ANEXO II: EJEMPLOS

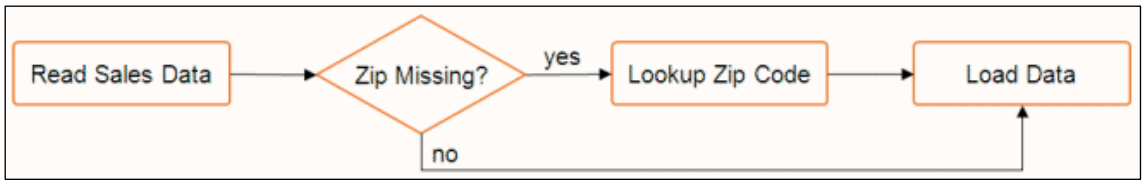
1. Mis primeras transformaciones

Como ya se ha visto, la perspectiva de Data Integration de Spoon permite crear dos tipos básicos de archivos de diseño: transformaciones y trabajos.

- Las transformaciones se utilizan para describir los flujos de datos para ETL, como leer desde una fuente, transformar datos y cargarlos en una ubicación de destino.
- Los trabajos se utilizan para coordinar actividades de ETL, como definir el flujo y las dependencias, para qué orden se deben ejecutar las transformaciones o prepararse para la ejecución mediante la comprobación de condiciones tales como las siguientes: "¿está mi archivo fuente disponible?" o "¿existe una tabla en mi base de datos?".

Este ejercicio guiará al alumno a través de la construcción de su primera transformación con PDI, introduciendo conceptos comunes a lo largo del camino.

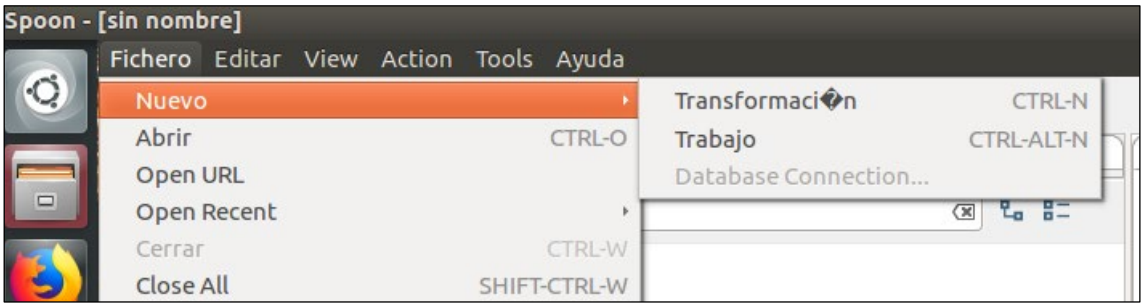
El escenario del ejercicio incluye un archivo plano de datos de ventas (**Fichero_Ejemplo_Entrada.xlsx**), que cargará en una base de datos para que se puedan generar listas de correo. A varios de los registros del cliente les faltan códigos postales que deben resolverse antes de cargarlos en la base de datos. La lógica se ve así:

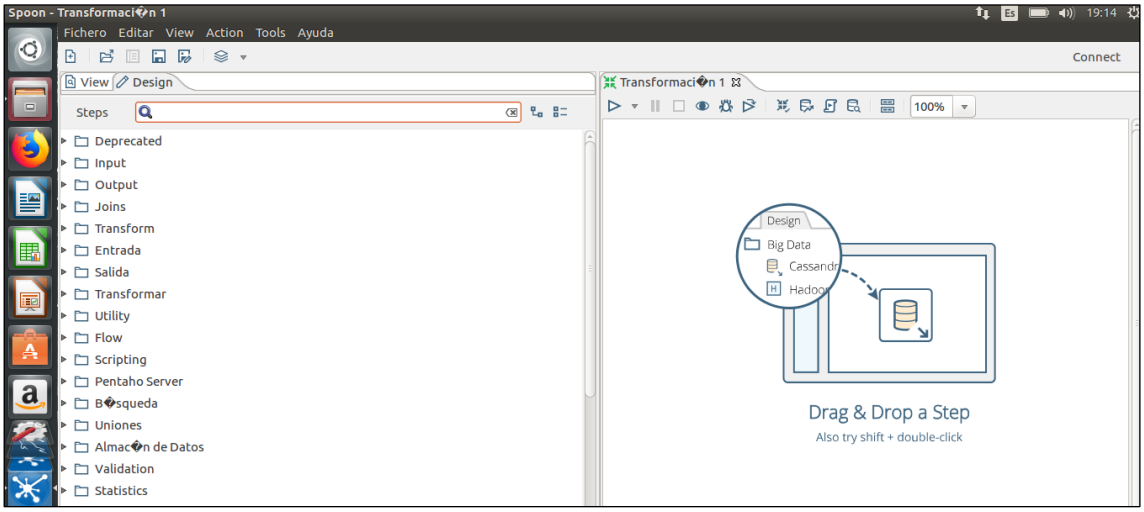


1.1. Recuperar los datos de un archivo plano

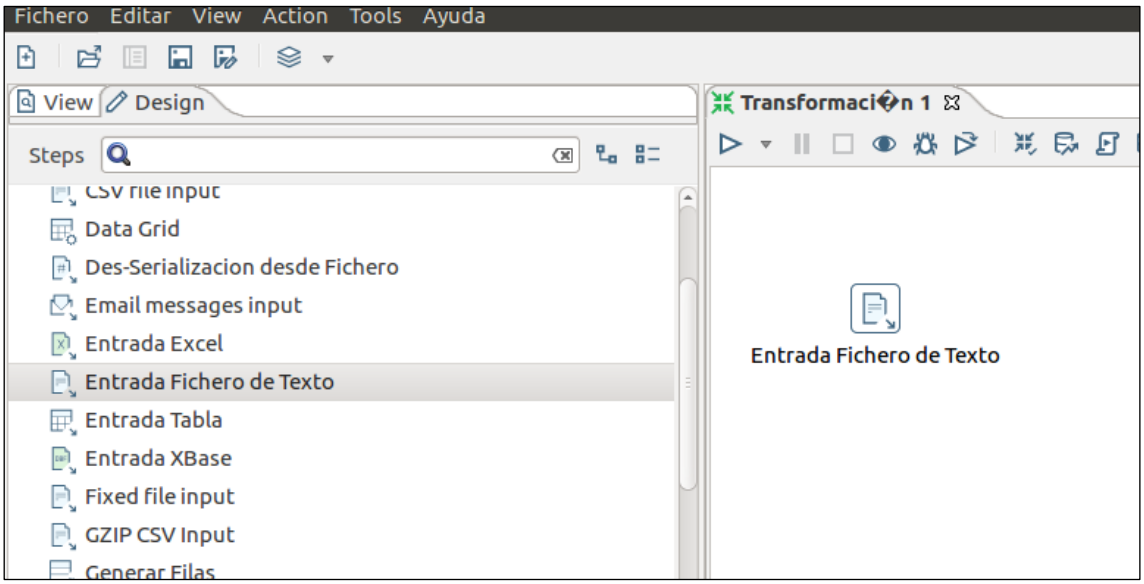
Han de seguirse las instrucciones que se facilitan a continuación para recuperar los datos de un archivo plano:

1) Seleccionar *Fichero > Nuevo > Transformación*, en la esquina superior izquierda de la ventana Spoon, para crear una nueva transformación:

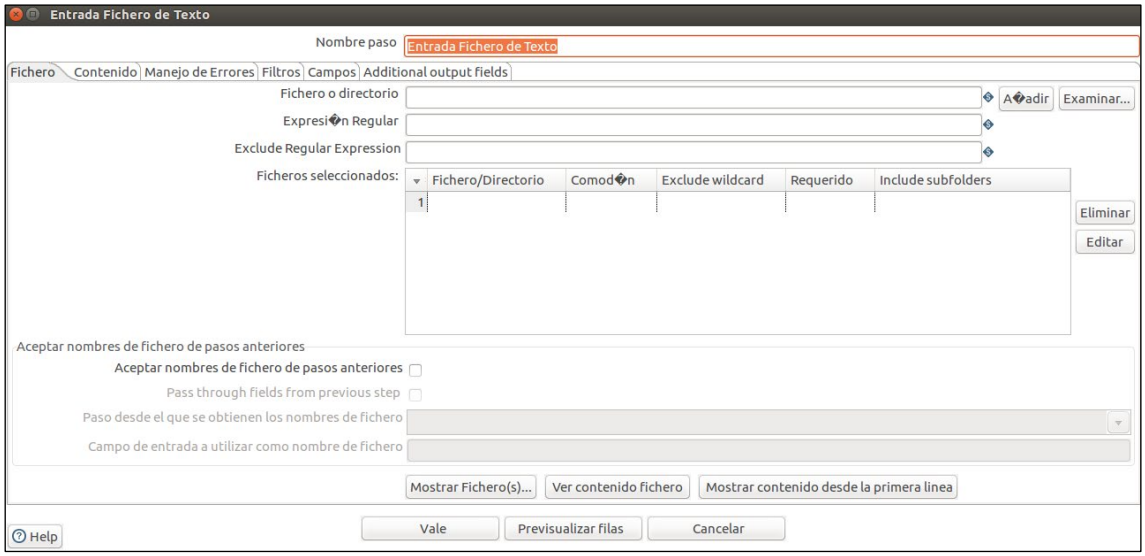




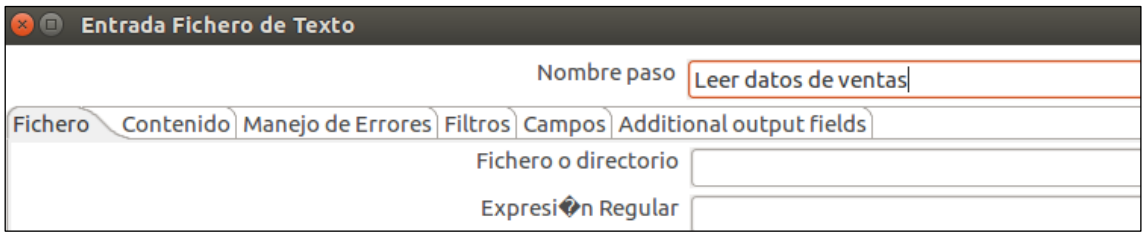
2) Debajo de la pestaña *Diseño* (*Design*), hay que expandir el nodo de entrada; luego, se selecciona y arrastra un paso de ingreso de *Entrada Fichero de Texto* (*Text File Input*) al lienzo.



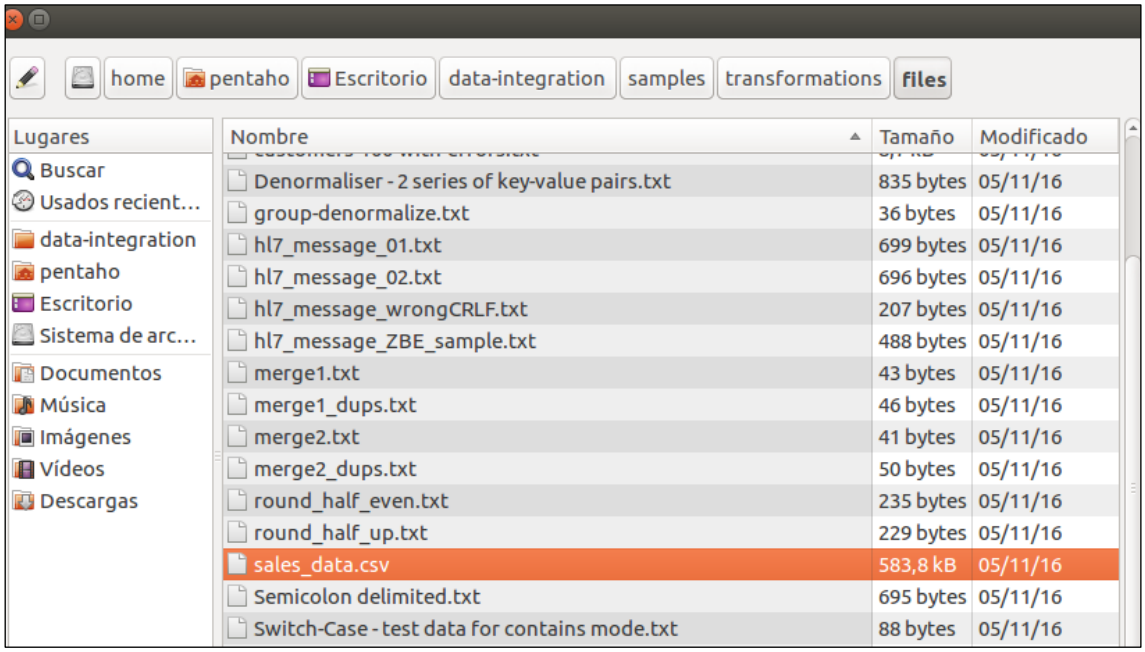
3) Se hace doble clic en el paso *Entrada Fichero de Texto*. La ventana de entrada del archivo de texto aparece. Esta ventana permite establecer las propiedades de este paso.



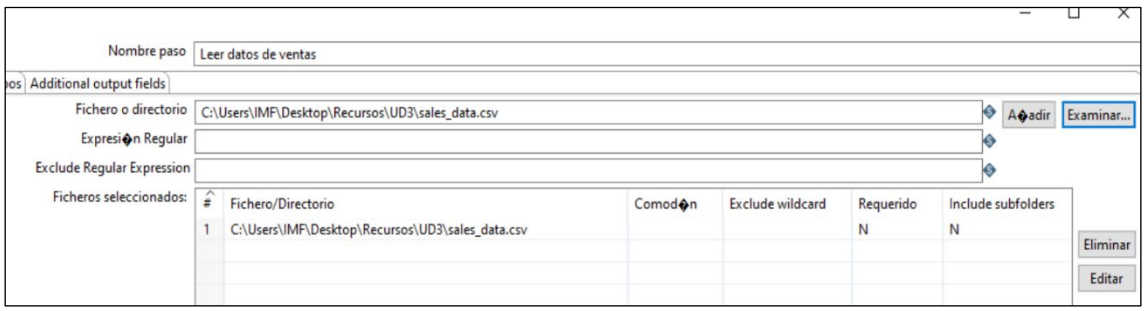
4) En el campo *Nombre paso*, hay que escribir “Leer datos de ventas”. Esto cambia el nombre del paso del fichero de texto a “Leer datos de ventas”.



5) Se hace clic en *Examinar…* para buscar el archivo de origen: "**sales_data.csv**".



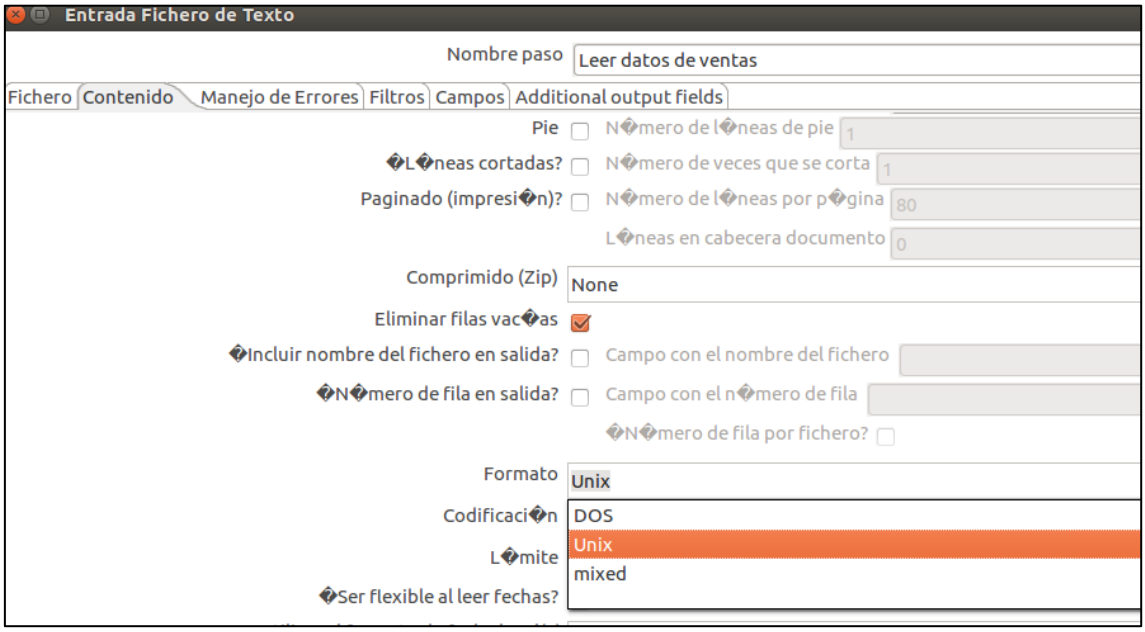
El botón *Examinar…* aparece cerca de la esquina superior derecha de la ventana, cerca del campo *Fichero o directorio*. Hay que hacer clic en *Aceptar*. La ruta al archivo fuente aparece en el campo *Fichero o directorio*.



6) Hay que hacer clic en *Añadir*. La ruta del archivo aparece en *Ficheros seleccionados*.

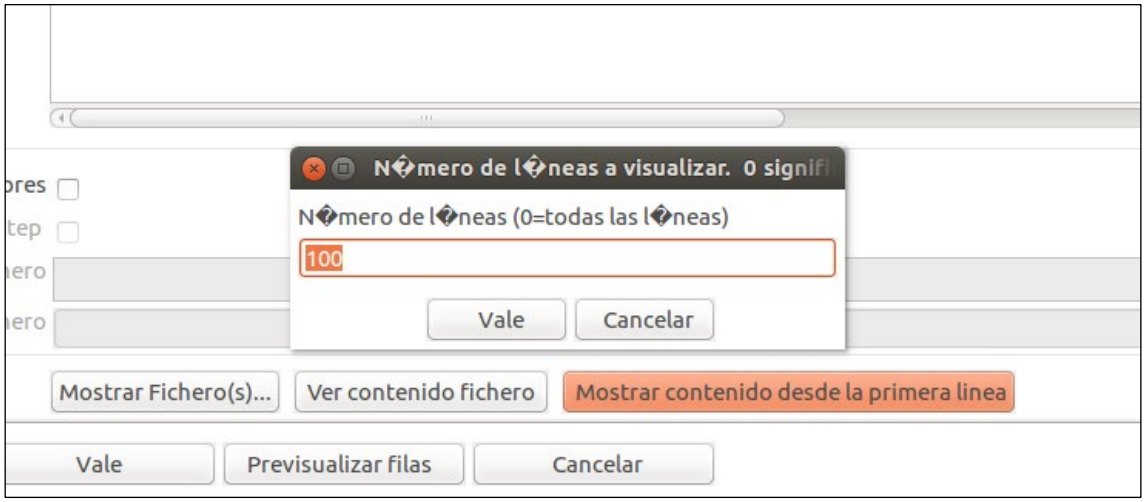
7) Para ver el contenido del archivo de muestra, hay que realizar los siguientes pasos:

- Hacer clic en la pestaña *Contenido* y, a continuación, configurar el campo *Formato*: **Unix**.



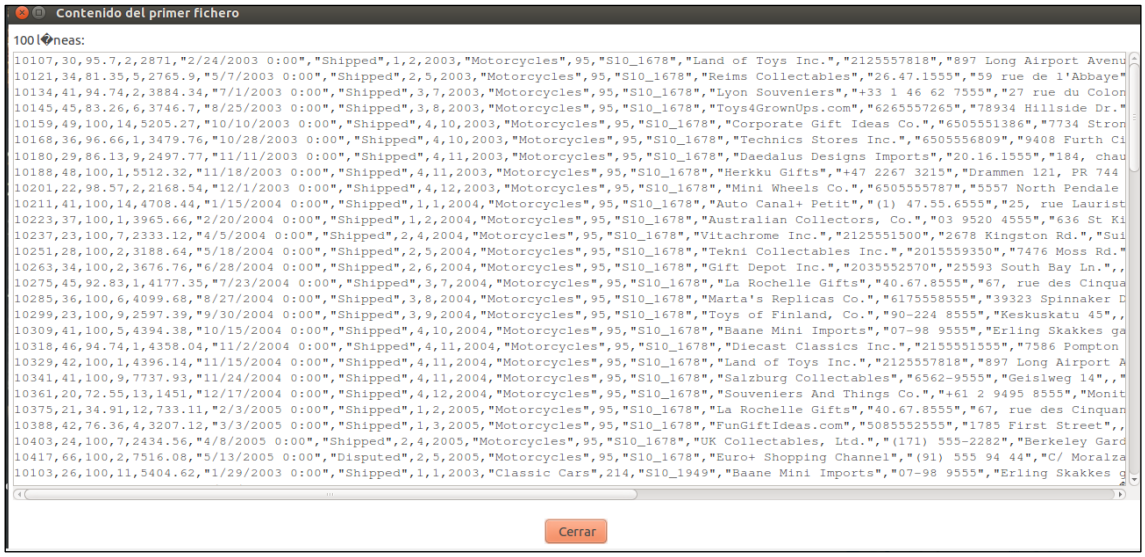
- Se hace clic en la pestaña *Fichero* y se clica también en *Mostrar contenido desde la primera línea*, que aparece en la parte inferior de la ventana.

- Aparecerá el mensaje *Número de líneas a visualizar* (*Nr. of lines to view*). Hay que hacer clic en el botón *Vale* para admitir el valor predeterminado.



- El contenido de la primera ventana del archivo muestra el archivo. Se debe examinar ese archivo de entrada para ver cómo está delimitado, qué carácter delimitador de texto se utiliza y si hay una fila de encabezado presente o no. En el ejemplo, el archivo de entrada está delimitado por comas (,), el delimitador de texto es unas comillas (“) y contiene una sola fila de encabezado que encierra nombres de campo.

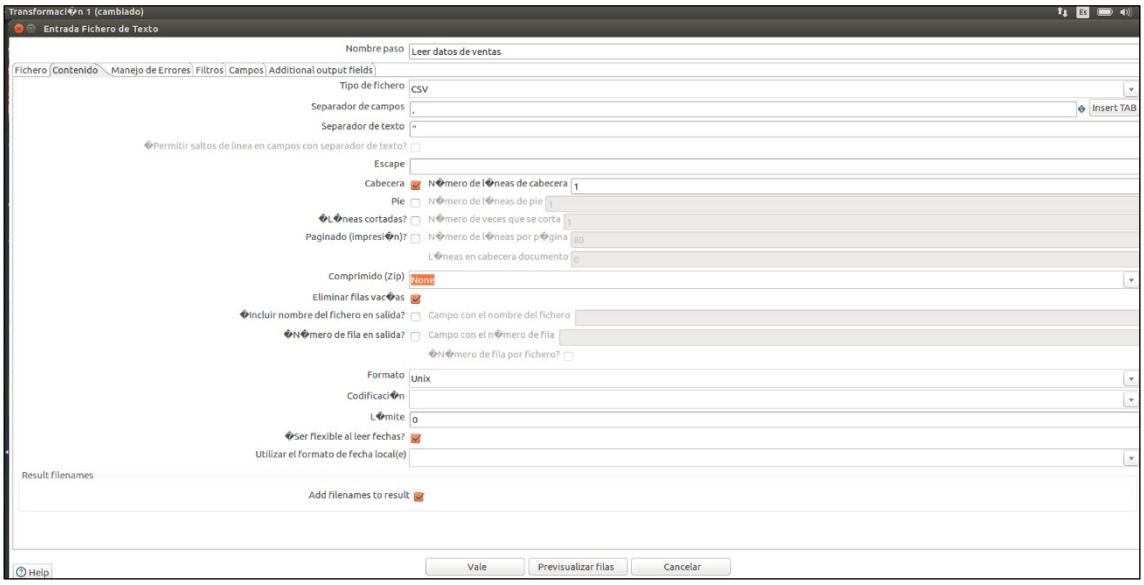
- Se hace clic en el botón *Cerrar* para cerrar la ventana.



8) Para proporcionar información sobre el contenido, hay que realizar los siguientes pasos:

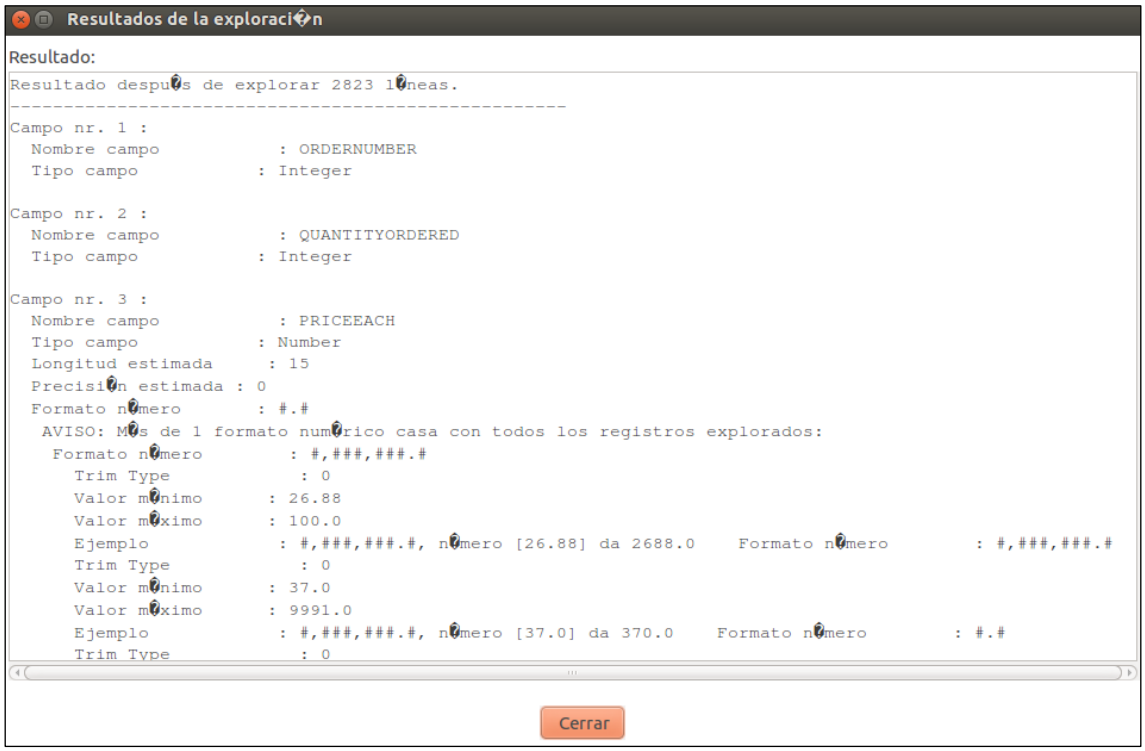
- Hacer clic en la pestaña *Contenido*. Los campos que aparecen debajo de la pestaña *Contenido* permiten definir cómo se formatea la información.

- Hay que asegurarse de que el separador está configurado en coma (,) y de que el delimitador esté configurado con comillas (“). También hay que habilitar el encabezado: hay una línea de filas de encabezado en el archivo.

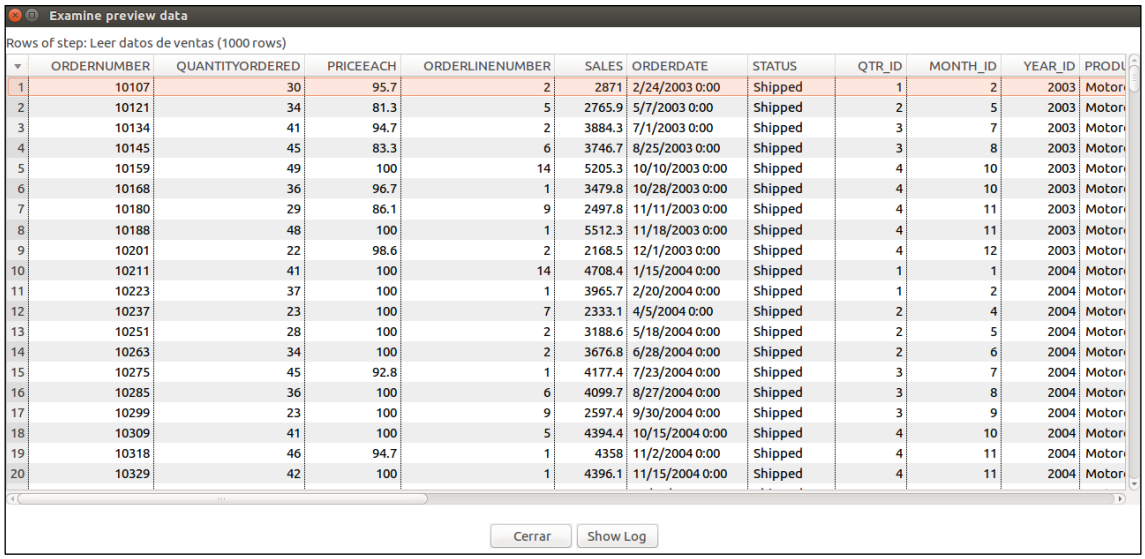


- Se hace clic en la pestaña *Campos* y, después, en *Traer campos*, para recuperar los campos de entrada del archivo de origen. Cuando aparezca la ventana *Número de líneas de la muestra* (*Nr. of lines to sample*), hay que ingresar *0* en el campo y luego hacer clic en *Vale*.

