**ETL PROJECT**

**GROUP 4: MARK BURTON, JAMMY LO, SCOTT FRAZIER, HOA ROACH**

**Topic:** Scraping the data of hockey players in skater position from National Hockey League (NHL) website. The relevant data of players, player stats and teams will be pulled from website, transformed and then loaded to cloud database for public access

**Data Source:**

Hockey players data: <http://www.espn.com/nhl/statistics/player/>

Hockey teams and abbreviation: <https://www.kaggle.com/martinellis/nhl-game-data>

Hockey teams: <https://en.wikipedia.org/wiki/National_Hockey_League>

**Extract:**

Our original data sources included gathering data the NHL section on the official ESPN website. From this website we found a data table consisting of a plethora of data pertaining to the many available statistics for the top players that are currently in the NHL. This data was then scraped to our jupyter notebook by utilizing html, and xml after finding the source code for the table on the ESPN website in the html inspect tool.

The hockey team’s information is easily found in Wikipedia website. There are total of 31 teams. We then convert the table to csv file for importing and transforming the data.

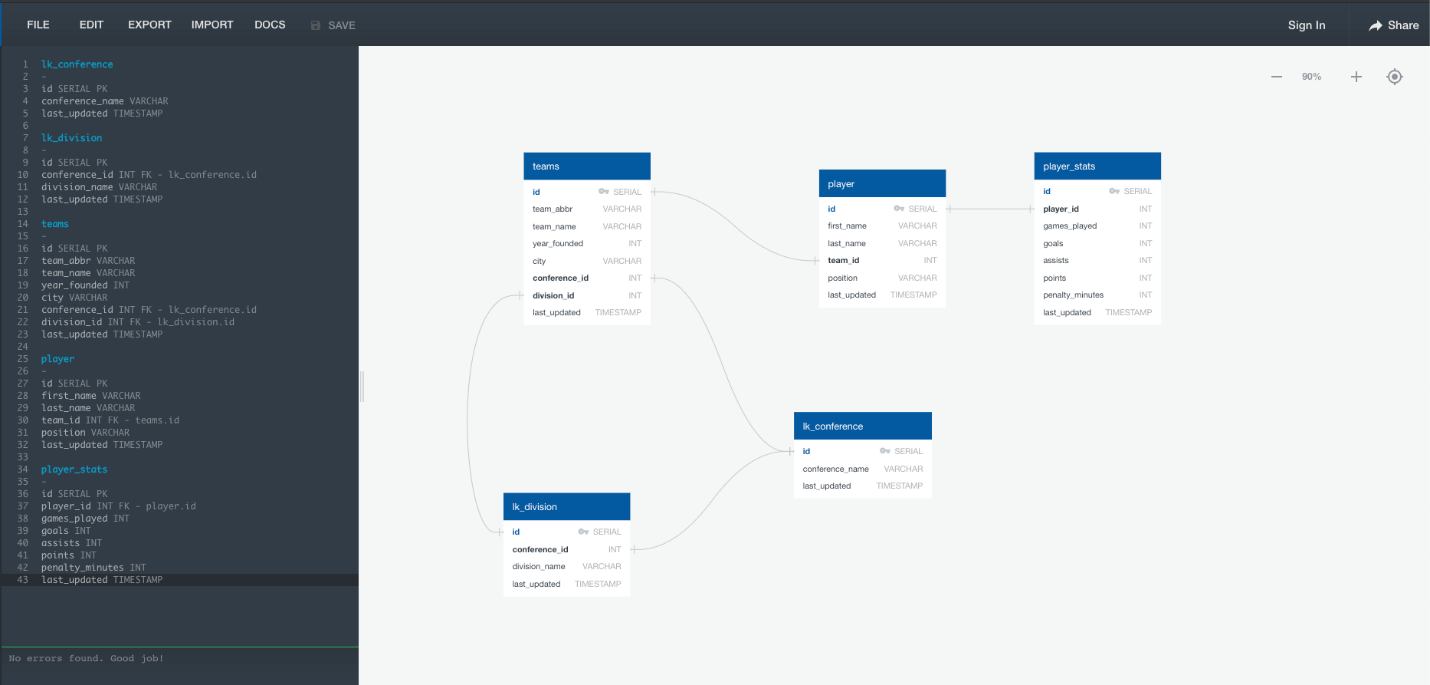
**Transform:** what data cleaning or transformation was required.

