User Manual

This manual provides step-by-step instructions to download, set up, and run the GDELT news application, including its ETL pipeline and full-stack web interface.

Prerequisites

Before running the application, ensure the following software is installed**:**

For Windows:

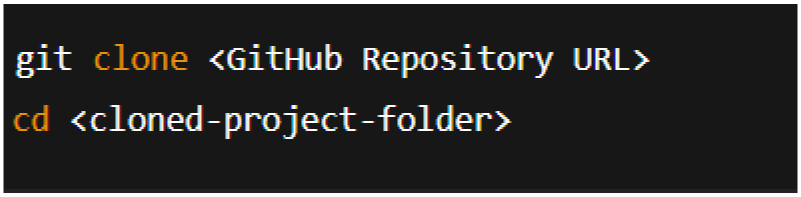
- Install Apache Spark
- Install Python (v3.11 or higher)
- Install Java JDK
- Install Node.js
- Install PostgreSQL and pgAdmin
- Install Git
- Optionally install VS Code or any code editor

For Linux/macOS:

- Install PySpark (instead of Spark binary)
- Other prerequisites same as above

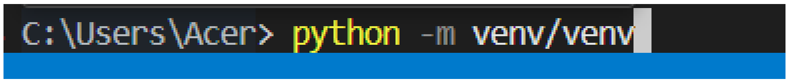
Environment Setup

1. **Clone the project from GitHub**:



```
git clone <GitHub Repository URL>
cd <cloned-project-folder>
```

2. **Create and activate a Python virtual environment**:



```
C:\Users\Acer> python -m venv/venv
```

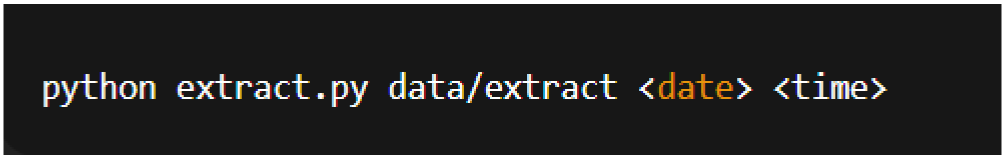
3. **Activate the Venv**
   For windows use : venv/Scritps/activate

For linux use :

Source venv/bin/activate

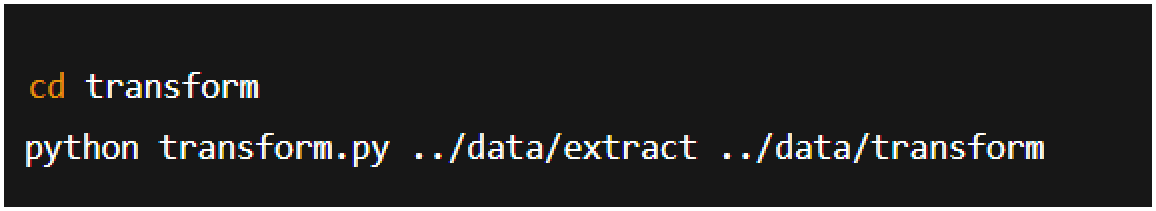
ETL Pipeline Steps

1. **Run the Extract script**:



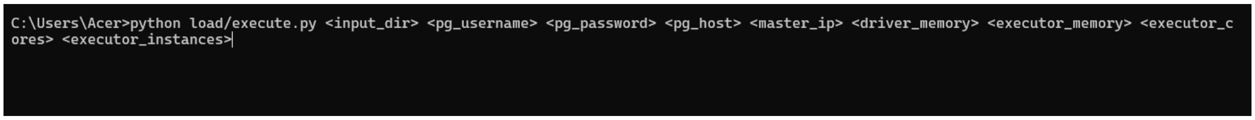
```
python extract.py data/extract <date> <time>
```

2. **Run the transform script**



```
cd transform
python transform.py ../data/extract ../data/transform
```

