TI2736-B: Assignment 4
Big Data Processing

Due date: 09.12.2015 @ 11.59pm

Please submit your report and code via Blackboard (please do **not** copy & paste your code directly into the report). Make sure to include your name and student number in your report.

The exercises in this assignment are mostly about Pig & Pig Latin. If you use Cloudera's Hadoop distribution, Pig is already installed and you can start right away.

For the exercises in this assignment, you need two datasets which are available on Blackboard: **data_assignment4.zip**. You will find two data files in the zip archive (and one README): `Salaries.csv` and `Batting.csv`. Both are real-world datasets. The former contains data about the salaries of baseball players in America between 1871 to 2013. The latter contains various statistics about the players.

To start writing scripts, download the dataset(s) to a local directory within your virtual machine (if you use CDH). Open a terminal, move to the directory you stored the datasets in and type `pig -x local`. This opens the interactive shell (called Grunt) in which you can test and type out your Pig Latin scripts. Note that starting the shell from the same directory as your data is just for convenience, as it saves you typing effort when loading data from file. To avoid a possible reoccurring error in the Grunt shell, type as first command `set io.sort.mb 5;` (this setting ensures that Pig does not use too much memory for sorting).

1. [*Salaries.csv*] Write a Pig Latin script that outputs the names of all teams where in 1985, every player has a salary above 100000

2. [*Salaries.csv*] Write a Pig Latin script that outputs all the teams with the corresponding average salary in 1998.

3. [*Salaries.csv*] Write a Pig Latin script that outputs for each league the amount of teams in 1999.

4. [*Batting.csv*] Write a Pig Latin script that outputs the top 10 players with the most hits in 1988. Output the playerID, and the amount of hits they made.

5. [*Batting.csv*] Write a Pig Latin script that outputs the player that batted (G_batting, 7th column) most games in 1980. Output the playerID and the amount of games the player batted.

6. [*Batting.csv*] Write a Pig Latin script that outputs the player in the ML1 team with the most runs in 1960. Output the playerID and the amount of runs they made.

7. [*Salaries.csv* + *Batting.csv*] Write a Pig Latin script that outputs players with a salary above 500,000 in 2001 who have more than 50 homeruns. Output the players, their amount of homeruns and their salary.
   Secondly, write a plain Hadoop job (or a chain of Hadoop jobs) that outputs the same information. What can you say when comparing the runtime between the Pig Latin script and your Hadoop job (or job chain)?

8. [*Batting.csv*] Write an *EvalFunc* UDF that takes two parameters as input, the games played (G) and the number of homeruns (HR) by a player. The UDF returns the player's percentage of homeruns per game. Write a Pig Latin script that uses this UDF and outputs the top 10 players with the highest homerun/game percentage. Submit the script, the UDF source code (1 file) and your result (a single number).
   Note: you can only solve this exercise after the Monday lecture!