



Data Engineering Project: Real-Time Top News Pipeline & Dashboard (USA & Canada)

Project Overview:

I built a fully automated data pipeline and dashboard that fetches, processes, and visualizes the top 10 news headlines from the USA and Canada—using only open-source tools, all running locally on my MacBook Pro.



Tech Stack

- Python (requests, pandas, SQLAlchemy)
- PostgreSQL (local database)
- Streamlit (interactive dashboard)
- NewsAPI (free news data source)



What I Built

1. Automated Data Ingestion:

- Fetches the latest top headlines from both the USA and Canada using the NewsAPI.
- No manual downloads—data is pulled live via API.

2. Data Transformation & Cleaning:

- Cleans and standardizes the news data (e.g., flattens nested fields, handles missing values, standardizes column names).

3. Data Storage:

- Loads the cleaned data into a local PostgreSQL database for persistence and querying.

4. Interactive Visualization:

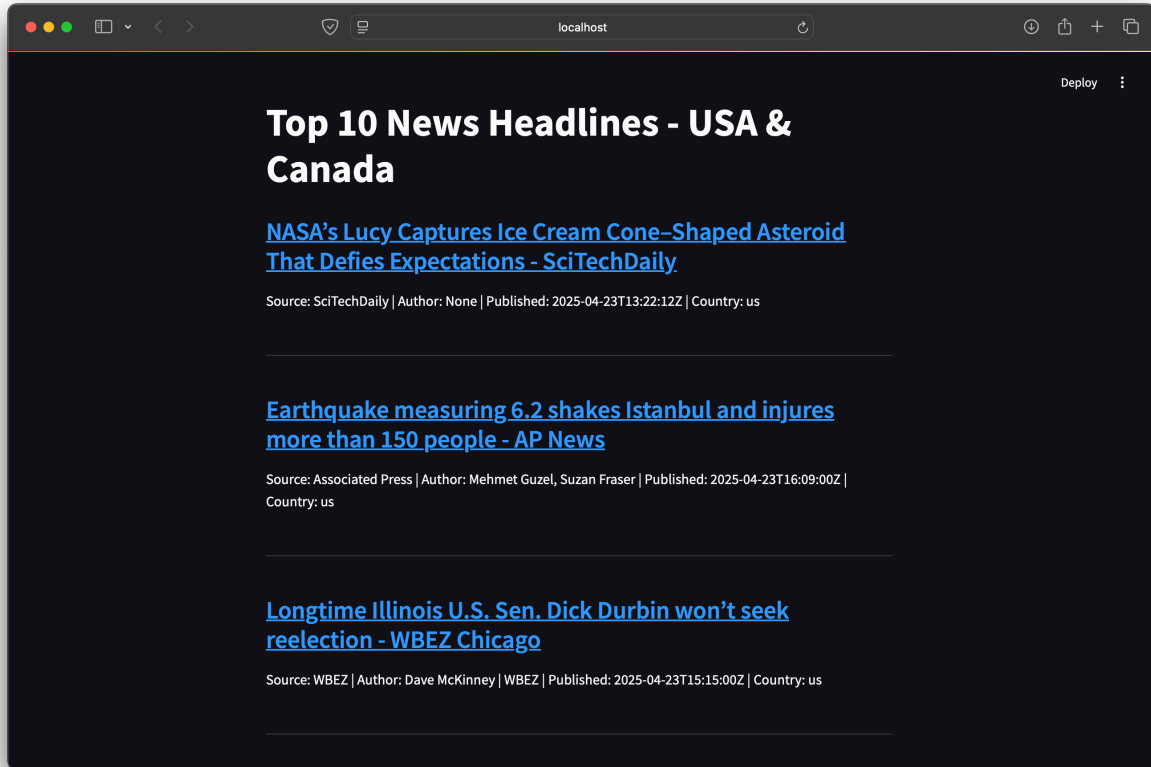
- Launches a Streamlit dashboard that displays the latest news headlines, sources, authors, and publication times, with clickable links to the original articles.

5. One-Click Pipeline:

- The entire workflow runs from a single script—just one command to fetch, process, store, and visualize the latest news.



Demo Screenshot



Key Features & Skills Demonstrated

- **API Integration:** Automated retrieval of live news data.
- **ETL Pipeline:** Real-world Extract, Transform, Load workflow with robust error handling.
- **Database Management:** Schema alignment and troubleshooting with PostgreSQL.
- **Data Visualization:** User-friendly, interactive dashboard with Streamlit.
- **Open Source & Local:** No paid services, no cloud dependency, fully reproducible on any Mac.



Project Motivation

I wanted to demonstrate my ability to design and implement a full-stack data pipeline using only free, open-source tools. This project simulates a real-world data engineering scenario—automating ingestion, transformation, storage, and visualization of dynamic external data.



Code & Documentation

- Check out the full code and setup instructions on GitHub:
- GitHub Repository Link (replace with your actual repo URL)



Let's Connect!

If you're interested in data engineering, data science, or building robust data pipelines, let's connect! I'm always open to new opportunities and collaborations.

#DataEngineering #Python #PostgreSQL #Streamlit #APIs #ETL #OpenSource
#PortfolioProject