**ETL Project:**

Collection of Climate Metrics that can impact Wildfires

**Members:**

Luis Santana

Luis Casas

Lourdes Rodriguez

**Purpose (use) of data procurement:**

Assess impact of human activity over earth’s climate, in particular, how oil production relates to increasing temperatures and observed trends in forest wildfires. Focused data gathering on two countries: Brazil and the USA, for years 2000-2013.

**Links/Data Sources:**

1. <https://www.indexmundi.com/energy/?product=oil&graph=production>
2. <https://www.kaggle.com/srikantsahu/co2-and-ghg-emission-data#emission%20data.csv>
3. <https://www.kaggle.com/berkeleyearth/climate-change-earth-surface-temperature-data#GlobalLandTemperaturesByCountry.csv>
4. http://queimadas.dgi.inpe.br/queimadas/portal-static/estatisticas\_paises/
5. https://www.ncdc.noaa.gov/societal-impacts/wildfires/ytd/0?params[]=acres&params[]=fires&params[]=apf

**Extraction/Transformation/Loading Process:**

1. csv files were pulled from:

* CO2 emissions: <https://www.kaggle.com/srikantsahu/co2-and-ghg-emission-data#emission%20data.csv>
* Global Temperatures: <https://www.kaggle.com/berkeleyearth/climate-change-earth-surface-temperature-data#GlobalLandTemperaturesByCountry.csv>

Used Jupyter Notebooks with Panda dataframes to load csv files and manipulate/transform data to desired table structures.

* [CleanCsv.ipynb](https://github.com/lourdesrm/ETL-project/blob/master/Data_source/CleanCsv.ipynb)
* [temp\_historical.ipynb](https://github.com/lourdesrm/ETL-project/blob/master/Data_source/temp_historical.ipynb)

Resulting dataframes saved on following csv files:

* CleanEmissionData.csv
* MaxTemperature.csv

1. Web scraping was performed on the following websites:

* World Oil Production: <https://www.indexmundi.com/energy/?product=oil&graph=production>
* Wildfire data on Brazil: <http://queimadas.dgi.inpe.br/queimadas/portal-static/estatisticas_paises/>
* Wildfire data on USA: [https://www.ncdc.noaa.gov/societal-impacts/wildfires/ytd/0?params[]=acres&params[]=fires&params[]=apf](https://www.ncdc.noaa.gov/societal-impacts/wildfires/ytd/0?params%5b%5d=acres&params%5b%5d=fires&params%5b%5d=apf)

Used Splinter and Beautiful soup to scrape chosen websites for desired climate and man-made related metrics. Used Jupyter Notebooks with Panda dataframes to generate desired table structures.

* [web-scraping.ipynb](https://github.com/lourdesrm/ETL-project/blob/master/Data_source/web-scraping.ipynb" \o "web-scraping.ipynb)

Resulting dataframes saved on following csv files:

* fire\_count\_brazil.csv
* fire\_count\_us.csv
* global\_oil\_production.csv

Gathered all generated csv files and used Jupyter Notebooks to merge them into a single main dataframe. This dataframe was used to load the SQL data base.

* [web-scraping.ipynb](https://github.com/lourdesrm/ETL-project/blob/master/Data_source/web-scraping.ipynb)

**Chosen Database:**

SQL was chosen as our data-base warehouse due to the relational nature of our main dataframe. Postgres used for manipulation and structure of database

Schema of tables in final DB:

A screenshot of a computer

Description automatically generated