- Si aparece la ventana *Resultados de la exploración*, hay que hacer clic en *Cerrar* para cerrar la ventana.

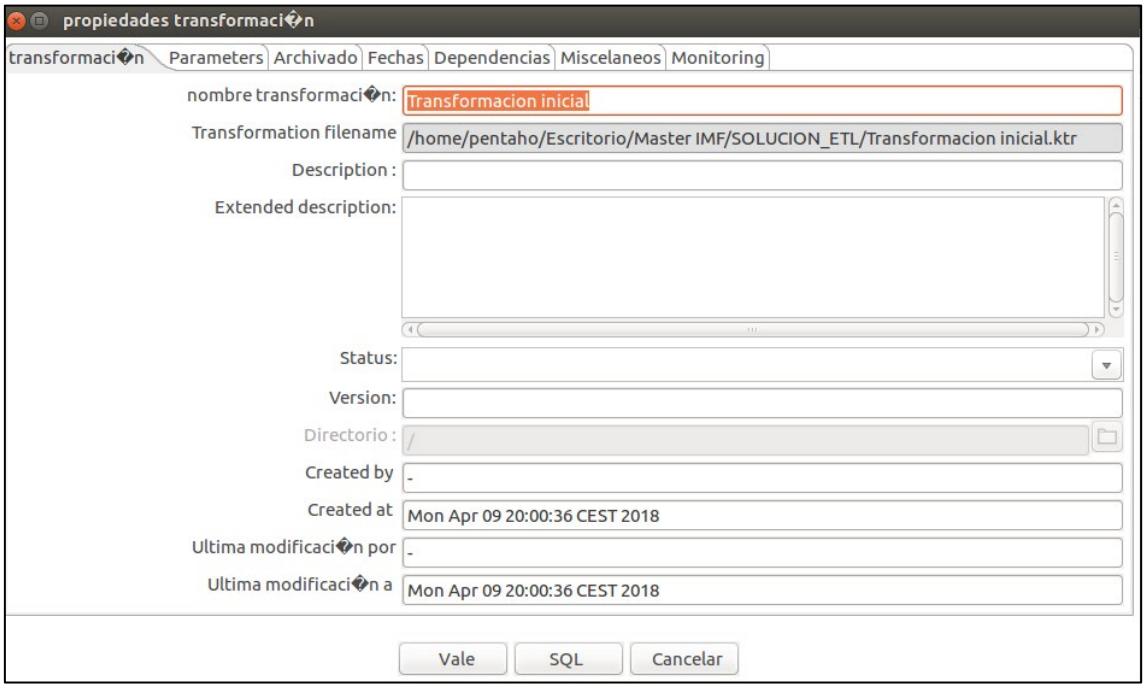


- Para verificar que los datos se leen correctamente, hay que llevar a cabo las siguientes acciones:
  - Hacer clic en la pestaña *Contenido* y, a continuación, en *Previsualizar filas*.
  - En la ventana *Introduce el tamaño de la previsualización*, hay que hacer clic en *Vale*. Aparece la ventana *Examine preview data*.
  - Se deben revisar los datos y luego hacer clic en *Cerrar*.



	ORDERNUMBER	QUANTITYORDERED	PRICEEACH	ORDERLINENUMBER	SALES	ORDERDATE	STATUS	QTR_ID	MONTH_ID	YEAR_ID	PRODU
1	10107	30	95.7	2	2871	2/24/2003 0:00	Shipped	1	2	2003	Motor
2	10121	34	81.3	5	2765.9	5/7/2003 0:00	Shipped	2	5	2003	Motor
3	10134	41	94.7	2	3884.3	7/1/2003 0:00	Shipped	3	7	2003	Motor
4	10145	45	83.3	6	3746.7	8/25/2003 0:00	Shipped	3	8	2003	Motor
5	10159	49	100	14	5205.3	10/10/2003 0:00	Shipped	4	10	2003	Motor
6	10168	36	96.7	1	3479.8	10/28/2003 0:00	Shipped	4	10	2003	Motor
7	10180	29	86.1	9	2497.8	11/11/2003 0:00	Shipped	4	11	2003	Motor
8	10188	48	100	1	5512.3	11/18/2003 0:00	Shipped	4	11	2003	Motor
9	10201	22	98.6	2	2168.5	12/1/2003 0:00	Shipped	4	12	2003	Motor
10	10211	41	100	14	4708.4	1/15/2004 0:00	Shipped	1	1	2004	Motor
11	10223	37	100	1	3965.7	2/20/2004 0:00	Shipped	1	2	2004	Motor
12	10237	23	100	7	2333.1	4/5/2004 0:00	Shipped	2	4	2004	Motor
13	10251	28	100	2	3188.6	5/18/2004 0:00	Shipped	2	5	2004	Motor
14	10263	34	100	2	3676.8	6/28/2004 0:00	Shipped	2	6	2004	Motor
15	10275	45	92.8	1	4177.4	7/23/2004 0:00	Shipped	3	7	2004	Motor
16	10285	36	100	6	4099.7	8/27/2004 0:00	Shipped	3	8	2004	Motor
17	10299	23	100	9	2597.4	9/30/2004 0:00	Shipped	3	9	2004	Motor
18	10309	41	100	5	4394.4	10/15/2004 0:00	Shipped	4	10	2004	Motor
19	10318	46	94.7	1	4358	11/2/2004 0:00	Shipped	4	11	2004	Motor
20	10329	42	100	1	4396.1	11/15/2004 0:00	Shipped	4	11	2004	Motor

- Hacer clic en *Vale* para guardar la información que se introdujo en el paso.
- Para guardar la transformación, hay que seguir las siguientes instrucciones:
  - Seleccionar *Fichero* > *Guardar* para guardar la transformación.
  - La ventana de propiedades de la transformación aparece. En el campo *Nombre transformación* se debe escribir "Transformación inicial". Hay que tener en cuenta que la ventana de propiedades de la transformación aparece porque está conectado a un repositorio. Si no estuviera conectado al repositorio, aparecería la ventana estándar para guardar.
  - En el campo *Directorio*, hay que hacer clic en el icono de la carpeta.
  - Expandir los directorios hasta seleccionar la carpeta en la que se desea guardar la transformación. La transformación se guarda en el repositorio de Pentaho.
  - Hacer clic en *Vale* y cerrar la ventana *Propiedades de transformación*.

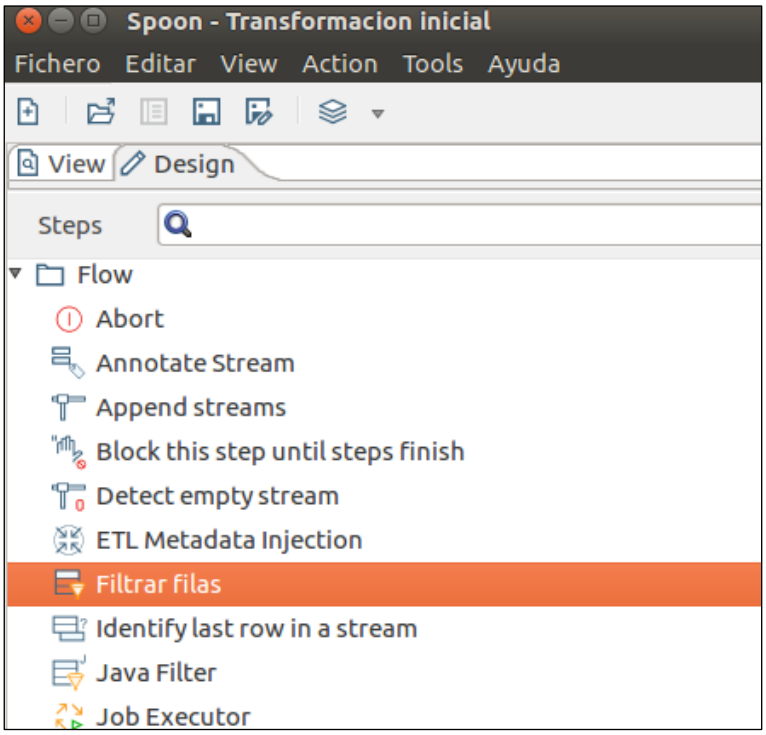


1.2. Filtrar registros con códigos postales perdidos

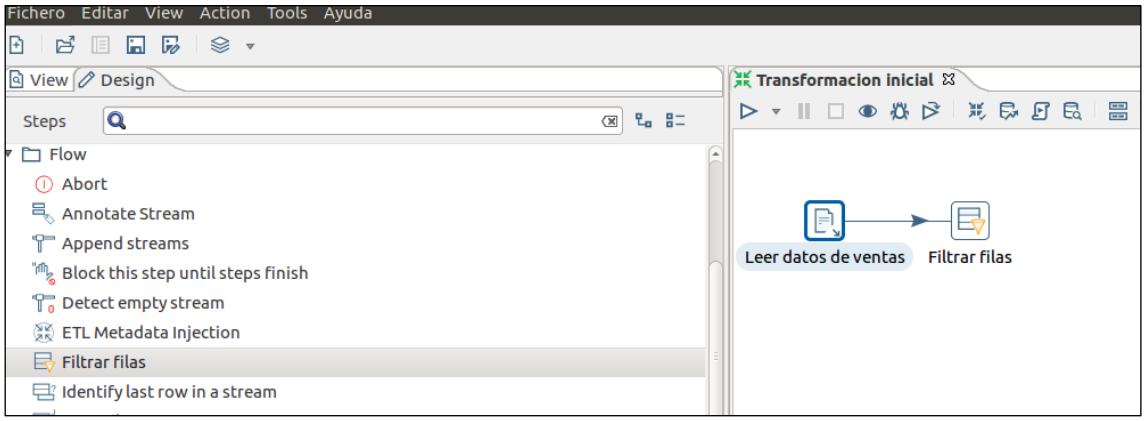
Después de completar el paso anteriormente detallado, *Recuperar datos desde un archivo plano*, se puede proceder a agregar el siguiente paso a la transformación.

El archivo fuente contiene varios registros en los que faltan códigos postales. Hay que emplear el paso de transformación *Filtrar filas* para separar esos registros y que puedan ser resueltos en un ejercicio posterior.

1) Agregar el paso *Filtrar filas* a la transformación. En la pestaña *Diseño*, seleccionar *Flow > Filtrar filas*.

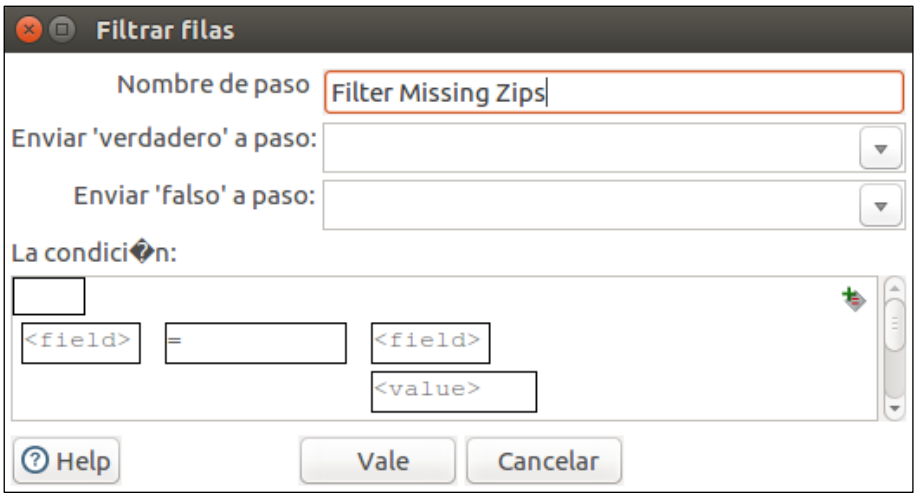


2) Crear un salto entre el paso *Leer datos de ventas* y el paso *Filtrar filas*. Los saltos se utilizan para describir el flujo de datos en la transformación. Para crear el salto, hay que hacer clic en el paso *Leer datos de ventas*, luego presionar la tecla *<MAYUS>* y dibujar una línea en el paso *Filtrar filas*.

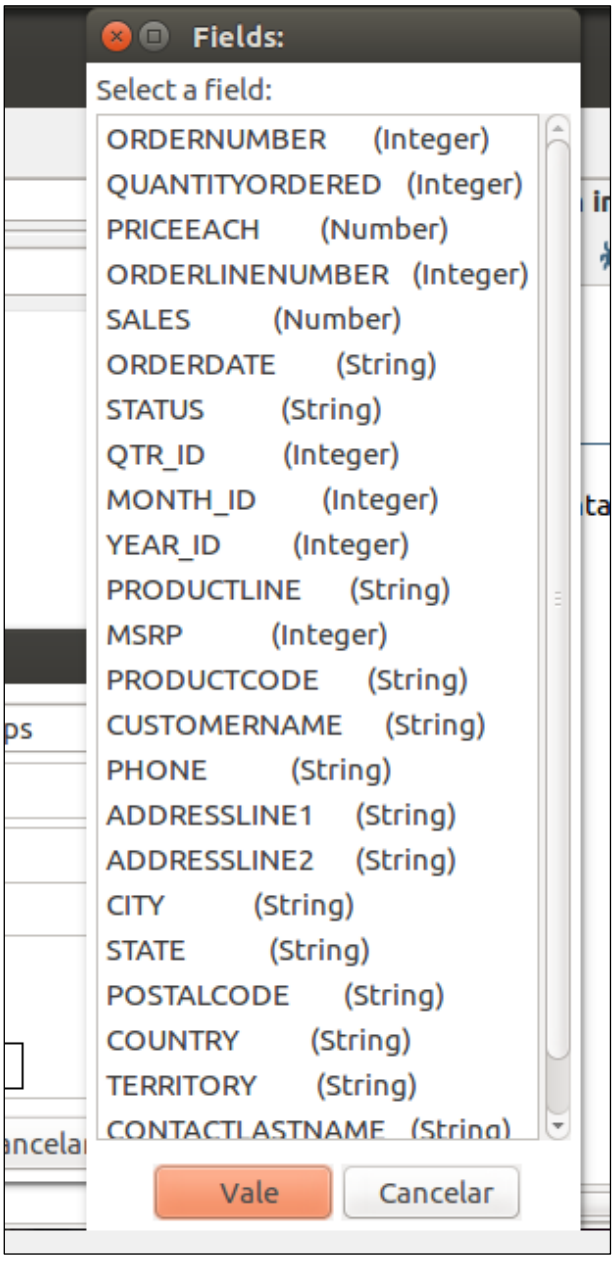


3) Hacer doble clic en el paso *Filtrar filas*. Aparece la ventana *Filtrar filas*.

