs imported cars", "id": 1
2 "against OpenAI to go forward", "url": "https://www.npr.org/2025/03/26/nx-s1-5288157/new-york-times-openai-copyright-case-goes-forward", "date": "2025-03-27", "source": "NPR", "summary": "New York Times OpenAI copyright case goes forward", "id": 2
3 "military base in Greenland", "url": "https://www.npr.org/2025/03/26/nx-s1-5341505/greenland-pitiful-space-military-base", "date": "2025-03-27", "source": "NPR", "summary": "Greenland pitiful space military base", "id": 3
4 "its now have a fix from the IRS", "url": "https://www.npr.org/2025/03/26/nx-s1-5341550/irs-ev-ta-credit-2024-solution", "date": "2025-03-27", "source": "NPR", "summary": "IRS ev ta credit 2024 solution", "id": 4
5 "stand trial over alleged coup attempt", "url": "https://www.npr.org/2025/03/26/nx-s1-5341507/bolsonaro-brasil-trial-coup-attempt", "date": "2025-03-27", "source": "NPR", "summary": "Bolsonaro Brasil trial coup attempt", "id": 5
6 "up new case against administration", "url": "https://www.npr.org/2025/03/26/nx-s1-5341540/baoberg-trump-signal-national-security-atlantic-laws", "date": "2025-03-27", "source": "NPR", "summary": "Baoberg Trump Signal National Security Atlantic Laws", "id": 6
7 "r plans group chat", "url": "https://www.npr.org/2025/03/26/nx-s1-5341496/how-npr-covers-itself", "date": "2025-03-27", "source": "NPR", "summary": "How NPR covers itself", "id": 7
8 "s", "url": "https://www.npr.org/2025/03/26/nx-s1-5341495/how-npr-covers-itself", "date": "2025-03-27", "source": "NPR", "summary": "When NPR talks with her daughter about injustice", "id": 8
9 "she talks with her daughter about injustice", "url": "https://www.npr.org/2025/03/26/nx-s1-5311607/amanda-knox-free", "date": "2025-03-27", "source": "NPR", "summary": "Amanda Knox free", "id": 9
10 "ked by the Trump administration?", "url": "https://www.npr.org/2025/03/26/nx-s1-5339507/what-is-state-secrets-privilege-trump-administration", "date": "2025-03-27", "source": "NPR", "summary": "What is state secrets privilege Trump administration", "id": 10
11 "helps gardeners use less plastic and peat", "url": "https://www.npr.org/2025/03/26/nx-s1-5334478/gardening-soil-blocking", "date": "2025-03-27", "source": "NPR", "summary": "Gardening soil blocking", "id": 11
12 "ical for advanced chips and medical devices", "url": "https://www.npr.org/2025/03/26/nx-s1-5340687/trump-cuts-nist-atomic-spectra-lab-advanced", "date": "2025-03-27", "source": "NPR", "summary": "Trump cuts NIST atomic spectra lab advanced", "id": 12
13 "a Free Europe Radio liberty", "url": "https://www.npr.org/2025/03/26/nx-s1-5341321/trump-radio-free-europe-radio-liberty-restraining-order", "date": "2025-03-27", "source": "NPR", "summary": "Trump radio free europe radio liberty restraining order", "id": 13
14 "anning 'ghost guns'", "url": "https://www.npr.org/2025/03/26/nx-s1-5341404/supreme-court-ghost-guns", "date": "2025-03-27", "source": "NPR", "summary": "Supreme court ghost guns", "id": 14
15 "the leaked Signal chat as more details emerge", "url": "https://www.npr.org/2025/03/26/nx-s1-5341359/intelligence-leaders-signal-house-hearings", "date": "2025-03-27", "source": "NPR", "summary": "Intelligence leaders signal house hearings", "id": 15
16 "tists say it boosts the elders' health, too", "url": "https://www.npr.org/sections/shots-health-news/2025/03/26/g/55822/volunteer-retirees", "date": "2025-03-27", "source": "NPR", "summary": "Volunteer retirees", "id": 16
17 "n Capitol Hill", "url": "https://www.npr.org/2025/03/26/nx-s1-5339951/mr-phs-congress-hearing", "date": "2025-03-27", "source": "NPR", "summary": "Mr phs congress hearing", "id": 17
18 "summarized most sensitive data, but can't say why", "url": "https://www.npr.org/2025/03/26/nx-s1-5339842/doe-data-access-privacy-act-social-security", "date": "2025-03-27", "source": "NPR", "summary": "Doe data access privacy act social security", "id": 18
19 "a U.S. representative on Signal messages", "url": "https://www.npr.org/2025/03/26/nx-s1-5339796/atlasic-releases-texts-signal-war-plans", "date": "2025-03-27", "source": "NPR", "summary": "Atlasic releases texts signal war plans", "id": 19
