## Extraction

### As a group we took interest in a Green Card & H1B (2014-2018) dataset from kaggle. The Dataset contains the last five years of H-1B visa (temporary work visa) & PERM (permanent residency application, also known as Green Card). Dataset has been collected between October 1 through September 30 (fiscal year) by US Department of Labor. WE DOWNLOADED an Excel file for each fiscal years (FY) between FY14 and FY18.

A second dataset we thought would dovetail nicely with the Green Card data was Human Development Data we extracted from the United Nations Human Development Report. Specifically, we downloaded CSV files detailing Unemployment, Human Development Index (HDI), and Education Index by country for each year between 1990 and 2017.

## Transformation

The Green Card data sets for each fiscal years were big, we had to individually parse the data, extract it into separate files called “FY\_\_ data\_clean” before merging the new data sets together.

We began examining our data sources and following a review of the Green Card data we decided to keep 31 of the 125 parameters in the dataset (removing 94). We went through each of the 125 columns and determined if they will help us answer our questions. Some of the questions we wanted to answer were where in the United States applications were being filed in, the nationality of applicants, the industry they work in, how much they were being paid. We also found that FY15, FY16, FY17, and FY18 had the same parameters, while FY14 data file was incomplete, as such we decided to eliminate FY14 from our analysis. In addition, the data in the HDI table was by calendar year, while the green card data was in fiscal year. So, we extracted month and year in separate columns for “case received date” and “decision date”.

The original files for FY15, FY16, FY17, and FY18 totaled 227.7 MB. However, after we imported each fiscal year table, transformed the data, and then extracted each fiscal year into 4 different files, the total file size equals 69.7 MB with 432819 rows x 35 columns

We also merged HDI, Education Index and Unemployment percentage csv files. We had to delete empty rows in each file before merging them together.

*Load*

We chose to use SQL postgres to load our data because we are more comfortable with using SQL but, we also thought a relational database will be a good fit since our data is structured. We created 2 tables one with the combined FY data “fy\_combined” and another with the combine human development data “indicator\_data “