Recurso complementario

Ejemplo 2: transformación y limpieza de datos

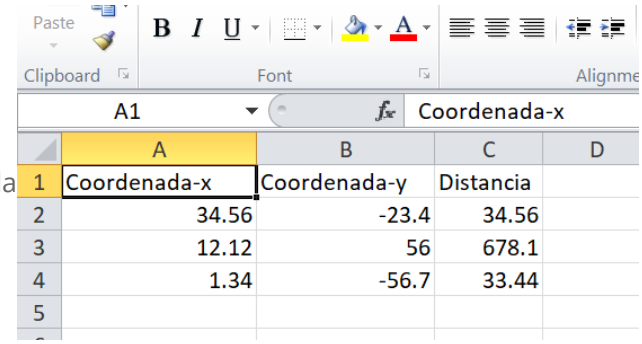
El segundo ejemplo de aplicación consiste en desarrollar un proceso *ETL* para leer un archivo de tipo Excel (XLS o XLSX) y volcar los datos dentro de un archivo de tipo XML. Esto implica que la transformación de Pentaho deberá leer las filas y columnas de la hoja de cálculo y generar las etiquetas XML en el formato de salida. Por esto, se creará una transformación de Pentaho Data Integration (*kettle*), a través de la interfaz *Spoon*, siguendo el estilo del ejercicio anterior, pero basado en los nuevos pasos.

Para construir la nueva transformación (que llamaremos **transformación02**), se deben cumplir los siguientes pasos:

1. Crea un archivo de Excel en algún directorio vacío, con nombre **excel2.xls**, que contenga los datos mostrados en la figura 1. Observa que los datos están **ubicados en tres columnas y al inicio de la hoja de cálculo (celda A1)**.

Figura 1.

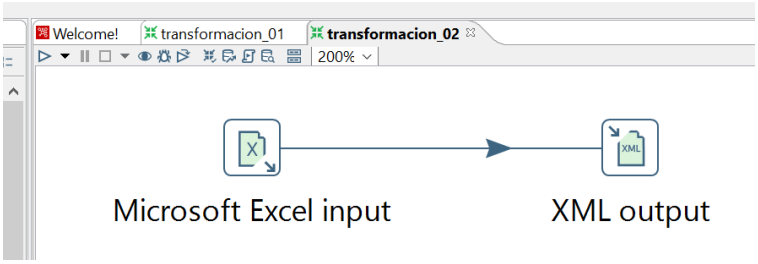
Hoja de Microsoft Excel con los datos a ser utilizados en la transformación



2. Construye la transformación de la figura 2, con los pasos y saltos correspondientes.

Figura 2.

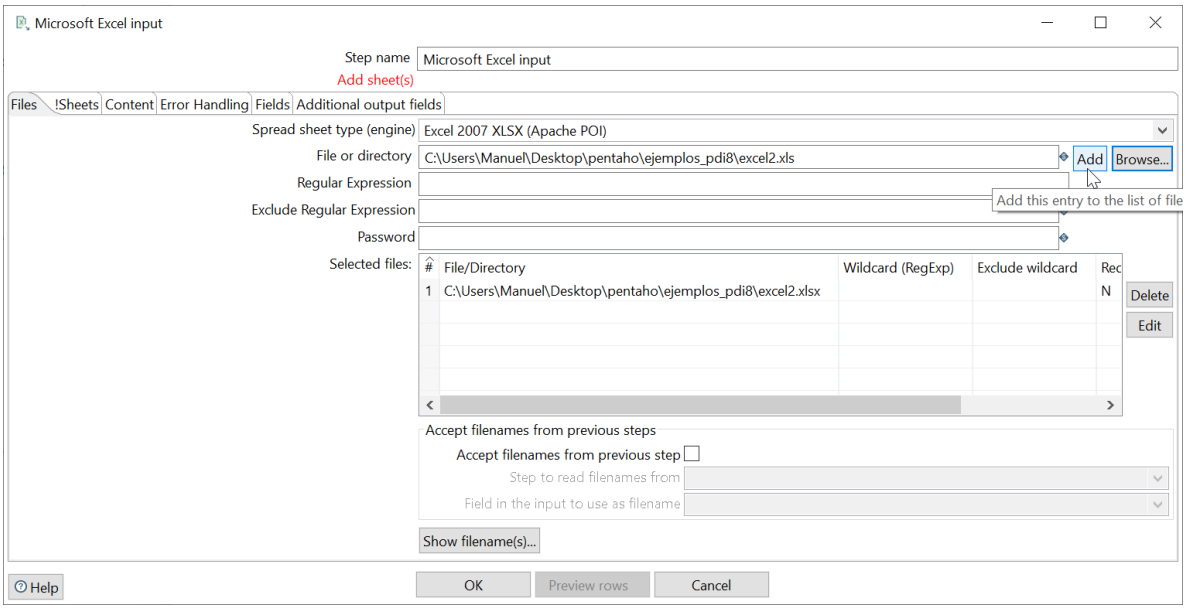
Esquema de transformación para el ejercicio 2



3. Configura el primer paso, de nombre **Microsoft Excel input**, con los parámetros mostrados en la figura 3.

Figura 3.