20 "bers against Pacers", "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 game winner at buzzer lifts Los Angeles", "id": 20
21 "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", "date": "2025-03-27", "source": "Aljazeera", "summary": "Trump reiterates US must have Greenland ahead of JD Vance visit", "id": 21
22 "f new AI-equipped suicide drones", "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": "Israeli military kills Hamas spokesman as Gaza assault continues", "id": 22
23 "reports of VP Riek Machar's arrest", "url": "https://www.aljazeera.com/news/2025/3/27/un-warns-of-conflict-in-south-sudan-amid-reports-of-vp-r", "date": "2025-03-27", "source": "Aljazeera", "summary": "Un warns of conflict in South Sudan amid reports of VP R", "id": 23
24 "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": "Aljazeera", "summary": "Russia Ukraine war list of key events day 1127", "id": 24
25 "s Syria", "url": "https://www.aljazeera.com/news/liveblog/2025/3/27/live-israels-relentless-bombardment-kills-26-palestinians-in-gaz", "date": "2025-03-27", "source": "Aljazeera", "summary": "Israels relentless bombardment kills 26 Palestinians in Gaza", "id": 25
26 "21 Latakia", "url": "https://www.aljazeera.com/news/2025/3/27/live-israels-relentless-bombardment-kills-26-palestinians-in-gaz", "date": "2025-03-27", "source": "Aljazeera", "summary": "Israels relentless bombardment kills 26 Palestinians in Gaza", "id": 26
27 "22 bombing Yemen hit Washington", "url": "https://www.aljazeera.com/podcasts/2025/3/27/signal-gate-hon-trump-officials-chat-on-bombing-yemen-hi", "date": "2025-03-27", "source": "Aljazeera", "summary": "Gate hon trump officials chat on bombing Yemen hit", "id": 27
28 "23 nt Rumyeasa Ozturk at Tufts", "url": "https://www.aljazeera.com/news/2025/3/26/trump-administration-arrests-turkish-student-at-tufts-revokes-vi", "date": "2025-03-27", "source": "Aljazeera", "summary": "Trump administration arrests Turkish student at Tufts revokes visa", "id": 28
29 "24 ch", "url": "https://www.aljazeera.com/program/inside-story/", "date": "2025-03-27", "source": "Aljazeera", "summary": "Donald Trump downplay", "id": 29
30 "25 ers found in Lithuania", "url": "https://www.aljazeera.com/news/2025/3/26/nato-chief-says-four-us-soldiers-dead-in-lithuania-in-training", "date": "2025-03-27", "source": "Aljazeera", "summary": "Four US soldiers dead in Lithuania in training", "id": 30
31 "26 ngoin protests over immigral", "url": "https://www.aljazeera.com/news/2025/3/26/istambul-elects-asian-interim-major-amid-ongoing-protests-over", "date": "2025-03-27", "source": "Aljazeera", "summary": "Asian interim major amid ongoing protests over", "id": 31
32 "27 Bolsonaro for coup attempt", "url": "https://www.aljazeera.com/news/2025/3/26/brazils-supreme-court-announces-it-will-try-bolsonaro-for-cou", "date": "2025-03-27", "source": "Aljazeera", "summary": "Brazilian Supreme Court announces it will try Bolsonaro for coup attempt", "id": 32
33 "28 uproar over the missing in Mexico", "url": "https://www.aljazeera.com/news/longform/2025/3/26/no", "date": "2025-03-27", "source": "Aljazeera", "summary": "Mexico missing", "id": 33

```

Do you mind taking a quick feedback survey?

In 1, Col 1 Spaces: 4 UTF-8 CR LF {} JSON Lines

- **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. At the top, there's a menu bar with File, Object, Tools, Edit, View, Window, Help. Below it is a toolbar with icons for New, Open, Save, Print, etc. The main window has a title bar 'Welcome scraping_db/postgres@PostgreSQL 17*' and a tab bar with 'Query' and 'Query History'. A code editor pane contains the following SQL query:

```

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

```

Below the code editor is a data grid titled 'Data Output' with columns: url, date, source, and summary. The data shows 18 rows of news articles from NPR, each with its URL, publication date, source (NPR), and a brief summary. The first few rows are:

	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/nx-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 officials said they would now exempt people who apply for Medicare and disability benefits, as well as supplement...
