BASEBALL HACKS™

Tips & Tools for Dissecting and Analyzing Statistics

O'REILLY®

Joseph Adler

HACK #10 Get a MySQL Database of Player and Team Statistics

Get a free database of historical baseball data from the Internet (covering every major league game from 1871 through today) in MySQL format.

If you don't have Microsoft Access (and you don't want to buy it), or if you are a more experienced database user, you might prefer to use MySQL. The web site *http://www.baseball-databank.org* offers the same database that the Baseball Archive web site offers, but as a MySQL dump file.

Even if you have Microsoft Access, I recommend that you try MySQL, because it's faster, more standards compliant, more flexible, and relatively painless once it's up and running. It's also easier to use with other software, which is important for many hacks in this book. For instructions on getting MySQL, see "Get and Install MySQL" **[Hack #8]**. For suggestions on how to make it easier to work with MySQL, see "Get a GUI for MySQL" **[Hack #18]**. Finally, see "Use SQL to Explore Game Data" **[Hack #16]** for more information about the SQL language.

After you have MySQL installed, here's how to get the files and load them into your database.

Step 1: Download the File

You can get the file you need from *http://www.Baseball-DataBank.org*. From the web site, just download the file labeled "Database in MySQL form" and save it to your local disk. (You can download the 2004 database file from *http://www.baseball-databank.org/files/BDB-sql-2005-08-02.sql.zip*. Or, you can check for a more current version and use it instead. Just be sure to change the filename in the instructions I'm outlining here.) This database is produced by a volunteer effort led by Sean Forman and Peter Kreutzer.

If you are using Mac OS X, Linux, or another system with standard Unix commands installed, download the file using the following command:

```
% curl http://www.baseball-databank.org/files/BDB-sql-2005-08-02.sql.zip\
> BDB-sql-2005-08-02.sql.zip
  % Total    % Received % Xferd  Average Speed   Time    Time     Time
Current
                                 Dload  Upload   Total   Spent    Left
Speed
100 6495k  100 6495k    0     0   587k      0  0:00:11  0:00:11 --:--:--
623k
```

Step 2: Decompress the File

The file is distributed as a single zipped file. You need to decompress this file to use it, just like the Access version described earlier. On Windows, you can double-click the file and your compression program should allow you to open the archive and extract the files. On Mac OS X, I decompressed this to the folder using this command:

```
~ % unzip BDB-sql-2005-08-02.sql.zip
```

Step 3: Create the Database

You need to create a database in MySQL before you can import the files. You can do this in two steps. First, give yourself permission to access the database. Then run the command to create the database. Here are the commands I used to create this database and the responses from MySQL:

```
 1  ~ % mysql -p -u root
 2  Enter password:
 3  Welcome to the MySQL monitor.  Commands end with ; or \g.
 4  Your MySQL connection id is 22 to server version: 4.0.21-standard
 5
 6  Type 'help;' or '\h' for help. Type '\c' to clear the buffer.
 7
 8  mysql> GRANT ALL ON bbdatabank.* TO 'jadler'@'localhost' IDENTIFIED BY
    'P@ssw0rd';
 9  Query OK, 0 rows affected (0.07 sec)
10
11  mysql> CREATE DATABASE bbdatabank;
12  Query OK, 1 row affected (0.00 sec)
13
14  mysql> quit
15  Bye
```

In line 8, notice that I created a database called *bbdatabank* and granted access to a user named jadler with the password P@ssw0rd. That's my username and an example password. You should change this to your username and pick a password that you like.

You're welcome to change the name of the database, but you will find that most hacks in this book refer to this database as *bbdatabank*. If you pick another name, make sure you modify my examples to reflect the name of your database.

You can also create a database using a GUI-driven tool like MySQL Administrator. See "Get a GUI for MySQL" **[Hack #18]** for instructions on where to get such programs.

Step 4: Import the Database

The file I unzipped was called *BDB-sql-2005-08-02.sql*. You can import this in a single step using this command (you will need to use the same username and password you used in the previous step):

```
~ % mysql -u jadler -p -s bbdatabank < BDB-sql-2004-12-02.sql
Enter password:P@ssw0rd
```

The < sign means "read the filename to the right and send it to the program to the left."

The -s option stops MySQL from providing feedback. (Trust me. You don't want feedback. The script manually inserts thousands and thousands of lines. If you omit this, MySQL prints "Query OK, 1 row affected (0.00 sec)" to the screen after each record is inserted, which means that MySQL prints this message a few hundred thousand times.) This command will finish within a couple of minutes on a typical computer.

