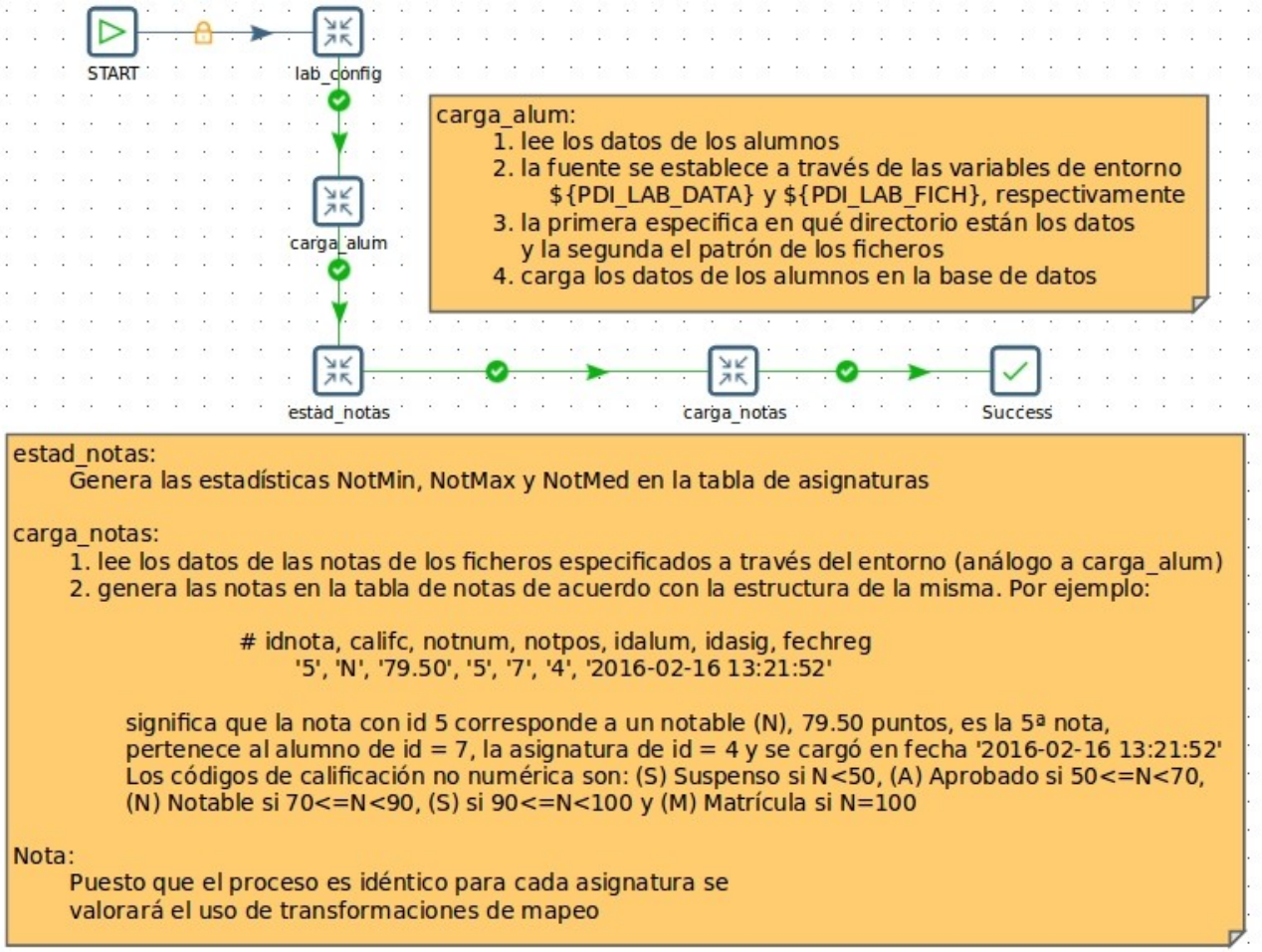
Procesos ETL con PDI

Este proyecto consiste en la implementación usando la herramienta *Spoon* de la suite *Pentaho* de un proceso ETL que simula la automatización del proceso de carga de las calificaciones de 4 asignaturas en una base de datos MySQL. Se suministra el siguiente material de apoyo:

1. un dataset ejemplo para usar como prueba durante el desarrollo de la ETL
2. el script DDL que crea la base de datos relacional destino de la carga
3. el trabajo principal ***main-notas.kjb*** que controla todo el proceso
4. una transformación ***lab_config.ktr*** que carga la configuración de un fichero ***lab.conf*** permitiendo de este modo parametrizar absolutamente el proceso
5. una transformación ***test_lab_config.ktr*** que permite probar la configuración previa

En la imagen adjunta se muestra la pantalla principal de la ETL (*main-notas.kjb*) en *Spoon*



