

Project 5: Working with Pig

DS 730

Overview

In this project, you will be working with Pig. You will be writing a Pig script to solve several problems. You are welcome to solve the problems using the local Pig system or by using AWS.

Project Tasks

We will be using three different files for this project. The first two are from the Pig Programming activity so you should know them well:

- **Batting.csv**
- **Master.csv**

The third file is **Fielding.csv** and deals with a player's fielding statistics for any given year. You can get all of these files from the online course.

You must write a Pig script to solve each of the following problems. You are welcome to create a Pig script for each problem or you can do all of them in one large Pig script. If there is a tie for any of the questions (e.g. Problem 3 may have multiple weights that are second most common), you should print out all of them.

Problem #	Description
1	Output the birth city of the player who had the most at bats (AB) in his career.
2	Output the top three birthdates that had the most players born. I am only looking for day and month combinations. For instance, how many were born on February 3 rd , how many were born on March 8 th , how many were born on July 20 th ... print out the top three dates.
3	Output the second most common weight.
4	Output the team that had the most errors in 2001.
5	Output the name of the player who had the most errors in all seasons combined.

Problem #	Description
6	<p>A player who hits well and doesn't commit a lot of errors is obviously a player you want on your team. Who were the top 3 players from 2005 through 2009 who maximized the following criterion:</p> $\left(\frac{\text{number of hits } (H)}{\text{number of at bats } (AB)} \right) - \left(\frac{\text{number of errors } (E)}{\text{number of games } (G)} \right) ?$ <p>The above equation might be skewed by a player who only had three at bats but got two hits. To account for that, <i>only consider players who had at least 40 at bats and played in at least 20 games</i>. Be aware that:</p> <ul style="list-style-type: none"> • Some players played for multiple teams during that 5-year span. • A player could have played multiple positions during that span. See, for instance, buchabr01 in 2004 who played LF, OF and RF in the same season. You may need a UDF for this problem.
7	<p>Sum up the number of doubles and triples for each birthCity/birthState combination. Output the top 5 birthCity/birthState combinations that produced the players who had the most doubles and triples.</p>
8	<p>Output the birthMonth/state combination that produced the worst players. The worst players are defined by the lowest of:</p> $\frac{\text{number of hits } (H)}{\text{number of at bats } (AB)}$ <p>To ensure one player does not skew the data, <i>make sure that at least three people came from the same state and were born in the same month</i>.</p> <p>For this problem, the year does not matter. For examples, a player born in December, 1970 in Detroit and a player born in December, 1982 in Detroit are in the same group because they were both born in December and are from Detroit.</p>

Submitting Your Work

When you are finished, create a file called **output.txt** and store your answers to each problem in that file. You should create a pig script for each question and store your code in those files so that they can be easily tested. Be sure you are reading in from the correct **/home/hduser/pigtest/*.csv** location in your code. Submit a .zip file containing these items to the **Project 5 dropbox**:

- Your code stored in Pig scripts
- The output file, which should be named **output.txt**