Seminario sobre GeoKettle

Primera actividad

Enunciado del problema.

Disponemos de un ***archivo de texto muy grande*** (pongamos 600Mbytes) con ***números reales con separador decimal punto y demasiados ceros decimales*** como los que se muestran:

17-06-2013;1.338000000000000000;1.546000000000000000;VALLADOLID ;VALLADOLID ;47016 ;4718610030;4718610

11-01-2013;1.374000000000000000;;ÁLAVA ;LOPIDANA ;01196 ;0105906003;0105906

10-01-2013;1.398000000000000000;;ALBACETE ;ABENGIBRE ;02250 ;0200101001;0200101

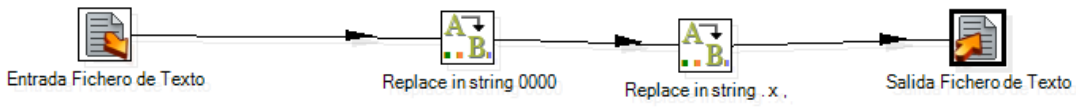
Para poderlos procesar bien con otras herramientas es necesario ***quitar ceros decimales*** a las cifras de números reales y ***cambiar el separador decimal punto (.) por la coma (,)***.

Solución inmediata:

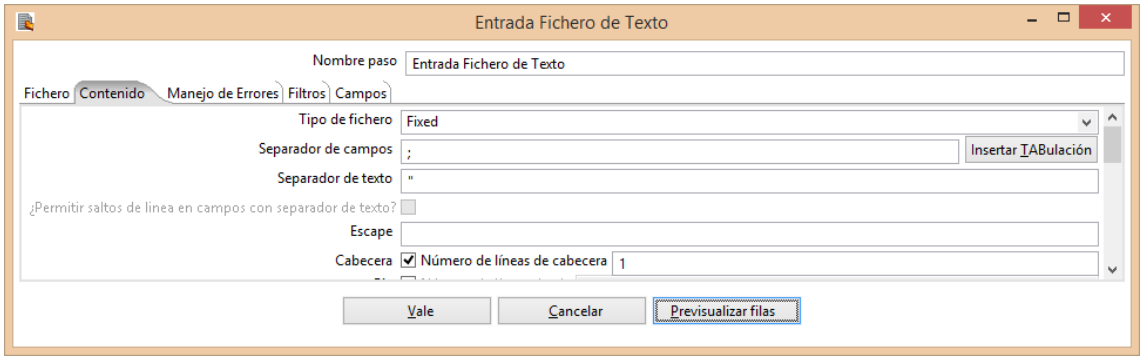
Utilizar un editor de texto y usar la herramienta sustituir, por ejemplo 14 ceros por nada y el punto por la coma. Si el equipo tiene suficiente memoria y potencia quizá lo pueda hacer en un tiempo razonable, en caso contrario el proceso se hace muy pesado.

Solución basada en GeoKettle:

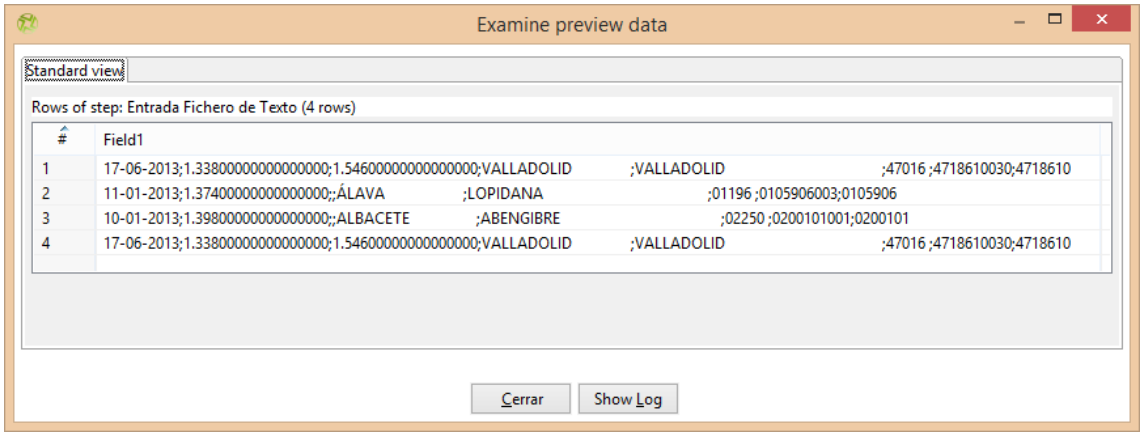
Crear una transformación simple que abra un fichero de texto interpretando toda una fila como un String (cadena de caracteres), reemplace los 14 ceros por nada, reemplace el punto por la coma y escriba la nueva cadena de caracteres en un nuevo archivo de texto.

