College Football Database

IST 659 Project

Jonah Witt

Section 72410

**Summary**

The goal of this project is to create a database to store data on NCAA Football teams. This is a personal database that will not be used by any member of the teams involved. Due to this, I am the only stakeholder in the project. This database will eventually serve as a basis to run predictive analysis. A large variety of data and statistics will be collected on each team. Initially, data will be collected from the 2017 Tennessee Volunteers. Next, the scope will be expanded to every team in the Southeastern Conference for the 2017 Season. After this project is complete, the database will expand to include other conferences and seasons. However, this project will just include teams from the SEC during the 2017 season. The data will include categorical data on the teams, conferences, and players as well as quantitative statistics, by game, for each player.

In order for the database to be relevant, it must be able to answer the following questions:

1. Who had the most rushing yards for the Tennessee Volunteers in 2017?
2. What team in the SEC had the kicker with the highest Field Goal Percentage in 2017?
3. What teams in the SEC won 6 or more games in 2017?
4. Who is the coach of the team that won the National Championship in 2017?
5. Did the Florida Gators beat the Georgia Bulldogs in 2017?

**Conceptual Model**

Since this is a personal database for recreational analysis, I am the only stakeholder in the project. The business rules collected for the database are as follows:

* A Team can belong to only one Conference per Season.
* A Team can only be a National Champion if that team has a Win result in the National Championship Game type.
* A Team can only be a Conference Champion if that team has a Win result in the Conference Championship Game type.
* A Team can have multiple Coaches in a Season, but may only have one Coach per Game.
* A Player is assigned one Position and one Class per Season.
* A Player can play for only one Team per Season.
* A Coach can coach for only one Team per Season.
* A Game can only have one winner and one loser.
* The Team with the most points is the winner of the Game.
* There is only one National Champion per Division per Season.
* There is only one Conference Champion per Conference per Season.
* Only two teams may play in the same Bowl Game per Season.
* A team can only play a game against another Team once per Game Type per Season.

Again, none of the data collected is intended to be used by any of the teams included in the database. Data will be collected from multiple sources and input into the database. Below are some examples of unrefined data that will be input into the database.

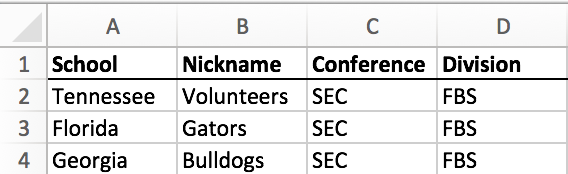


Figure 1: Sample Categorical Data on Teams

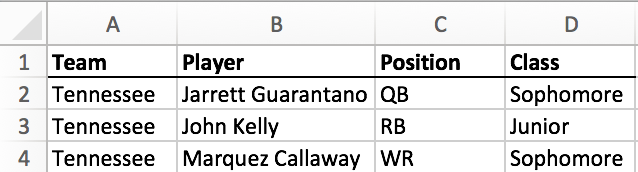


Figure 2: Sample Categorical Data on Players

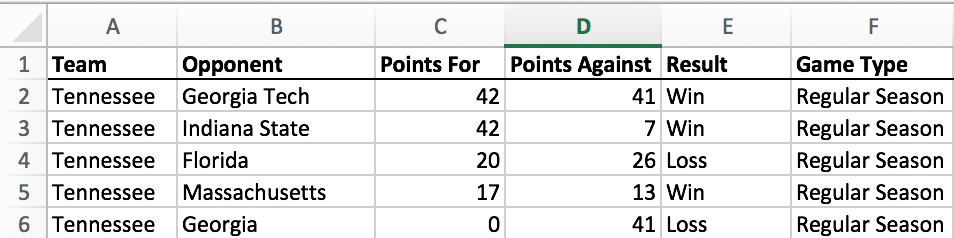


Figure 3: Sample Data from Games

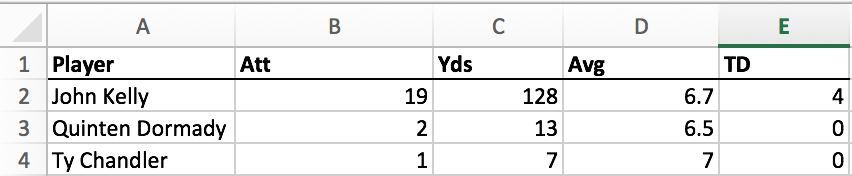


Figure 4: Sample Example of Rushing Statistics for a Game

Here is a glossary of the entities, attributes, and relationships for the database:

|  |  |
| --- | --- |
| **Entity** | **Attribute** |
| Division | Name [ru] |
| Conference | Name [ru] |
| Team | School [ru]  Nickname [r]  Wins [d]  Losses [d]  Conference Championships [d]  National Championships [d] |
| Season | Year [r]  Wins [d]  Losses [d]  Conference Championship [d]  National Championship [d]  Bowl Game [d] |
| Game | Opponent [r]  Points For [r]  Points Against [r]  Result [r]  Game Type [r] |
| Coach | Name [r]  Wins [d]  Losses [d]  First Season [r]  Final Season  Status [r]  National Championships [d]  Conference Championships [d] |
| Player | Name [r]  Position [r]  Class [r] |
| Stats | Pass Completions Pass Attempts Pass Completion Percentage Passing Yards Passing Yards Per Attempt Adjusted Passing Yards Per Attempt Passing Touchdowns Passing Interceptions Passing Efficiency Rating Rushing Attempts Rushing Yards Rushing Yards Per Attempt  Rushing Touchdowns Receptions Receiving Yards Receiving Yards Per Reception Receiving Touchdowns Plays From Scrimmage Yards From Scrimmage  Yards From Scrimmage Per Play Touchdowns From Scrimmage Solo Tackles Assisted Tackles Total Tackles Tackles For Loss Interceptions Interception Return Yards Interception Return Yards Per Interception Interception Return Touchdowns Passes Defended Fumbles Recovered Fumble Recovery Return Yards Fumble Recovery Return Touchdowns Fumbles Forced Kickoff Returns Kickoff Return Yards Kickoff Return Yards Per Return Kickoff Return Touchdowns Punt Returns  Punt Return Yards Punt Return Yards Per Return Punt Return Touchdowns Extra Points Made Extra Point Attempts Extra Point Percentage Field Goals Made Field Goal Attempts Field Goal Percentage Kicking Points Punts Punting Yards Punting Yards Per Punt |
| **Relationships** | |
| Each Division contains 1 or more Conference, each Conference belongs to 1 and only 1 Division  Each Conference has 1 or more Team, each Team belongs to zero or 1 Conference  Each Team participates in 1 or more Season, each Season has 1 or more Team  Each Season contains 1 or more Game, each Game belongs to 1 and only 1 Season  Each Game is coached by 1 and only 1 Coach, each Coach coaches 1 or more Game  Each Game contains 1 or more Stats, each Stats belongs to 1 and only 1 Game  Each Player has zero or more Stats, each Stats belongs to 1 and only 1 Player | |

