3380 Group 16
Yirong Wang & Fidelio Ciandy

Summary of The Dataset

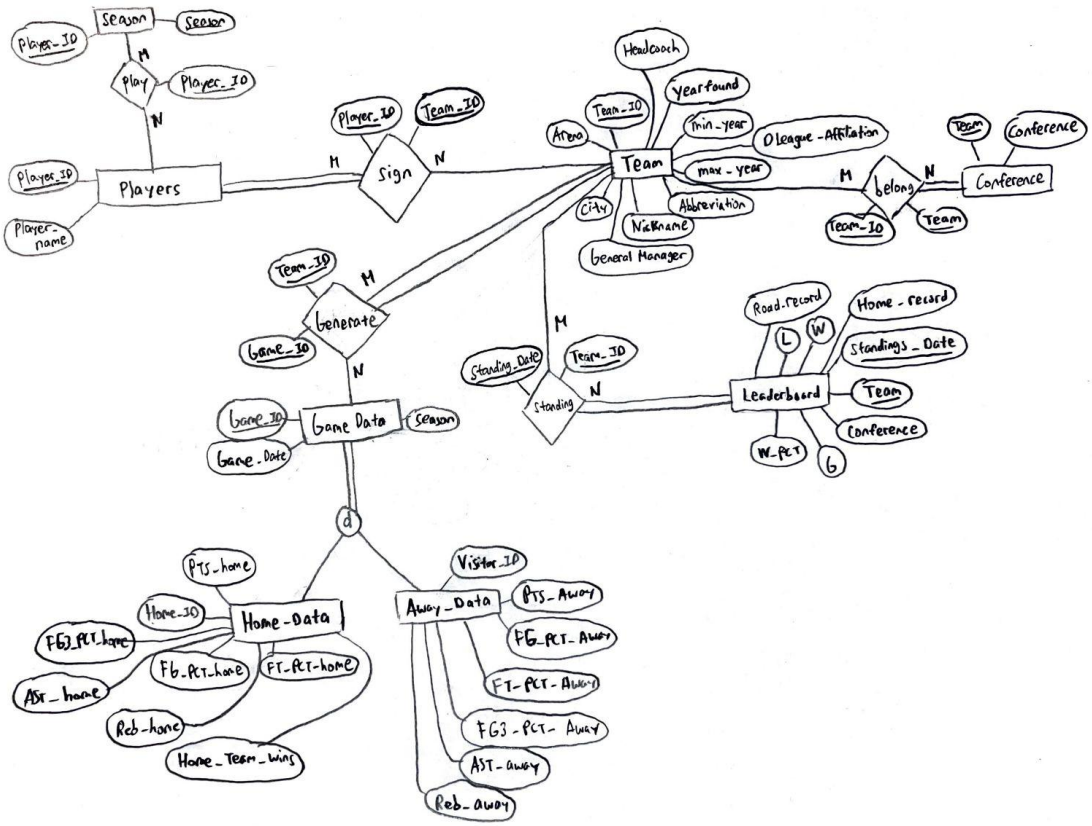
(https://www.kaggle.com/datasets/nathanlauga/nba-games)

This dataset is about NBA games from 2004 until December 2020. There are a total of 5 tables from the original dataset, but we only chose 4 tables to work with. The number of columns and rows from each table are listed below:

- player.csv with 4 columns and 7229 rows,
- games.csv with 19 columns and 25k rows,
- team.csv with 13 columns and 31 rows,
- ranking.csv with 11 columns and 200k rows.

The Player table describes the player name, player id and the season that specific player plays. The Games table describes the game statistics when a specific team plays with another team. The Team tables describe the team background. The Ranking table shows the standing of each team in the NBA leaderboard.

ER Model


Season
Season
Player_ID
Play
Player_ID
M
N
Player_ID
Players
Player_name
Player_ID
Team_ID
M
Sign
N
Headcoach
Year-found
Team_ID
Min_year
Area
League-Affiliation
Team
max_year
City
Nickname
Abbreviation
General Manager
M
belong
N
Team
Conference
Team
Conference
Team_ID
Team
Team_ID
M
Generate
Game_ID
N
Game Data
Game_ID
Season
Game_Date
d
Standing_Date
Team_ID
M
Standing
N
Road_record
L
W
Home_record
Standings_Date
Leaderboard
Team
Conference
W_PCT
G
PTS_home
Home_ID
Home-Data
FG3_PCT_home
FG_PCT_home
FT_PCT_home
AST_home
Reb_home
Home_Team_wins
Visitor_ID
Away_Data
PTS_Away
FG_PCT_Away
FT_PCT_Away
FG3_PCT_Away
AST_Away
Reb_away

Final Relational Model

players.csv

0NF:

R(season, player_ID, player_name)

1NF: No change, no Multi Valued Attribute (MVA) on the table.

