**MLB Baseball Database**

**Christopher Doherty, Oliver Nabavian**

**Bangerz Data Collective**

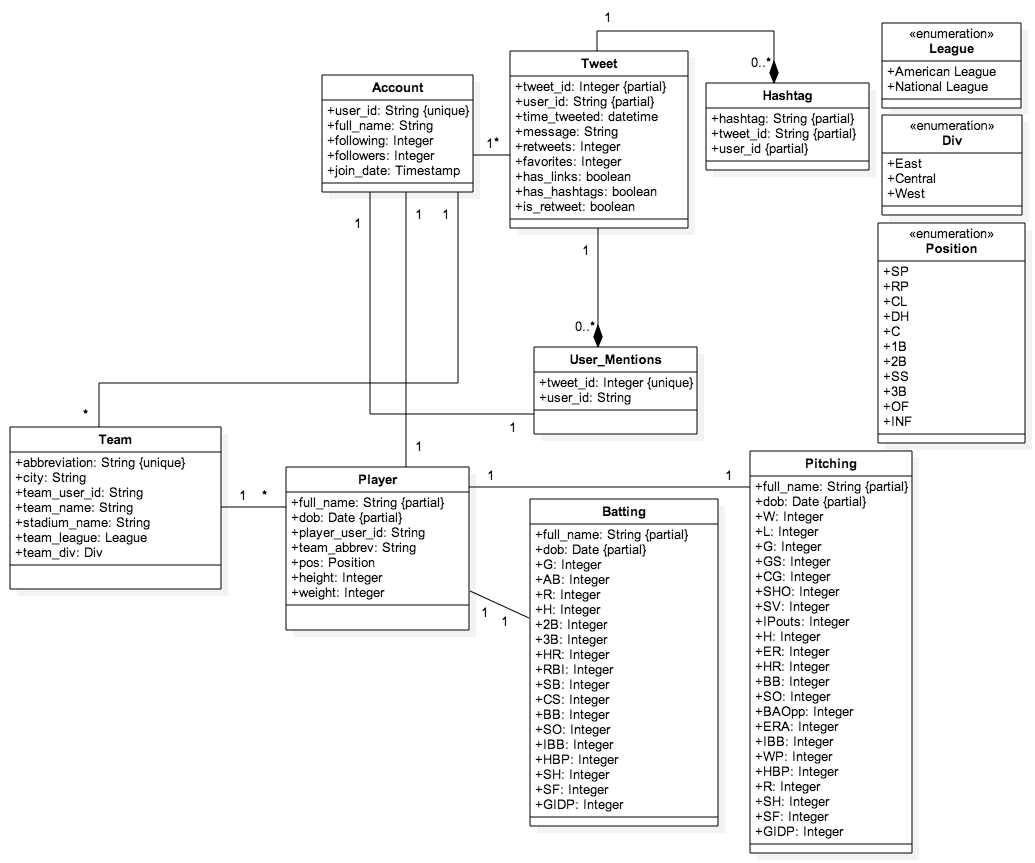
**4/29/15**

The MLB Baseball database provides you with 2014 MLB stats for current players. All teams are linked to their respective Twitter data and all players are also linked to Twitter, provided that player has a twitter account. Currently, the database provides the last 20 Tweets from each account in the database, but as the accounts in the database tweet more, then periodic updates will be provided.

Included in this data collective are a variety of scripts; they are as follows:

* **database\_project.py**
  + The main script for the project. Gets player names and twitter handles from baseball-reference then interacts with a database file to match up with only current players(players who have played within the last year). The player info is then inserted into the baseball database. Team information is then grabbed from mlb.com and parsed and inserted into the database.
* **player\_stats.py**
  + A script that query's Lahman's Baseball database(<http://www.seanlahman.com/baseball-archive/statistics/>) to pull player stats from the database. Will be updated to manipulate data and insert into our database.
* **twitter.py**
  + This script takes all players that have been determined to have a twitter handles and downloads their last 20 tweets and inserts them into the database.
* **twitter\_player\_handles.py**
  + This script matches player who have played in the last 12 months with a list of players who have twitter handles provided by baseball-reference(<http://www.baseball-reference.com/friv/baseball-player-twitter-accounts.shtml>).
* **hashtags.py**
  + This script parses all downloaded tweets and scrapes the hashtags from them. It then inserts these tags into the database’s *Hashtag* table.
* **user\_mentions.py**
  + This script does the same as above but inserts user mentions into the *User\_Mentions* table.

**Database Schema**



Here is the schema for our database. The *Team* and *Player* tables hold demographic information about teams and players respectively and are linked by the team abbreviation. The *Batting* and *Pitching* tables contain statistical information about players and are linked to the *Player* table by their *full\_name* and *dob* to guarantee that each player record is unique, as players can have the same name but are unlikely to have the same birthday. The *Account* table holds information about a Twitter account and is linked to *Player* and *Team* by *user\_id*. The *Tweet* table contains information about each Tweet and is linked to the *Account* table by the *user\_id* of the user that tweeted it. Finally, *Hashtag* and *User\_Mentions*  contain information about Tweets with Hashtags and User Mentions in them, and are linked to *Tweet* by the *tweet\_id.*

Below are a couple of use cases to see what this database is capable of:

