Project Ideas - Sandeep Sharma

Table of Contents

OPTION A: News & Media Coverage Aggregator

Staying informed in today's world about current events or trends across different websites is challenging. Project aims to automate gathering articles, images and videos from various websites, store them in **Hadoop for fault tolerant storage and analyze them in RDMS** to extract insight about sentiment, topic and coverage.

Why It's Meaningful?

1. **Real-world Relevance**: News/media analytics is a huge field—journalists, researchers, and social scientists often need aggregated data.
2. **Handling Multiple Types of Data**: Text (article content), images (thumbnails), and possibly embedded videos.
3. **Demonstrates End-to-End Pipeline**: From **data collection** (crawling) → **HDFS** → **Database** → **Analytics**.
4. **Feasible**: Plenty of news sites allow limited scraping of headlines, article summaries, images, etc.

STEP 1: 3 Dynamic Websites for Scraping

1. **CNN** or **BBC:** - Images and embedded videos also headlines, links to articles could be scrapped.
2. **Youtube** or **Vimeo:** Specific channel / commentary, URL, thumbnails
3. **Region Site**

Goal to gather –

a. Article Title
b. Date / Time
c. Author / Source
d. Image or Video Link

STEP 2: Planning of Database

5+ tables. Here's an example:

1. **Articles**

- article_id (PK)
- title
- publish_date
- author_id (FK to Authors)
- source_id (FK to Sources)
- content (could be a text field)

2. **Authors**
    - author_id (PK)
    - author_name
3. **Sources (websites or publishers)**
    - source_id (PK)
    - source_name (e.g., “CNN”, “YouTube Channel XYZ”)
    - source_url
4. **Media (images and videos)**
    - media_id (PK)
    - article_id (FK to Articles)
    - media_type (e.g., “image”, “video”)
    - media_url (URL to the image or video)
    - local_path (if you download it)
5. **Topics/Tags (for classification)**
    - topic_id (PK)
    - topic_name (e.g., “Politics”, “Sports”, “Tech”)
6. **Article_Topics (a join table if articles can have multiple topics)**
    - article_id (FK to Articles)
    - topic_id (FK to Topics)

STEP 3: Business Questions

1. Which topics / sports are most covered across the sources?
2. Which source (website/publisher) posts the most frequent updates?
3. How many articles include videos vs. just images?
4. Who are the top 5 authors (by article count)?
5. Which topics often coincide (co-occurrence) in the same article?
6. How many Articles published per week?
7. Average length of article and reads.

OPTION B: Environmental Data Aggregator (can be done to sports or drugs)

Collect, store, and analyse air quality metrics and environmental updates (articles, images, videos) from multiple sources to understand pollution trends, emissions, and relevant news.

STEP 1: Possible Websites

1. **Government Air Quality Site:** For example, the U.S. EPA (Environmental Protection Agency) or local environment agencies that publish daily/hourly air quality data.
    - They often provide data like AQI (Air Quality Index), PM2.5/PM10 levels, and more.
    - Some have JSON/CSV endpoints, some have dynamic HTML.
2. **Global/Regional Environment News Website:** Something like UN Environment Program news feed or an environment-focused blog that includes images/videos about climate events, deforestation, wildlife, etc.
3. **NASA Earth Observatory (or similar)** for satellite images, maps, or short video content explaining climate phenomena.
    - NASA sites often have images with descriptive text.
    - Alternatively, you can find YouTube channels from official environment organizations or NGOs to scrape video metadata (title, publish date, description, etc.).

Check robots.txt or official APIs (sometimes these agencies provide an open API).

Why It's Meaningful

- Tracking pollution trends is critical for public health.
- Combining raw data (AQI, PM2.5) with news/media provides insights into how major events (wildfires, industrial accidents) correlate with pollution spikes.
- Demonstrates an end-to-end pipeline with structured (numerical) and unstructured (articles, images) data.

2. Database Schema (5+ Tables)

Here's one example:

1. **Locations**
    - location_id (PK)
    - city_name
    - country_name
    - region (optional)
2. **Measurements** (Actual air quality data)
    - measurement_id (PK)
    - location_id (FK)
    - measurement_date
    - aqi (Air Quality Index)
    - pm25 (fine particulate matter)
    - pm10
    - co (carbon monoxide), etc.
3. **Sources** (the websites/APIs you crawled)
    - source_id (PK)
    - source_name
    - source_url
4. **Articles** (for environment news or blog posts)
    - article_id (PK)
    - title

- publish_date
- content (text)
- source_id (FK to Sources)

5. **Media** (images or videos from NASA or environment news)
    - media_id (PK)
    - article_id (FK to Articles)
    - media_type (e.g., “image”, “video”)
    - media_url (full URL)
    - local_path (if downloaded)

STEP 3: Business Questions

1. How does the average AQI vary by city or region over time?
2. Which events (news articles) coincide with sudden changes in AQI?
3. Which source provides the most frequent content or updates?
4. How many articles include images vs. videos?
5. Which city shows the most improvement or deterioration over a selected period?
6. Top Polluted Cities

OPTION C: Environmental Project **(Focused on Climate / Pollution Data)**

Environmental data—such as air quality indicators, greenhouse gas emissions, or satellite imagery—is scattered across multiple official sources. Organizations like NASA, NOAA, and local environmental agencies offer open data, but it’s fragmented in different formats. This project aims to **consolidate** measurements (numerical data), **imagery** (satellite or ground photos), and **video** (educational or analysis clips) into a single data pipeline. We’ll store this data in Hadoop for fault tolerance, then design a relational database to answer key questions about pollution trends, climate anomalies, and major events (e.g., hurricanes, wildfires).”