4) En el campo *Nombre de paso*, hay que escribir "Filter Missing Zips", por ejemplo.

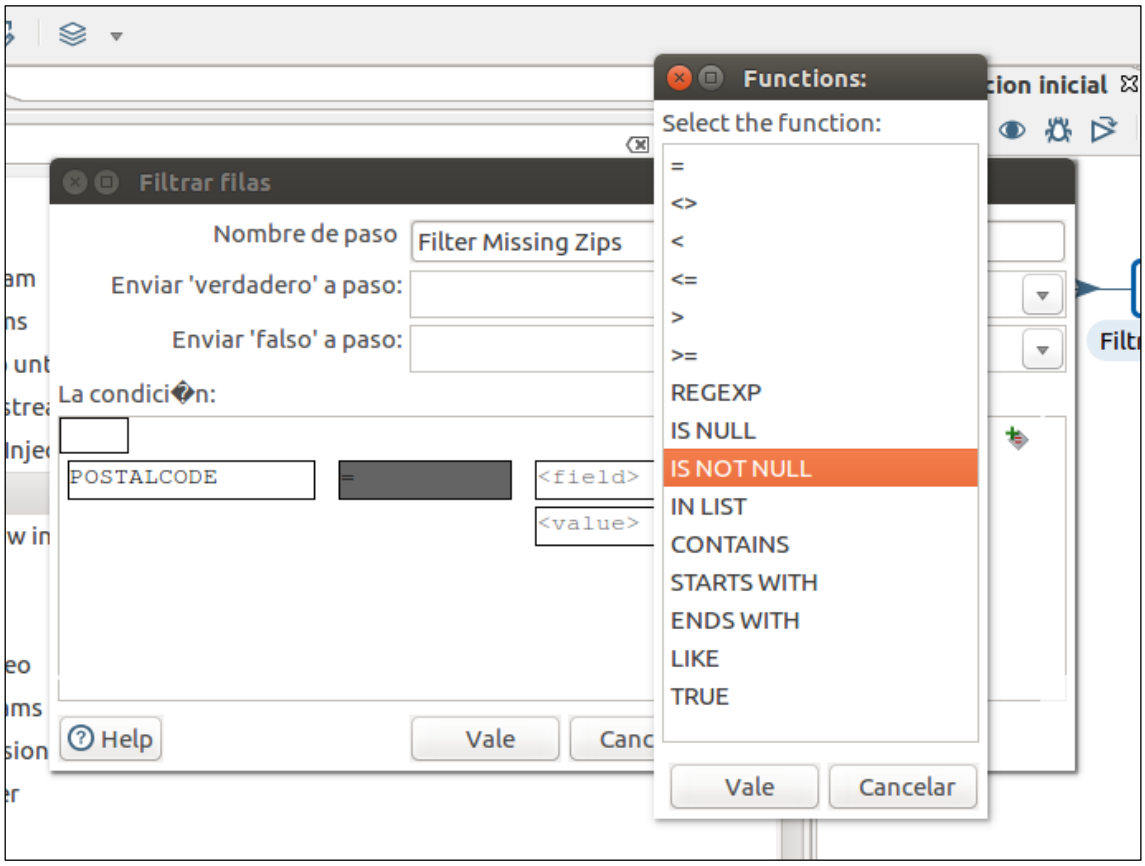


5) Hacer clic en *<field>* (campo), debajo de la condición. La ventana *Fields* aparece. Estas son las condiciones que pueden seleccionarse:



6) En la ventana *Fields:* hay que seleccionar *POSTAL CODE (String)* y hacer clic en *Vale*.

7) Hacer clic en el operador de comparación (establecido en "=" de manera predeterminada) y seleccionar *IS NOT NULL* de *Functions:* en la ventana que aparece.



8) Hacer clic en *Vale* para cerrar la ventana *Functions:* y volver a *Filtrar filas*.

9) Hacer clic en *Vale* para salir de la ventana *Filtrar filas*.

10) Téngase en cuenta que habrá que volver a este paso más adelante para configurar Enviar datos verdaderos a paso y Enviar datos falsos a ajustes de paso, después de agregar los pasos de destino a la transformación.

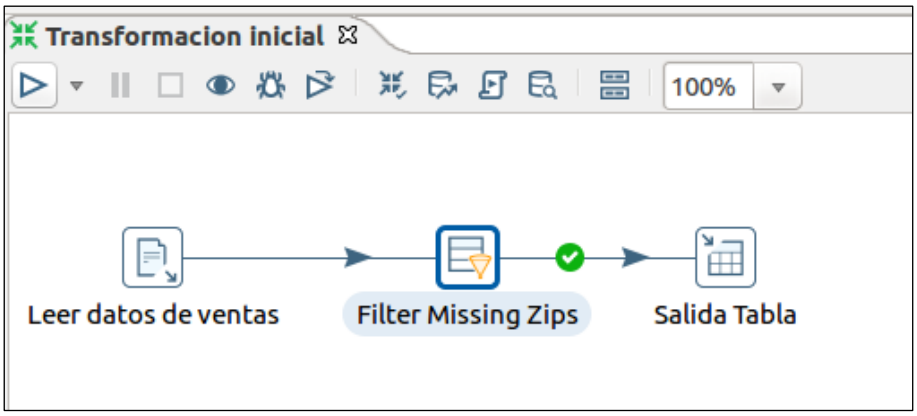
11) Seleccionar *Fichero > Guardar* para guardar la transformación.

1.3. Cargar datos en una base de datos relacional

Una vez completado el paso anterior, *Filtrar registros con códigos postales perdidos*, se puede comenzar a trabajar con todos los registros que salen del paso *Filtrar filas*, donde *POSTAL CODE (string)* no era nulo (la condición verdadera), y cargarlos en una tabla de base de datos. Síganse estas instrucciones:

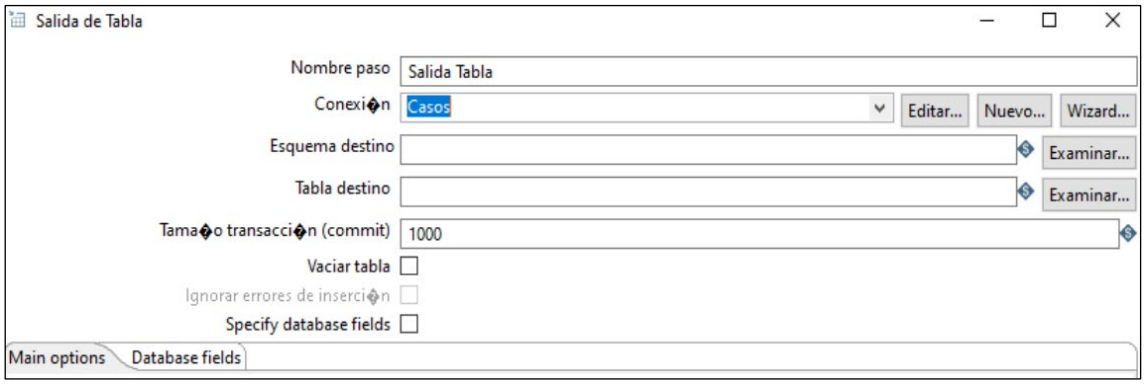
1) Expandir el contenido del grupo *Salida* que aparece debajo de la pestaña *Diseño*.

2) Hacer clic en arrastrar el paso *Salida tabla* a la transformación. Hay que crear un salto entre el paso *Filtrar filas*, anteriormente diseñado, y el nuevo paso *Salida* tabla. En el cuadro de diálogo que aparece hay que seleccionar *Result is TRUE*.



3) Hacer doble clic en el paso *Salida de tabla* para abrir las propiedades de edición.

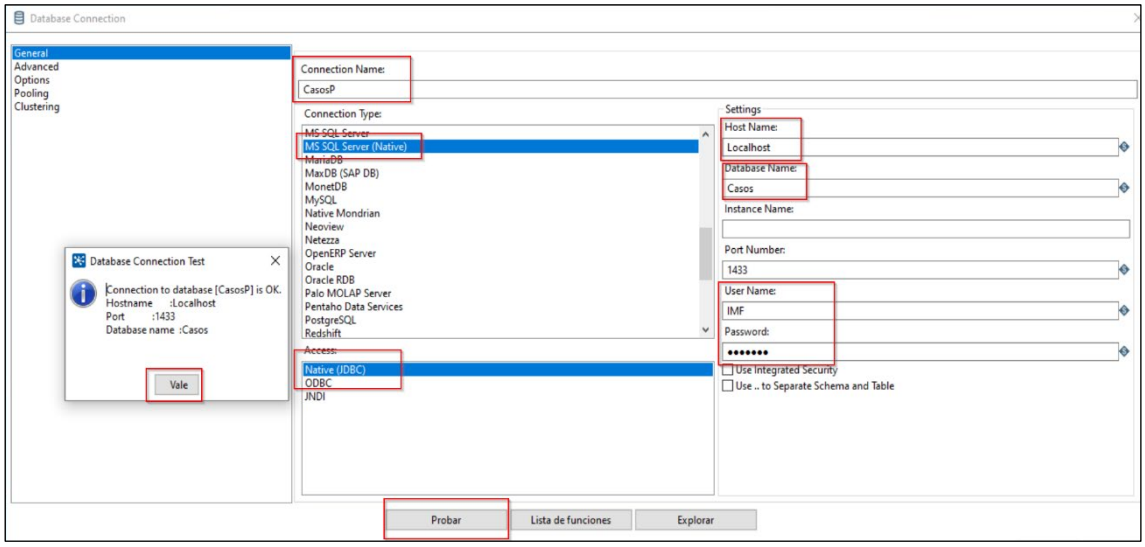
4) Cambiar el nombre del paso de *Salida de tabla* para escribir en la base de datos.



5) Hacer clic en *Nuevo…*; aparece al lado del campo *Conexión*. Se debe crear una conexión a la base de datos. Aparece el cuadro de diálogo *Database Connection* a la base de datos.

6) Proporcionar la configuración para conectarse a la base de datos. Es muy importante asegurarse de que Pentaho Business Analytics Server se está ejecutando.

Field	***Setting***
Connection Name:	Sample Data
Connection Type	MS SQL Server (Native)
Access	Native (JDBC)
Host Name	localhost
Database Name	Casos
Port Number	1433
User Name	IMF
Password	IMF1234



7) Hacer clic en *Probar* para asegurarse de que las entradas son correctas. Aparece un mensaje de éxito. Hacer clic en *Vale*.

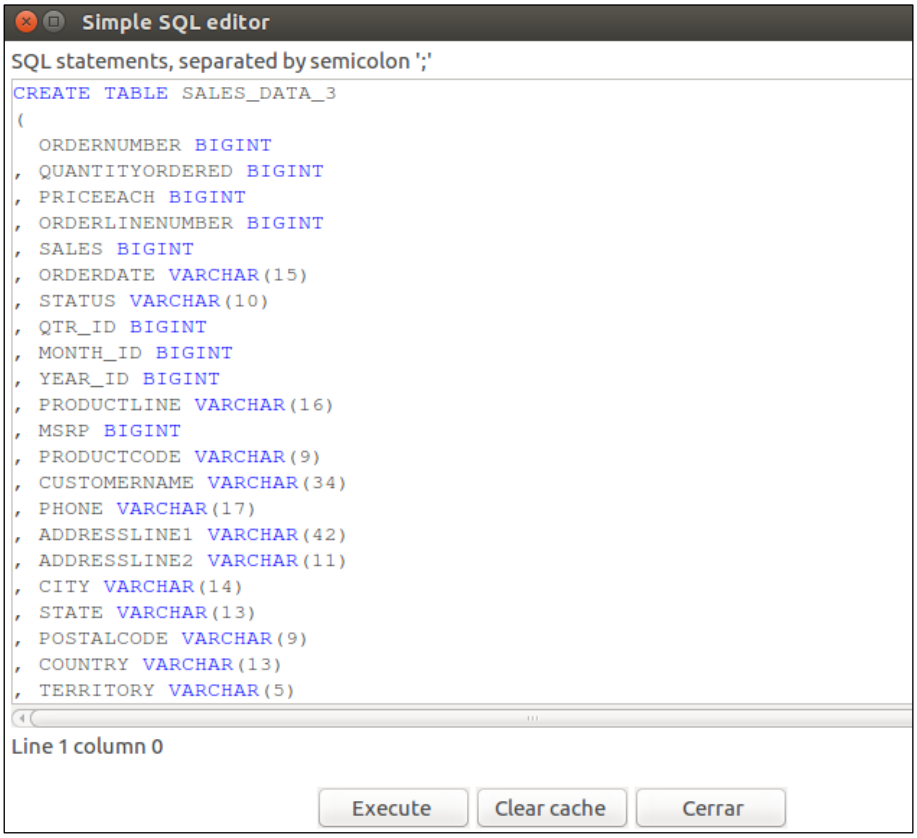
8) Si se obtiene un error al probar la conexión, hay que asegurarse de haber proporcionado la información de configuración correcta, tal y como se describe en la tabla, y de que la base de datos de muestra se está ejecutando.

9) Hacer clic en *Ok* para salir de la ventana *Database Connection*.

10) Escribir "SALES_DATA" en el campo de texto *Tabla destino*.

11) Como esta tabla no existe en la base de datos de destino, será necesario usar el *software* para generar el lenguaje de definición de datos (DDL) con el que crear la tabla y ejecutarla. Los DDL son los comandos SQL que definen las diferentes estructuras en una base de datos, como *CREATE TABLE*.

   - En la ventana *Salida de tabla*, hay que habilitar la propiedad *Vaciar tabla*.
   - Hacer clic en el botón *SQL* (aparece en la parte inferior del cuadro de diálogo *Salida de tabla*), para generar el DDL y poder crear la tabla de destino. Solo se debe realizar una única vez.
   - La ventana del editor SQL simple, *Simple SQL editor*, aparece con las instrucciones SQL necesarias para crear la tabla, si la tabla no existe. Pero si ya se ha ejecutado una vez y, por lo tanto, existe la tabla, mostrará opciones de redefinición de tabla.



- Hacer clic en *Execute* para ejecutar la instrucción SQL.
- Aparecerá la ventana *Resultados de las sentencias SQL*. Hay que examinar los resultados y, a continuación, hacer clic en *Vale* para cerrar los resultados de la ventana de declaraciones SQL.
- Hacer clic en *Cerrar* en la ventana del editor de SQL simple.
- Hacer clic en *Vale* para cerrar la ventana *Salida de tabla*.

12) Guardar la transformación.

