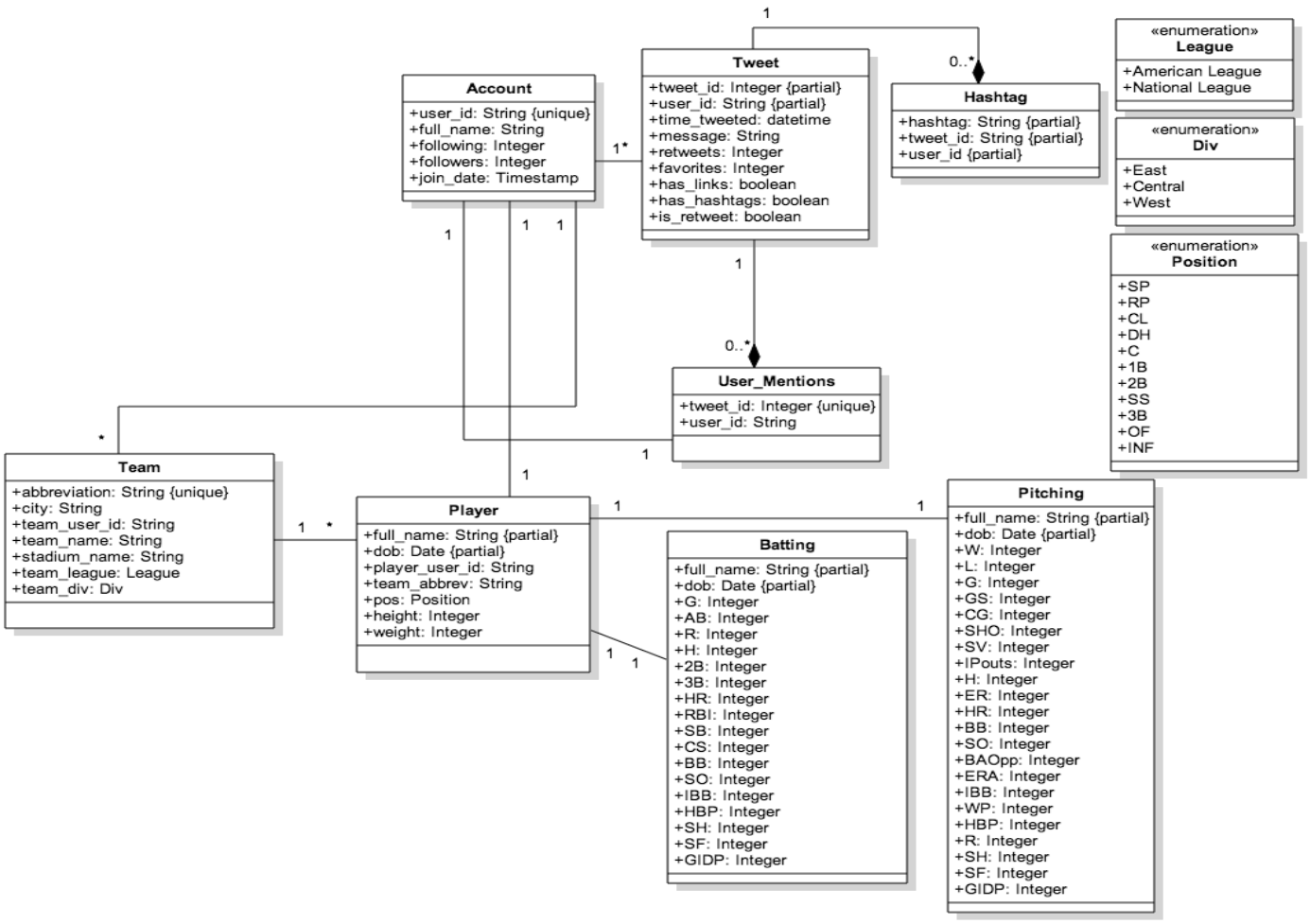
MLB Baseball Database
Christopher Doherty, Oliver Nabavian
Bangerz Data Collective
4/29/15

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:

```
/*
Use Case 1
Description: See which Players have the most followers
Actor: User
Precondition: Players must have Twitter accounts
Steps: Find all players with Twitter accounts and then find the most followed(Top 15)
Actor action: Request to see Players with Twitter accounts
System Responses: Return list of 15 players with full_names and Twitter handles
Post Condition: User will be given name and handle of most followed players
Alternate Path:
Error: User input is incorrect
*/
SELECT a.full_name, a.user_id, a.followers
FROM (Account a LEFT JOIN Player p ON p.player_user_id = a.user_id)
ORDER BY a.followers DESC
LIMIT 15;
/* Output
+-----------------------------------+-----------------+-----------+
| full_name                         | user_id         | followers |
+-----------------------------------+-----------------+-----------+
| Nick Swisher                      | NickSwisher     |   1736163 |
| ダルビッシュ有(Yu Darvish)          | faridyu         |   1252441 |
| Brandon Phillips                  | DatDudeBP       |   1005157 |
| 田中将大/MASAHIRO TANAKA           | t_masahiro18    |    844487 |
| David Ortiz                       | davidortiz      |    804894 |
| Luis montes Jiménez               | Chapomontes10   |    662539 |
| Jose Bautista                     | JoeyBats19      |    648481 |
| Mike Trout                        | Trouty20        |    629613 |
| Brian Wilson                      | BrianWilson38   |    605029 |
| Justin Verlander                  | JustinVerlander |    585342 |
| Miguel Cabrera                    | MiguelCabrera   |    565891 |
| Robinson Cano                     | RobinsonCano    |    490128 |
| Bryce Harper                      | Bharper3407     |    460107 |
| Matt Kemp                         | TheRealMattKemp |    434733 |
| CC Sabathia                       | CC_Sabathia     |    416583 |
+-----------------------------------+-----------------+-----------+
*/
```

```
/*
Use Case 2