Figure 5: Entity-Relationship Diagram Glossary

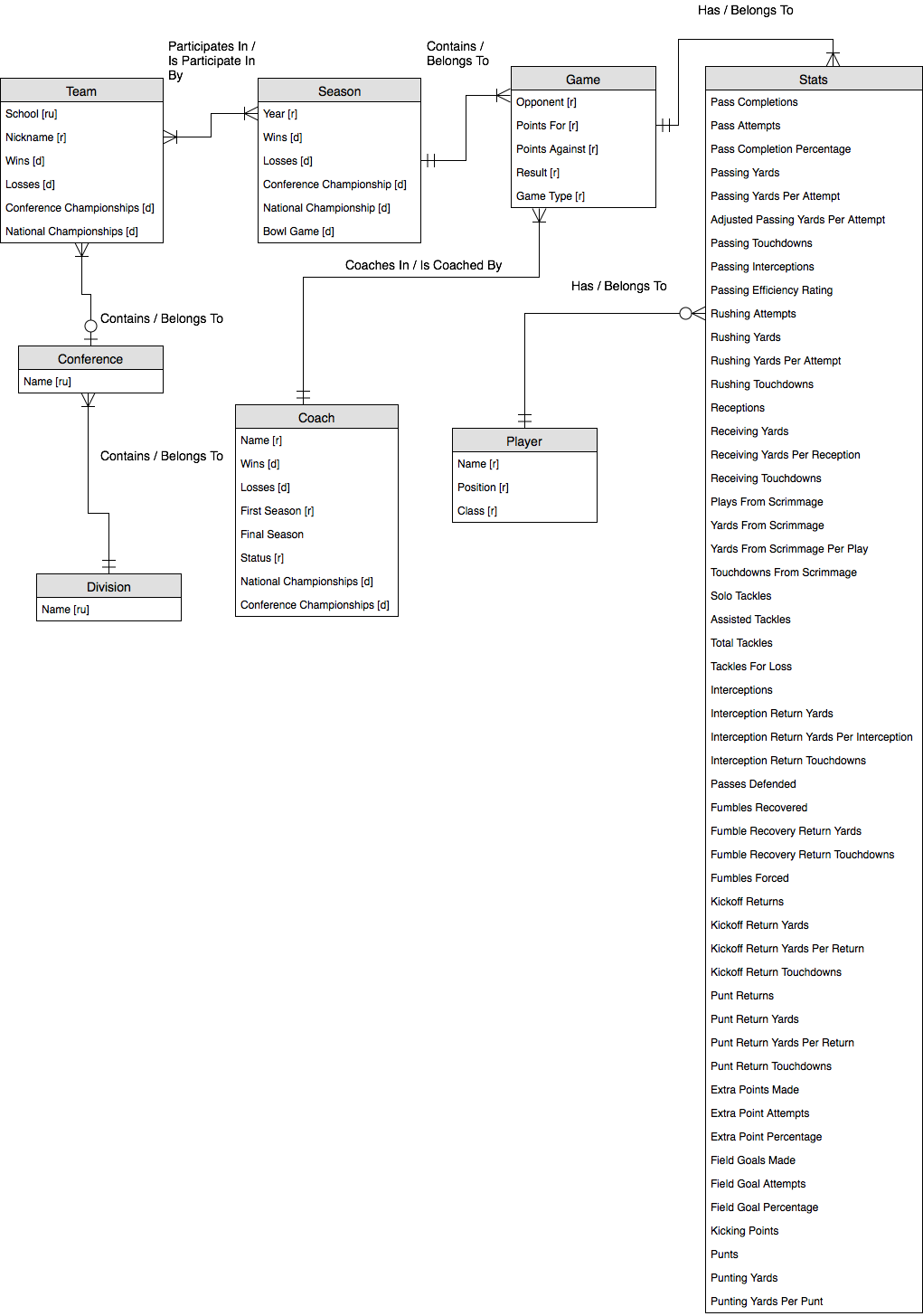


Figure 6: Entity-Relationship Diagram

**Normalized Logical Model**

The conceptual model was decomposed into relations in 3NF as follows:

* division (division id, name)
* conference (conference id, name, *division id*)
* team (team id, school, nickname, wins, losses, conference championships, national championships, *conference id*)
* season (season id, year, wins, losses, conference championship, national championship, bowl game)
* game (game id, opponent, points for, points against, result, game type, *coach id*, *season id*)
* coach (coach id, first name, last name, wins, losses, first season, final season, status, national championships, conference championships)
* player (player id, first name, last name, position, class)
* game\_stats (game\_stats id, pass completions, pass attempts, pass completion percentage, passing yards, passing yards per attempt, adjusted passing yards per attempt, passing touchdowns, passing interceptions, passing efficiency rating, rushing attempts, rushing yards, rushing yards per attempt , rushing touchdowns, receptions, receiving yards, receiving yards per reception, receiving touchdowns, plays from scrimmage, yards from scrimmage, yards from scrimmage per play, touchdowns from scrimmage, solo tackles, assisted tackles, total tackles, tackles for loss, interceptions, interception return yards, interception return yards per interception, interception return touchdowns, passes defended, fumbles recovered, fumble recovery return yards, fumble recovery return touchdowns, fumbles forced, kickoff returns, kickoff return yards, kickoff return yards per return, kickoff return touchdowns, punt returns, punt return yards, punt return yards per return, punt return touchdowns, extra points made, extra point attempts, extra point percentage, field goals made, field goal attempts, field goal percentage, kicking points, punts, punting yards, punting yards per punt, *game id*, *player id*)

Once the relations were normalized they were put into the logical model. In this step, several attributes were added and data types were selected. For each entity, a surrogate key was added as the primary key. Surrogate keys were chosen because some entities (players and coaches) could have the same name. Since surrogate keys were going to need to be used, I chose to streamline that across each entity to keep the database consistent. In addition, the name attributes for Player and Coach were decomposed into first names and last names in case they needed to be accessed or sorted individually.

The data type for each attribute was chosen on the following criteria:

* Name attributes were chosen as varchar to accommodate names of different character sizes. Size for each attribute was chosen based on the estimated amount of characters needed for each type of attribute.
* Year attributes were chosen as char with a character size 4 because no arithmetic will need to be done on the year attributes.
* All of the statistics that are counts were chosen as int type so that arithmetic can be done on them.
* All of the statistics that were averages or percentages were chosen as decimals to show the precise value of each entity.
* Wins and losses were chosen as int type so that they can be kept as counts.
* Championships for every entity except for Season were chosen as int so that they can be kept as counts.
* Championships for Season were chosen as varchar with size 3 to contain “Yes” or “No”
* Coach Status was chosen as varchar with size 7 because that is the largest size value it will hold (“Current”).
* Player position and class were chosen as char with size 2 because each value is an abbreviation with two letters.
* Points for and against were chosen as int so that arithmetic can be done on them.
* Result was chosen as varchar with size 4 since that is the largest size value it will hold (“Loss”).
* Game type was chosen as varchar with size 25 as an estimate of the largest value that it will hold.

Below is the completed database diagram:

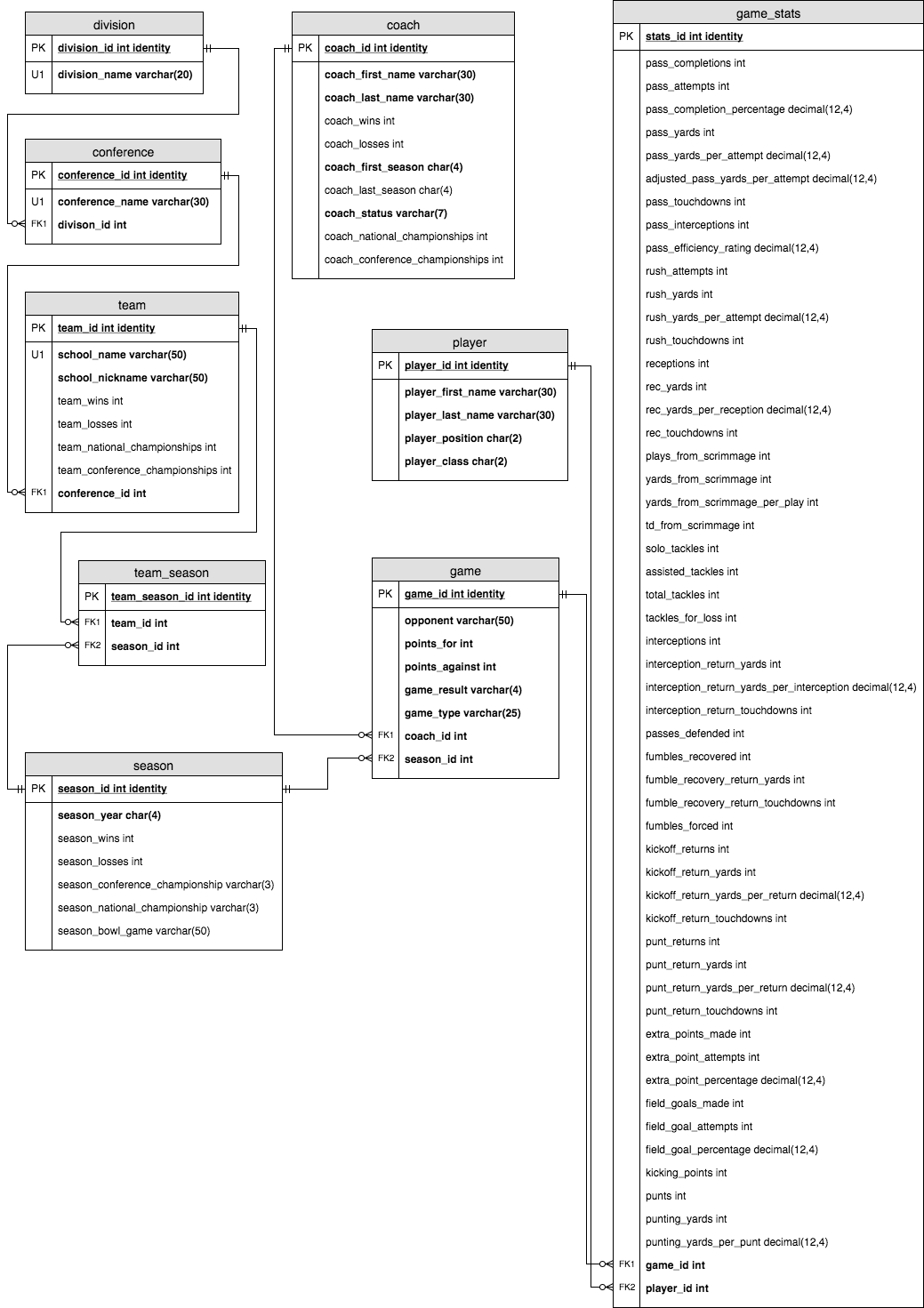


Figure 7: Database Diagram