Tutorial Talend Open Studio for Big Data

Cómo crear un proceso para automatizar la normalización tablas de indicadores del banco mundial

1) Descargar los datos fuente

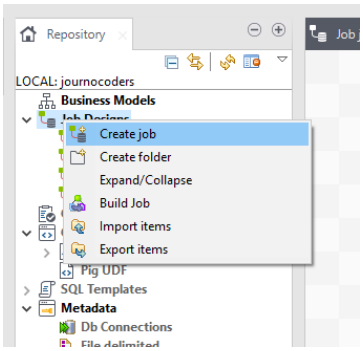
Descargar del Banco Mundial las tablas de todos los indicadores que deseemos procesar y comparar como archivos Excel.

Para el ejercicio usaremos:

Población de refugiados por país o territorio de asilo

Población de refugiados por país o territorio de origen

2) Abrir Talend y crear un nuevo job.

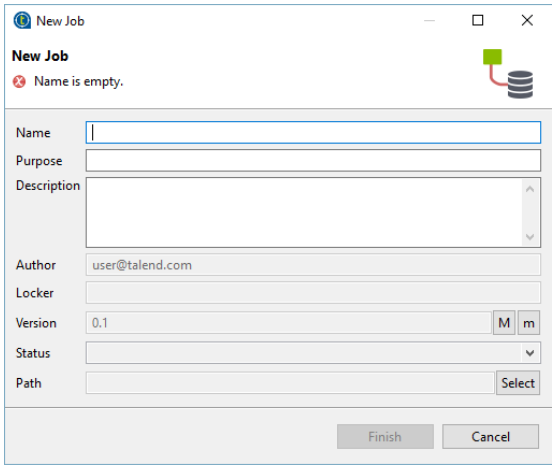


En el panel de la izquierda: Repositorio, en la rama de Job Designs podemos dar clic derecho para que se despliegue un menú de opciones. La primera opción es para crear un nuevo Job.

El nuevo Job necesita al menos un nombre que debe cumplir con los requisitos para nombrar variables en el Lenguaje Java.

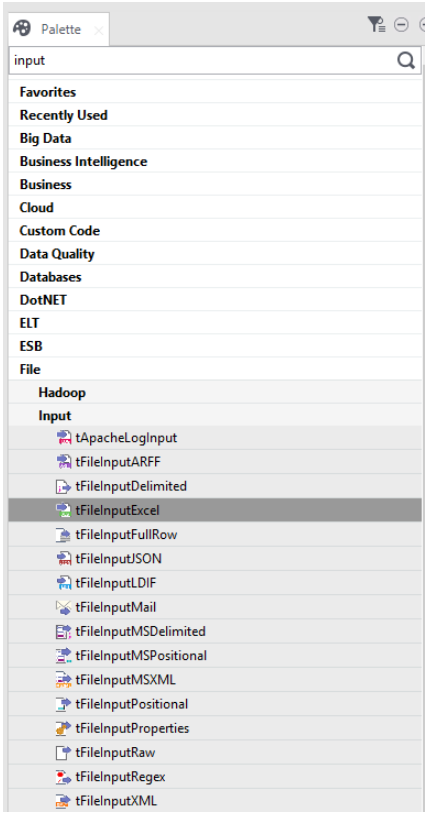
- Alfanumérico pero NO debe comenzar con números.

- No debe contener espacios en blanco
- Se puede usar el caracter _
- Distingue Mayúsculas y minúsculas



3) Extraer los datos.

Buscamos y arrastramos de la paleta al área de trabajo el componente para leer los datos de un archivo Excel: tFileInputExcel