Step 5: Check That Everything Is There

Now, let's check that the database has loaded. Start the MySQL program and type show tables;. You should see something like this:

```
mysql> show tables;
+-----------------------+
| Tables_in_bbdatabank2 |
+-----------------------+
| Allstar               |
| AwardsVotes           |
| AwardsWinners         |
| Batting               |
| Fielding              |
| FieldingOF            |
| Managers              |
| ManagersHalf          |
| Master                |
| Pitching              |
| Salaries              |
| Teams                 |
| TeamsFranchises       |
| TeamsHalf             |
| Transactions          |
+-----------------------+
15 rows in set (0.01 sec)
```

If you want to check more carefully that everything is there, you can count the number of rows in each table. The Baseball DataBank web site includes information on the number of rows in each table. You can find this at *http://www.baseball-databank.org/files/tables.txt*. To check the number of rows in each table, you can use a SQL query to count the number of rows:

```
mysql> select count(*) from batting;
+----------+
| count(*) |
+----------+
|    85978 |
+----------+
1 row in set (0.03 sec)
```

The Contents of the Database

The current version of the Baseball DataBank database contains data through 2005; you can download it from *http://www.baseball-databank.org/files/BDB-sql-2005-12-30.sql.zip*. Currently, the Baseball DataBank database contains over 20 tables. The structure of this database is almost identical to the structure of the Baseball Archive database (see "Get an Access Database of Player and Team Statistics" **[Hack #9]** for more information).

While I was writing this book, the structure of the database changed twice. For this book, I used the August 2005 version, which used a slightly different database structure than the current version. (The August 2005 database was normalized to remove duplicate keys and better utilize the features of a relational database.) You can download a SQL script to convert the current Baseball DataBank database to the correct format, plus a copy of the database that works with the code in this book, at *http://www.oreilly.com/catalog/baseballhks*.

Here is a short description of the database structure used in this book:

`Master`
: This table contains biographical information about each player, including their full name, birth date, country of origin, and batting and throwing hands. Each player has a unique ID (called `idxLahman`) that is referenced from other tables in the database.

`Batting`
: This table contains batting statistics for each player, on each team, during each season. Rows are uniquely identified by `idxBatting`, or by `idxLahman`, `idxTeams`, and `stint`.

`Fielding`
: This table contains fielding statistics for each player, on each team, during each season. Rows are uniquely identified by `idxFielding`, or by `idxLahman`, `idxTeams`, and `stint`.

`FieldingOF`
: This table tells you how much time outfielders spent in each fielding position. It is referenced by `idxFielding`.

`Pitching`

This table contains pitching statistics for each player, on each team, during each season. Rows are uniquely identified by `idxPitching`, or by `idxLahman`, `idxTeams`, and `stint`.

`Teams`

This table contains information on each team for each season, including aggregate batting statistics, pitching statistics, the team record, and postseason performance. Each line is uniquely referenced by a field called `idxTeams`. (You need to join the `batting`, `fielding`, or `pitching` tables with the `Teams` table to determine the year.)

`TeamsHalf`

This table shows win–loss records for each team, midway through each season.

`TeamsFranchises`

This table includes the full name of each team, indexed by the `idxTeamsFranchises` field.

`Allstar`, `AwardsVotes`, `AwardsWinners`, `Managers`, `ManagersHalf`, `Salaries`, *and* `Transactions`

These tables contain what you would expect; I don't use any of these in this book, so I'm not going to describe them in depth.

The Baseball Archive database contains a similar set of tables but with slightly different indexes. Here are the key differences in how the tables are indexed:

`Master`

The unique key is called `playerID`.

`Batting`, `Pitching`, *and* `Fielding`

Rows are uniquely identified by `playerID`, `teamID`, `yearID`, and `stint`.

`Teams`

Rows are uniquely identified by `teamID` and `yearID`.

Hacking the Hack

Here are a few tweaks to help you get the most from the career databases.

Annual updates. These databases are updated annually with new statistics. These web sites might offer files containing incremental updates (new players and annual statistics). If not, just create a new database with the year name (for example, *bbdatabank2004*), download the entire new version, and import it into this new database. (Replace `bbdatabank` in the previous commands with bbdatabank2004.) This way, you'll be able to keep any work you did with the old tables.

Getting baseball statistics as text files. Finally, you can download baseball statistics as (comma-separated) text files. This is a good choice if you plan to use another database or a statistical analysis program such as R **[Hack #34]**. You can download the current version of the text files from *http://baseball1.com/statistics*.

A History of the Player and Team Databases

The Baseball Archive and Baseball DataBank databases have a long history. As you might imagine, it took a lot of work to assemble a database of player statistics going back to 1871. The story of the free databases begins with Pete Palmer. During the 1970s and 1980s, Pete compiled paper records for baseball games into a computer database and used this database to produce the statistics in the book *Total Baseball* (Total Sports). CMC Corporation of Portland, Oregon, released a CD-ROM containing all the statistics from the 1993 edition of *Total Baseball*.

According to Pete Palmer, Sean Lahman took the data from this CD and used it to build the files available at *http://www.baseball1.com*. Over time, Sean and other people have worked to keep this database up-to-date, and they have supplemented the original database with tables describing awards, all-stars, and other information. Somewhat amazingly, Lahman does not allow his database to be redistributed by others. (If you intend to share data with others, I recommend that you use the Baseball DataBank data instead because there are restrictions on redistributing Lahman's information.)

The same files are used at *http://www.baseballreference.com*, *http://www.espn.com*, and even *http://www.mlb.com*. It's a shame that Pete Palmer is not more widely acknowledged for his work. However, these databases are a great resource for baseball fans, and I'm glad this information is freely available.