12	https://www.npr.org/2025/03/26/nx-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-circuit-alien-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...
14	https://www.npr.org/2025/03/26/nx-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-s1-5268157/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-plutnik-space-military-base	2025-03-27	NPR	Vice President JD Vance will travel to Greenland this week, including a stop at Plutnik Space Base, the U.S. Defense...
17	https://www.npr.org/2025/03/26/nx-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 step.
18	https://www.npr.org/2025/03/26/nx-s1-5341507/bolsonaro-brasil-trail-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.

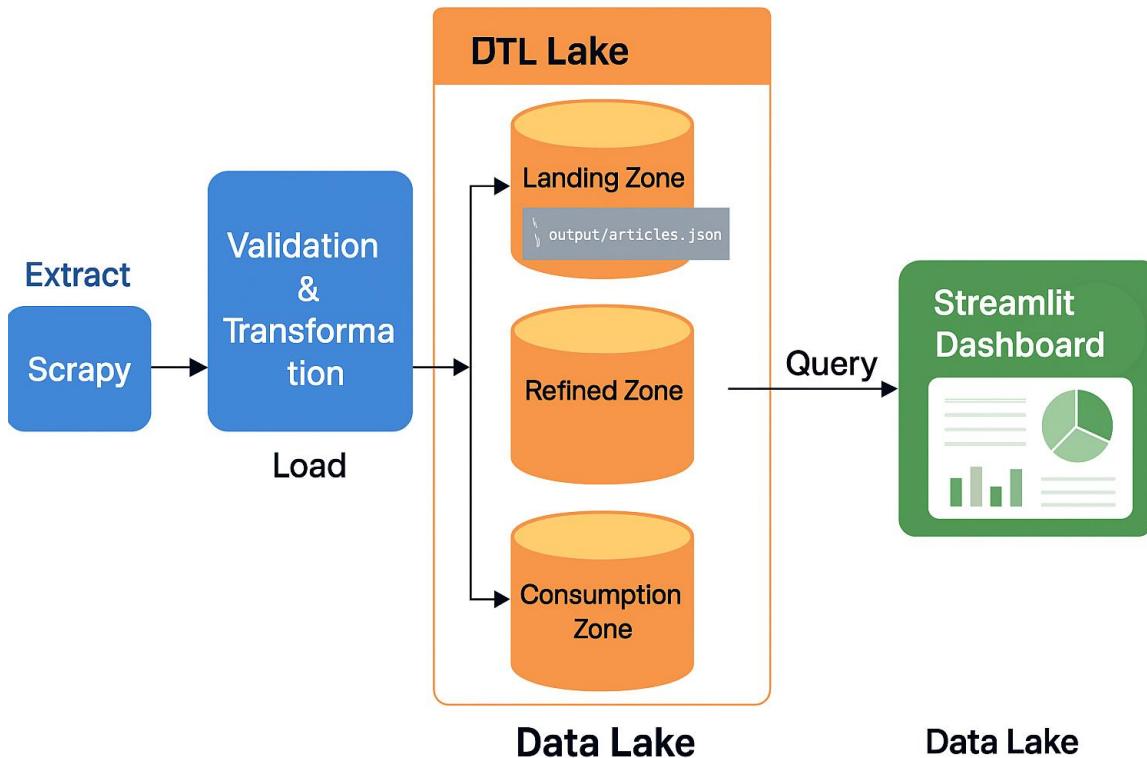
Con la siguiente estructura:

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

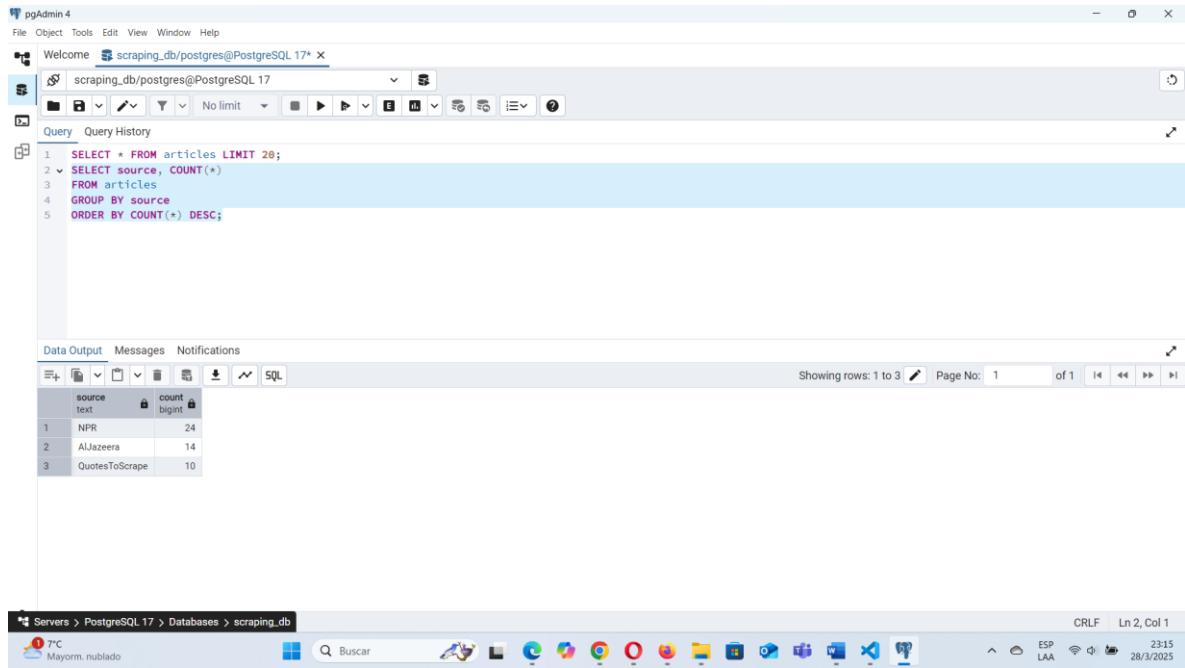
- Total, de artículos scrapeados (con `st.metric()`)