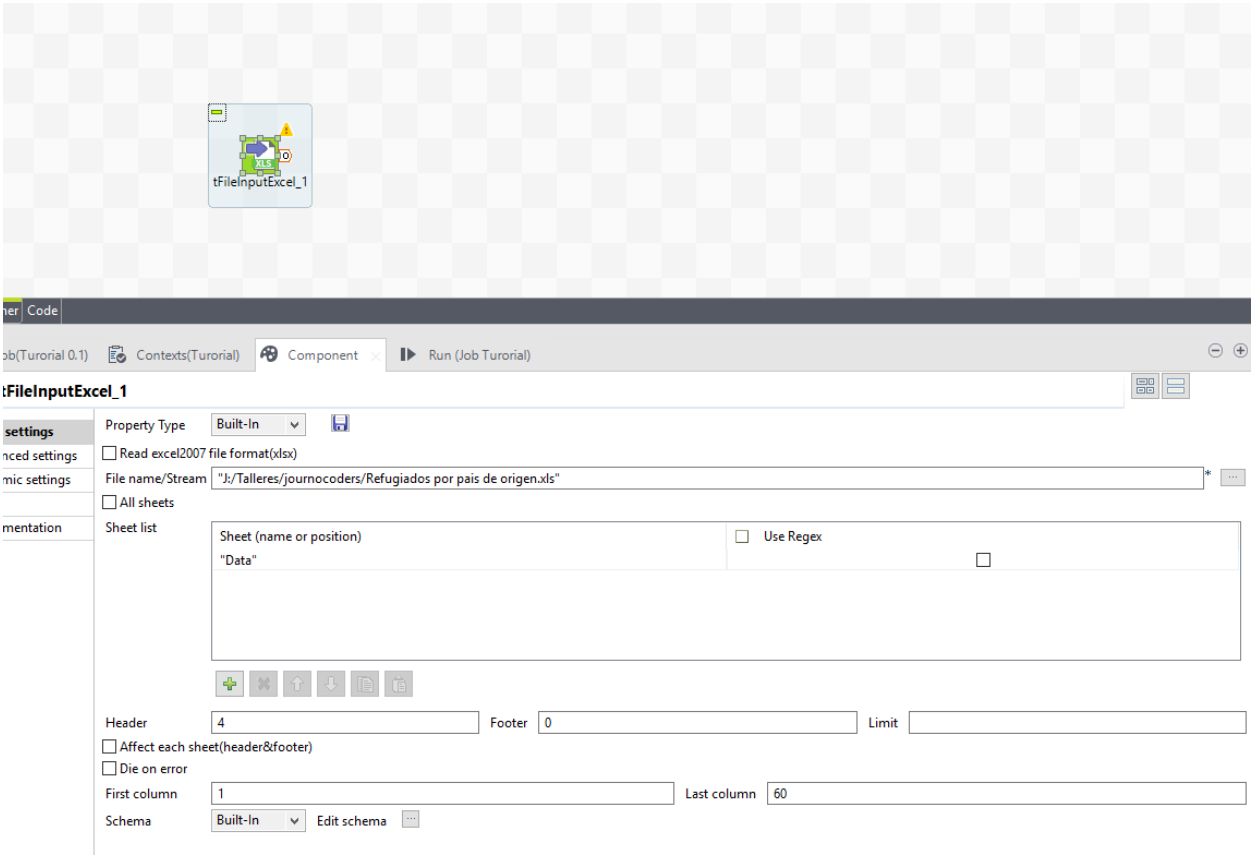


Cuando nos colocamos encima del componente, en el tab Componente en la parte inferior se mostrarán los parámetros de configuración del componente.

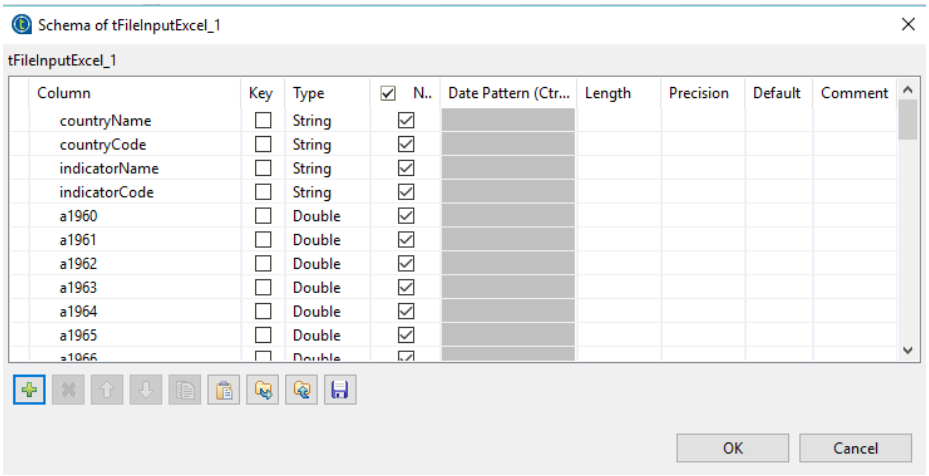
de

Ahí debemos definir:

1) La ruta del archivo Excel a extraer (File name)
2) El nombre de la hoja a extraer (Sheet list)
3) La posición de Inicio del dataset (Header)
4) La última columna a leer (Last column)



Con el botón “Edit Schema” se despliega una ventana donde podemos definir los campos de la tabla y sus metadatos:

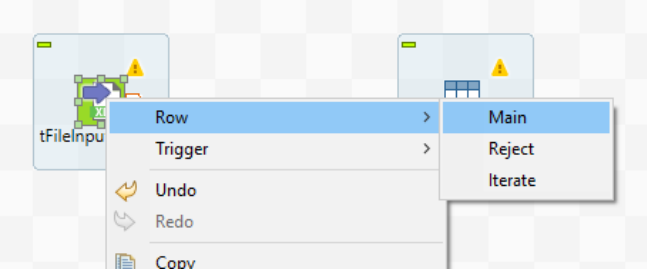


4) Transformar los datos

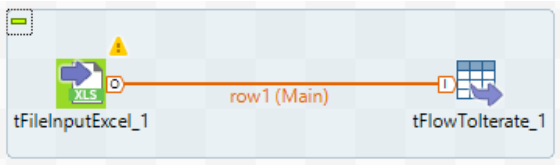
Luego, debemos arrastrar otro componente que se llama TFlowToIterate en la categoría de **Orchestration,**

Ahora vamos a unir esos dos componentes por medio del flujo de datos principal.

Para esto pinchamos con el botón derecho sobre el componente tFileInputExcel y seleccionamos el flujo de datos principal, como muestra la imagen:



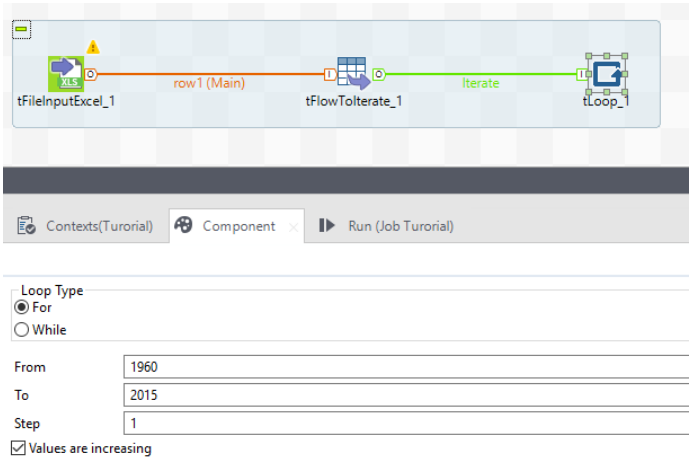
Sin tocar antes nada más, ahora pinchamos el componente tFlowToIterate, de esta manera los componentes quedan ligados:



Y lo que estamos diciendo es: el componente tFileInputExcel va a extraer la tabla del Excel y va a pasar los datos al componente tFlowToIterate.

El componente tFlowToIterate sirve para aplicar un proceso a CADA UNO de los registros en el conjunto de datos que recibe.

A continuación, arrastramos el componente tLoop al área de trabajo y lo conectamos al tFlowToIterate por medio del link **iterate**.

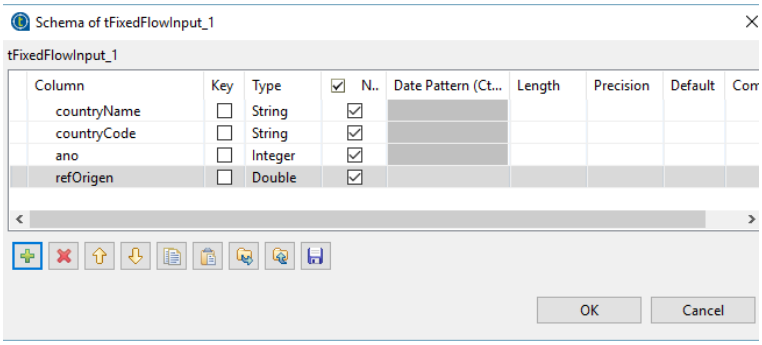


Este componente nos va a servir para aplicar un conjunto de instrucciones varias veces según un valor que se utiliza como "índice", el cual va a comenzar con el valor 1960, que es el primer año de la data hasta 2015.

