**Extract:**

For our ETL Project we utilized 4 csv files extracted from the following sources:

1. <https://www.kaggle.com/martinellis/nhl-game-data#game.csv>
2. <https://www.kaggle.com/martinellis/nhl-game-data#team_info.csv>
3. <https://www.kaggle.com/martinellis/nhl-game-data#player_info.csv>
4. <https://www.kaggle.com/martinellis/nhl-game-data#game_skater_stats.csv>

**Transform:**

We imported the above datasets into Jupyter Notebook and used Pandas to transform the datasets.

game\_df :

data cleaning:

1. Removed 10 unwanted columns from the dataframe.

data filtering:

1. Filtered the “type” column to only include rows with “R” where “R” is short for regular season games. This step effectively removed all “P” playoff games from the dataset.
2. Filtered the “season” column to only include rows with “20172018” where “20172018” represents the 2017-2018 season. This step effectively removed all other seasons from the dataset.

dataframe merging:

1. Merged the game\_df with the team\_df, left\_on = away\_team\_id and right\_on = team\_id so that the Away Team values would be their respective names, instead of the away\_team\_id.
2. Merged the game\_df with the team\_df, left\_on = home\_team\_id and right\_on = team\_id so that the Home Team values would be their respective names, instead of the home\_team\_id.

Additional data cleaning:

1. Removed 3 unwanted columns from the new\_game\_df dataframe.

team\_df:

data cleaning:

1. Combined the “shortName” and “teamName” columns into a single column “Team.”
2. Removed 5 unwanted columns from the dataframe.
3. Dropped 2 unwanted teams from the dataframe, because they did not exist during the 2017-2018 season.
4. Replaced 2 team names in order to assign each their real “Team” name.

player\_info\_df:

data cleaning:

1. Combined the “firstName” and “lastName” columns into a single column “Name.”
2. Removed 6 unwanted columns from the dataframe.
3. Renamed columns
4. Modified the csv delimiter, changing it from “|” to “,” in the player\_info\_df output file.

game\_skater\_stats\_df:

data cleaning:

1. Kept only games in new\_game\_df, so that only games during the season we are interested were kept.
2. Only kept columns we were interested in: game\_id, player\_id, team\_id, penaltyMinutes
3. Renamed columns
4. Modified the csv delimiter, changing it from “|” to “,” in the skater\_info\_df output file.

**Load:**

The following tables were created in PostgreSQL database for the above Transformations to be loaded into.

games:

game\_id int Primary Key,

season int,

type varchar,

away\_team\_id int,

home\_team\_id int,

outcome varchar

players:

player\_id int Primary Key,

name varchar,

primary\_position varchar

teams:

team\_id int Primary Key,

team varchar

skater\_stats:

game\_id int,

player\_id int,

team\_id int,

penalty\_minutes int,

Primary Key (game\_id,player\_id)

game\_results:

game\_id int,

position varchar,

home\_team varchar,

home\_team\_pim int,

away\_team varchar,

away\_team\_pim int,

outcome varchar,

avg\_pim\_by\_position\_game float,

Primary Key (game\_id,position)