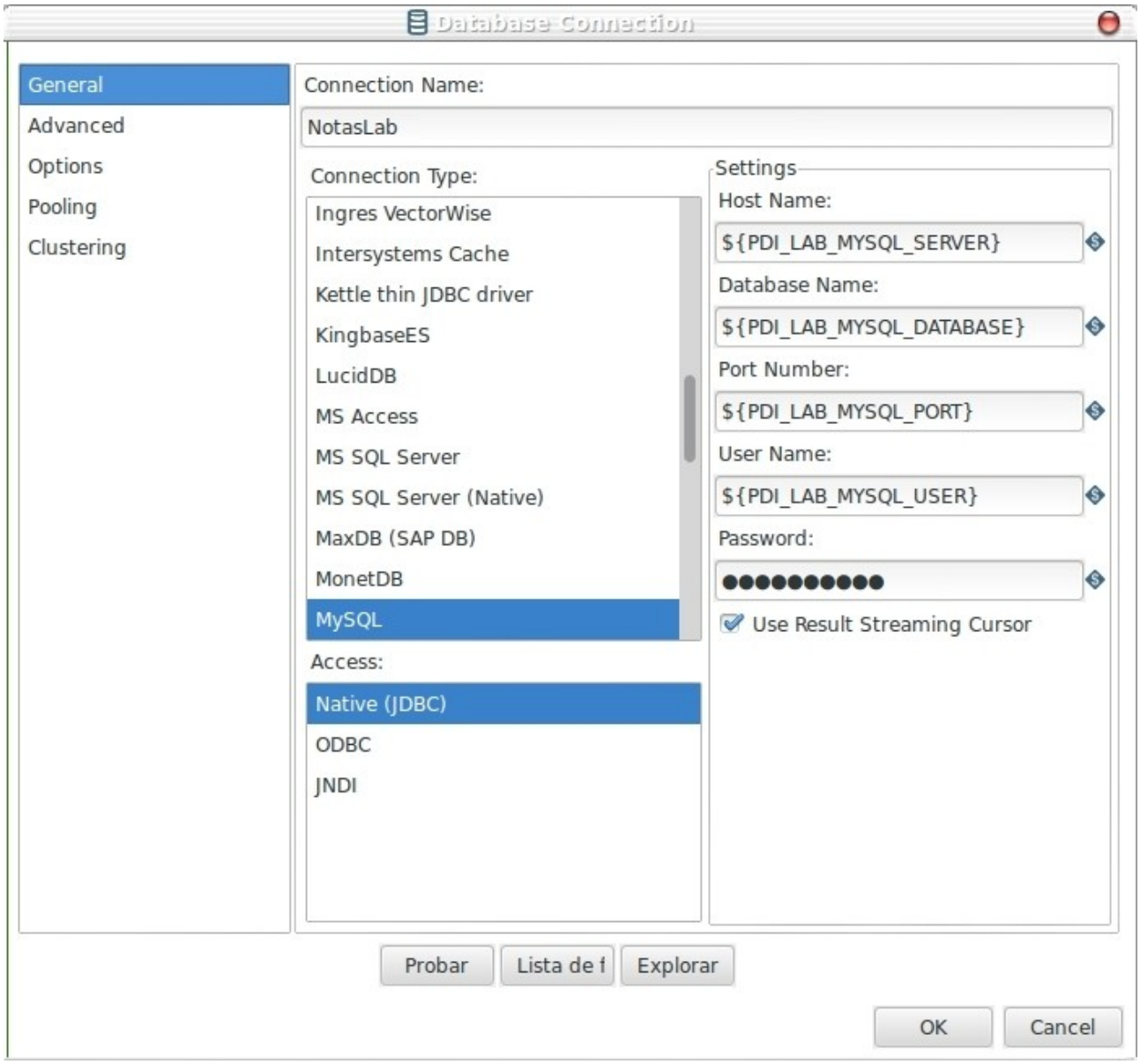
Se deberán implementar obligatoriamente las transformaciones ***carga_alum.ktr***, ***estad_notas.ktr*** y ***carga_notas.ktr*** que, respectivamente, cargan alumnos, calculan estadísticas de las asignaturas y cargan notas en la base de datos

En el primer caso, se debe resolver en una única transformación mientras que en los dos restantes se trata del mismo proceso para cada una de las cuatro asignaturas en el dataset. Por ende, se requiere utilizar transformaciones de mapeado. En cualquier caso, todas las transformaciones que se implementen deberán estar en la misma carpeta que las tres anteriores y se referenciarán dentro de Spoon de manera relativa. Por ejemplo:

${Internal.Transformation.Filename.Directory}*/el_nombre_que_sea.ktr*

Lógicamente, se entregarán todas las transformaciones que sean pertinentes para salvar las diferencias de estructura o de formato entre los datos crudos y lo que se almacena en la base de datos. Por ejemplo, en los datos de los alumnos el nombre y los apellidos están en un único campo mientras que en la base de datos se almacenan en dos campos separados

Otro aspecto fundamental es la definición en la herramienta *Spoon* de una conexión a base de datos compartida por todas las transformaciones. Esta conexión se denominará ***NotasLab*** y se definirá utilizando los parámetros cargados del fichero de configuración según se muestra en la figura:



Ojo porque en el campo password, aunque no se muestra, también se usa una variable de entorno

Se pide:

1. **Parte Obligatoria (85%)**
   a) Implementar la ETL que automatiza la carga de datos y el cálculo de las estadísticas
   b) Ejecutar el proceso contra el servidor de la asignatura utilizando la infraestructura dispuesta a tal efecto. En particular, los resultados de la ETL deberán almacenarse en la BD denominada G17*XXP1* donde G17*XX* son los 5 caracteres del código de equipo
   c) La ETL debería contemplar cargas sucesivas de datos en las que, como es obvio, el único requisito es que se respete el formato de los ficheros de entrada
2. **Parte Opcional (15%)**
   a) Diseñar un modelo multidimensional partiendo de la base de datos ***pdi_notas***
   b) Desplegar el diseño realizado en el servidor de la asignatura. En particular, los resultados deberán almacenarse en la BD denominada G17XXP2 donde G17XX son los 5 caracteres del código de equipo
   c) Para la realización de este ejercicio se recomienda fehacientemente tomar como referencia de diseño el **Modelo Multidimensional *Sakila*-*Star***. En la evaluación no se van a considerar actualizaciones de dimensiones Tipo II (*Slowly Changing*) pero, no obstante, deberá tenerse en cuenta en el diseño documentando convenientemente esta posibilidad. Es, por ello, que será preciso utilizar el paso de Búsqueda/Actualización de la categoría Almacén de Datos (Data Warehouse)