Aprovechamiento estadístico de los registros administrativos

BID
Banco Interamericano de Desarrollo

TABLA DE CONTENIDO

EJERCICIOS PRÁCTICOS DEL MÓDULO 2

Fuente: Segui, Federico (2019). Guía práctica sobre el uso estadístico de registros administrativos - Métodos y herramientas para la integración y explotación de registros administrativos con fines estadísticos.

Este ejercicio práctico tiene como propósito que puedas poner en práctica lo visto durante la sección de profundización del módulo 2.

NOTA: Este ejercicio se basa en los ejercicios 1_1 y 1_2 del video tutorial Pentaho (parte 2) de la sección de profundización, pero es un ejercicio diferente.

1. Ejercicios introductorios, conceptos básicos, transformación de variables

Aplicación práctica de conceptos básicos de extracción y carga de datos, transformaciones simples, filtros básicos y diseño de flujos de datos simples. Uso de la interfaz PDI Spoon para familiarizarse con los elementos que la componen.

Ejercicio 1

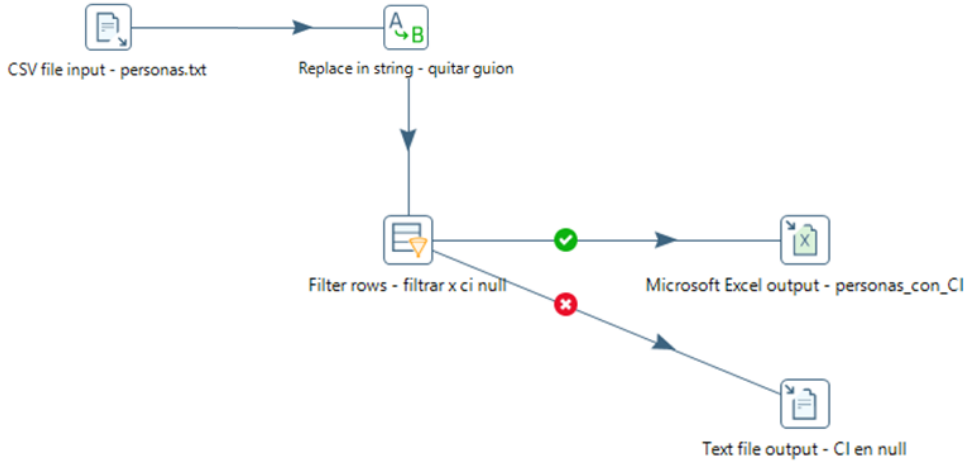
Lectura (extracción) de datos de diferentes fuentes y formatos, depuración de datos (transformación) y escritura (carga) de datos en diferentes formatos.

Se trata de un ejercicio muy simple como para empezar a familiarizarse con la interfaz del software. El archivo de entrada tiene muy pocas filas (casos) y unas pocas variables (columnas), así se pueden ver rápidamente cuáles son los datos “sucios” o inválidos. La idea de este primer ejercicio es “limpiar” los datos de una columna, quitando los caracteres inválidos y filtrar filas o casos para descartar o procesar en otro momento aquellos registros que datos en blanco o nulos. En siguientes ejercicios (y cursos más avanzados) se trabajará con archivos con más filas y columnas donde se irá complejizando el proceso de depuración y transformación de datos.

En Pentaho PDI Spoon crear una transformación que como step de entrada lea el archivo de texto personas.txt (descargarlo de la plataforma del curso), agregar a continuación un step para quitar (o reemplazar por el carácter vacío) el guión “-“ del campo CI (cédula de identidad), filtrar las filas que tengan CI distinto de null y si se cumple esto (conector true) generar una salida en formato Excel con el nombre del archivo de salida: personas_con_CI. En caso contrario (conector false), generar una salida en formato texto con el nombre del archivo de salida: CI_null. Ubicar ambos archivos de salida en su carpeta de trabajo.