- Tabla de artículos por fuente (`st.dataframe()`)
- Gráfico de barras (`st.bar_chart()`)

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.



The screenshot shows the pgAdmin 4 interface. At the top, there's a menu bar with File, Object, Tools, Edit, View, Window, Help. Below it is a toolbar with various icons. The main area has a title bar "Welcome scraping_db/postgres@PostgreSQL 17*". Underneath is a query editor window containing 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;
```

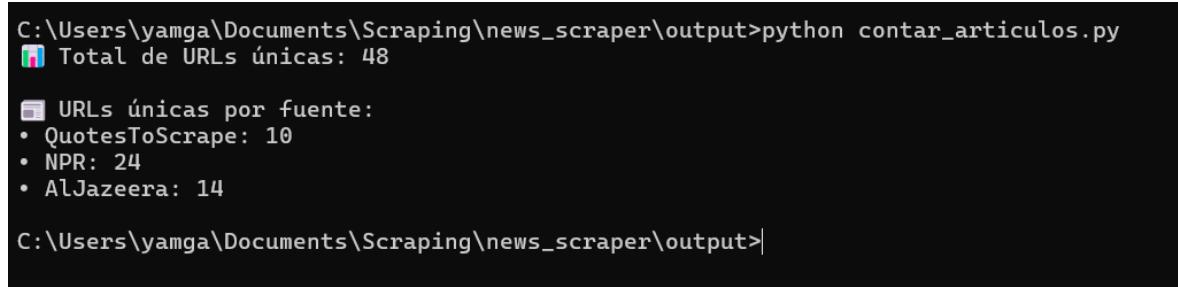
Below the query editor is a "Data Output" tab showing the results of the last query:

source	count
NPR	24
AlJazeera	14
QuotesToScrape	10

At the bottom of the pgAdmin window, there's a status bar with "Showing rows: 1 to 3" and a page number "1 of 1".

Below the pgAdmin window, the Windows taskbar is visible, showing the system tray with icons for battery, signal, and date/time (28/3/2025).

