ETL Project Write Up – Eric Staron, Grant Hewlett, Mo Naghibi

- **E**xtract:
    - Our project team performed extraction from Data World and government census data. The NBA data downloaded from Data World was originally in an xls format and was converted to csv prior to loading into Jupyter Notebook for cleaning.
    - The Census data was downloaded as a csv and prior to loading the Jupyter Notebook for cleaning
- **T**ransform:
    - NBA data: we uploaded a csv of a database of NBA players up to 2014 with career stats and birthdates. The data was a mix a strings and integers. Our plan was to analyze the popularity of NBA player names in two specific years as compared to the census data name popularity by birth year. In order do this we had to parse the NBA data to be usable. First we had to do basic cleaning a drop duplicates and blanks from the data source. Then we had to split the player name column into two separate columns, “first_name” and “last_name”. We did this with a string split and lambda function. To get the birth year broken out we did used the exact same string split function to drop the month and day from the year string. Once we had our three cleaned columns, we formatted them as a new data frame and exported that data frame as a CSV.
    - Census data: Once in Jupyter notebook we imported pandas, sql alchemy, and pymysql. We read the head of the file and created a new data frame for the birth year, 1950. We then dropped unnecessary columns so that our data frame only included first names, number of names, and birth year. We did the exact same process to get filter on birth year, 1990 and then combined each of those two data frames as a single data frame.
    - After cleaning and formatting both data sets, we uploaded the NBA data frame csv to the jupyter notebook and set up the connection to MySQL

- **L**oad:
    - Prior to loading the combined dataset that was created by cleaning, filtering and joining within Pandas we set up the schema in MySQL.
    - We created two database tables in order to import our data frames, one table NBA, which included player names and birthyear, and Names table contain data from census include names, popularity and the year.
    - We created a view, and join the tables based on birth year to see the popularity of NBA player names in given years in a relational database.