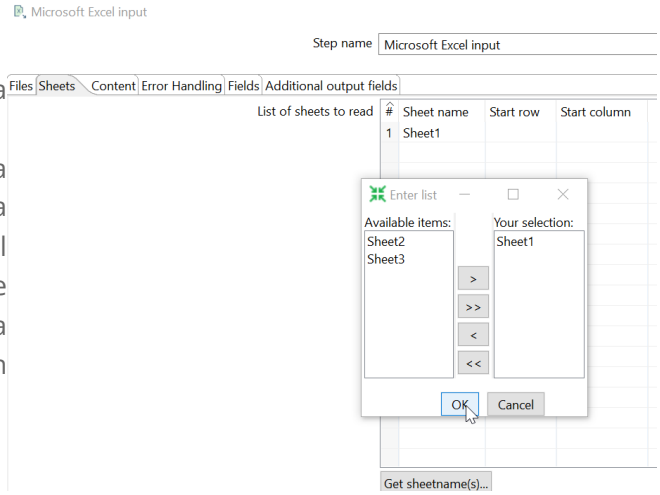
*Configuración general del paso **Microsoft Excel input** para leer los datos desde la hoja de cálculo*



4. En la pestaña *Sheets* del mismo componente, selecciona la hoja activa dentro del archivo de Excel que contiene los datos a ser leídos por la transformación (figura 4).

Figura 4.

Selección de la hoja activa a leer desde el archivo de Excel

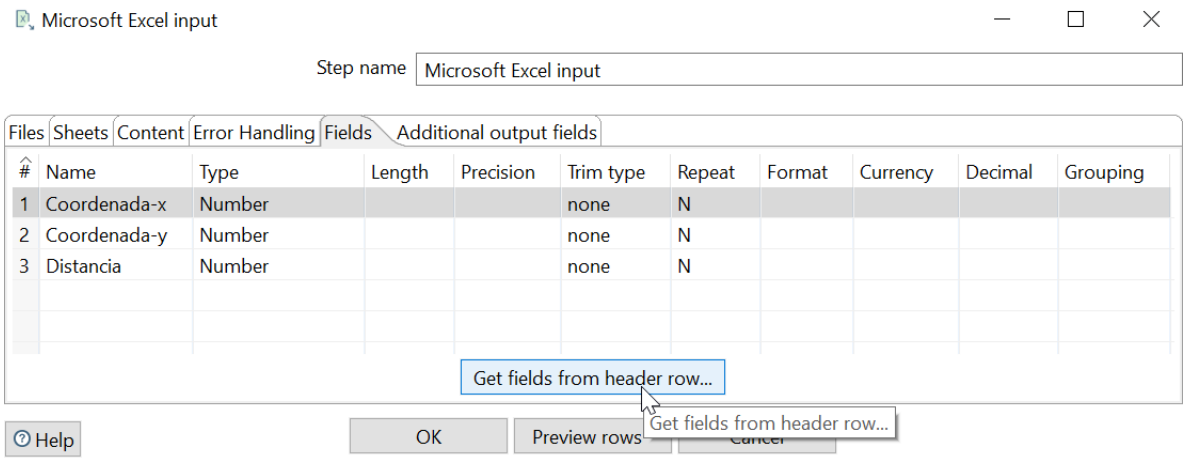


5. Ve a la pestaña *Fields* e indica los campos a leer desde el archivo de Excel (figura 5); también puedes probar presionando el botón **Get fields from header row...**, si es que deseas hacerlo automáticamente. Se

entiende que los nombres de campos deben estar incluidos en la primera fila de la hoja de cálculo.

Figura 5.

