**Учебный проект: Анализ новостных статей**

**Задача:**  
Собрать, обработать и сохранить данные из новостных сайтов, чтобы анализировать самые популярные темы.

**ETL-пайплайн:**

1. **Extract (извлечение данных)**
   * Используем API новостного агрегатора (например, [NewsAPI](https://newsapi.org/) или [Mediastack](https://mediastack.com/)) или парсим сайты с BeautifulSoup.
   * Загружаем статьи по заданным темам (например, "технологии", "наука").
2. **Transform (обработка данных)**
   * Очищаем текст (удаляем HTML, стоп-слова, лишние символы).
   * Анализируем текст:
     + Подсчет самых частых слов.
     + Определение тональности (можно использовать TextBlob или VADER).
     + Определение главных тем (LDA или просто TF-IDF).
3. **Load (загрузка в хранилище)**
   * Сохраняем в PostgreSQL:
     + Таблица articles (заголовок, источник, дата, текст, тема).
     + Таблица stats (частотность слов, тональность, дата).
   * Альтернативный вариант — складывать обработанные данные в MinIO в формате JSON/Parquet.

**Расширения:**

* Добавить отправку данных в Dashboard (например, через Streamlit или Metabase).
* Доработать с Airflow (автоматизировать обновление данных).

**Почему этот проект полезен?**

* Ты научишься работать с API и парсингом.
* Попрактикуешься с текстовой аналитикой.
* Закрепишь работу с базами данных и MinIO.
* В будущем можно легко адаптировать под Airflow.