Recurso complementario

Ejemplo: automatización del flujo *ETL* mediante trabajos en Pentaho

En este primer ejemplo se revisarán los pasos básicos para crear un trabajo de *Kettle* que transfiera todos los datos de una tabla relacional, desde un motor de bases de datos hacia otro. En este caso, el paso de **ingesta de datos** (es decir, el paso de *extracción* del procesamiento *ETL*) simplemente leerá los datos desde las tablas de datos del archivo **FoodMart.mdb** en Microsoft Access y la parte de **carga de datos** los insertará en tablas relacionales dentro de MariaDB, sin efectuar transformaciones explícitas de datos. Para ello, se seguirán las etapas a continuación:

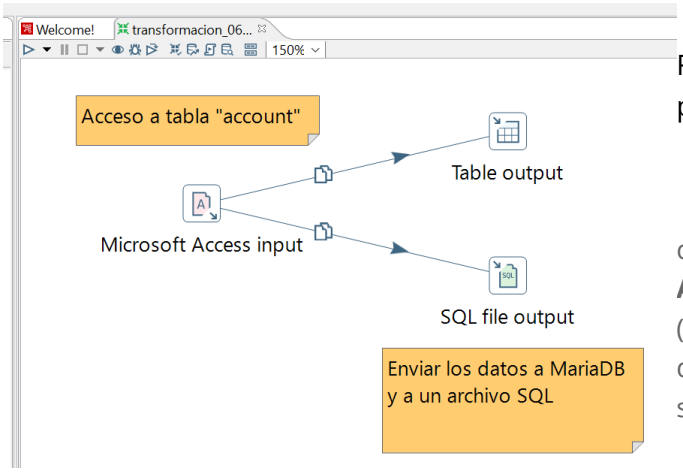
a) Crea una nueva transformación en *Kettle* (figura 1), con el nombre de **transformacion**_06a (el nombre no importa mucho, fuera de permitirle recordar cuál es el orden en que se aplicarán dentro del trabajo...).

b) Agrega los pasos indicados en la figura 1 y realiza las configuraciones indicadas más abajo. El paso MS Access input permitirá leer los datos desde una tabla (la tabla **account**) del archivo **FoodMart.mdb**; el segundo paso **Table output** te ayudará a crear una nueva tabla **account**, dentro de MariaDB, para enviar los datos del archivo; y, el tercer paso **SQL file output** creará un archivo de script de SQL en un directorio de disco con sentencias *insert* de SQL para contar con un respaldo que se pueda ejecutar en cualquier otro motor de base de datos para regenerar los datos fácilmente.

NOTA: cuando crees ambos saltos (*hops* o flechas de flujo), aparecerá un cuadro de diálogo preguntando si deseas **distribuir** o **copiar** los datos. Elige **copiar**, ya que eso permitirá enviar todos los datos hacia ambos pasos; si eliges **distribuir**, una mitad de los datos irá al paso **Table output**, mientras que la otra mitad irá al paso **SQL file output**, evitando que ambos pasos obtengan todos los datos de la tabla relacional, como se pretendía inicialmente.

Figura 1.

Diagrama de la transformación 06a a crear

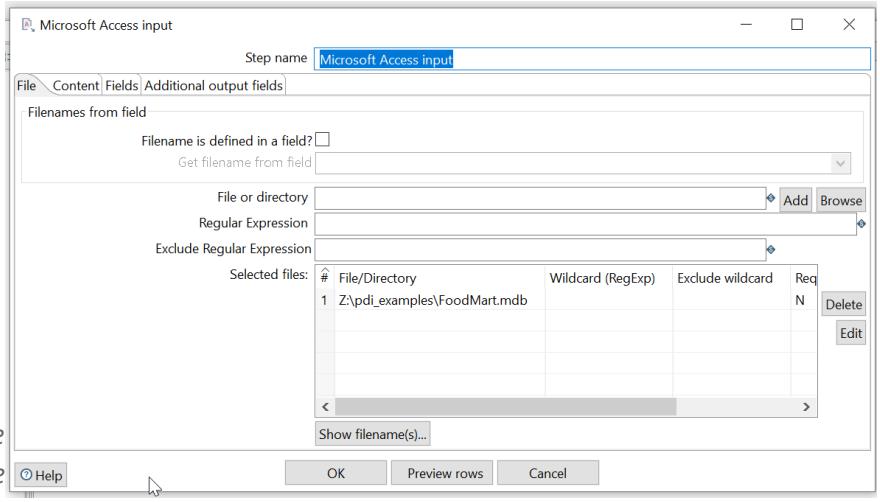


Fuente: elaboración propia.