2NF:

Season (season, player_ID)

Players(player_ID, player_name)

3NF: No change, no transitivity.

BCNF: No change, all determinant is a superkey

Games.csv

0NF:

R(Game_ID, Game_Date, Home_Team_ID, Visitor_Team_ID, season, PTS_home, FG_PTS_home, FG_PCT_home, FT_PCT_home, FG3_PCT_home, AST_home, REB_home, PTS_away, FG_PCT_away, FT_PCT_away, FG3_PCT_away, AST_away, REB_away, HOME_TEAM_WINS).

1NF: No change, no MVA.

2NF: No change, no partial key dependency.

3NF:

Games(Game_ID, Game_Date, Home_Team_ID, Visitor_Team_ID, season)

Home_Data(Game_Date, Home_Team_ID, PTS_home, FG_PTS_home, FG_PCT_home, FT_PCT_home, FG3_PCT_home, AST_home, REB_home, HOME_TEAM_WINS)

Away_Data(Game_Date, Visitor_Team_ID, PTS_away, FG_PCT_away, FT_PCT_away, FG3_PCT_away, AST_away, REB_away)

BCNF: No change, all determinants are a superkey

Teams.csv

0NF:

Team(Team_ID, Min_Year, Max_Year, Abbreviation, Nickname, year_founded, City, Arena, Arena_Capacity, Owner, General_Manager, Head_coach, Dleague_Affiliation)

1NF: No change, no MVA

2NF: No change, no partial key dependency

3NF: No change, no transitivity

BCNF: No change. All determinants are a super key.

Ranking.csv

0NF:

R(Standings_Date, Team, Season_ID, Conference, Game_played, Game_Won, Game_Lost, Win%, Home_Record, Road_Record)

1NF: No change, no MVA

2NF:

Leaderboard(Standing_Date, Team, Season_ID, Game_played, Game_won, Game_Lost, Win%, Home_Record, Road_Record)

Conference(Team, Conference).

3NF: No change, no transitivity

BCNF: No change. All determinants are a super key.

Post merge and post normalization:

Season (Season, player_ID)

Player (player_ID, player_name)

Games(Game_ID, Game_Date, Home_Team_ID, Visitor_Team_ID, season)

Home_Data(Game_Date, Home_Team_ID, PTS_home, FG_PTS_home, FG_PCT_home, FT_PCT_home, FG3_PCT_home, AST_home, REB_home, HOME_TEAM_WINS)

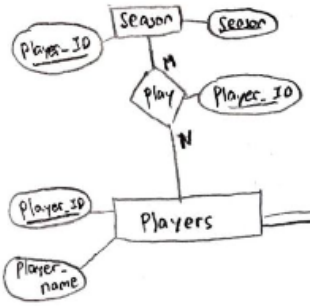
Away_Data(Game_Date, Visitor_Team_ID, PTS_away, FG_PCT_away, FT_PCT_away, FG3_PCT_away, AST_away, REB_away)

Team(Team_ID, Min_Year, Max_Year, Abbreviation, Nickname, year_founded, City, Arena, Arena_Capacity, Owner, General_Manager, Head_coach, Dleague_Affiliation)

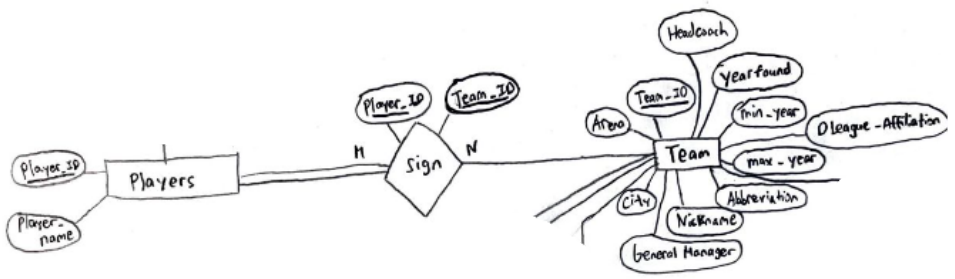
Leaderboard(Standing_Date, Team, Season_ID, Game_played, Game_won, Game_Lost, Win%, Home_Record, Road_Record)

Conference(Team, Conference).

