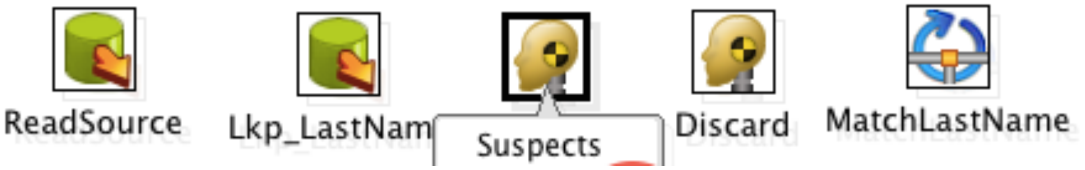
CS412 – Mini-MP2: Preprocessing Data

This mini-MP asks you to use Pentaho Kettle (Spoon) software.
Download (~800MB): http://community.pentaho.com/projects/data-integration/
Launch: http://wiki.pentaho.com/display/EAI/02.+Spoon+Introduction
To be able to use mysql database, you need to copy mysql-connector-java-5.1.32-bin.jar to the lib folder of your kettle installation folder:
http://dev.mysql.com/downloads/connector/j/

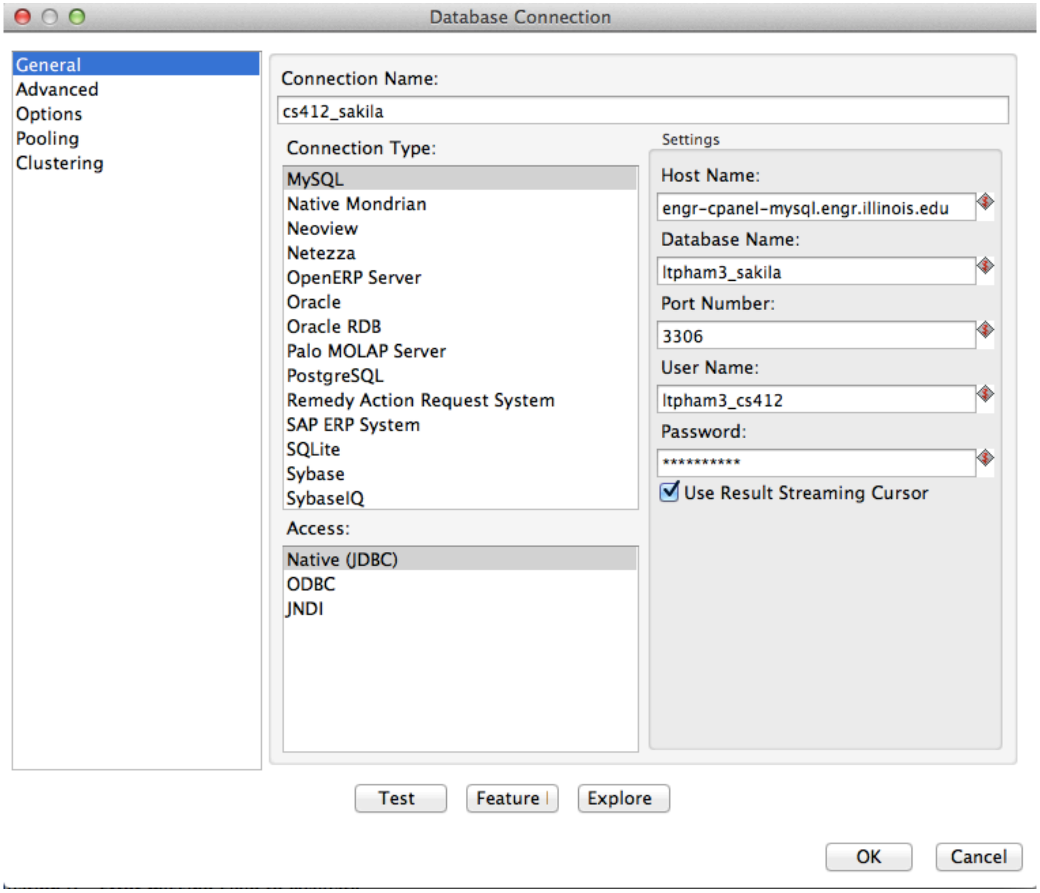
We will use Sakila sample database from mysql. We are particularly interested in the Customer table. We uploaded it to our database, so you can use the database online, which means you don't need to install mysql server.

Here, we want you to build the simplest workflow that output pairs of last names that look similar, but they currently belong to different customers, possibly because of mistakes of Sakila employees when inputting the data.

To get started, please open file cs412.ktr in Spoon. You can find the file in the homework folder. After opening it, you will see the following components:



- ReadSource: It is incomplete, you must specify Connection. You may create a new one like the picture below. Password is cs412kevin



Connection Name: cs412_sakila

Connection Type: MySQL

Access: Native (JDBC)

Host Name: engr-cpanel-mysql.engr.illinois.edu

Database Name: ltpham3_sakila

Port Number: 3306

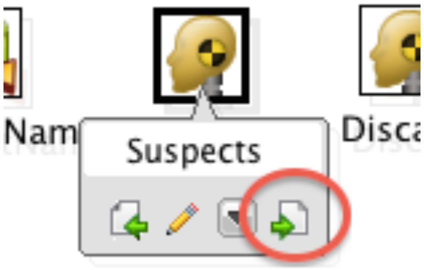
User Name: ltpham3_cs412

Password: **********

☑ Use Result Streaming Cursor

- Lkp_Lastname: It is also incomplete, you also need to specify Connection. You can reuse the created one.
- MatchLastName: fuzzy join lastnames in ReadSource with lastnames in Lkp_Lastname. You feel free to choose the algorithms, the thresholds, etc. **Please try with both "Get closer value" marked and unmarked, and report the difference.**
- Suspects: a component to contain the list of suspect pairs.
- Discard: a component to contain the list of remaining pairs.
- Please look for "Filter Rows" component in the design tab of the left lane., and drag it to the canvas. Rename it to "Select Suspects" and write the necessary filter so that the "true rows" will go to "Suspects" and the "false rows" will go to Discard.

Please organize all the components above into a workflow, and create arrows to connect the components. You may do that by clicking on the source component, then clicking on the tiny icon as shown below, and drag the arrow to the destination component.



Please report the picture of the final workflow and a screenshot of the first 10 suspect pairs, each of which contain 2 last names that look similar to each other.

Note: During the demonstration, I also play with data profiling. You may try it, but it is not mandatory:
http://wiki.pentaho.com/display/EAI/Kettle+Data+Profiling+with+DataCleaner

Hint: If you don't know what to do, you may watch our demonstration at the end of Friday Sep 5th lecture.

Have fun!