3. R**un the load Script**

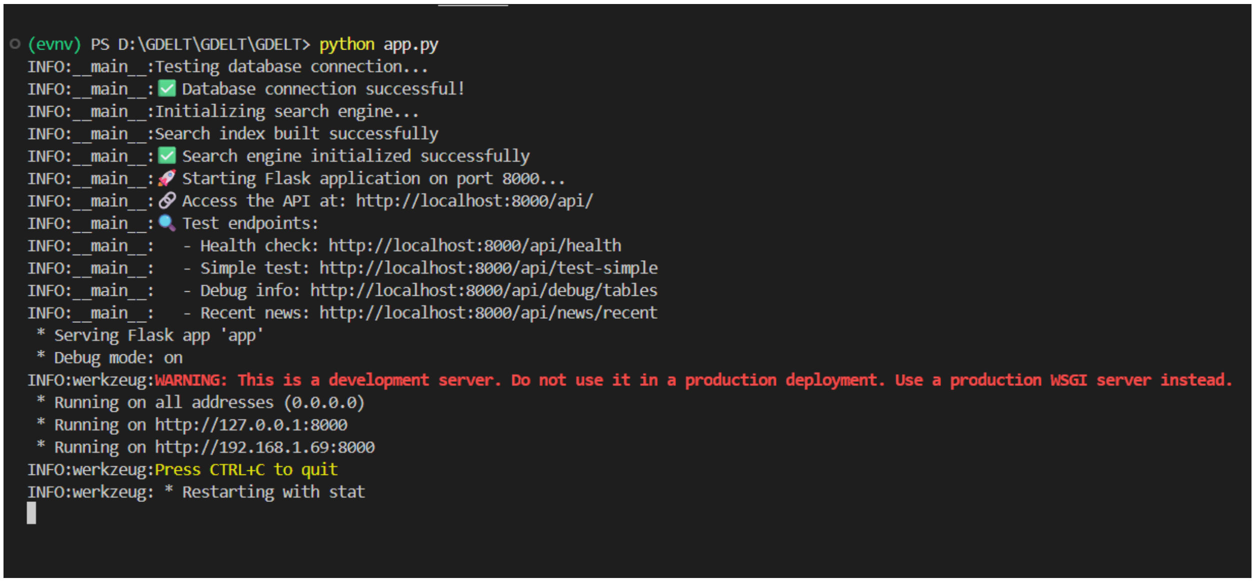


```
C:\Users\Acer>python load/execute.py <input_dir> <pg_username> <pg_password> <pg_host> <master_ip> <driver_memory> <executor_memory> <executor_c
ores> <executor_instances>
```

Run the Web Application

Backend (Flask API)

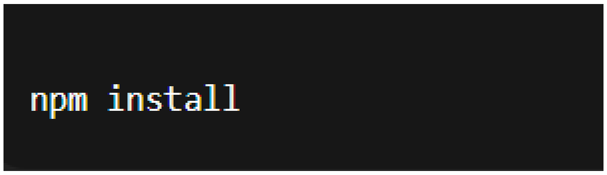
1. From the root directory, run:



```
○ (evnv) PS D:\GDELT\GDELT\GDELT> python app.py
  INFO:__main__:Testing database connection...
  INFO:__main__:✅ Database connection successful!
  INFO:__main__:Initializing search engine...
  INFO:__main__:Search index built successfully
  INFO:__main__:✅ Search engine initialized successfully
  INFO:__main__:🚀 Starting Flask application on port 8000...
  INFO:__main__:🔗 Access the API at: http://localhost:8000/api/
  INFO:__main__:🔍 Test endpoints:
  INFO:__main__:   - Health check: http://localhost:8000/api/health
  INFO:__main__:   - Simple test: http://localhost:8000/api/test-simple
  INFO:__main__:   - Debug info: http://localhost:8000/api/debug/tables
  INFO:__main__:   - Recent news: http://localhost:8000/api/news/recent
   * Serving Flask app 'app'
   * Debug mode: on
  INFO:werkzeug:WARNING: This is a development server. Do not use it in a production deployment. Use a production WSGI server instead.
   * Running on all addresses (0.0.0.0)
   * Running on http://127.0.0.1:8000
   * Running on http://192.168.1.69:8000
  INFO:werkzeug:Press CTRL+C to quit
  INFO:werkzeug: * Restarting with stat
  ▌
```

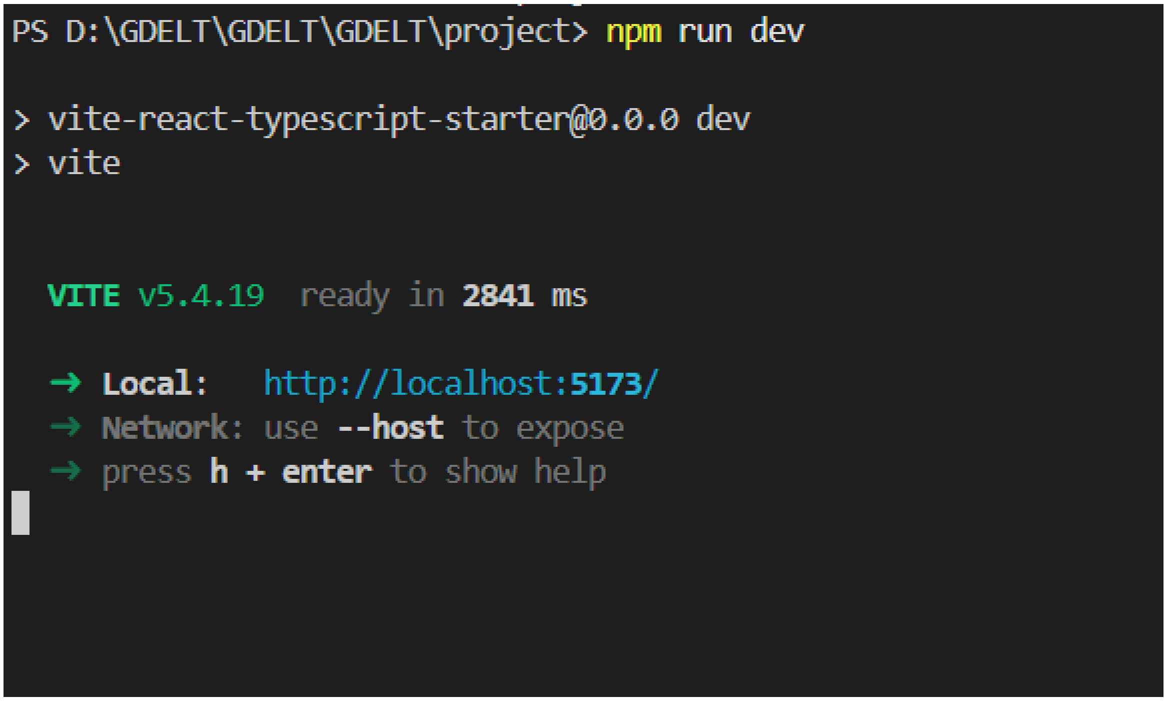
Frontend (React)

