INST 327 Section 0203 – Database Design and Modeling

Final Project Report

5/10/2021

Team 7

Charlie Smith, Matthew Santos, Nicholas Urquhart, Daniel Rong,
Nabil Siddiqui

Final Project Report

Introduction

Big data has been transforming the field of information technology. Technological advances in our ability to gather, store, and analyze data has improved our understanding in many aspects of the world. By collecting data on every bit of information that is out there, we are able to make connections that we never otherwise would have seen. The service our database provides fans and team managers alike to compare the stats of MLB league players. Users are given the opportunity to view the accolades of players and their favorite teams.

Baseball has a large amount of data that teams have tracked ranging from player performance data, to team yearly stats and standings. As more and more analysis is done on these data sets, our understanding of the game will improve. And so for this project, we gathered data on Major League Baseball statistics from 1871 to 2019. However, for the sake of time constraints and proper presentation, we collectively decided to show a sample of the potential of our database through showing a mix of both real and hypothetical data.

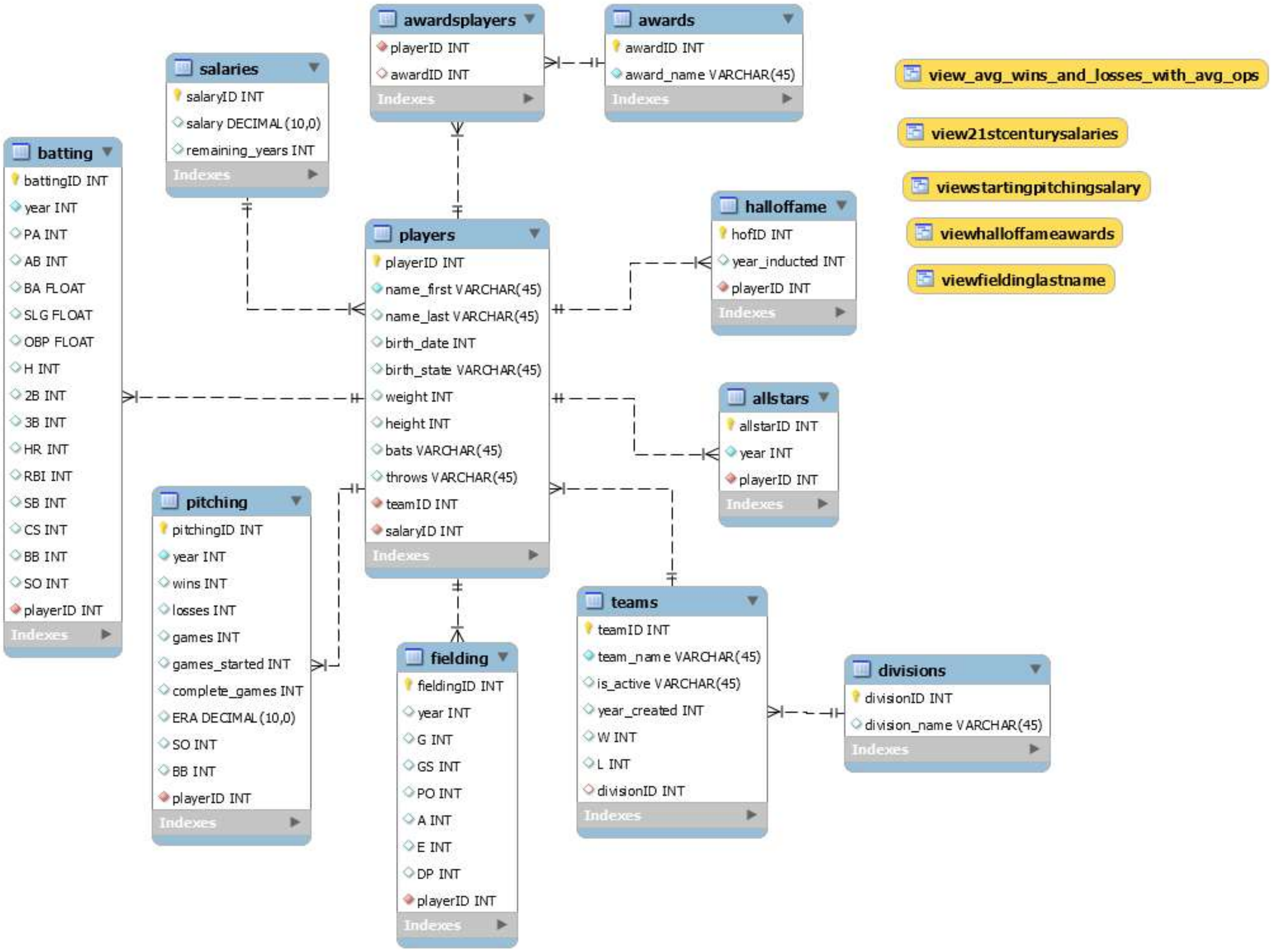
Database Description

Physical Database

Our database includes 11 tables. The main table is the players table. This table references nine other tables. Six of these relationships are one to many: awardsplayers, hallfoffame, allstars, fielding, pitching, and batting, and two are many to one: salaries and teams. The awardsplayers table has a many to one relationship with the awards table. The teams table has a many to one relationship with the divisions table.