c) En el paso **Microsoft Access input** configura (figuras 2, 3 y 4) la ruta de archivo desde donde se leerán los datos (archivo **FoodMart.mdb**); el nombre de la tabla desde donde se leerán los datos y los campos a leer.

Figura 2.

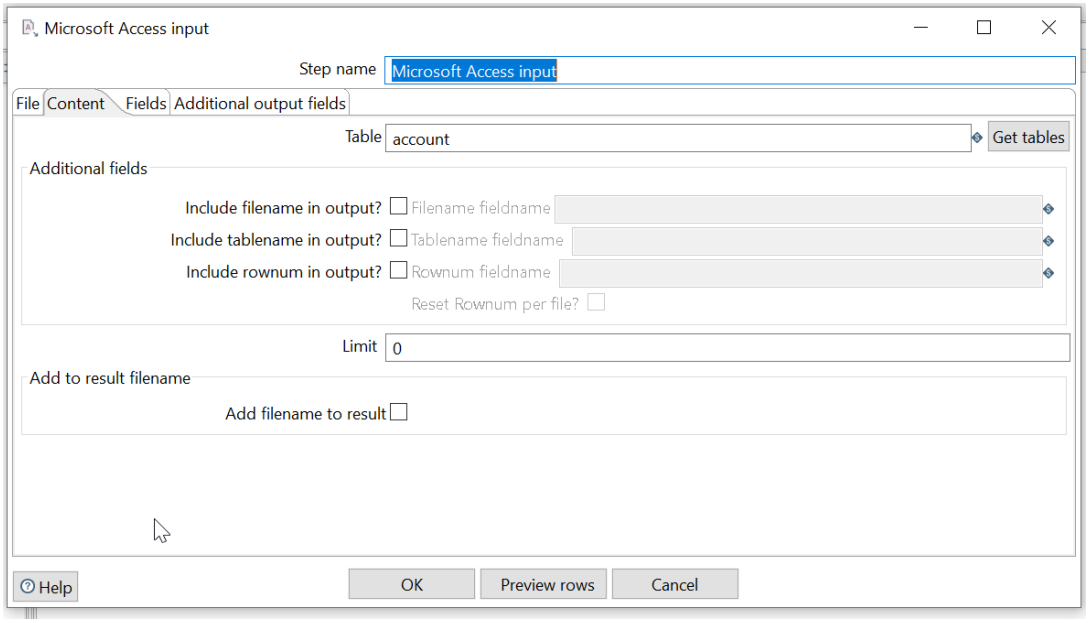
Configuración del paso Microsoft Access input



Fuente: elaboración propia.

Figura 3.

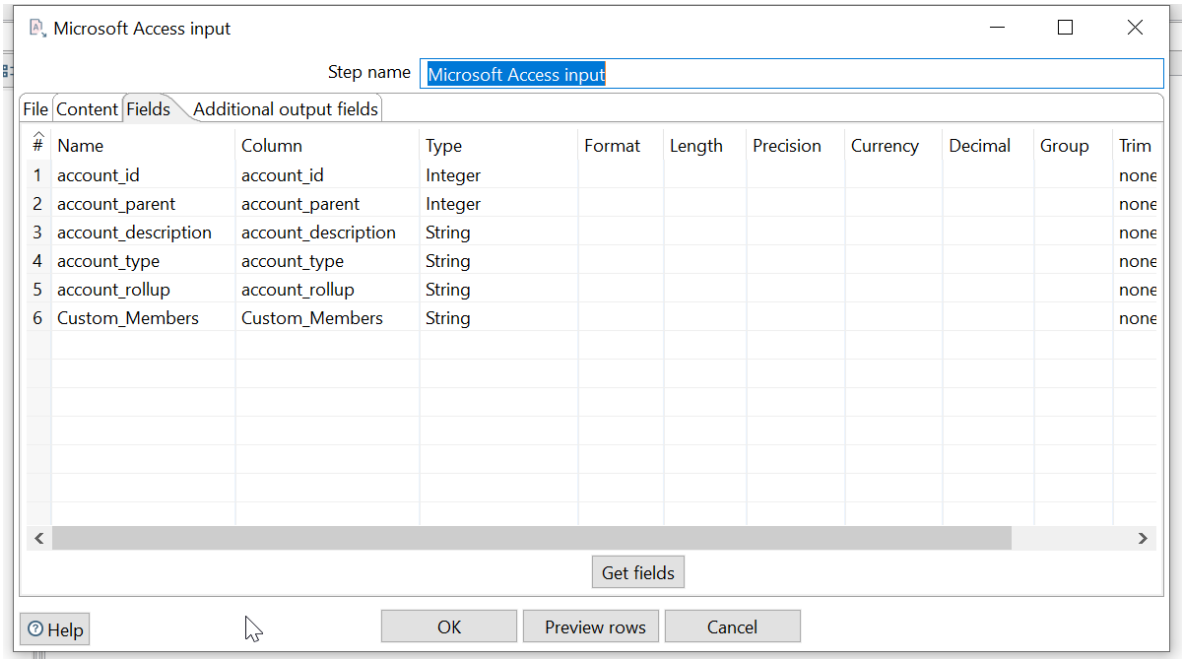
Configuración del nombre de tabla de base de datos a leer

Fuente: 

elaboración propia.

Figura 4.

Configuración de columnas a leer desde la base de datos

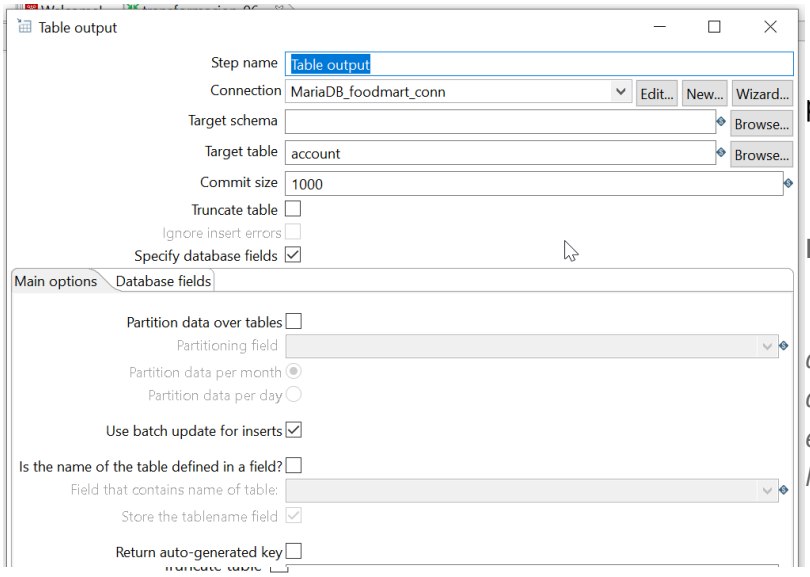


Fuente: elaboración propia.

d) En el paso **Table output** (figura 5) especifica la conexión de base de datos de *Kettle* a utilizar para enviar los datos hacia MariaDB; indica la tabla de base de datos **account** a rellenar con datos; agrega los campos de base de datos a leer/escribir (figura 6) y, además, presiona el botón **SQL**, en la parte inferior, para crear la tabla dentro de MariaDB (figura 7).

Figura 5.

