**Technical report for ETL-Project**

**NBA Data / Team 4**

**William Clewett, Pushpraj (Raj) Rana, and Paul Murray**

**Extract**

For our project we used data about the NBA from the following sites:

<https://www.kaggle.com/drgilermo/nba-players-stats?select=Seasons_Stats.csv>

[www.basketball-reference.com](http://www.basketball-reference.com)

<https://data.world/bgp12/nbancaacomparisons/workspace/file?filename=players.csv>

In total, we started with eight CSV files. Seven of the files are from Basketball Reference, while our largest data file, season statistics was from Data World. Basketball Reference is a free site that stores a large amount of NBA data for public use. Every table on the site can be downloaded as a CSV, and what was most valuable to us was that every players and team has their own page. We used a piece of each player and team’s page to create primary keys. This concept was key to our web scraping script for the NBA Awards table. Below are the CSVs and their source, please note “Players Data” and “Players School and Size” are the same dataset:

Franchise Data – [www.basketball-reference.com](http://www.basketball-reference.com)

NBA Results – [www.basketball-reference.com](http://www.basketball-reference.com)

NCAA Results – [www.basketball-reference.com](http://www.basketball-reference.com)

Players Data – <https://data.world/bgp12/nbancaacomparisons/workspace/file?filename=players.csv>

Players Draft – [www.basketball-reference.com](http://www.basketball-reference.com)

Players School and Size – <https://data.world/bgp12/nbancaacomparisons/workspace/file?filename=players.csv>

Season Stats – <https://www.kaggle.com/drgilermo/nba-players-stats?select=Seasons_Stats.csv>

Team Abbreviations– [www.basketball-reference.com](http://www.basketball-reference.com)

**Transform**

To begin we did a small amount of trimming to our datasets in Microsoft Excel. For example, on the Players Data and Players School and Size dataset we removed NCAA and NBA Statistics, since they were mostly incomplete and we already had game statistics within our Season Stats dataset. After our Excel work, we implemented the rest of our transformations within Jupyter Notebook files. As mentioned earlier, a small amount of web scraping was done for the NBA Awards dataset to retrieve primary keys for our players. The primary keys for players is shown below:



The last section of each player’s URL is a primary key setup by Basketball Reference. In this case it is “abdelal01”. Primary keys for each team were done in a similar fashion. In this case every team that has ever played an NBA game has a three letter abbreviation. Our last set of primary keys were assigned programmatically by looping over each school in our dataset with an “S” and an increasing four digit number. For example, “S0166” is the University of La Salle in our database.

Our largest join, occurred between our final Players Dataset and the Season Stats. We used an Inner Join to make sure that any player not in our Players Dataset would not be listed in our final Season Stats Dataset. Season Stats were joined with a subset of the Players Dataset, player\_id and school\_id, on name. All other datasets were joined in similar fashion to add either the player id, the school id, or the teams abbreviation.

**Load**

To load our cleaned CSVs into Postgres, we created a SQL script based on our ERD. We had to make sure that the primary and foreign keys were ordered correctly to allow Postgres to properly create our relational database. We named our database “nba\_stats\_db”, and imported all of our data via the Import/Export wizard once the table creation was complete.