1.4. Recuperando datos del archivo de búsqueda

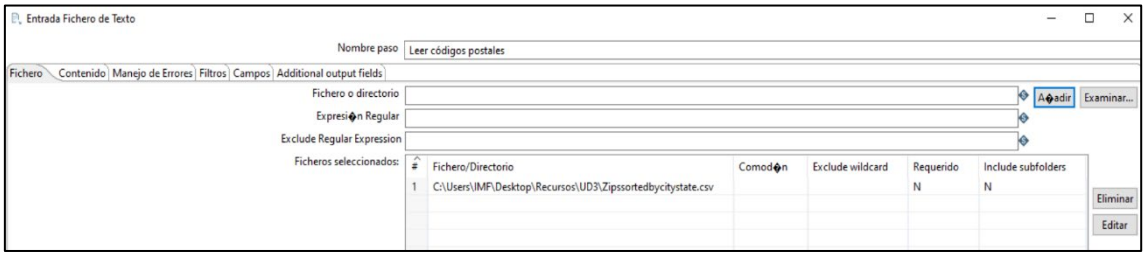
Después de cargar los datos en una base de datos relacional, ya se pueden recuperar datos del archivo de búsqueda. Se proporciona un segundo archivo de texto que contiene una lista de ciudades, Estados y códigos postales, que se usará ahora para buscar los códigos postales de todos los registros que faltan —la rama *falsa* del paso *Filas de filtro*—. Primero, se usará un paso de ingreso de archivo de texto para leer desde el archivo de origen, luego se usará un paso de búsqueda de transmisión para traer los códigos postales resueltos a la transmisión. A continuación, se describen los pasos que hay que seguir para ello:

1) Agregar un nuevo paso de *Entrada Fichero de Texto* a la transformación. En este paso, se recuperarán los registros del archivo de búsqueda. No hay que agregar un salto todavía.

2) Abrir la ventana del paso *Entrada Fichero de Texto*. A continuación, hay que introducir “Leer códigos postales” en la propiedad de nombre de paso.

3) Hacer clic en *Examinar…* para buscar el archivo fuente “**Zipssortedbycitystate.csv**”.

4) Hacer clic en *Añadir*. La ruta al archivo aparece debajo de *Archivos seleccionados*.



5) Para ver los contenidos del archivo de muestra, hay que:
   - Hacer clic en la pestaña *Contenido*. A continuación, hay que configurar el campo *Formato* en Unix.
   - Hacer clic en la pestaña *Fichero* nuevamente y después en *Ver contenido fichero*, que aparece en la parte inferior de la ventana.
   - Aparece la ventana *Número de líneas a visualizar*. Hay que hacer clic en el botón *Vale* para admitir el valor predeterminado.
   - El contenido de la primera ventana del archivo muestra el archivo. Hay que examinar ese archivo de entrada para ver cómo está delimitado, qué carácter de recinto se utiliza y si hay una fila de encabezado presente o no. En el ejemplo, el archivo de entrada está delimitado por comas (,), el carácter del recinto son comillas (“) y contiene una sola fila de encabezado que encierra los nombres de los campos.
   - Hacer clic en el botón *Cerrar* para cerrar la ventana.

6) En la pestaña *Contenido*, hay que cambiar el carácter *Separador de campos* por una coma (,) y confirmar que la configuración del *Separador de texto* sean unas comillas (“). También hay que asegurarse de que la opción *Cabecera* esté seleccionada.

7) En la pestaña *Campos*, hay que hacer clic en *Traer campos* para recuperar los datos del archivo CSV.

8) Aparece la ventana *Número de líneas a visualizar*. Hay que introducir *0* en el campo y luego hacer clic en *Aceptar*.

9) Si aparece la ventana *Resultado de la exploración*, hay que hacer clic en *Cerrar* para cerrarla.

10) Hacer clic en *Previsualizar filas* para asegurarse de que las entradas son correctas. Cuando se solicite *Introduce el tamaño de la previsualización*, hay que hacer clic en *Vale*. También hay que revisar la información en la ventana y luego hacer clic en *Cerrar*.

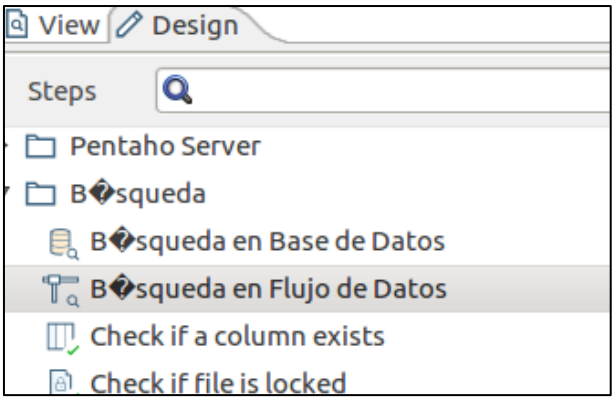
11) Hacer clic en *Vale* para salir de la ventana *Entrada Fichero de Texto*.

12) Guardar la transformación.

1.5. Resolviendo Missing Zip Code Information

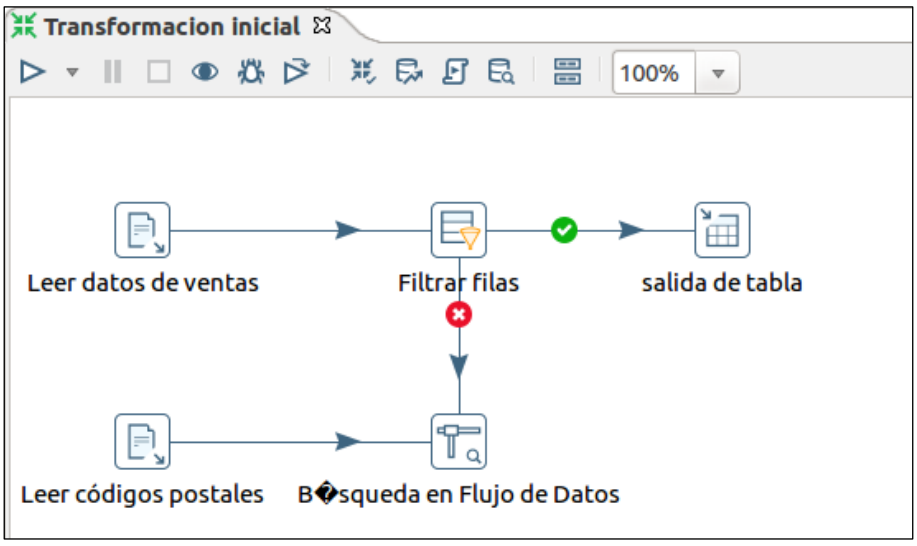
Después de recuperar los datos del archivo de búsqueda, es posible comenzar a resolver los códigos postales que faltan.

1) Agregar un paso de búsqueda de flujo a la transformación. Para ello, hay que hacer clic en la pestaña *Design* (*diseño*) y, a continuación, expandir la carpeta *Búsqueda* y elegir *Búsqueda en Flujo de Datos*.



2) Dibujar un salto desde el paso *Filtrar filas* hasta el paso *Búsqueda en Flujo de Datos*. En el cuadro de diálogo que aparece, hay que seleccionar *Result is FALSE*.

3) Crear un salto desde el paso Leer códigos postales al paso *Búsqueda en Flujo de Datos*.

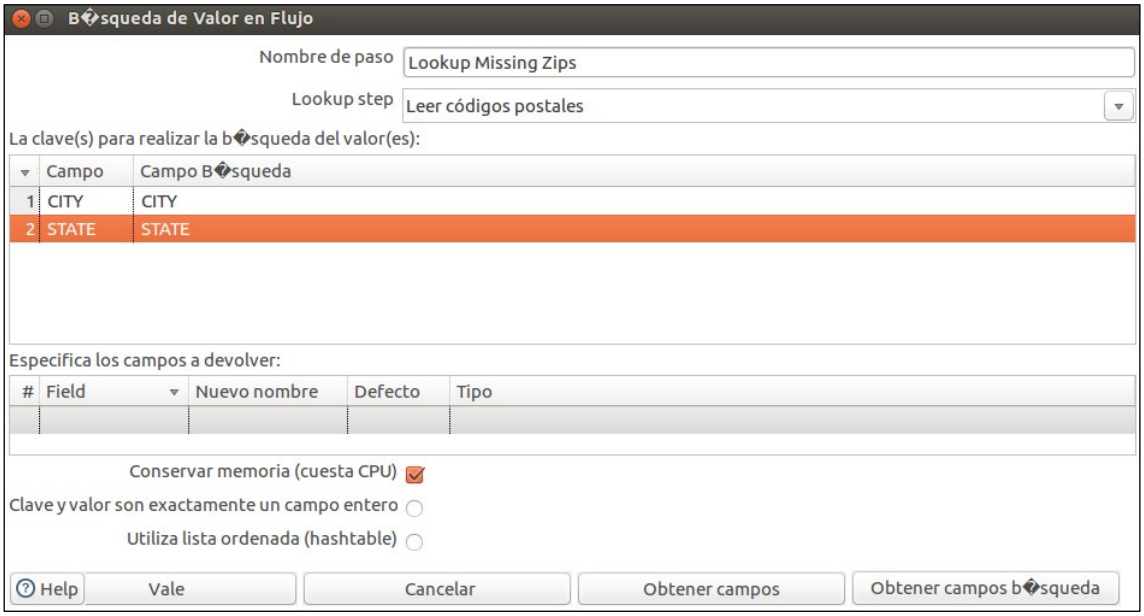


4) Hacer doble clic en el paso *Búsqueda en Flujo de Datos* para abrir la ventana *Búsqueda en Flujo de Datos*.

5) Cambiar el nombre de Búsqueda en Flujo de Datos por Lookup Missing Zips.

6) En el cuadro desplegable *Lookup step*, hay que seleccionar *Leer códigos postales* como paso de búsqueda.

7) Definir los campos *CITY* y *STATE* en la(s) clave(s) para buscar la tabla de valores. En la fila n. 1, hay que hacer clic en el menú desplegable de la columna *Campo* y seleccionar *CITY*. A continuación, hay que hacer clic en la columna *Campo de Búsqueda* y seleccionar *CITY*. En la fila n. 2, hay que hacer clic en el campo desplegable en la columna *Campo* y seleccionar *STATE*. A continuación, hay que hacer clic en la columna *Campo de Búsqueda* y seleccionar *STATE*.

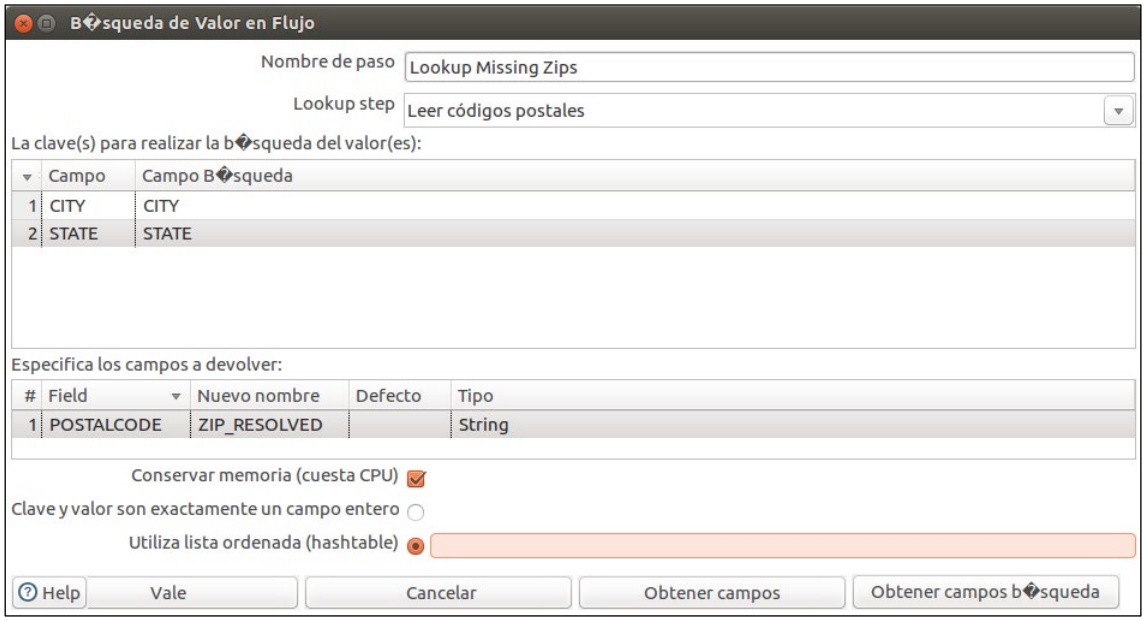


8) Hacer clic en *Obtener campos de búsqueda*.

9) *POSTALCODE* es el único campo que se desea recuperar. Para eliminar las líneas *CITY* y *STATE*, hay que hacer clic con el botón derecho en la línea y seleccionar *Borrar filas seleccionadas*.

10) En el campo *Nuevo nombre*, hay que asignar a *POSTALCODE* un nuevo nombre de *ZIP_RESOLVED* y asegurarse de que el tipo esté configurado en cadena (*string*).

11) Habilitar Usar lista ordenada (hashtable).

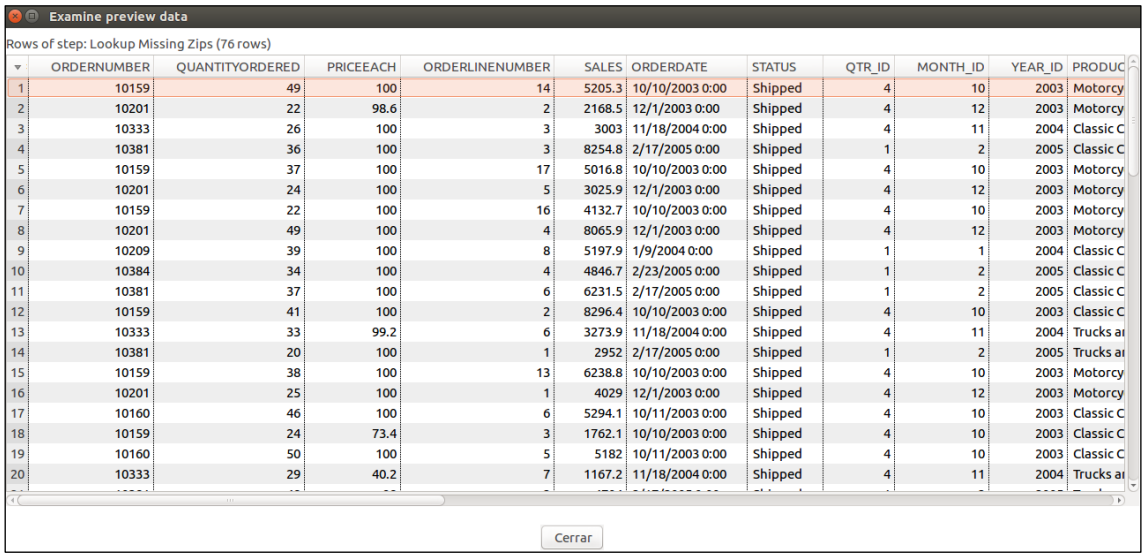


12) Hacer clic en *Vale* para cerrar el cuadro de diálogo *Búsqueda de Valor en Flujo*.