Logical Design

Below is an image of our ERD.



Sample Data

Through the use of web scraping methods, we conducted searches from existing databases to compile to ours. These databases are:

- ESPN
- MLB stats
- Baseball-Reference

- Lahman’s baseball database

It is important to note that much of our original plans to use this data required much more time than what we intentionally planned to reach this point of the database. Though the data used in our database is exhibitive of real statistics, the databases we scoured were mainly used as inspiration for our final product. Each major statistic such as batting, fielding, or pitching has one season an individual player has played and they might not align with other major statistical tables. So, we collected randomized data of players that might not associate properly with real life players that are included in our database.

Views/Queries

Query Name	**VIEWS(5)**	**FILTERING (3)**	**AGGREGATION(2)**	**LINKING (1)**	**SUBQUERY (1)**	**JOINS AS VIEWS (4)**
All-star salaries of the 21st century	✓	✓				✓
Awards and hof winners	✓	✓	✓	✓	✓	✓
Avg wins and losses per division, with average OPS per division	✓		✓			✓
Starting Pitchers’ Salaries	✓					✓
Fielding Stats Based on Last Name A	✓	✓				✓
Total	**5**	**3**	**2**	**1**	**1**	**5**

The following list describes what each query we wrote for our database displays:

Query 1: Creates a view of players and their salaries in years after 2001.

Query 2: Creates a view of players in the hall of fame and all awards won with the year.

Query 3: Creates a view of average on base plus slugging stat per division as well as their average wins/losses.

Query 4: Creates a view of all the pitchers' salaries along with the amount of games and starts they played that season.

Query 5: Creates a view with all the fielding stats. This query is an example of the type of queries that can be done with the three major statistical tables of our database: batting, fielding, pitching.

Changes from original design

While making our database, we ran into a few problems. For one, we discovered that there was simply too much data for us to sift through with all the different aspects of baseball. As a result, we decided to change our database to be one that is focused on gathering individual player stats from a mix of eras. Our database represents a depiction of how the data will be organized and viewed by users. Compared to the original ERDs and pseudocode, our team had finished normalization while varying between both one-to-one and one-to-many relationships between our tables. We also changed our data collecting methods entirely as we collected minimum data regarding real life players. Though these players have played in the MLB in prior years, stats and accolades are all hypothetical and only represent the realistic potential of our database.

By reviewing our initial proposal's potential tables and entities, our team had concluded that the original primary key names were not distinguishable from the multiple tables' primary keys we planned to work with. Our ERD shows a drastic change in the names and the many relationships of the tables that fit more toward the scope for our project. We had originally intended to exclude the awards table, but since our database focuses on the collection of individual players' careers, we decided that it was significant enough to implement along with the all-stars and hall of fame tables. Another major change from the original database was the exclusion of the managers table. As we progressed through the project, we decided to narrow down the original dataset to remove managers and only include players - this change was implemented mainly to standardize our database to avoid clutter. We concluded that since our audience included managers, it is unnecessary to include their stats in our database.

Database ethics

We believe that database ethics was not relevant to the project topic. All of our data collected is about how well players and teams have performed during a season. All of which is web-scraped public information for those that keep an eye out for it during the public games. If we were to possibly expand our database to include biometric data such as sleep patterns or diet, then we might start stepping into more unreasonable territories.

Collecting biometric data could be extremely useful for tracking general player health and wellness throughout their career. As they are athletes, their physical wellness is of utmost importance. However, this can raise some issues when it comes to the use of data. Namely, who does the data belong to? The player or the team? And how critical is privacy when it comes to this data? Should it be readily available for the public to see or should only team managers be able to see it. At some point, teams will need to decide what is okay to track and what isn't.

Lessons learned

Outside of just the field of SQL, we have developed skills that can carry onto many different aspects of database management. By creating this baseball database, we were able to create complex relationships between data we have collected as a team. While we have progressed through this project, however, we quickly learned that the scope of the project should remain simple until we reach that goal and expand on it. Our main issue was collecting abundant amounts of data and working backwards from them and basing our project on that. We ultimately should have decided on much narrower and realistic scopes.

Potential Future Work

Obviously, a potential for future work would be to add more data into the database. As of now, we have a fraction of the entire MLB history in our database. For example, 36 players exist in our database, but over 18,000 people have played in the MLB. Now, that is a large number, and probably something we don’t have the knowledge yet to complete, so we might instead focus on a particular era.

Another potential for future work would be the addition of more statistical columns. Baseball is a numbers game, just about every action and event has a statistic attached to it. Therefore, the range of statistics we could include is virtually indefinite. We could even include statistics involving advanced metrics. With analytics, we could include statistics that involve mathematical formulas. We could even use statcast, which was first introduced in 2015. Statcast tracks things such as exit velocity from the bat, and a hitter’s launch angle. These sorts of stats could be very appealing to those interested in advanced stats.

The potential of more statistical columns also brings the potential of including stored procedures and triggers. Many advanced stats, especially analytics, are just mathematical formulas composed of normal baseball stats. For example,OPS, which stands for On Base Plus Slugging, simply adds the common statistics OBP (On Base Percentage) and SLG (Slugging). We could easily add a trigger to our database that whenever a new player is added to the batting column, their OPS is automatically calculated from OBP and SLG and added as a column.