Project Phase 2

Edmund Lo, Zexi Lv

Design Decisions

For our team relation, we made some minor changes after Phase 1. In our Phase 1 schema, we decided to generate a random number to be the team ID (tID). However, we decided to change the tID to a three letter abbreviation of the team's city and name. This abbreviation is used in the original dataset and as a result, we do not have to generate an int to the tID of each team. This change was made so that the team could be easier to reference and identify when looking at data from other relations like Player, Coach or Game. Other changes to this relation include making a user-defined data type called Percentage which is a float in the limits of 0 to 1. This data type is used for the W/L%, FG%, 3P%, 2P% and FG% attributes. We also had to change the data types of most of the other stats to FLOAT, instead of INT, because they had up to three decimals.

For our player relation, we changed the data types to Percentage or FLOAT from INT as necessary. Also, in the original dataset for the Players, there are rows containing the combined stats of a player who had been on multiple teams during the season. Since we want to analyze the effect that each player has on each team, we decided to remove those combined stats from the dataset and from importing them into our schema.

Cleaning Process

For our team relation, we used the Team Per Game Stats from the 2020-2021 season. First, we removed the rank column and added a column called tID with the three letter abbreviations of the team city and name. These abbreviations were taken from the original Players dataset. Next, we removed the asterisks from the team names. The asterisks were used to indicate if a team made the playoffs, however they are unnecessary for us. Next, we split the team names into its city and name. Lastly, we merged the W, L, and W/L% columns from the Conference Standings tables from the 2020-2021 season with our updated table. The tables were merged based on their names.

For our player relation, we used the Player Per Game Stats from the 2020-2021 season. First we removed the player identifiers (\achiupr01) from the ends of the player names. Next we split their first and last name into two columns. Next we removed the rows where the player's team is TOT. This row is the combined stats of a player who had been on multiple teams during the season. Next, we removed the Rank column and created our own pID column by generating an integer for each player starting from 1. Lastly we substituted 0 for any cell that had a null value. This substitution was made because some cells were empty because the player had taken zero

of those shots. For example, if a player had never shot a 3 pointer, their 3P% would be null. We changed this to 0 because of our NOT NULL constraint and because it makes sense that their 3P% would be 0.

For the game relation, data from the 2020-21 NBA Schedule and Results was used. The original data was split in different locations, grouped by month (e.g one page on the site had all games played in December, another had games in January, etc.). This data was copied into one file since we did not see any need for separating the games by when they were played. The date the game was played had the day of the week in the entry, and was not compatible with the SQL date datatype. We removed this to be able to format the attribute as a date properly. Similar to the player relation, we added an game ID (gID) as an integer counting from 1 up. Lastly, the tID was obtained by looking up the names of the teams in the team relation.

Finally, for the coach relation, data from the 2020-21 NBA Coaches table was used. Again, we split the coaches' names into first and last name attributes. As well, a unique integer ID (cID) for each coach was added by counting from 1 up. Then, some of the attributes we used from the data were the number of playoff games, wins, and losses each coach was a part of. In the case that the coach's team had not made it to playoffs, these values in the original data were left empty. For the purposes of our investigation, we decided to change these values to 0 to maintain our NOT NULL constraint on these attributes.