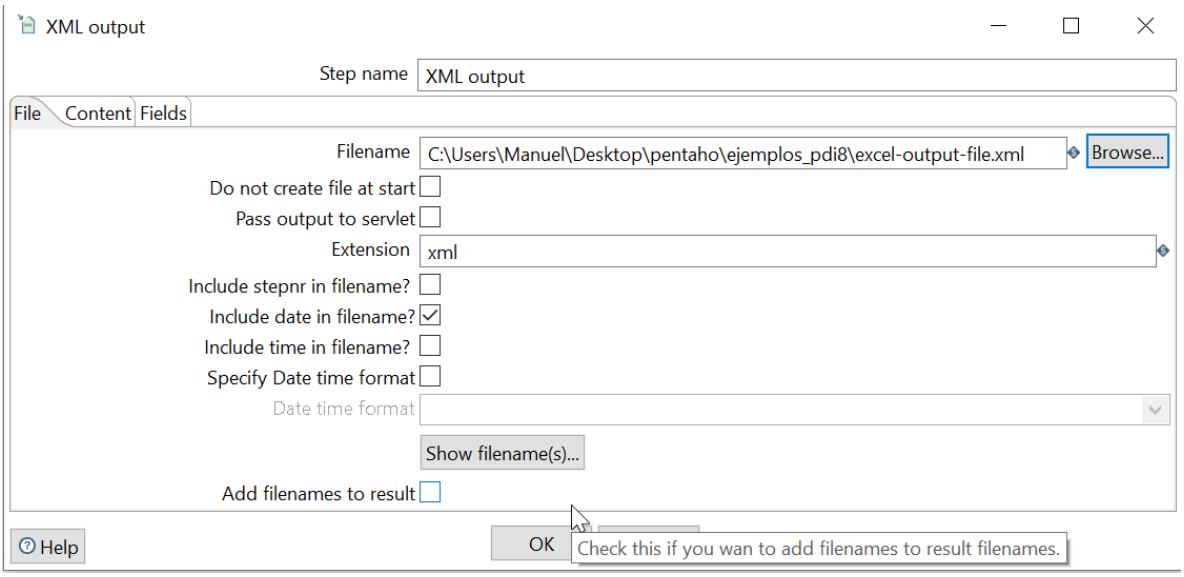
Adquisición de nombres de columna desde el archivo de Excel



6. Ahora, configura el segundo paso, llamado **XML output**, según lo indicado en la figura 6.

Figura 6.

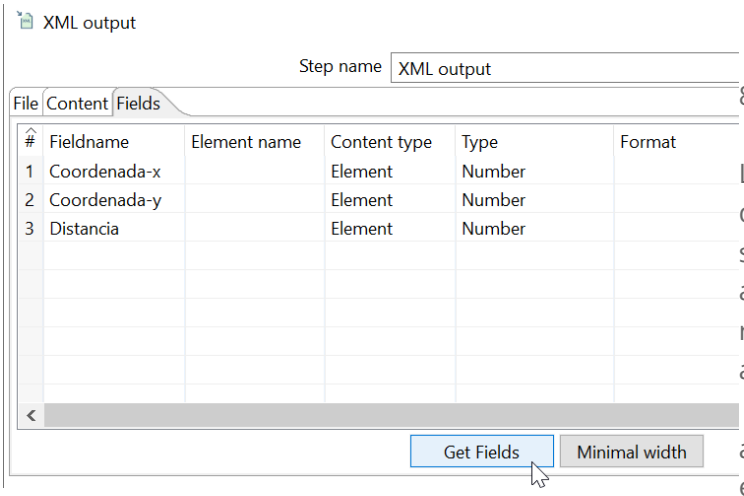
*Configuración general del componente **XML output***



7. También ve a la pestaña *Fields* del componente y obtén los campos que provee el primer componente de **Microsoft Excel input** (figura 7). Prueba también a presionar el botón **Get Fields** para hacer la operación en forma automática.

Figura 7.

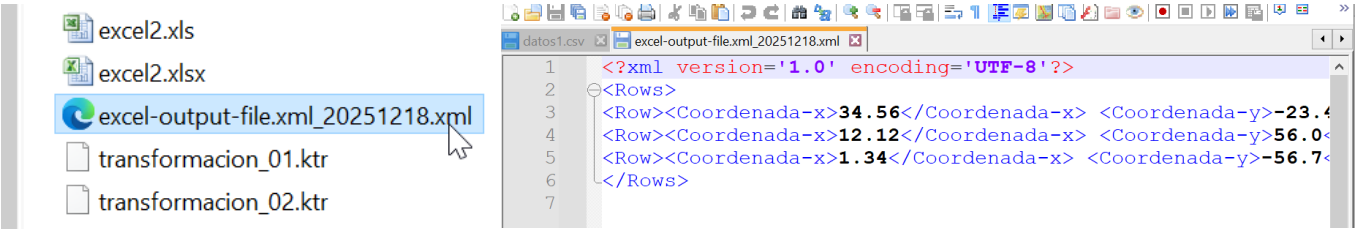
Lectura de los campos a escribir dentro del archivo de salida en XML



8. Ejecuta la transformación. Luego, ve al directorio de salida para el archivo XML y revisa que esté allí el archivo. Finalmente, ábrelo con un editor de texto y ve su contenido (figura 8).

Figura 8.

Archivo XML de salida con su contenido visible en un editor de texto



Como cierre, este ejercicio evidencia la relevancia de los procesos *ETL* en la gestión y preparación de datos, mostrando de manera aplicada cómo Pentaho Data Integration permite integrar, transformar y estructurar información proveniente de fuentes heterogéneas, como archivos Excel, hacia formatos estandarizados como XML. La correcta configuración de cada paso de la transformación no solo garantiza la integridad y calidad de los datos, sino que también fortalece competencias clave en análisis y manejo de información, fundamentales para la toma de

decisiones en contextos organizacionales y tecnológicos, reforzando así el aprendizaje práctico de los conceptos revisados en la unidad.