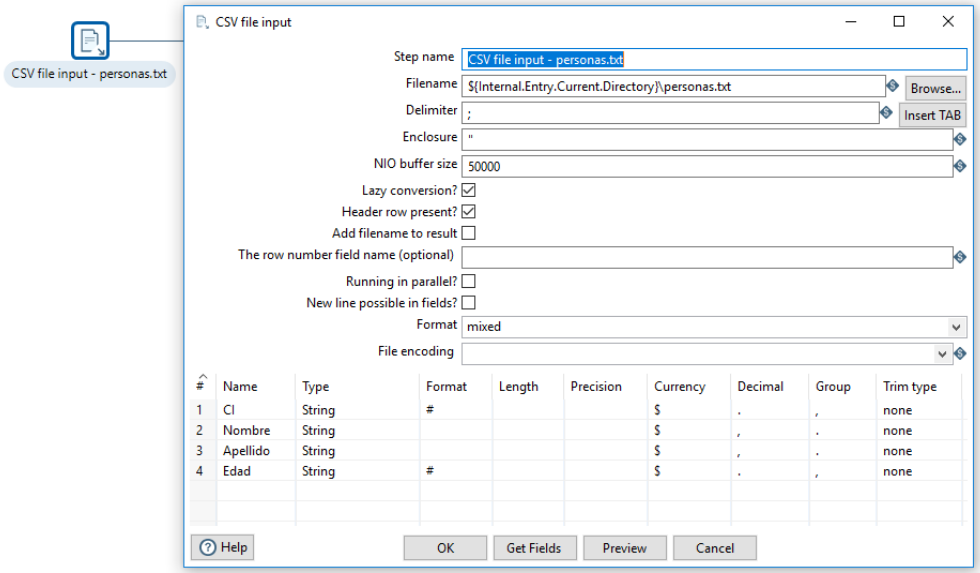
Guardar la transformación en su carpeta de trabajo con el nombre ejercicio1.ktr (recuerde subirlo a la plataforma del curso al finalizar el ejercicio).

Ejecutar la transformación y verificar el resultado en el archivo de salida.

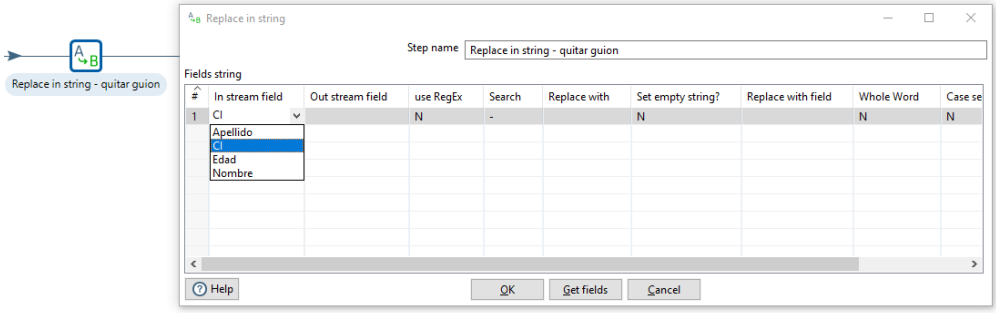


Pasos para crear y ejecutar la transformación:

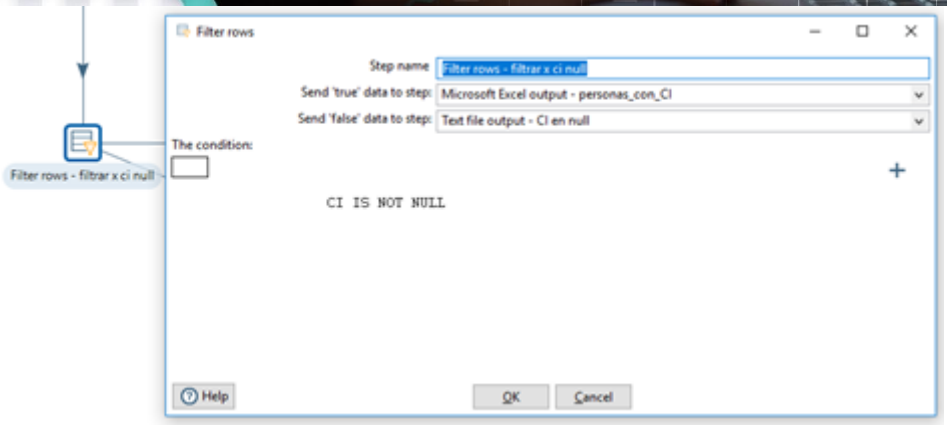
1) En la Barra de herramientas principal, hacer clic en el icono (New file) y seleccionar Transformation.
2) Guardar la transformación. En la Barra de herramientas principal, hacer clic en el icono (Save current file) y en la ventana que se abre indicar ubicación (su carpeta de trabajo en disco) y nombre del archivo de la transformación (ejercicio1). Clic en Guardar (Save).
3) En el Panel de Exploración de la izquierda, en la pestaña Design (Diseño) hacer clic en la carpeta Input (Entrada) para abrirla y ver los steps (pasos) disponibles. Agregar al lienzo el step CSV file input (doble clic sobre el step o arrastrarlo con el mouse hasta el lienzo).
4) Una vez ubicado en el lienzo el step CSV file input hacer doble clic sobre éste para abrirlo y configurarlo.
5) En la ventana de CSV file input indicar como Filename el nombre y ubicación del archivo (hacer clic en Browse… para buscar y seleccionar el archivo personas.txt). Como Delimiter indicar el carácter punto y coma ";". Hacer clic en Get Fields para leer el cabezal del archivo con los nombres de los campos/columnas. Indicar el tipo de datos (Type) de todos los campos como texto (String), dejar las columnas Length (largo) y Precision (cantidad de decimales) en blanco para todos los campos. Establecer el formato (Format) como "#" (numeral) para los campos CI y Edad (pues son campos numéricos enteros). Hacer clic en Preview para verificar que los datos son leídos correctamente, cerrar la ventana de Preview y hacer clic en Ok para terminar.



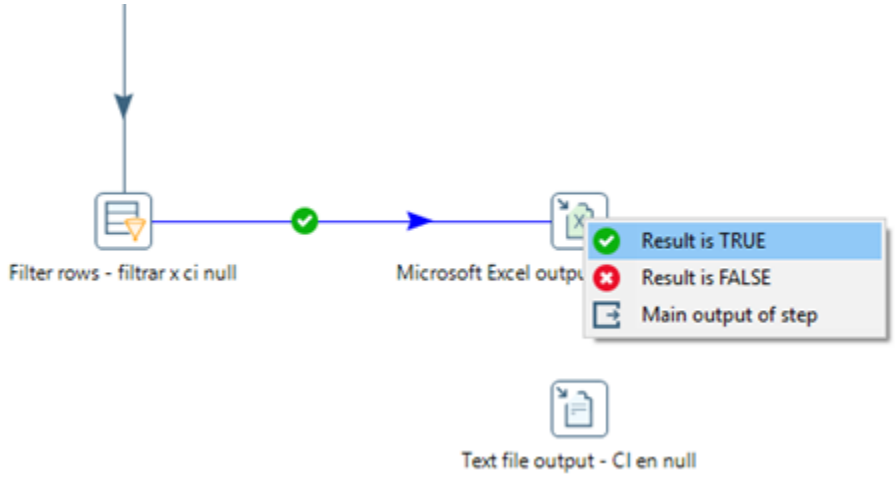
6) Agregar un step Replace in String (Remplazar texto) desde la carpeta Transform (pestaña Design del Panel de Exploración a la izquierda) y conectarlo al anterior step. Hacer doble clic sobre el step para configurarlo.
7) En la ventana de configuración de Replace in string seleccionar el campo CI en la columna correspondiente a In stream field (campo de entrada del flujo de datos). En la columna Search (Buscar) ingresar el carácter a buscar, en este caso el guión "-" y en la columna Replace with (Remplazar con) dejar en blanco para que el guión se remplace por vacío o nada (en este ejercicio el guión será quitado, por eso no se indica nada en Replace with). Clic en Ok para finalizar la configuración del step.