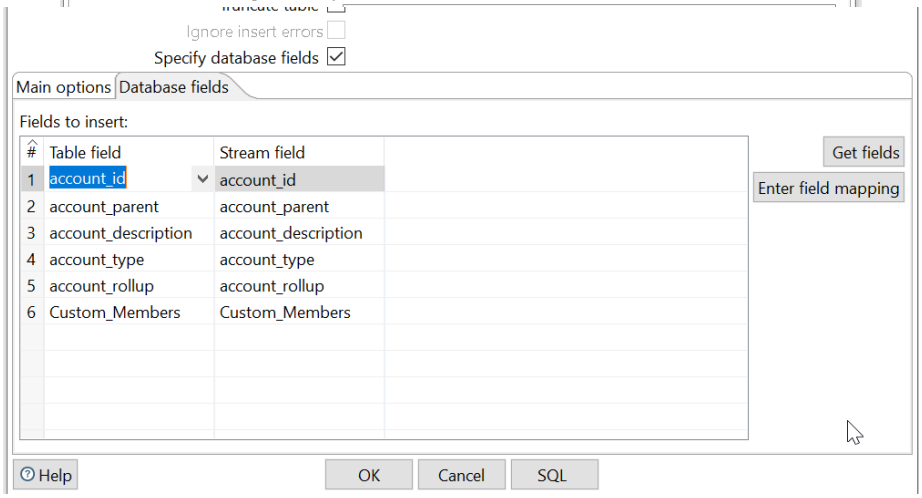
Configuración del paso Table output



Fuente: elaboración propia.

Figura 6.

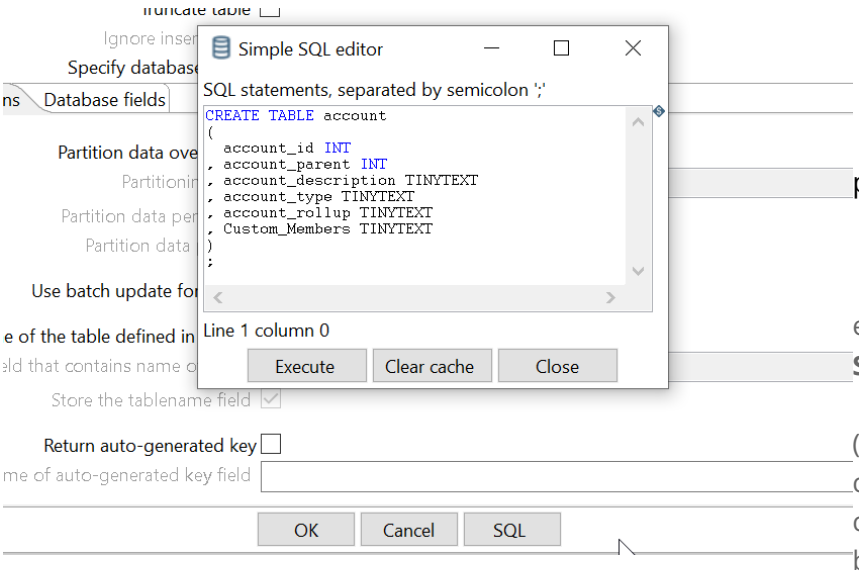
Especificación de las columnas de datos a escribir en MariaDB



Fuente: elaboración propia.

Figura 7.

*Al presionar el botón SQL, aparecerá el script de SQL para crear la tabla en MariaDB. Presiona el botón **Execute** para crear la tabla y asegurarse de que esté presente al ejecutar la transformación*

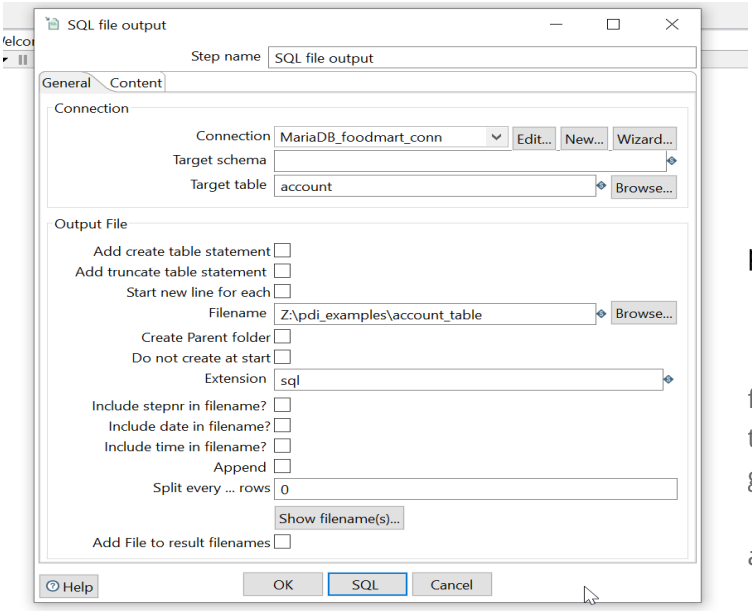


Fuente: elaboración propia.

e) En el paso **SQL file output** (figura 8) configura la conexión de base de datos de *Kettle* (que se creó al inicio de la semana), indica la tabla relacional que se usará en las sentencias SQL de salida e indique dónde se almacenará el script SQL de salida a crear.