Description: Get a Players Twitter handle with their 2014 hits
Actor: User
Precondition: Player must have a Twitter account to be included
Steps: Find all players with Twitter accounts and then find each players hits
Actor action: Request to see Players with Twitter accounts
System Responses: Return list of all players on Twitter with their 2014 hits
Post Condition: User will be given name and handle as well as hits of players
Alternate Path:
Error: User input is incorrect
*/
SELECT p.full_name, p.player_user_id, b.H
FROM ((Player p INNER JOIN Batting b ON b.full_name=p.full_name AND b.dob=p.dob)
LEFT JOIN Account a ON a.user_id=p.player_user_id)
WHERE b.h > 0 AND NOT p.player_user_id="NULL"
ORDER BY b.H DESC;
/* Output (Truncated)
+----------------------+------------------+-----+
| full_name            | player_user_id   | H   |
+----------------------+------------------+-----+
| Jose Altuve          | @JoseAltuve27    | 225 |
| Miguel Cabrera       | @MiguelCabrera   | 191 |
| Ian Kinsler          | @IKinsler3       | 188 |
| Robinson Cano        | @RobinsonCano    | 187 |
| Ben Revere           | @BenRevere9      | 184 |
| Denard Span          | @thisisdspan     | 184 |
| Adam Jones           | @SimplyAJ10      | 181 |
| Howie Kendrick       | @HKendrick47     | 181 |
| Hunter Pence         | @hunterpence     | 180 |
| Jose Abreu           | @79JoseAbreu     | 176 |
| Dee Gordon           | @FlashGJr        | 176 |
| Jonathan Lucroy      | @JLucroy20       | 176 |
| Freddie Freeman      | @FreddieFreeman5 | 175 |
| Jose Reyes           | @lamelaza_7      | 175 |
| James Loney          | @theloney_s      | 174 |
| Mike Trout           | @Trouty20        | 173 |
| Andrew McCutchen     | @TheCUTCH22      | 172 |
| Albert Pujols        | @PujolsFive      | 172 |
| Charlie Blackmon     | @Chuck_Nazty     | 171 |
| Buster Posey         | @BusterPosey     | 170 |
| Alexei Ramirez       | @ImTheRealAlexei | 170 |
| Nelson Cruz          | @ncboomstick23   | 166 |
| Alcides Escobar      | @alcidesescobar2 | 165 |
| Yasiel Puig          | @YasielPuig      | 165 |
| Christian Yelich     | @ChristianYelich | 165 |
| Erick Aybar          | @aybarer01       | 164 |
...
*/
```