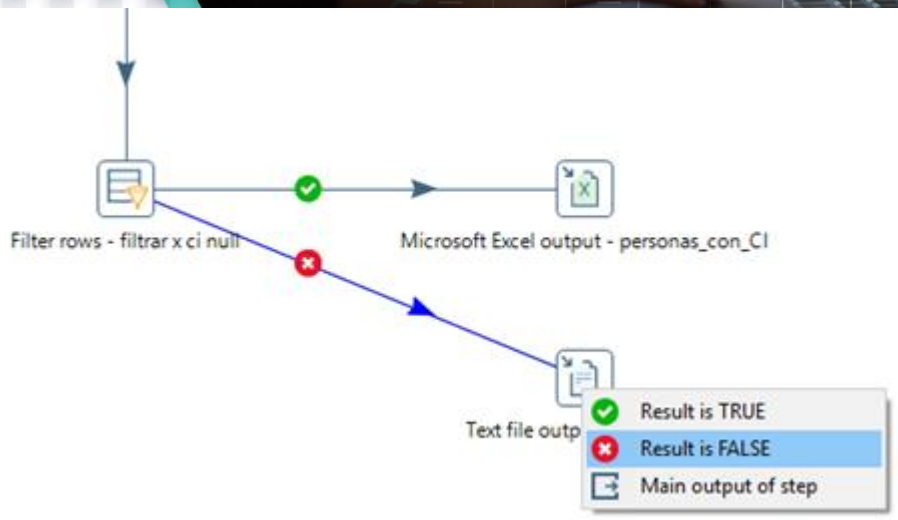
8) Agregar un step Filter rows (Filtrar filas) desde la carpeta Flow (pestaña Design del Panel de Exploración a la izquierda) y conectarlo al anterior step. Hacer doble clic sobre el step para configurarlo.
9) En la ventana de configuración de Filter rows seleccionar el campo CI en la opción correspondiente a <field> (campo de la condición). En la opción de la función de comparación (donde está el signo "=") seleccionar IS NOT NULL (no es nulo). Clic en Ok para finalizar la configuración del step.



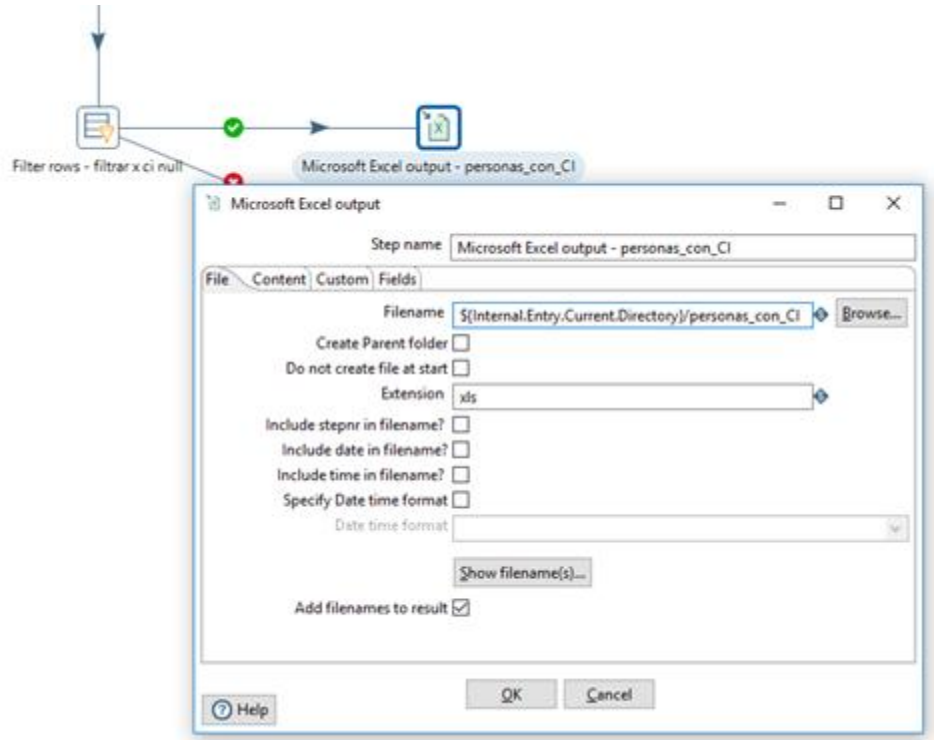
10) Agregar dos step de salida de los datos: un Microsoft Excel Output y otro Text file output.
11) Desde el step Filter rows conectar al step Microsoft Excel Output y en el menú que aparece al soltar el conector en el step de destino, seleccionar la opción Result is TRUE (resultado de la condición de filtro = Verdadero, es decir, cuando el valor del campo CI no es nulo, tiene un dato).