13) Guardar la transformación.

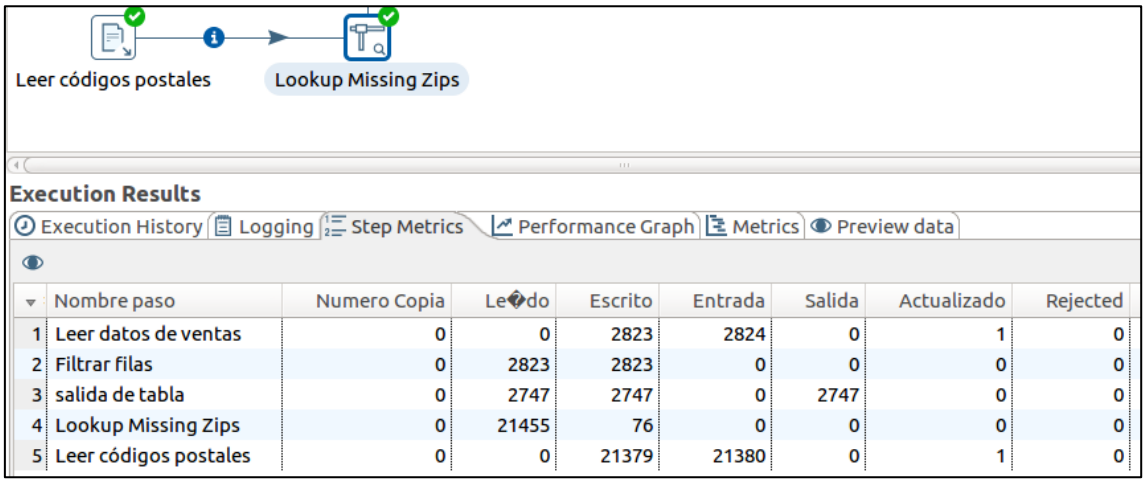
14) Para previsualizar los datos, hay que realizar lo que se describe a continuación:

- En el lienzo, hay que seleccionar el paso *Lookup Missing Zips* y luego hacer clic con el botón derecho. Desde el menú que aparece, selecciónese *Preview*.
- En la ventana de diálogo *Transformation debug dialog*, hay que hacer clic en *Quick Launch* para obtener una vista previa de los datos que fluyen a través de este paso.
- Aparece la ventana *Examine preview data*. Téngase en cuenta que el nuevo campo, *ZIP_RESOLVED*, se ha agregado a la ruta que contiene los códigos postales resueltos.



	ORDERNUMBER	QUANTITYORDERED	PRICEEACH	ORDERLINENUMBER	SALES	ORDERDATE	STATUS	QTR_ID	MONTH_ID	YEAR_ID	PRODUC
1	10159	49	100	14	5205.3	10/10/2003 0:00	Shipped	4	10	2003	Motorcy
2	10201	22	98.6	2	2168.5	12/1/2003 0:00	Shipped	4	12	2003	Motorcy
3	10333	26	100	3	3003	11/18/2004 0:00	Shipped	4	11	2004	Classic C
4	10381	36	100	3	8254.8	2/17/2005 0:00	Shipped	1	2	2005	Classic C
5	10159	37	100	17	5016.8	10/10/2003 0:00	Shipped	4	10	2003	Motorcy
6	10201	24	100	5	3025.9	12/1/2003 0:00	Shipped	4	12	2003	Motorcy
7	10159	22	100	16	4132.7	10/10/2003 0:00	Shipped	4	10	2003	Motorcy
8	10201	49	100	4	8065.9	12/1/2003 0:00	Shipped	4	12	2003	Motorcy
9	10209	39	100	8	5197.9	1/9/2004 0:00	Shipped	1	1	2004	Classic C
10	10384	34	100	4	4846.7	2/23/2005 0:00	Shipped	1	2	2005	Classic C
11	10381	37	100	6	6231.5	2/17/2005 0:00	Shipped	1	2	2005	Classic C
12	10159	41	100	2	8296.4	10/10/2003 0:00	Shipped	4	10	2003	Classic C
13	10333	33	99.2	6	3273.9	11/18/2004 0:00	Shipped	4	11	2004	Trucks a
14	10381	20	100	1	2952	2/17/2005 0:00	Shipped	1	2	2005	Trucks a
15	10159	38	100	13	6238.8	10/10/2003 0:00	Shipped	4	10	2003	Motorcy
16	10201	25	100	1	4029	12/1/2003 0:00	Shipped	4	12	2003	Motorcy
17	10160	46	100	6	5294.1	10/11/2003 0:00	Shipped	4	10	2003	Classic C
18	10159	24	73.4	3	1762.1	10/10/2003 0:00	Shipped	4	10	2003	Classic C
19	10160	50	100	5	5182	10/11/2003 0:00	Shipped	4	10	2003	Classic C
20	10333	29	40.2	7	1167.2	11/18/2004 0:00	Shipped	4	11	2004	Trucks a

- Hacer clic en *Cerrar* para cerrar la ventana.
- Si aparece la ventana *Selecciona el paso a previsualizar*, hay que hacer clic en el botón *Cerrar*.
- Téngase en cuenta que los resultados de la ejecución, que aparecen en la parte inferior de la ventana de Spoon, muestran las métricas actualizadas en la pestaña *Métricas* del paso.



	Nombre paso	Numero Copia	Le?do	Escrito	Entrada	Salida	Actualizado	Rejected
1	Leer datos de ventas	0	0	2823	2824	0	1	0
2	Filtrar filas	0	2823	2823	0	0	0	0
3	salida de tabla	0	2747	2747	0	2747	0	0
4	Lookup Missing Zips	0	21455	76	0	0	0	0
5	Leer códigos postales	0	0	21379	21380	0	1	0

1.6. Completando la transformación

Después de completar la información perdida en el origen del código postal, la última tarea es limpiar el diseño del campo en su secuencia de búsqueda. La limpieza hace que coincida con el formato y el diseño de la otra secuencia que va al paso *Salida de tabla*. Se debe crear un paso *Seleccionar valores* para cambiar el nombre de los campos en la secuencia, eliminar campos innecesarios, etc.

1) Agregar un paso *Selecciona/Renombra valores* a la transformación, expandiendo la carpeta *Transformar* y eligiendo *Selecciona/Renombra valores*.

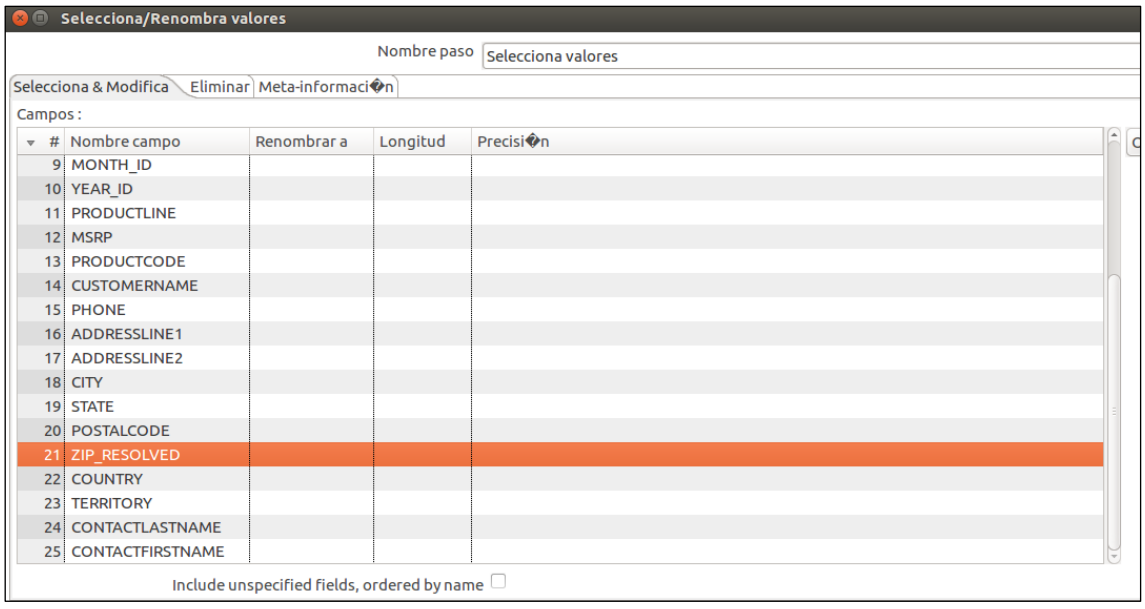
2) Crear un salto desde Lookup Missing Zips al paso Selecciona/Renombra valores.

3) Hacer doble clic en el paso *Selecciona/Renombra valores* para abrir el cuadro de diálogo de propiedades.

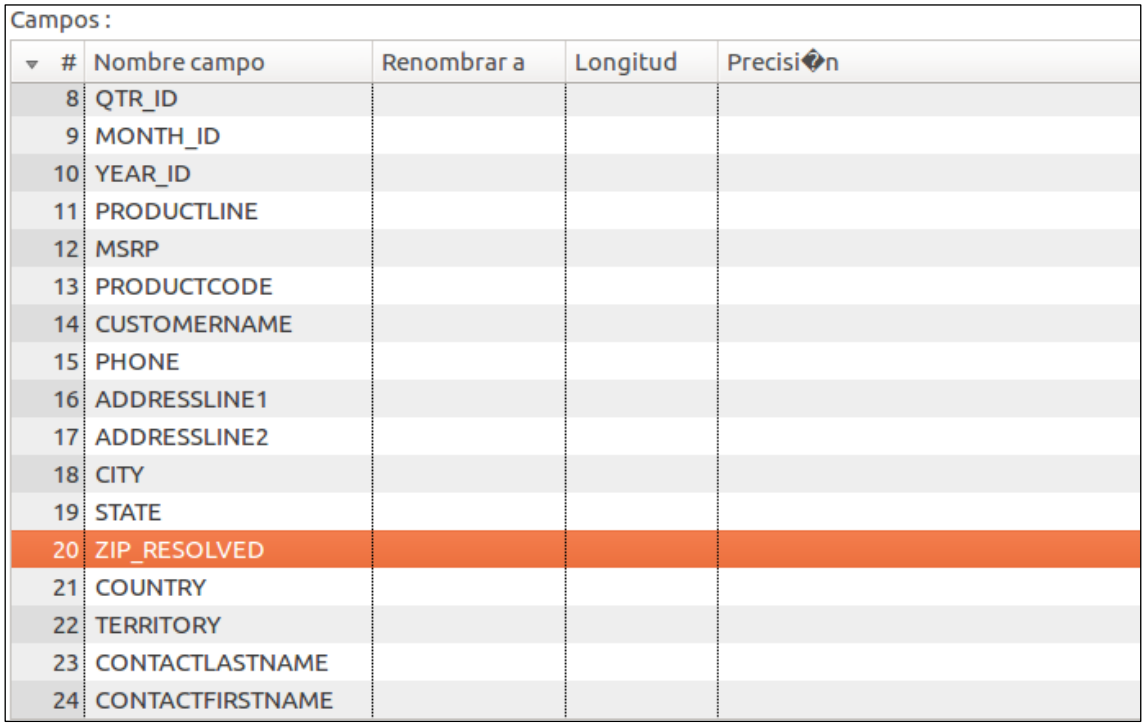
4) Cambiar el nombre del paso Selecciona/Renombra valores a Seleccionar valores.

5) Hacer clic en *Obtener campos a seleccionar* para seleccionar y recuperar todos los campos, y comenzar a modificar el diseño de la ruta.

6) En la lista *Nombre campos*, hay que buscar la columna *#* y hacer clic en el número del campo *ZIP_RESOLVED*. Hay que usar *<CTRL> <UP>* para mover *ZIP_RESOLVED*, que aparece justo debajo del campo *POSTALCODE* (el que todavía contiene valores nulos).



7) Seleccionar el campo *POSTALCODE* antiguo en la lista (línea 20) y eliminarlo.



Campos :

#	Nombre campo	Renombrar a	Longitud	Precisi&#65533;n
8	QTR_ID			
9	MONTH_ID			
10	YEAR_ID			
11	PRODUCTLINE			
12	MSRP			
13	PRODUCTCODE			
14	CUSTOMERNAME			
15	PHONE			
16	ADDRESSLINE1			
17	ADDRESSLINE2			
18	CITY			
19	STATE			
20	ZIP_RESOLVED			
21	COUNTRY			
22	TERRITORY			
23	CONTACTLASTNAME			
24	CONTACTFIRSTNAME			

8) El campo *POSTALCODE* original se formateó como una cadena de nueve caracteres. Debe modificarse el nuevo campo para que coincida con el formulario. Para ello, hay que hacer clic en la pestaña *Meta-Información*.

9) En la primera fila de los campos, hay que hacer clic en la columna *Nombre campo* y seleccionar *ZIP_RESOLVED* para modificar la tabla metadatos para la sección.

10) Hay que escribir “POSTALCODE” en la columna *Renombrar a*. Debe seleccionarse *String* en la columna *Tipo*, y escribir “9” en la columna *Longitud*. Hay que hacer clic en *Vale* para salir del cuadro de diálogo de editar propiedades.

Selecciona/Renombra valores

Nombre paso: Selecciona valores

Selecciona & Modifica | Eliminar | Meta-informaci&#65533;n

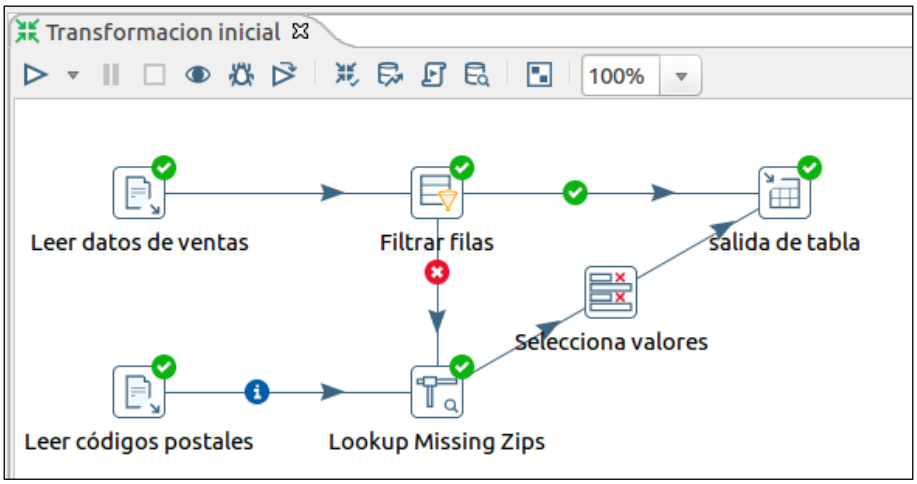
Campos a modificar meta informaci&#65533;n :

Nombre campo	Renombrar a	Tipo	Longitud	Precisi&#65533;n	Binary to Normal?	Format	Date Format Leni
1 ZIP_RESOLVED	POSTALCODE	String	9				

11) Dibujar un salto desde el paso *Seleccionar valores* al paso *Salida de tabla*.

12) Cuando se solicite, hay que seleccionar *Main output of step*.

13) Guardar la transformación.



1.7. Ejecutar la transformación

La integración de datos proporciona varias opciones de implementación. La ejecución de una transformación explica estas y otras opciones disponibles para su ejecución. Esta parte final de este ejercicio para crear una transformación se centra exclusivamente en la opción de ejecución local.

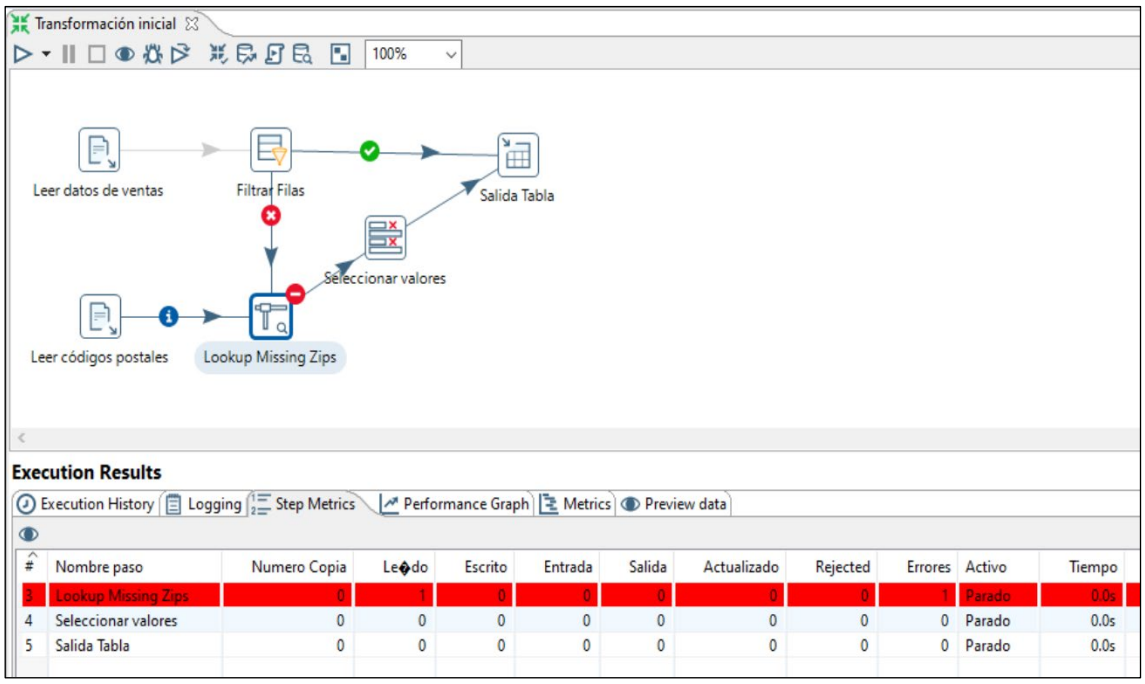
1) En la ventana del cliente de PDI, hay que seleccionar *Action* > *Ejecutar*.

2) Aparece la ventana *Ejecutar una transformación*. Hay que mantener la opción *Ejecución local* predeterminada de Pentaho para este ejercicio. Se utilizará el motor Pentaho nativo y se ejecutará la transformación en la máquina local. Consúltese *Ejecutar configuraciones* si se tiene interés en configuraciones que usen otro motor, como Spark, para ejecutar una transformación.

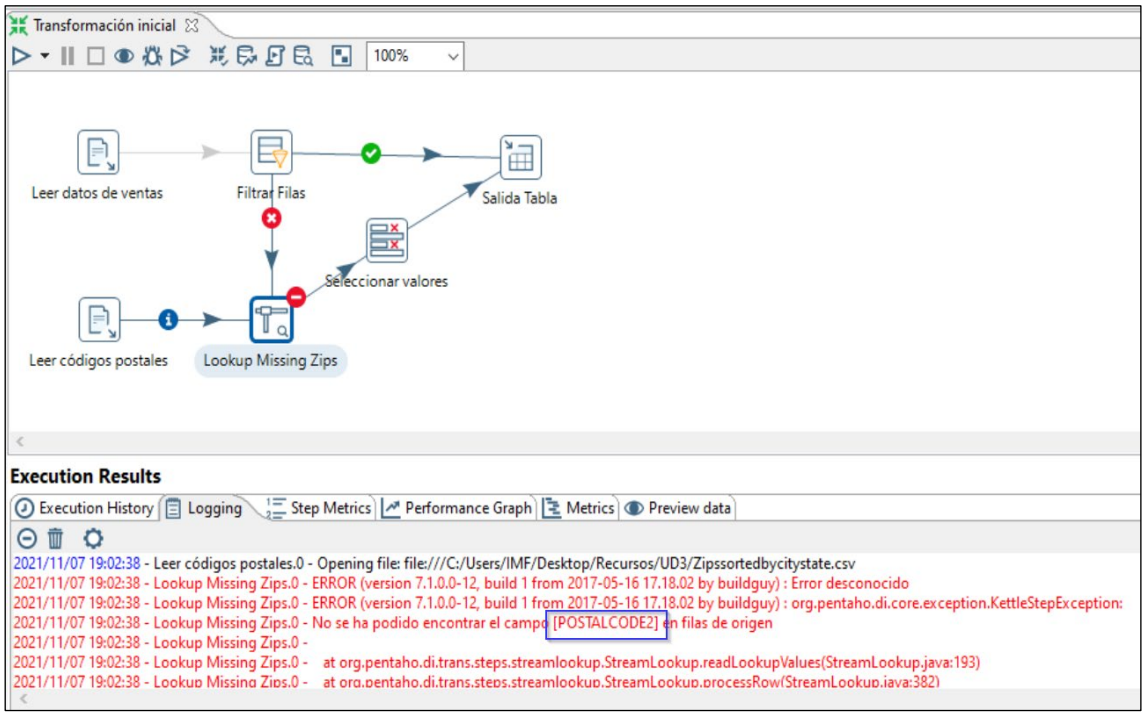
3) Hacer clic en *Ejecutar*. La transformación se ejecuta. Al ejecutar la transformación, el panel *Resultados de ejecución* se abre debajo del lienzo.

4) La sección *Execution Results* contiene varias pestañas diferentes que ayudarán a ver cómo se ejecutó la transformación, a detectar errores y a supervisar el rendimiento:

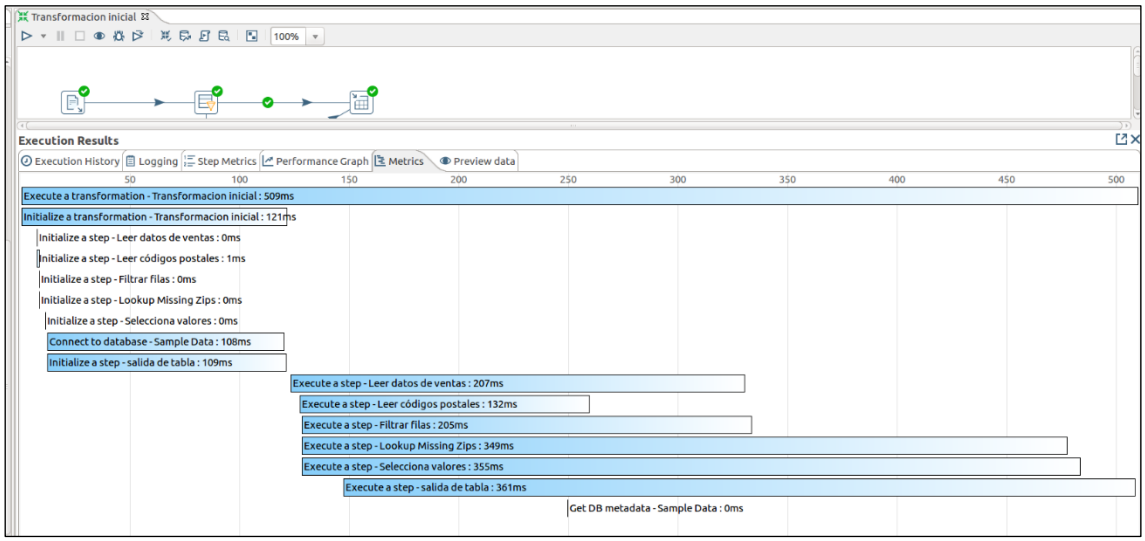
- La pestaña *Step Metrics* proporciona estadísticas para cada paso de la transformación, incluyendo cuántos registros se leyeron o escribieron, cuántos causaron un error, cuántos procesaron la velocidad (filas por segundo) y más. Esta pestaña también indica si se produjo un error en un paso de transformación.
- En este ejemplo no se introdujo voluntariamente ningún error, por lo que debería ejecutarse correctamente. Pero, si se hubiera producido un error, los pasos que causaron la transformación fallida aparecerían señalados en rojo.
- En el siguiente ejemplo, el paso *Lookup Missing Zips* provocó un error:



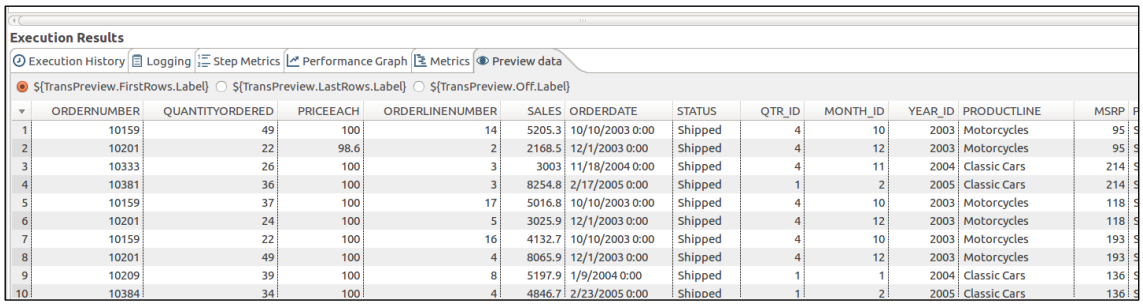
- La pestaña *Registro* (*Logging*) muestra los detalles de registro para la ejecución más reciente de la transformación. También permite profundizar para determinar dónde ocurren los errores. Las líneas de error están resaltadas en rojo. En el ejemplo siguiente, el paso *Lookup Missing Zips* provocó un error porque intentó buscar valores en un campo llamado *POSTALCODE2*, que no existía en la secuencia de búsqueda:



- La pestaña *Historial* (*Execution History*) proporciona acceso a las métricas de pasos e información de registro de ejecuciones previas a la transformación. Esta función solo funciona si se ha configurado la transformación para iniciar sesión en una base de datos, a través de la pestaña *Registro*, del cuadro de diálogo *Configuración de transformación*.
- El *Gráfico de rendimiento* (*Perfomance Graph*) permite analizar el rendimiento de los pasos en función de una variedad de métricas, lo que incluye cuántos registros se leyeron o escribieron, o causaron un error, la velocidad de procesamiento (filas por segundo), etc. Al igual que el historial de ejecución, esta función requiere que se configure la transformación para iniciar sesión en una base de datos, a través de la pestaña *Registro*, en el cuadro de diálogo *Configuración de transformación*.
- La pestaña *Métricas* (*Metrics*) permite ver un diagrama de Gantt después de que se hayan ejecutado la transformación o el trabajo. Esto muestra información como cuánto tiempo lleva conectarse a una base de datos, cuánto tiempo se dedica a ejecutar una consulta SQL o cuánto tiempo se tarda en cargar una transformación.



– La pestaña *Vista previa* (*Preview data*) muestra una vista previa de los datos.



	ORDERNUMBER	QUANTITYORDERED	PRICEEACH	ORDERLINENUMBER	SALES	ORDERDATE	STATUS	QTR_ID	MONTH_ID	YEAR_ID	PRODUCTLINE	MSRP
1	10159	49	100	14	5205.3	10/10/2003 0:00	Shipped	4	10	2003	Motorcycles	95
2	10201	22	98.6	2	2168.5	12/1/2003 0:00	Shipped	4	12	2003	Motorcycles	95
3	10333	26	100	3	3003	11/18/2004 0:00	Shipped	4	11	2004	Classic Cars	214
4	10381	36	100	3	8254.8	2/17/2005 0:00	Shipped	1	2	2005	Classic Cars	214
5	10159	37	100	17	5016.8	10/10/2003 0:00	Shipped	4	10	2003	Motorcycles	118
6	10201	24	100	5	3025.9	12/1/2003 0:00	Shipped	4	12	2003	Motorcycles	118
7	10159	22	100	16	4132.7	10/10/2003 0:00	Shipped	4	10	2003	Motorcycles	193
8	10201	49	100	4	8065.9	12/1/2003 0:00	Shipped	4	12	2003	Motorcycles	193
9	10209	39	100	8	5197.9	1/9/2004 0:00	Shipped	1	1	2004	Classic Cars	136
10	10384	34	100	4	4846.7	2/23/2005 0:00	Shipped	1	2	2005	Classic Cars	136

