ETL Project

Vincent Heningburg

Paul Bernert

**Abstract:**

Inspiration, motivation of doing an ETL project in general, justification for why we’re doing the things we’re doing, why some contents are necessary, etc.

**Extract:**

The source for our two datasets comes from Kaggle, and the URL for this information is: <https://www.kaggle.com/n2cholas/competitive-pokemon-dataset>. The two datasets that come from this url are: pokemon-data.csv and move-data.csv. The dataset “pokemon-data” includes information related to each pokemon, such as their combat stats, type-information, and a “tier-list”, which basically means the general consensus on how strong a pokemon is relative to other pokemon in competitive matches. The dataset also includes an entry for each pokemon for their moves list. The “move-data” dataset explains combat states related to a particular move.

The two files were in two separate formats, despite both being the “.csv” extension. The moves dataset was in conventional CSV, while the Pokemon dataset was semi-colon-separated. This is because some of the categories have entries stored as lists. The information in the lists are separated by commas, so the columns cannot also be separated by commas (hence why it’s stored as semi-colons).

**Transform:**

The .csvs were imported as a Pandas DataFrame, where the list entries were stored as strings. The first step was to parse these strings and turn them back into lists. The next step was to pull in the columns, but because not all of the information in the “pokemon-dataset” was relevant for our project, we decided to remove several of the columns. The original dataset contains 12 columns, and we decided to remove four of the columns that weren’t relevant. Because some of the columns still contained information that should be stored separately (such as Pokemon Type 1 and Type 2 currently being stored in the same column), we needed to add three new columns. We replaced the move column which stored every move the pokemon could potentially learn (which could be over 100 moves) to be one random move from that list. We renamed every category that had spaces in the column name (such as “Special Attack”) to no longer have spaces (“Special\_Attack”).

For the “moves-dataset” we removed several of the columns that weren’t relevant (from 9 columns down to 6) and renamed the “Name” column to “Move” and the “Type” column to “Move\_Type”.

The last transformations necessary was to merge the two transformed dataframes on the “Move” column.

**Load:**

The final combined dataframe was loaded into a Postgres database with a table consisting of: the following information: Name, HP, Attack, Defense, Special\_Attack, Special\_Defense, Speed, Type\_1, Type\_2, Move, Move\_Type, Category, PP, Power, Accuracy.