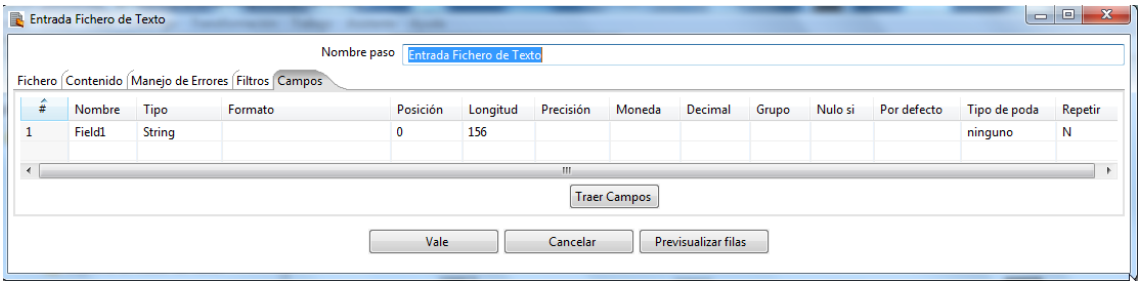


La primera etapa de la transformación, además de conectar con el fichero de los datos iniciales, hay que indicar que toda la línea se lee como una cadena de texto y se le nombra como *Field1*. Para ello tenemos que evitar que el tipo de Fichero se reconozca como un csv para lo que indicaremos que el “Tipo de Fichero” es Fixed.

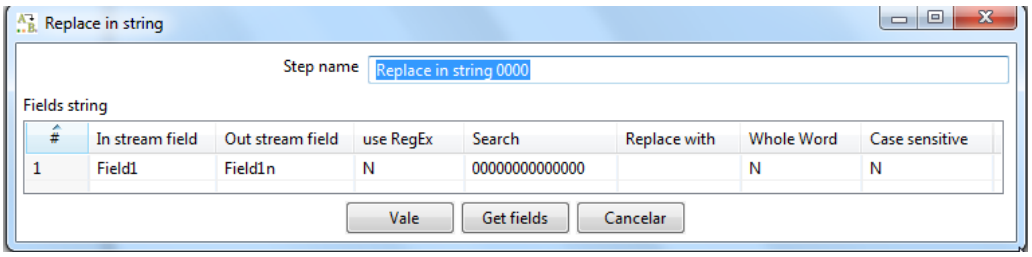


En “campos” definiremos un único Campo Field1 con la posición y longitud adecuadas para que al pulsar previsualizar filas se recupere toda la información de la fila. Ver ilustración.

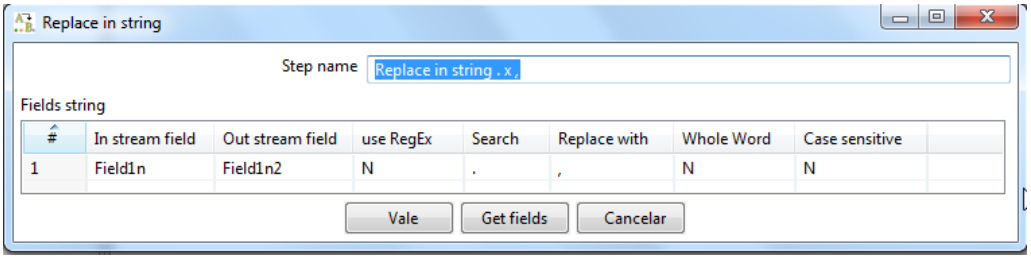




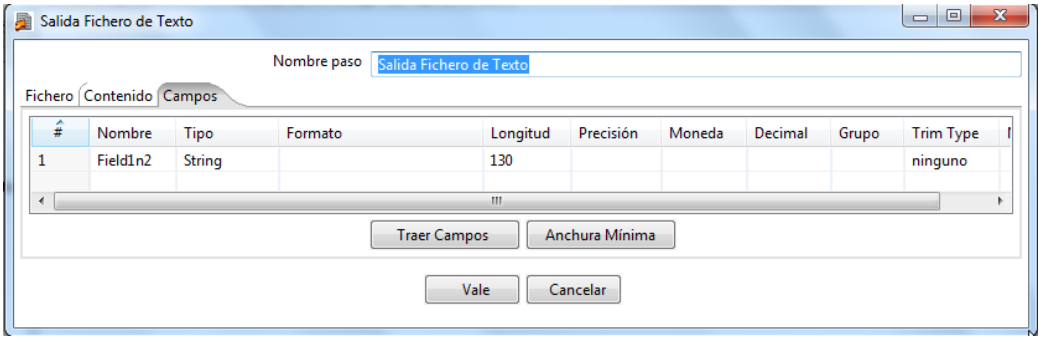
En la primera transformación lo que hay que indicar es cómo se desea llamar a la cadena de datos resultante *Field1n* y los parámetros de cambio *Search* y *Replace with*. Ver ilustración.



En la segunda transformación, de modo similar se cambia el punto por la coma. Ver ilustración.



En último lugar se guarda cada cadena de caracteres *Field1n2* en el archivo de salida. Ver ilustración.


Salida Fichero de Texto
Nombre paso
Salida Fichero de Texto
Fichero
Contenido
Campos
Nombre Tipo Formato Longitud Precisión Moneda Decimal Grupo Trim Type
1 Field1n2 String 130 ninguno
Traer Campos
Anchura Mínima
Vale
Cancelar