

Moneyball with Pig and Hive  
Russ Lankenau  
rlankenau@maprtech.com

The provided data contains various information about Major League Baseball between 1921 and 2011. The data is freely available on the web from [retrosheet.org](http://retrosheet.org)<sup>1</sup>. The majority of the data is in the form of event records, i.e. plays. Each play describes an event that occurs when a batter steps up to the plate. This could be an out, a hit, or some other event such as a runner stealing a base. In addition, there are other types of data such as team rosters, game information, player lists, team lists, and reconstructed data for games where no clear record exists.

info	Information about a single game (e.g. time of day, weather, location)
start	A starting player, including name, ID, position, home/away, and batting order
sub	A player substituted in during play
play	A single event during a game
com	A comment inserted by the recorder
id	The game ID, consisting of year, month, day, and game number for the day
version	The file version
data	Additional data about a player (e.g. number of errors during the game)

Table 1: Record Types

The records are comma-separated, and are intermixed within the event files. Each event file contains the home games for a single team for a single year. American League teams are listed in .EVA files, and National League teams are listed in .EVN files.

	Player ID	Player Name		Home(1) or Away(0)		Batting Order		Position
<i>start,</i>	<i>grudm001,</i>	<i>"MarkGrudzielanek",</i>		<i>0</i>	,	<i>1</i>	,	<i>6</i>

  

	Inning	Player ID	Count	Pitches		Play				
<i>play,</i>	<i>1,0</i>	<i>,aaro101,</i>	<i>12</i>	<i>,BF1CX,</i>	<i>S</i>	<i>7</i>	/	<i>L7S</i>	.	<i>3-H;1-2</i>
					Play Type	Fielder(s)		Description		Runner movement

Figure 1: Sample Records

The provided sample code<sup>2</sup> extracts one very basic metric, a count of plays for each player, grouped by ball location and result. Using the sample as a guide, can we extract other interesting data from this data set?

- What is the relationship between player names and IDs? (e.g. one-to-one?)
- Which player has the most career RBIs?
- Can we tell which players' performance drops off most in later innings?
- Which players are the most consistent between home games and away games?
- 
- 

<sup>1</sup><http://www.retrosheet.org/game.htm>

<sup>2</sup><https://github.com/rlankenau/mapr-moneyball>

<sup>3</sup>Rarely, a pitchout will result in a strike or foul ball. These are recorded as Q and R respectively.

<sup>4</sup>n is 1, 2, or 3, corresponding to base thrown to. The base number is preceded by a '+' if the pickoff was thrown by the pitcher.

Table 2: Play Codes