ETL Project: Annual Fishing Data

Gizem Yali-Pepper, Khalid Ettalii, Omar Eltorai

# **Extract:** Data sources

The datasets provides data of annual nominal catches of more than 200 species of fish and shellfish in the Northeast Atlantic region, which are officially submitted by 20 [International Council for the Exploration of the Sea (ICES)](http://www.ices.dk/) member countries between 2006 and 2014. The unit metric used to measure the catch quantity is total live weight (“TLW”).

* Link: <https://www.kaggle.com/victorgenin/ices-fish-catch/version/1#RECO_Export_08-04-2016.csv>
* There are three primary files (csv and xlsx format) that were used:
  + country\_code.csv
  + fish\_catches.csv
  + fish\_name.xlsx

# **Transform:** Data transformation

Cleaning activities performed included primarily data dropping, transposing (using “melt” in pandas), aggregating, and merging. The majority of the data cleaning and transformation was performed using pandas.

For data dropping, we went through each file and determined which data fields were relevant and which were not relevant to this project. We dropped the fields that we determined were not relevant (shown to the right).

After filtering for only the relevant data, we then performed data transformation as needed. In the “fish\_catches.csv” data set, we transposed the data so that the annual data was not broken into separate columns, but was put into a single column, and we added a new column that corresponded with the year. We were able to do so using the ‘melt’ functionality in pandas to unpivot a DataFrame from wide format to long format.

After observing a discrepancy between the country data in the “fish\_catches.csv” and “country\_code.csv” data sets, we combined the two data sets in pandas to remove the catch data for the countries which were not included in the “country\_code.csv” set.

# **Load:** Final Production Database

Once the three individual data sets were cleaned, we then loaded the data into a relational database using PostgreSQL. The database was named “etl\_db”. Four tables were created in the schema, and included (column header shown in parentheses):

1. catch (country, species, year, tlw) – contains country code, fish code, year, and total live weight (tlw) caught
2. country (country, description) – contains the country code and name of the country
3. fish (species, scientific name, family) – contains the fish code and taxonomic details for the fish
4. fish\_name (species, name) – contains the fish code and common name for the fish

With the tables created, the data from the csv files were imported to populate the newly created tables. The populated tables in the database were then accessed in Python libraries (pandas, matplotlib, etc.) using the SQLAlchemy ORM.

The diagram below maps the data base and the relationships between the different tables.

**Entity Relationship Diagram (ERD)**

