ETL REPORT: NBA STATISTICS

Team members

Shruti

Shao

Aydan

Omar

**Objective and methodology**

The project intended to search for data sites with NBA players and teams information. The final purpose was to implement an NBA\_database prototype that will hold the following information:

1. Player’s information: Based on the NBA Sense API (<http://nbasense.com/nba-api/>). It contains player’s information (unique ID, first name, last name, position, team ID, player’s statistics (points per game, assists, etc.). Due to the length of the data and time limitations, we will only retrieve data for the 2018 season. Once the database and queries are set up, we can add other years starting 2016 (the API does not have data prior to 2016).
2. Team’s information: Also based on NBA Sense API (Sense API (<http://nbasense.com/nba-api/)>). It contains team’s information (unique team ID, team’s name, city, arena capacity)
3. Team’s logos: Scrapping the team’s logo URL from Wikipedia. This information will be added to the master team’s table in the database.
4. Add information on NBA MVP player’s (downloaded as an Excel file from <https://data.world/datatouille/nba-finals-and-mvps>) . Data goes back to 1950 and contains information on the MVP player for each year (name, team, position, nationality and status (whether the MVP’s team win the NBA championship, Divisional Championship or did not reach any of the final stages of the tournament).

The project included the following steps

1. Extracting data from API and creating Pandas data frames
2. Transforming the data in order to:
   1. Filtering the database to contain the fields that would be uploaded into the NBA\_database tables
   2. Remove duplicate records from the master player and team tables to have unique player’s and team’s ID
   3. Drop the none value rows from the table.
   4. By merging team table, and set the team ID to the MVP table. Thus the MVP table with primary key and will be ok to work with other time.
3. Design and creation of the NBA\_database using PostgressSQL

Database diagram and relations

We selected PostgressSQL or structured data to hold the information we

The database includes the following tables:

1. Master\_Player: Unique player’s ID and first and last name (2016-2018)
2. Team: Unique team’s ID, team name and additional information (2016-2018)
3. Player\_info: Player’s additional information and position
4. Player’s statistic’s
5. MVP : team\_id, player name, year , team name(1956-2018)