Extract, Transform, Load

ETL Project Final Report

Group Adam Bilski
Alan Riveros
Aja Tashjian

Project Scope:

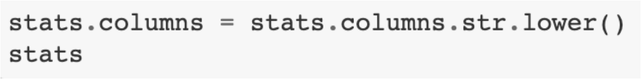
For this project we want to be able to find different sources of data that can potentially join with each other. Using tools such Python, Pandas, and Postgres, just to name a few, separate forms of data will be extracted, transformed, and then loaded into a final database of a new data set. The potential analytics of our data could be used to compare stats of basketball players with their current annual salary.

Extract

Using a Python and a Jupyter Notebook, we were able to scrape data from basketball-reference.com for the basketball players stats. We then downloaded .csv files on basketball salaries and the highest paid athletes from data.world and basketball-reference.com.

Transform

We loaded our data into pandas in order to capture only the data wanted as well as prep for adding to Postgres. Many columns in the Pandas data frames required naming and formatting changes and some required parsing out data in columns in order to keep a specific part of the data. For example, we discovered that Postgres automatically lower cased out column names, which required us to use code to uniform the syntax used. Any columns with extraneous data were dropped.



```
stats.columns = stats.columns.str.lower()
stats
```

Load

Once data was cleaned it was loaded into Postgres using Pandas into three tables:

- Players (basketball player names and their current salaries)
- Top_35 (basketball player names including their pay, winnings, endorsements and marital status)
- Stats (basketball player names and their current stats)

We decided to use CREATE VIEW strategy, named “first_view” to join the stats table into top_35. From there we created a second CREAT VIEW, names “second_view” to join in the Players data into the “first_view”. From this we were able to view data from all three of our data sources into one larger database that we can use. This allows for cross-referencing a basketball player’s stats with their earnings (or perhaps marital status).