Methods of Advanced Data Engineering

Data Report

Dushyant Supare

Summer Semester 2024

1. Question

Climate change is a well-established reality. As we all know, it causes several changes in the world, the most of which are negative. This research will investigate the link between climate change at the nation level and the number of internally displaced people inside a country as a result of it. The key impact point will be determining which nation was more affected and when it happened, between 2008 and 2013. This is required for the tackling of the underlying causes.

2. Data Sources

There are two datasets in this report.
Datasource 1: Kaggle
Metadata URL: https://www.kaggle.com/berkeleyearth/climate-change-earth-surface-temperature-data

Data Type: CSV

This dataset was uploaded to Kaggle by Berkley earth. This includes the information about surface temperatures. There are many files in this but the current focus in on country level so surface temperatures according to countries are selected. The temperatures are recorded since 1750 up to 2013.

Datasource 2: Our World In data
Metadata URL: https://ourworldindata.org/grapher/internally-displaced-persons-from-disasters?tab=csv

Data Type: CSV

This dataset contains statistics on the number of persons displaced internally. This information is country-specific and contains the overall number of individuals displaced every year by country. This information is relatively easy to understand.

3. Data Pipeline

3.1 Description

The data pipeline has been implemented using Python. Where the Kaggle API was used for the first data source, a zip file was obtained, which was then unzipped and placed in a specified folder (data folder). Because a downloadable CSV URL was not accessible for the second data source, it was obtained via Selenium automation and saved to the project folder. The fundamental transformations and data cleansing were completed.

3.2 Dataset fetching

The dataset fetching was the most challenging process as it required a lot of effort and investigation. The first dataset came from Kaggle, so the Kaggle API needed to be configured. This was accomplished by creating credentials. The Kaggle API function was called, which deals with logging into Kaggle using credentials stored on the local computer and searching for the dataset. This function is currently part of the Activity1.py file, but it will be decoupled in the future and moved to a distinct Python file.

Selenium was used to fetch the second dataset because there was no direct downloading URL provided. It required the setup of a Chrome session that would navigate to the specified URL and download the CSV. This procedure entailed visiting the webpage first, declining cookies, then proceeding to the download option, opening the popup, and lastly clicking on the download CSV option. Finally, the downloaded dataset is placed in the projects folder.

3.3 Cleaning and Transformation

As the source of the dataset differs, the cleaning and transformation steps are different.
For both the datasets there is one standard cleaning procedure that is removing the missing values and dropping duplicate rows.

<u>Cleaning and Transformation (Data Source 1):</u>

As the missing values and duplicate rows issue was taken care of, further data cleaning was done. After initial import it was seen that some misinterpreted data was reflecting. This was removed altogether by putting the condition on country name. Secondly as the dataset is huge, spanning from 1750 to 2013, only relevant data is needed. The data from 2008 to 2013 was taken. A small transformation on the Average temperatures was applied, by rounding them off to two digits.

<u>Cleaning and Transformation (Data Source 2):</u>

As this dataset was simple in nature. Non complicated transformations were needed. After an initial analysis it was found that the “Code” column was not needed as it was a short form representing the countries and can be discarded.

3.4 Problems Encountered

As there is a difference between two datasets, the dataset 1 having entries from 1750 to 2013 and dataset 2 having entries from 2008 to 2023, a common ground must be reached to perform this descriptive analysis, so the analysis would be from the year 2008 to 2013. The most common years between the two datasets.

Another problem was the unavailability of the direct link for CSV download for dataset 2, after a lot of exploration, such as finding direct link through chrome console, trying out the “owid-catalog” python library, and selenium automation option. The selenium automation option was chosen as it was easy and achievable in less amount of time.

3.5 Error Handling

Only relevant columns were selected in the dataset. It will be made sure that the analysis would be done based on years, and countries. And result would give a correlation between the surface temperatures and internal displacement.

4. Result and Limitations

4.1 Data Pipeline Output

Two SQLite files are made as an output of this data pipeline, one for each dataset. The data quality is ensured as the resultant data in both the SQLite files is accurate, consistent, complete, well timed and relevant.

4.2 Data Format

The simplicity and effectiveness of the SQLite file format have led to its selection. The database is housed in a single file in this self-contained, serverless database engine. Because SQLite files don't need to be installed or configured, they're a suitable option for the use case at hand. Because these files are simple to use, manage, and set up.

4.3 Potential Issues

Regardless, even when a static data source has been chosen to serve as the pipelines' input, unforeseen changes at the source might potentially cause interruptions in the pipeline and result in incorrect or no output.