Figura 8.

Configuración del paso SQL file output



Fuente: elaboración propia.

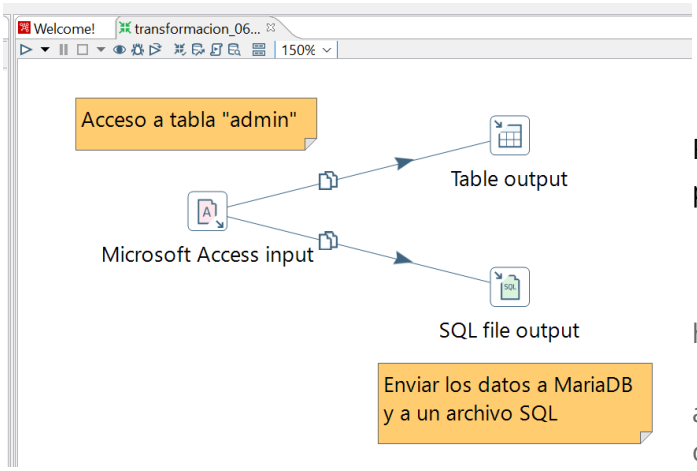
f) Ya teniendo todo configurado, guarda la transformación anterior y ejecútala.

Visualiza los datos en MariaDB (a través de HeidiSQL) y revisa si se copiaron de forma íntegra (todos con los mismos campos de datos).

g) Toma la transformación anterior y guárdala con otro nombre (**transformacion_06b**), ya que esta lidiará con el traspaso de datos desde la tabla **admin** hacia MariaDB (figura 9). En general, repite las mismas configuraciones de la primera tabla (instrucciones (b) hasta (e)) y revisa que los datos se copien de forma íntegra.

Figura 9.

Esquema de la transformación 06b

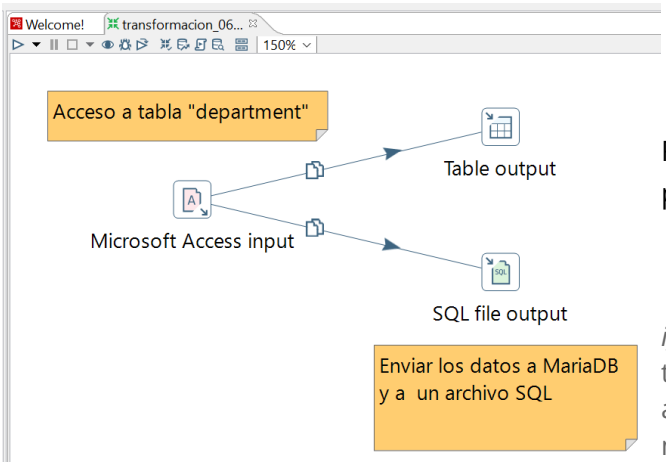


Fuente: elaboración propia.

h) Toma la transformación anterior y guárdala con otro nombre (**transformacion_06c**), ya que esta lidiará ahora con el traspaso de la tercera tabla de datos **department** hacia MariaDB (figura 10). En general, repite las mismas configuraciones de la primera y segunda tablas (instrucciones (b) hasta (e)) y revisa que los datos se copien de forma íntegra.

Figura 10.