1. Then on a new terminal ,Navigate to the frontend project directory:
2. Then install the dependencies



```
npm install
```

3. Start the development server:

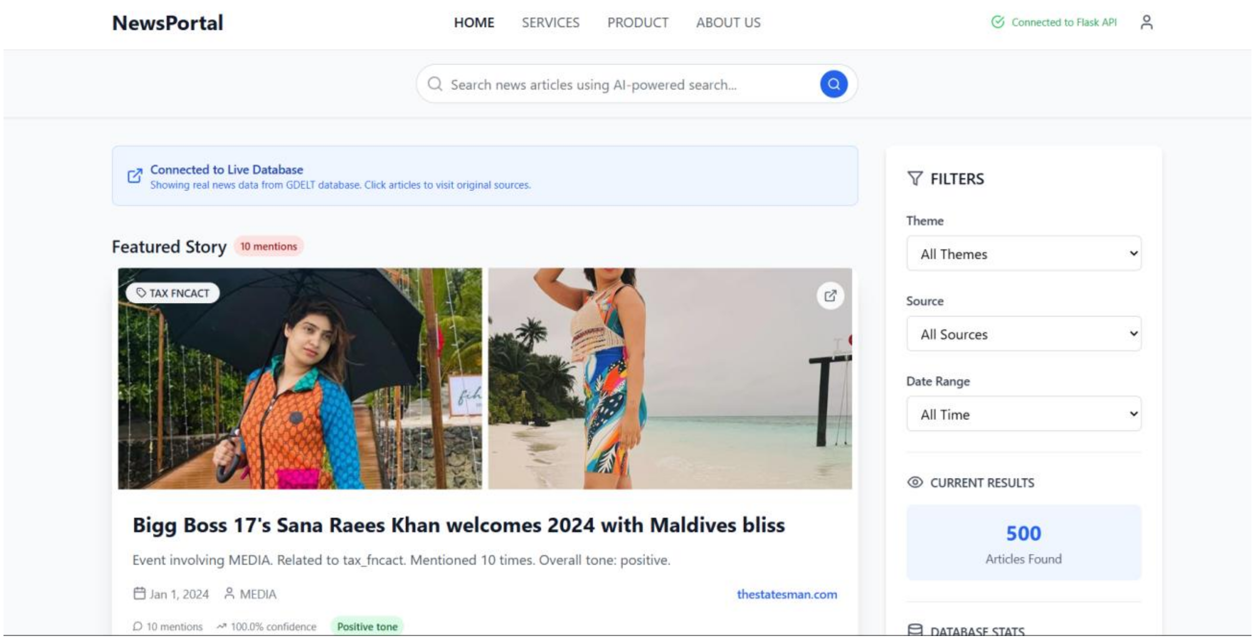


```
PS D:\GDELT\GDELT\GDELT\project> npm run dev

> vite-react-typescript-starter@0.0.0 dev
> vite


  VITE v5.4.19  ready in 2841 ms

  ➜  Local:   http://localhost:5173/
  ➜  Network: use --host to expose
  ➜  press h + enter to show help
```



Connected to Live Database
Showing real news data from GDELT database. Click articles to visit original sources.

Featured Story 10 mentions

TAX FNCACT

Bigg Boss 17's Sana Raees Khan welcomes 2024 with Maldives bliss

Event involving MEDIA. Related to tax_fncact. Mentioned 10 times. Overall tone: positive.

Jan 1, 2024 MEDIA thestatesman.com

10 mentions 100.0% confidence Positive tone

FILTERS

Theme: All Themes

Source: All Sources

Date Range: All Time

CURRENT RESULTS

500
Articles Found

DATABASE STATS