For the web scraping portion of this project, we cleaned our data by pulling in all rows and columns of data from a specific data table on the website we were scraping from. We then decided to create our final table by utilizing seven columns of data and dropping the remaining ten columns, because we would not be utilizing those in our final table. Once we had all of the desirable data, we continued to parse through the data by creating new columns such as a first name, last name, and position column from the original “Player” column which contained all of this information originally. To finish off the web scraping data table, we finally rearranged the data in the desired order that would make the most sense to anyone who viewed the data, such as putting the first and last name columns at the beginning of the table. The final data scraped from website then saved to 2 tables of player information and player stats. These tables then saved to csv files for transformation and upload.

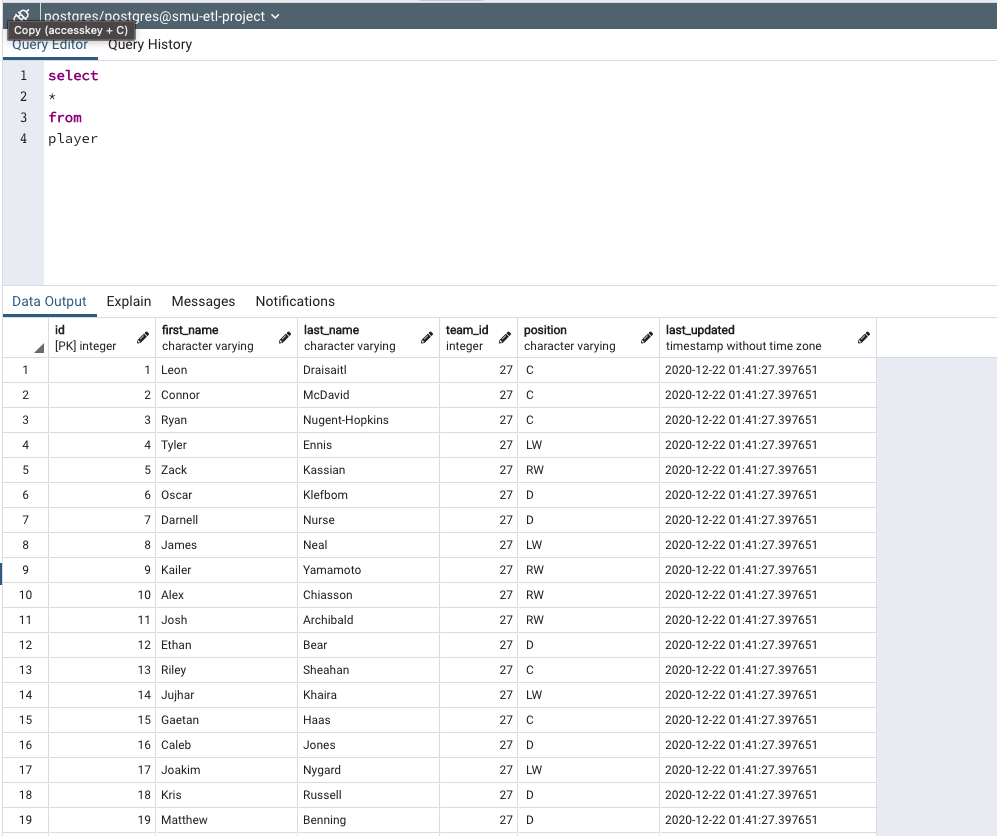
Most of the transformation work requires us to join different tables and rename the columns for appropriate data structures. Some of columns require extracting the certain characters from columns.