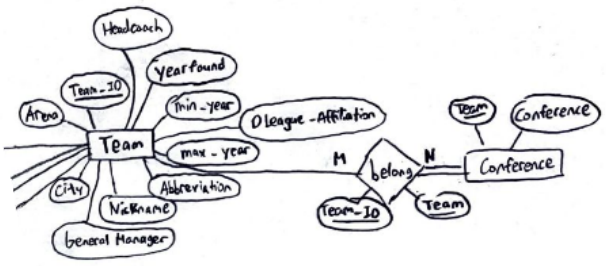
Justifications for participation/ cardinality constraints

- 

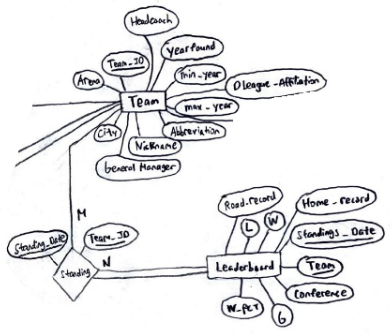
    - Players - Play - Season
        - It is a many to many relationship because players can play many seasons and each season could be played by many players.
        - It is not a total participation because there could be a chance that a player does not play for that specific season. So, it is just a partial participation.



-
    - Player - Sign - Team
        - It is a many-to-many relationship because many players can sign to different teams and each team can sign many players.
        - It is a total participation because a team must have players. Also, players must be signed to a team to be considered a player. Therefore, it is a total participation.

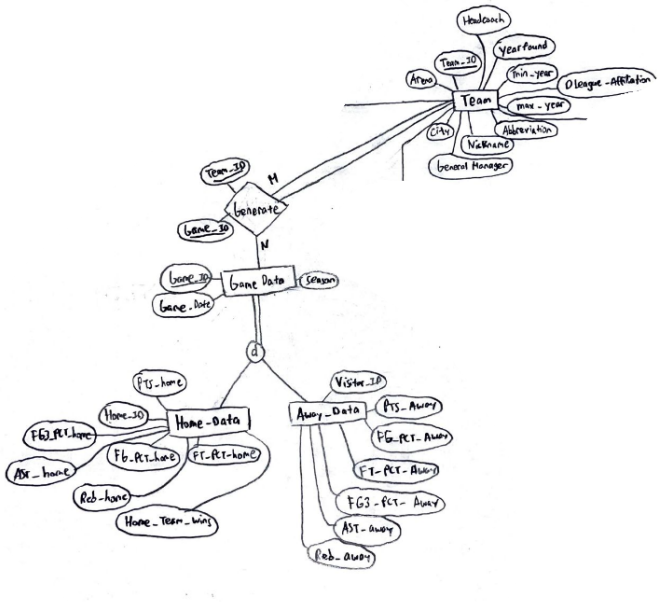


-
    - Team - belong - Conference
        - Team belongs to a Conference is a many to many relationship because teams can be in a west or east conference. Also, the conference can be filled with many teams. So, it is a many to many relationship.
        - It is a total participation because a team MUST be in a west or east conference.



-
    - Team - Standing - Leaderboard

- Many teams could have a stand in the leaderboard. Also, the leaderboard is populated by many teams. Therefore, it is a many to many relationship.
- It is a total relationship because to be considered a team, they must be in the leaderboard. It does not matter if a team is at the bottom of the leaderboard. They have to be in the NBA leaderboard or rank.



-
  - Team - Generate - Game data
    - It is a many-to-many relationship because Teams can generate a lot of data and much data will be generated in the same team.
    - It is total participation because when a team plays they must have game data and game data must be generated when the team plays.
  - Game Data Disjoint
    - Game data is a disjoint set because game data can be divided into two parts. Where the first part is the home team data and the other part is the away team data.
    - It is also a total participation because game data must have a home data or an away data. It could not be anything else other than that.