Esquema de la transformación 06c

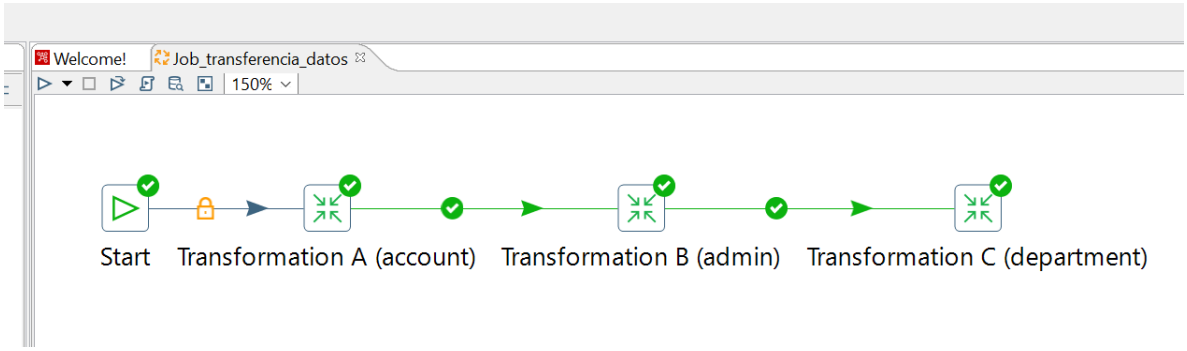


Fuente: elaboración propia.

i) Ahora, guarda las tres transformaciones anteriores. Crea un nuevo trabajo en *Spoon*, utilizando el menú **File -> New -> Job**, y guárdalo con algún nombre descriptivo. Agrega, desde el panel izquierdo, categoría **General**, un paso **Start** y tres pasos **Transformation**. Finálmente, únelos, según lo indicado en la figura 11.

Figura 11.

Esquema del trabajo a realizar en Spoon



Fuente: elaboración propia.

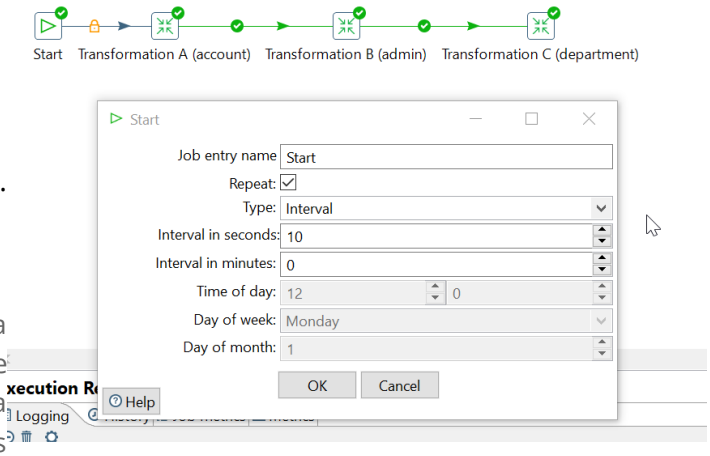
j) Haz doble clic en cada paso de transformación e indica la ruta del nombre de archivo para cada transformación creada en los pasos anteriores del ejercicio (transformaciones 06a, 06b y 06c). Guarda el trabajo y ejecútalo con el botón **Run** general. Verás que el trabajo ejecuta los tres pasos (es decir, las tres transformaciones) en secuencia; si no hay errores en ningún paso, al ir a HeidiSQL, podrás ver todos los datos ya traspasados automáticamente a la base de datos de MariaDB.

k) **ADICIONAL**: si haces doble clic en el paso **Start**, se abrirá un cuadro de diálogo como el de la figura 12. Si se observa, podrías seleccionar el cuadro **Repeat**, mediante el cual se puede lograr

que este trabajo se repita automáticamente cada cierto tiempo. Si configuras luego **Type** con el valor **Interval** y estableces el tiempo de repetición cada 10 segundos, al guardar el trabajo y presionar el botón **Run**, verás cómo el trabajo se realiza cada 10 segundos automáticamente, acumulando una gran cantidad de datos en las tablas de MariaDB (datos repetidos, eso sí).

Figura 12.

Configuración del trabajo de Kettle para que se ejecute cada 10 segundos, de forma ininterrumpida



Fuente: elaboración propia.

Desactiva la configuración de repetición y limpia las tablas relacionales en MariaDB. En todo caso, si lo deseas, prueba otros modos de repetición. Igualmente, considera que para que la repetición se realice automáticamente en ambientes de producción, se requerirá ejecutar Pentaho en modo *batch*, o bien, instalar el servidor Web Pentaho Server para manejar estas tareas de forma autónoma.