12) Desde el step Filter rows conectar al step Text file output y seleccionar la opción Result is FALSE (resultado de la condición de filtro = Falso, es decir, cuando el valor del campo CI es nulo, sin dato), en el menú que aparece al soltar el conector en el step de destino.

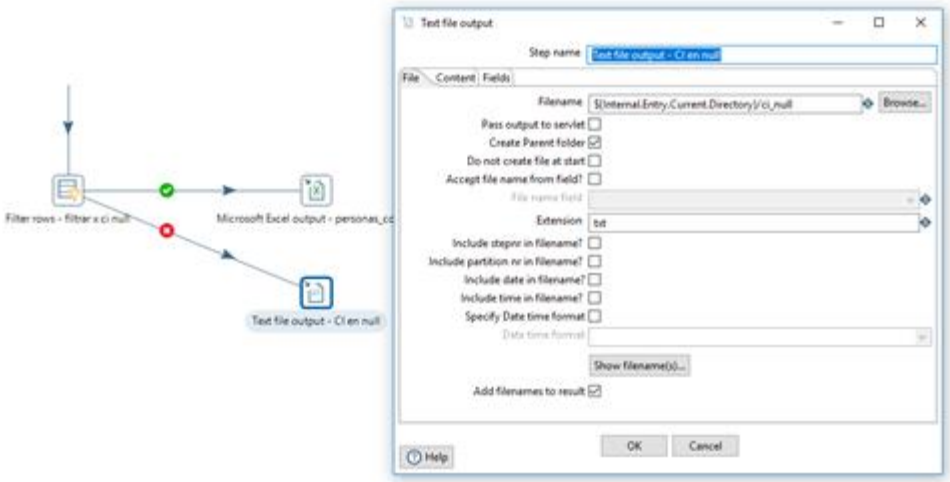


13) Hacer doble clic sobre el step Microsoft Excel Output para abrir la ventana de configuración de la salida a Excel.
14) En la ventana de Microsoft Excel Output indicar la ubicación y nombre del archivo de salida (personas_con_ci). Es posible seleccionar la ubicación haciendo clic en Browse…La extensión del archivo está indicada en Extension, por lo cual no es necesario poner la extensión .xls junto al nombre del archivo en Filename (de otro modo quedará repetido ".xls.xls"). Hacer clic en Ok para confirmar los datos ingresados.



15) Hacer doble clic sobre el step Text file output para abrir la ventana de configuración de la salida a un archivo de texto.
16) En la ventana de Text file output indicar la ubicación y nombre del archivo de salida (ci_null). Es posible seleccionar la ubicación haciendo clic en Browse…La extensión del archivo está

indicada en Extension, por lo cual no es necesario poner la extensión .txt junto al nombre del archivo en Filename (de otro modo quedará repetido ".txt.txt"). Hacer clic en Ok para confirmar los datos ingresados.



Ejecutar la transformación. En la Barra de herramientas de transformaciones hacer clic en el icono ▷ (Run). En la ventana Run Options que se abre hacer clic en Run y comenzará la ejecución de la transformación.