Data Sources for Trial Assignment (Pick any One)

Tasks:
1. Designing an ETL Pipeline
2. Architecture Proposal
3. Sample processing script executed on a small subset of the data set
4. Required to be scalable to process data streams in parallel

Source 1: FRED

Description: Federal Reserve Economic Data (FRED), the free economic database from the Federal Reserve Bank of St. Louis, aggregates data from 40 different data sources, including the Bureau of Labor Statistics, the Bureau of Economic Analysis, Census, OECD, World Bank, International Monetary Fund, and the Federal Reserve System
The task is to scrape a small subset of data for Interest Rates. They can be found here
URL: https://fred.stlouisfed.org/categories/22.

GuideLines:
Write an ETL pipeline or scraping script for collecting the time series data:
Make sure the metadata and actual time series are collected and stored separately for each time series dataset
Ensure that the datasets are properly json serialisable
Write Metadata, Time Series Data, Cleaning, Standardising Scripts.
Ensure the metadata and time series is connected using a unique id key
Designate a Unique id for each time series.
Propose a Tech Stack or Architecture of the system to execute the above workload:
For Tech Stack Choose Scraping tools like selenium, BS4, LXML etc and ensure python is used as the programming language.
For Architecture Proposals use AWS services and give more system design aspects.
required to be scalable to process data streams in parallel using multicore or parallel processing.
Use the repo provided and create a fork so the team can do a code review. Follow the Submission GuideLines mentioned in the repo

Source 2: AIS

Description: AIS vessel tracking has been the mariner’s most significant development in navigation safety since the introduction of the radar. AIS tracking system was originally developed as collision avoidance tool which enables commercial vessels to ‘see’ each other more clearly in any conditions and to improve the helmsman’s information about the surrounding environment. AIS does this by continuously transmitting vessels’ position, identity, speed and course, along with other relevant information, to all other AIS equipped vessels within range. Combined with a shore station, this system also offers port authorities and maritime safety bodies the ability to manage maritime traffic and reduce the hazards of marine navigation.

The task is to scrape the station coverage data from the following link: https://www.aishub.net/stations and also store it in a time series format

https://www.aishub.net/

Source 3: GDACS

The integrated GDACS website offers the following disaster information systems and online coordination tools:

GDACS Disaster Alerts, which are issued and disseminated to some 25,000 subscribers immediately following sudden-onset disasters. The automatic estimates and risk analysis–the basis of the alerts –are provided by the European Commission Joint Research Centre (JRC).

The Virtual OSOCC a restricted online platform for real-time information exchange and cooperation among all actors in the first phase of the disaster. Information updates from the affected country and international responders are moderated by a dedicated team. The Virtual OSOCC has some 19,000 registered users, and is managed by the United Nations Office for the Coordination of Humanitarian Affairs (OCHA).

Maps and satellite imagery from various providers are shared on the Virtual OSOCC trough the GDACS Satellite Mapping and Coordination System (SMCS). It provides a communication and coordination platform where organisations may monitor and inform stakeholders of their completed, current and future mapping activities during emergencies. This service is facilitated by the United Nations Institute for Training and Research (UNITAR) Operational Satellite Applications Programme (UNOSAT).

Task is to build a timeseries event store for disaster events occurred in India Using GDACS Disaster Alerts

https://gdacs.org/About/overview.aspx

https://gdacs.org/

Source 4: GDELT

GDELT: http://gdeltproject.org

GDELT stands for Global DataBase for Event, Language and Tone.
Supported by Google Jigsaw, the GDELT Project monitors the world's broadcast, print, and web news from nearly every corner of every country in over 100 languages and identifies the people, locations, organizations, themes, sources, emotions, counts, quotes, images and events driving our global society every second of every day, creating a free open platform for computing on the entire world.

The objective is to scrape the data and setup proper flow for collection and consumption of the latest data, along with historical data.
The details can be accessed from here:
https://blog.gdeltproject.org/gdelt-2-0-our-global-world-in-realtime/

Source 4: Projects/Tenders data

TED provides comprehensive public procurement information free of charge in an easy-to-use format to all interested parties. It is expected to increase market transparency, decrease transaction costs, and facilitate government accountability. It is a central, public, and open procurement platform that has been set up to contribute to achieving value for money in public procurement as well as increase integrity throughout the public sector.

The objective is to scrape the data and set up a proper flow for the collection and consumption of the latest data, along with historical data.
The data to be scraped can be accessed from here:
https://ted.europa.eu/TED/search/searchResult.do

OTHERS

Original Data sources from Resource Watch: https://resourcewatch.org/data/explore