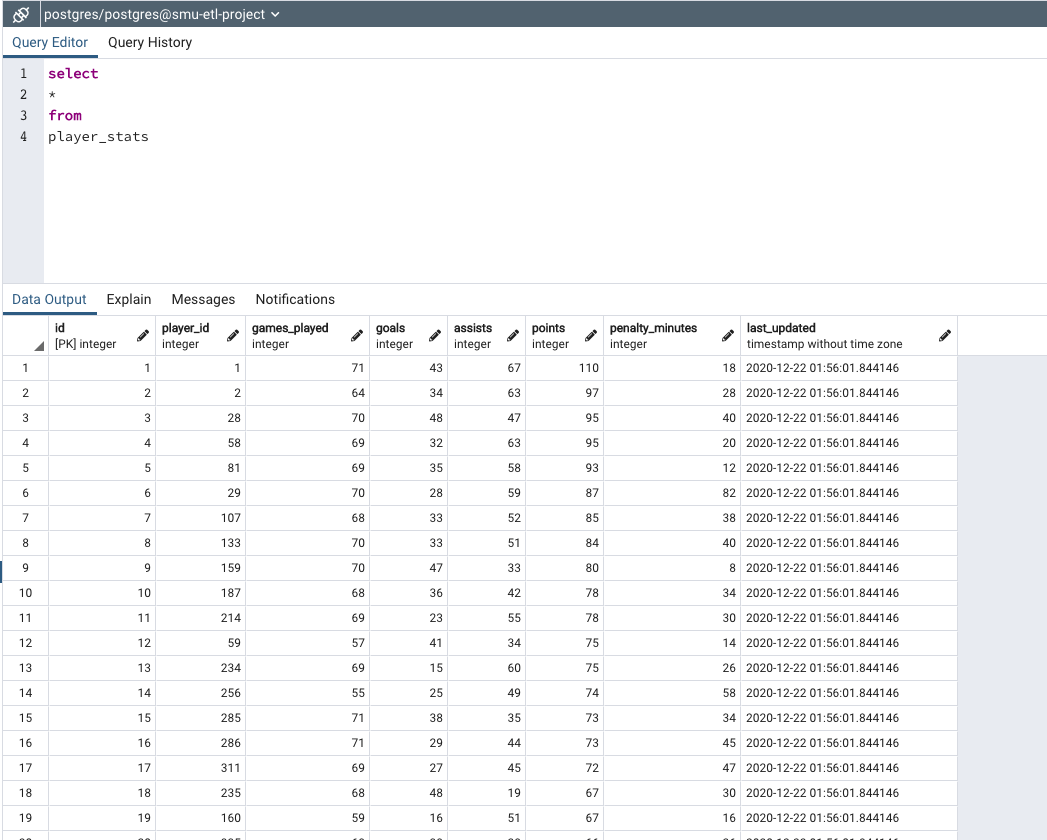
**Load:**

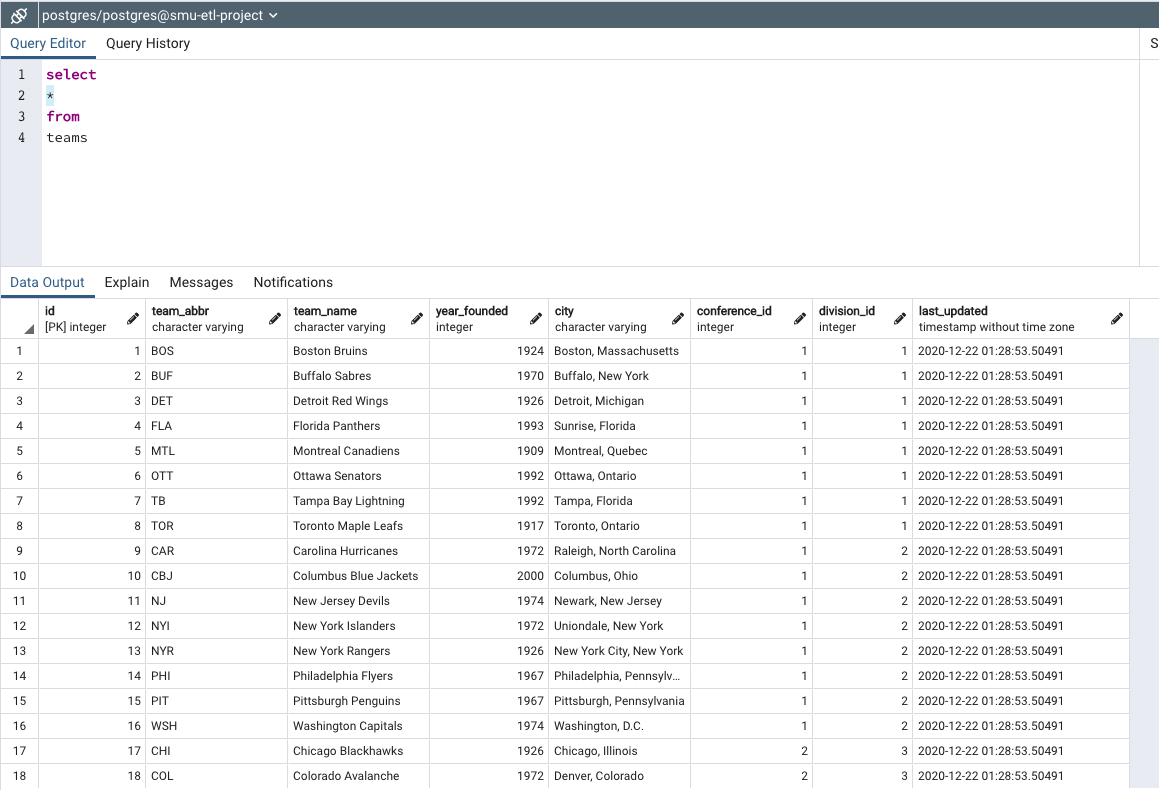
Final database that we created has the following structure represented in ERD diagram:

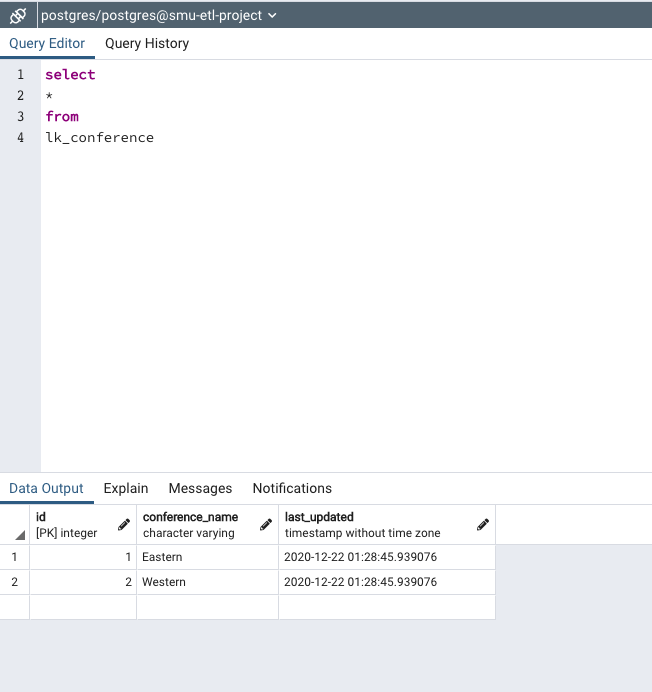


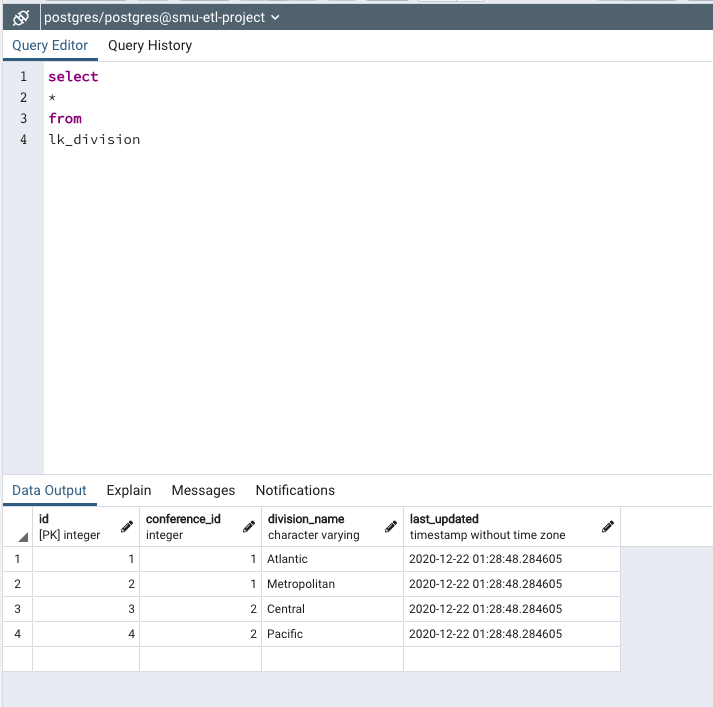
The tables are all uploaded and viewed in pgAdmin as follow:











We also worked on building the Flask API for all our database tables. From the home page, we add the links for each tables as easy access. Moreover, if you want to query the teams under specific division, you can dynamically operate it by add the division id on the URL.