2. Mis primeros trabajos

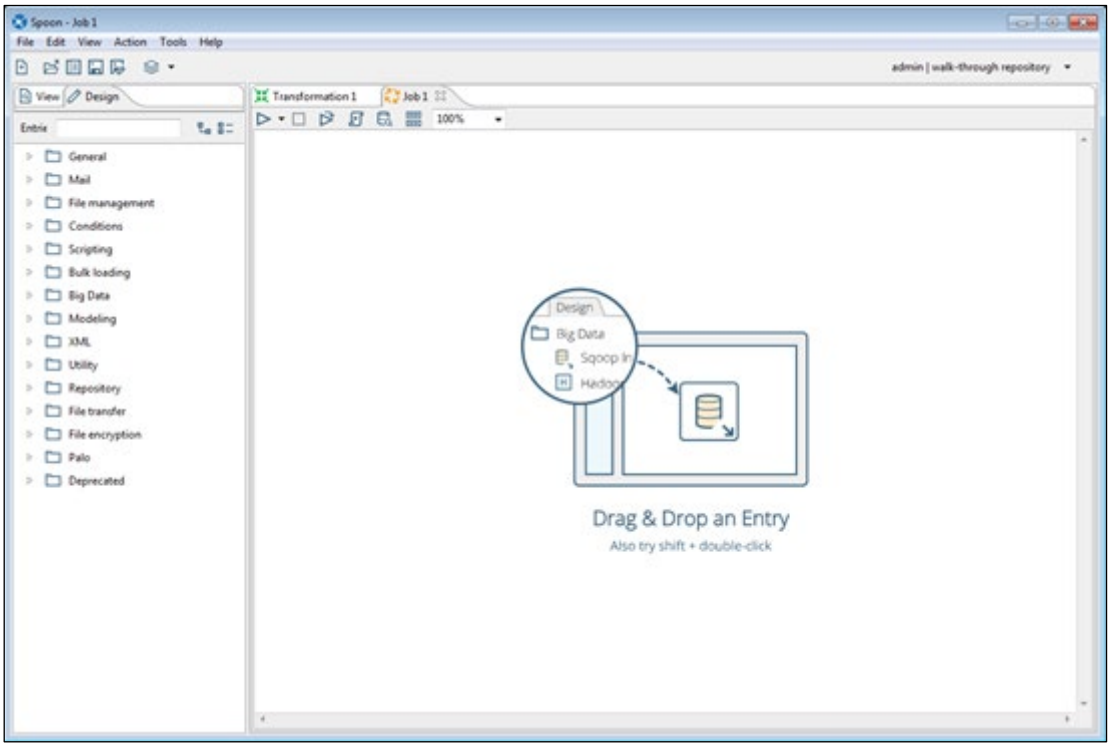
Tal y como se ha comentado, los trabajos se utilizan para coordinar actividades de ETL como las que se mencionan a continuación:

- Definir el flujo y las dependencias, para qué orden se deben ejecutar las transformaciones.
- Prepararse para la ejecución comprobando condiciones como “¿está mi archivo fuente disponible?” o “¿existe una tabla?”.
- Realización de operaciones de base de datos de carga masiva.
- Administración de archivos, cómo publicar o recuperar archivos mediante FTP, copiar archivos y eliminar archivos.
- Envío de notificaciones de éxito o error a través del correo electrónico.

Para este ejemplo, se puede imaginar que un sistema externo es responsable de colocar la entrada **"sales_data.csv"** en su ubicación de origen todos los sábados por la noche a las 9. Se pretende crear un trabajo que verifique que el archivo ha llegado y ejecute su transformación para cargar los registros en la base de datos. En un ejercicio posterior, se programará el trabajo para que se ejecute todos los domingos por la mañana, a las 9.

A continuación, se describen los pasos que hay que dar:

1) Ir a *Archivo > Nuevo > Trabajo*.



2) Expandir la carpeta *General* y arrastrar una entrada de trabajo de inicio al espacio de trabajo gráfico. La entrada del trabajo de inicio define dónde comenzará la ejecución.

3) Expandir la carpeta *Condiciones* y agregar una entrada de trabajo: *Archivo existente*.

4) Dibujar un salto desde la entrada *Iniciar trabajo* hasta la entrada de trabajo *Archivo existente*.

5) Hacer doble clic en la entrada de trabajo *Archivo existente* para abrir un cuadro de diálogo de propiedades de edición. Hacer clic en *Examinar* y configurar el filtro, que aparece en la parte inferior de la ventana, para todos los archivos. Hay que seleccionar **"sales_data.csv"**.

6) Hacer clic en *Aceptar* para salir de la ventana *Abrir archivo*.

7) Hacer clic en *Aceptar* para salir de la ventana *Comprobar si existe un archivo…*

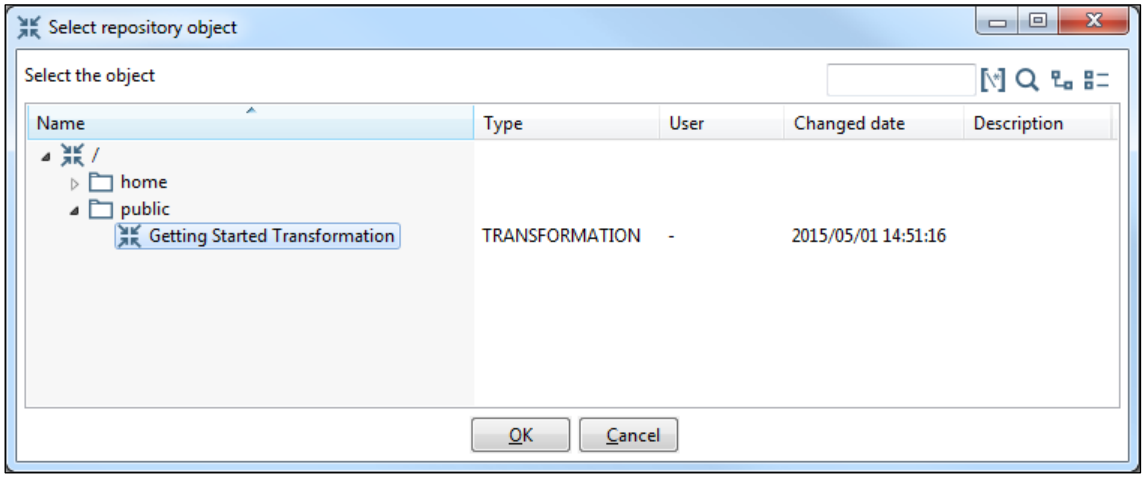
8) En Spoon, hay que expandir la carpeta *General* y agregar una entrada de trabajo *Transformación*.

9) Dibujar un salto entre las entradas de trabajo *Archivo existente* y *Transformación*.

10) Hacer doble clic en la entrada del trabajo *Transformación* para abrir el cuadro de diálogo de propiedades de edición.

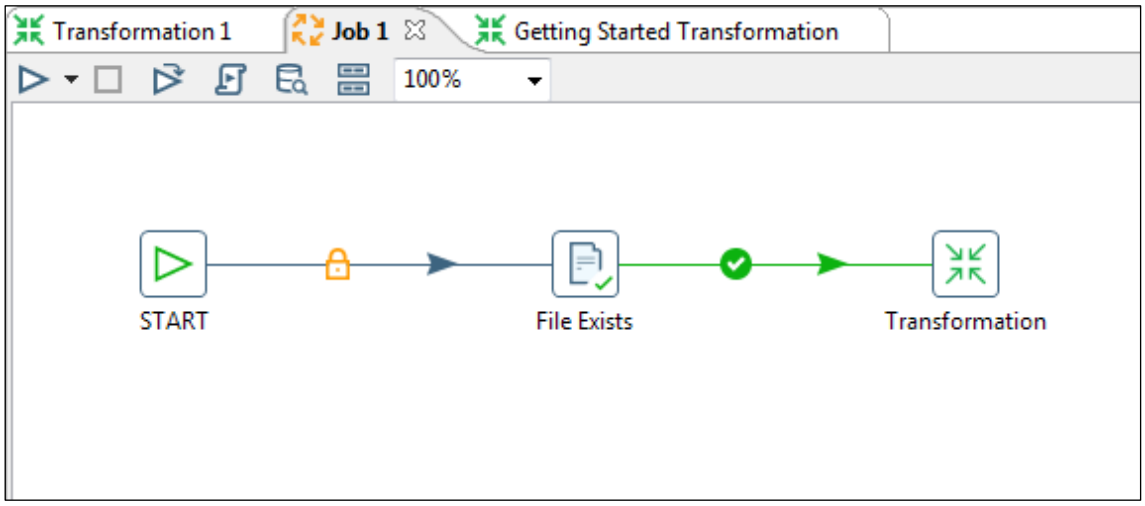
11) Hacer clic en *Examinar* para abrir la ventana *Seleccionar objeto de repositorio*. Buscar y seleccionar la transformación que se creó en el tutorial de transformación de PDI.

12) Expandir el árbol del repositorio para encontrar la transformación de muestra. Hay que seleccionarlo y hacer clic en *Aceptar*.



13) Guardar el trabajo como trabajo de muestra.

14) Hacer clic en el icono *Ejecutar*, en la barra de herramientas. Cuando aparezca la ventana *Opciones de ejecución*, hay que elegir el tipo de entorno *Local* y hacer clic en *Ejecutar*. El panel *Resultados de ejecución* se debería abrir, mostrando las métricas del trabajo y la información de registro para la ejecución del trabajo.

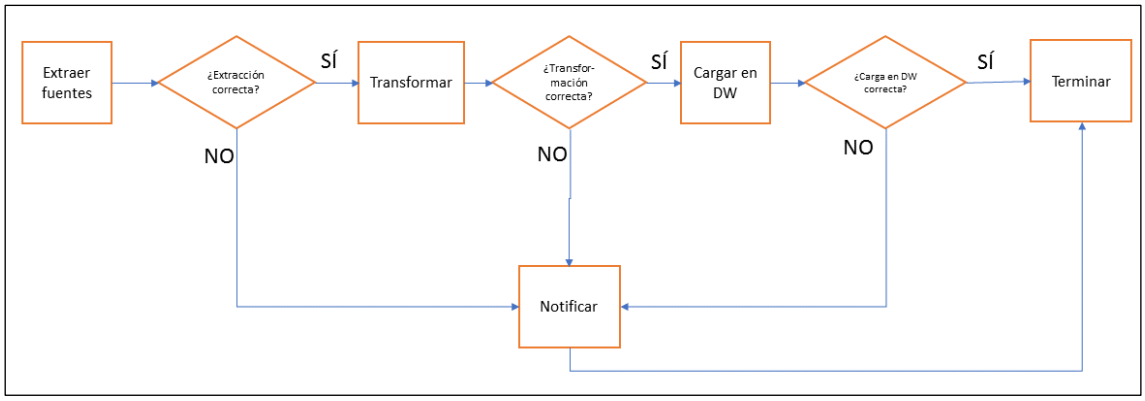


3. Trabajo *(job)* global de referencia de proceso ETL con Pentaho Data Integration para un proyecto estándar

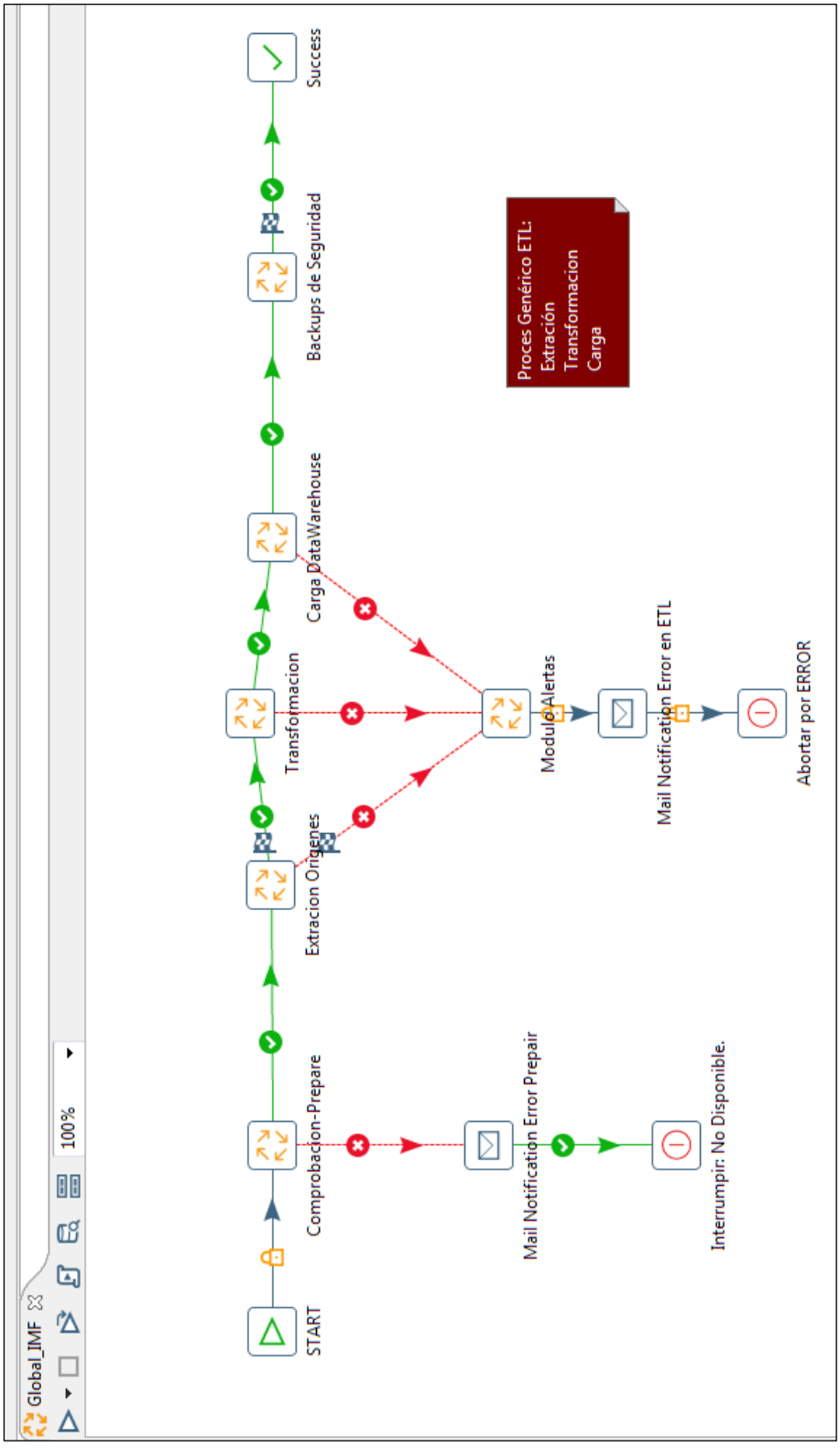
El propósito del siguiente caso es mostrar al alumno un ejemplo de referencia para un proceso ETL estándar, implementado con la herramienta ETL de Pentaho, Pentaho Data Integration.

El objetivo es diseñar e implementar con PDI el trabajo global necesario para realizar un proceso de extracción, transformación y carga, presente en todo proyecto de inteligencia de negocio.

Para ello, se partirá del diseño del diagrama funcional del flujo de datos:



Como se ha visto en los puntos anteriores, una de las grandes ventajas de Pentaho es que la definición de *workflow* es muy similar al diseño funcional que se puede realizar con un diagrama de flujos como el anterior, por lo que este trabajo global de proceso ETL completo se puede diseñar como se muestra en la figura siguiente:



Véase el archivo "**Global_IMF.kjb**".