STEP 1: 3 Possible Websites

1. **NASA Earth Observatory (Images / Videos)**
    - NASA often publishes **satellite images** (e.g., of wildfires, deforestation) and short explanatory videos.
    - can scrape the pages to collect image URLs, captions, dates, and possibly video links (YouTube or embedded).
    - Example: earthobservatory.nasa.gov
2. **NOAA (National Oceanic and Atmospheric Administration)**
    - NOAA has climate and weather data (temperature anomalies, CO2 levels, hurricane tracking, etc.).
    - Some of it is behind an API or CSV/JSON downloads; other pages might be dynamic HTML.
    - Example: www.climate.gov or www.ncei.noaa.gov
3. **Local Environment Agency** (or a 3rd party aggregator like OpenAQ)
    - For **air quality** measurements (PM2.5, PM10, O3, CO, etc.) for different cities.
    - Or if you prefer water quality from a specific region or municipality.
    - Usually provides **live or recent** data in a structured format, plus some dynamic HTML to scrape.

STEP 2: The Database

STEP 3: Database (5+ Tables)

1. **Locations**
    - location_id (PK)
    - location_name or city_name
    - country
    - region (optional)
2. **Measurements** (e.g., daily or monthly climate/air/water data)
    - measurement_id (PK)

- location_id (FK → Locations)
- measurement_date
- measurement_type (e.g., “AQI”, “Temperature”, “CO2 level”)
- value (numeric)

3. **Events** (e.g., hurricanes, wildfires, extreme weather)
   - event_id (PK)
   - event_name (e.g., “Hurricane Ian”)
   - start_date
   - end_date (optional if ongoing)
   - description (text describing the event)
4. **Media** (storing images/videos from NASA or NOAA)
   - media_id (PK)
   - event_id (FK, if relevant)
   - media_type (“image” or “video”)
   - media_url (URL to the resource)
   - local_path (if you download the file)
   - publish_date (date/time it was posted)
5. **Sources** (to track which website gave you the data)
   - source_id (PK)
   - source_name (e.g., “NASA Earth Observatory”)
   - source_url

STEP 3: Business Questions

1. Which locations have the highest average AQI (or temperature) over a given month?
2. How has the average temperature (or CO2 level) changed year-over-year for each region?
3. Which event (wildfire, hurricane, etc.) had the most media coverage (images + videos) from NASA?

4. **Frequency of extreme events**: Count how many events occurred in a given year (or region).
5. **Which source provides the most data entries?** (Comparing NOAA vs. NASA vs. the local agency)
6. How many images vs. videos do we have in our Media table overall?
7. Among the events with the highest severity (if you store severity) or largest impact, which region is impacted the most?

OPTION D: Medical/Pharmaceutical Project **(Focused on Clinical Trials / Drug Safety)**

Clinical trial data, drug approvals, and safety advisories are published on different official websites. Researchers and healthcare professionals struggle to track ongoing studies, newly approved drugs, and potential safety alerts. This project will unify structured data (trial results, drug names, sponsor info) with multimedia (company announcements, educational videos) from official or community sites into Hadoop. We'll then load it into a relational database to run queries that reveal top sponsors, safety alerts, and trial outcomes.

STEP 1: Three Possible Websites / Data Sources

1. **ClinicalTrials.gov**
    - Provides a vast database of **clinical trials** (study title, sponsor, status, summary).
    - can scrape or use their API.
    - The site is dynamic, but if the API is too big, you can focus on **a smaller subset** of trials (e.g., a particular condition or sponsor).
2. **OpenFDA**
    - Official platform that provides data on **drug adverse events**, **recalls**, etc.
    - There's an API with JSON responses (which is *not exactly scraping*, but still dynamic data retrieval).
    - Or you can parse some of their web pages if you want the "scraping" aspect.
3. **Pharmaceutical Company Websites or WHO**

- Some big pharma companies post **press releases** about new approvals or research updates (with images, sometimes videos).
- WHO (World Health Organization) might have dynamic pages for drug/vaccine guidelines with embedded media.
- Or a platform like EMA (European Medicines Agency) for drug approval reports.

(Again, check terms of service and robots.txt. Official sites typically allow or have open data sets for research, but confirm your usage.)

STEP 2: Example Database Schema (5+ Tables)

1. **Clinical_Trials**
   - trial_id (PK)
   - title
   - start_date
   - end_date (optional if not ended)
   - status (e.g., "Recruiting", "Completed", "Terminated")
   - condition (e.g., "Diabetes", "Cancer")
2. **Sponsors** (organizations running the trial)
   - sponsor_id (PK)
   - sponsor_name
   - country
3. **Trial_Sponsors** (join table if multiple sponsors per trial)
   - trial_id (FK → Clinical_Trials)
   - sponsor_id (FK → Sponsors)
   - **(Composite key)**
4. **Drug_Alerts** (data from openFDA or other recall/alert source)
   - alert_id (PK)
   - drug_name
   - alert_date
   - alert_type (e.g., "Recall", "Adverse Event")

- description (reason for the alert)

5. **Media** (images, videos, or official documents)
    - media_id (PK)
    - media_type (“image”, “video”, “pdf”)
    - media_url
    - local_path
    - publish_date
    - related_drug or related_trial_id (optional foreign key if relevant)

STEP 3: Business Questions

1. Which sponsors have the most clinical trials active right now?
2. Which drugs have the most safety alerts (recalls, adverse events)?
3. How many clinical trials are completed vs. recruiting vs. terminated?
4. Which countries host the most clinical trial sponsors?
5. How many media items (images/videos/pdfs) are associated with a particular trial or drug alert?
6. Show top 5 clinical trials by earliest start date (i.e., oldest ongoing trials).
7. Yearly trend of new alerts: how many drug alerts per year?