

Techcrunch Webscrape Data Pipeline

Data Sources

- TechCrunch RSS Feed
- Accessed via Firecrawl API



Ingestion

- Python scripts
- Firecrawl API call (POST request)
- Extracted raw markdown content
- GitHub Actions (daily scheduled pipeline)



Processing

- Convert markdown to HTML using markdown package
- Parse HTML with BeautifulSoup
- Filter valid <a> tags with:-title > 5 words, link starts with http
 - Infer published_at from URL (regex & datetime)
- Clean DataFrame: Strip whitespace, Drop duplicates, Normalize title/link fields
- Add metadata: source = 'TechCrunch'



Storage

- PostgreSQL RDS Instance
 - Schema: raw
- Tables: techcrunch_articles
Columns: title, link, published_at, source



Users

- Looker Studio Dashboards
- SQL Notebooks (Jupyter)
- Final Presentation Assets (PDF/Slides)