**ETL Project**

Team Members: Paul Nisenson and Aditi Rai (Team Two on Two)

Objective: Extract data from multiple sources, transform and load the data so it can be easily accessed/queried by our team.

Practical Question: How has the impact of international players changed in the NBA over the past 30 years?

We have access to multiple applications with relevant data:

1. Basketball-Reference.com (BKRef) – contains all historical and current NBA statistical data in tabular form organized by player, team, season etc.
2. Wikipedia “List\_of\_foreign\_NBA\_players” – page contains a table of each “international” player by nationality and birthplace.

While BKRef has the ability to query all historical data, you can only export approx. 200 rows of data at a time. We will attempt to scrape the data of each player by season and create a Python “For Loop” to pull this data for all seasons from 1980 – 2020. We will then join this data with the list of foreign players to be able to analyze players’ statistics by country.

Approach:

1. Locate/identify source data
   1. See above
2. Extract data into files/dataframe
   1. Use pandas to read html tables
   2. Loop through multiple pages of web data
   3. Save each year as a csv to be imported into database
3. Study / review data
   1. Look for data inconsistencies/missing data
   2. We’ve identified “\*” on player names in Wikipedia, for example
4. Clean data individually
   1. Removing duplicate rows
   2. Add years to each DataFrame
   3. Making sure keys match on both sets of data
5. Find a common value(s)/column(s)
   1. Player Name
6. Merge the datasets
7. Upload the merged data into a database
   1. MongoDB/SQLite – fast!
8. Provide a toolset to be able to query the data loaded
   1. Finished SQL Database
9. Plan to present some visualizations, time permitting