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_scraping\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_scraping\output>
```

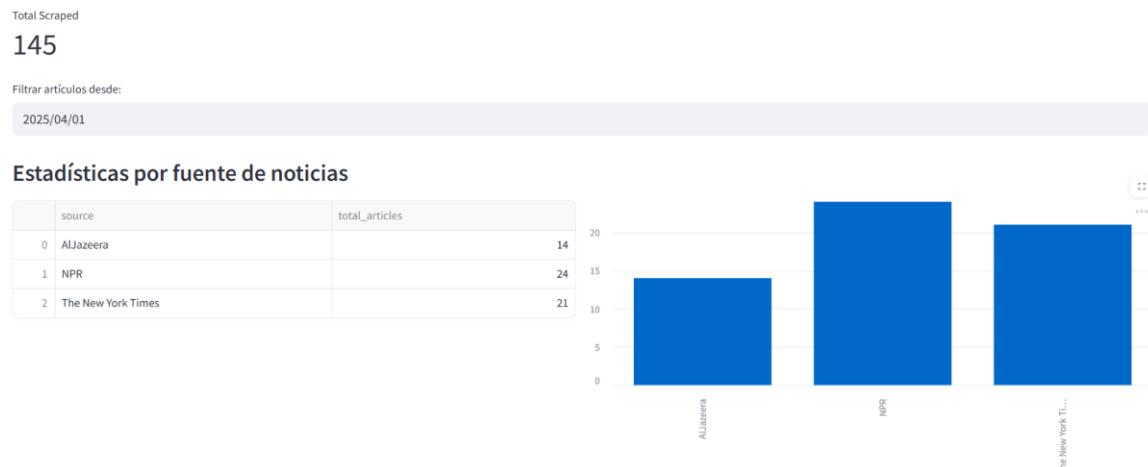
Y ambos arrojan los mismos resultados.

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

Screenshot of Microsoft Excel showing a table named "articles_postgres" with columns: id, title, url, date, source, and summary. The table contains approximately 140 rows of news articles from various sources like NPR, Al Jazeera, and The New York Times.

	A	B	C	D	E
1	id	title	url	date	source
2	1139	"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
3	1149	"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
4	1159	"I love you without knowing how, or when, or from where. I love you simply, without problems or pride." https://quotes.toscrape.com/page/3/		27/3/2025	QuotesToScrape
5	1169	"The more that you read, the more things you'll know. The more that you learn, the more places you'll go." https://quotes.toscrape.com/page/4/		27/3/2025	QuotesToScrape
6	1179	"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
7	1189	"There is nothing I would not do for those who are really my friends. I have no notion of loving people by heart." https://quotes.toscrape.com/page/6/		27/3/2025	QuotesToScrape
8	1199	"That's the problem with drinking, I thought, as I poured myself a drink. If something bad happens you dr https://quotes.toscrape.com/page/7/		27/3/2025	QuotesToScrape
9	1209	"If I had a flower for every time I thought of you...I could walk through my garden forever."		27/3/2025	QuotesToScrape
10	1219	"Anyone who has never made a mistake has never tried anything new."		27/3/2025	QuotesToScrape
11	1229	"The truth," Dumbledore sighed. "It's a beautiful and terrible thing, and should therefore be treated with https://quotes.toscrape.com/page/10/		27/3/2025	QuotesToScrape
12	1239	Climate change and overfishing threaten Vietnam's tradition of making fish sauce.		27/3/2025	NPR
13	1240	Exhibit takes visitors inside the annex where Anne Frank lived		27/3/2025	NPR
14	1241	As U.S. foreign aid grants get slashed, Greenland gets money for a dog race		27/3/2025	NPR
15	1242	Social life Security officials partially walk back plans for in-person verification		27/3/2025	NPR
16	1243	Appeals court sides with judge who blocked deportations under wartime authority		27/3/2025	NPR
17	1244	Trump announces new 25% tariff on imported cars and car parts		27/3/2025	NPR
18	1245	Judge allows 'New York Times' copyright case against OpenAI to go forward		27/3/2025	NPR
19	1246	What to know about Pituffik, the only U.S. military base in Greenland		27/3/2025	NPR
20	1247	EV buyers who missed out on their tax credits now have a fix from the IRS		27/3/2025	NPR
21	1248	Brazil's Supreme Court says Bolsonaro must stand trial over alleged coup attempt		27/3/2025	NPR
22	1249	Federal judge who drew Trump's anger picks up new case against administration		27/3/2025	NPR
23	1250	How Europeans are reacting to the Yemen war plans group chat		27/3/2025	NPR
24	1251	Why NPR covers itself when it's in the news		27/3/2025	NPR
25	1252	Why Amanda Knox returns to Italy — and how she talks with her daughter about injustice		27/3/2025	NPR
26	1253	What is the 'state secrets privilege' invoked by the Trump administration?		27/3/2025	NPR

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.

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.

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.

El pipeline ahora crea automáticamente la tabla en PostgreSQL y valida todos los datos.

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