**ETL Project Report**

**Group 3: Andrew Anastasiades, Tianyue Yang, Caroline Larry, Niraj Khatri, Diana Kennen**

**Extract: your original data sources and how the data was formatted (CSV, JSON, pgAdmin 4, etc).**

Our group used three data sources available from data.world. The first two data sources were generated by the Department of Agriculture. One was a csv with data related to US Imports of Aquaculture products. The other was a csv with data related to US Exports of Aquaculture products. The third data source was a csv with land temperatures by country over the last 150 years. One limitation of the data is that it solely contains US information, so if a country exported aquaculture to a country other than the US, it would not be captured in this data set.

* Department of Agriculture
  + Aquaculture Import Data (548k row csv)
  + <https://query.data.world/s/dwmlyhl77itcyz262x4qe2m33z5rsn>
* Aquaculture Export Data (368k rows csv)
  + <https://query.data.world/s/per2tntndcdwiabsumleonm3svcfl5>
* Global Land Temperature
  + <https://data.world/data-society/global-climate-change-data/workspace/file?filename=GlobalLandTemperatures%2FGlobalLandTemperaturesByMajorCity.csv>

**Transform: what data cleaning or transformation was required.**

Our group transformed the data via cleaning, joining, filtering, and aggregating. The cleaning process included:

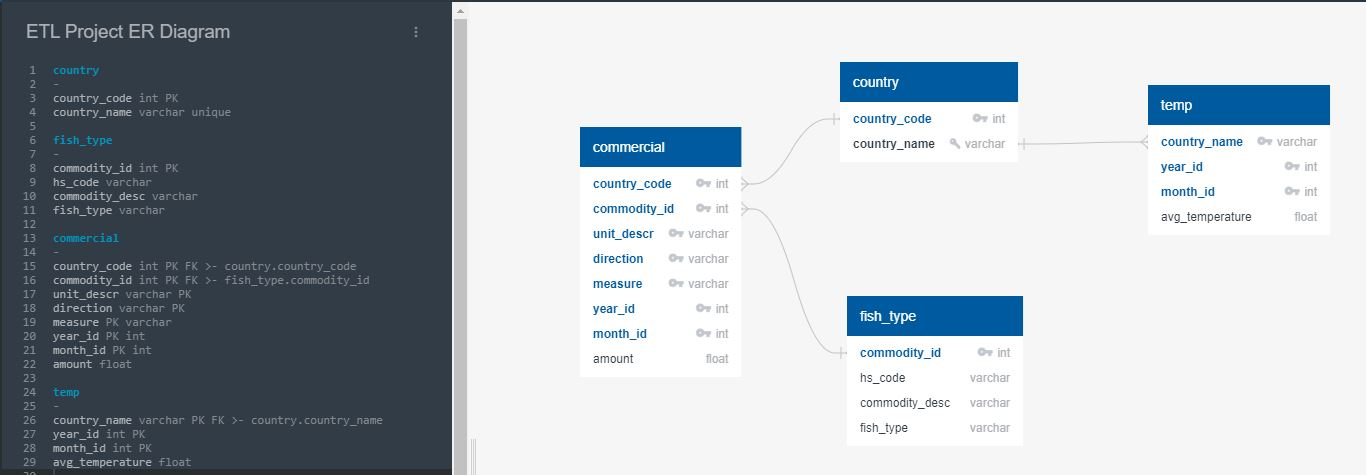
* Dropping source ID from the commercial dataframe
* Dropping temp uncertainty, city, latitude, longitude, and day from the temperature dataframe
* Splitting the attribute description into US Import/Export and Quantity or Value
* Renaming countries for compatibility across dataframes
* Consolidating East/West Germany, Zaire/DR Congo, and Serbia data
* Consolidate duplicate commodities and duplicate HS codes
* Categorizing commodities in fish type as within 8 basic types (ie. trout, salmon)

The US Import and Export dataframes were concatenated. Countries were filtered for compatibility with the temperature dataframe. Our group also aggregated the commercial data by country, fish type, measure, and direction. Temperature data was grouped by country, year, and month.

**Load: the final database, tables/collections, and why this was chosen.**

SQL was chosen as the location for the final database. Our data was organized and relational so we needed to choose a relational database. Four tables were created: commercial, country, temp, and fish type. The country table and the fish type include the unique codes and their corresponding names/ identifiers. We connected those tables to our commercial table and temperature table with a one to many relationship

Figure 1: ERD

Upon testing, for upload SQL, several additional cleaning steps had to take place due to missing HS Code values and duplicate keys.

Several sample queries were developed to demonstrate the types of questions that could be answered using the database.