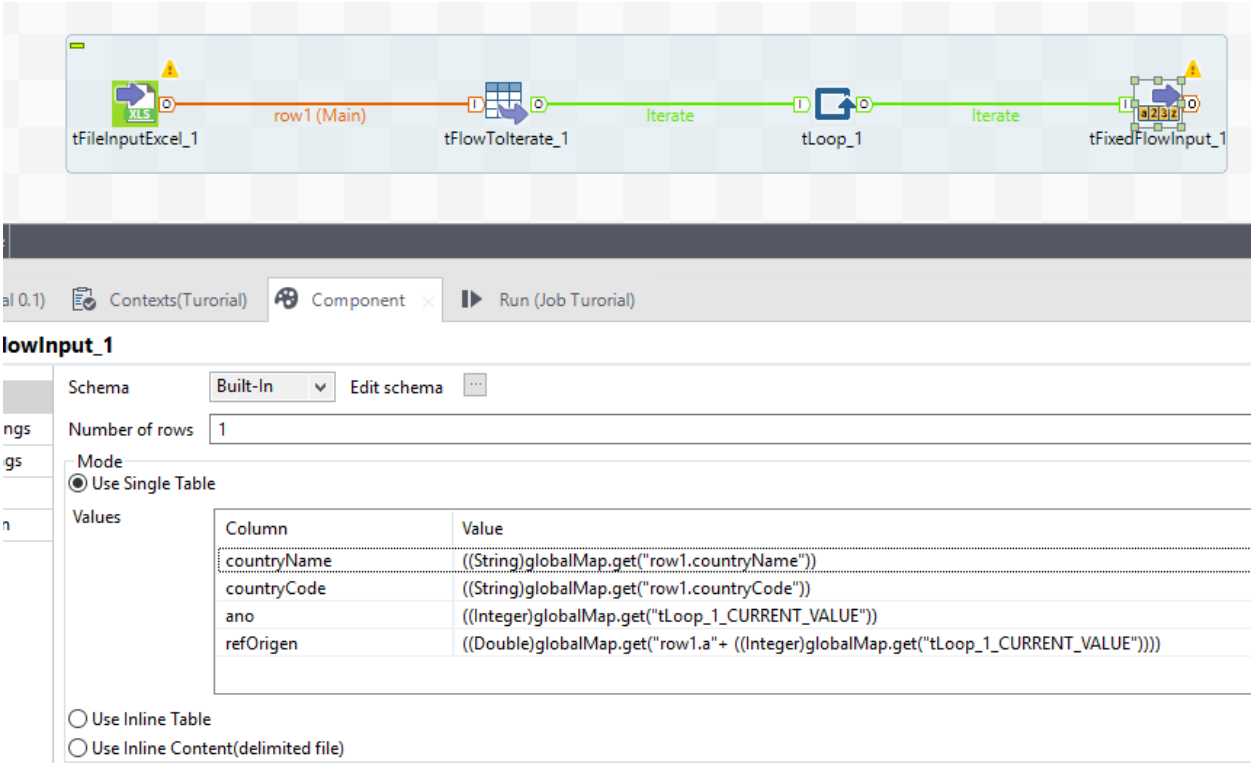
En otras palabras vamos a ejecutar un proceso para cada registro del Excel y para cada año desde 1960 hasta 2015

Seguido, vamos a arrastrar y a ligar un componente llamado tFixedFlowInput, que sirve para crear dinámicamente un flujo de datos según las columnas que definamos y el valor que les asignemos.

Primero vamos a definir el esquema para este componente (Clic sobre "Edit Schema")



Luego, establecemos el valor que va a tener cada columna de este nuevo set:

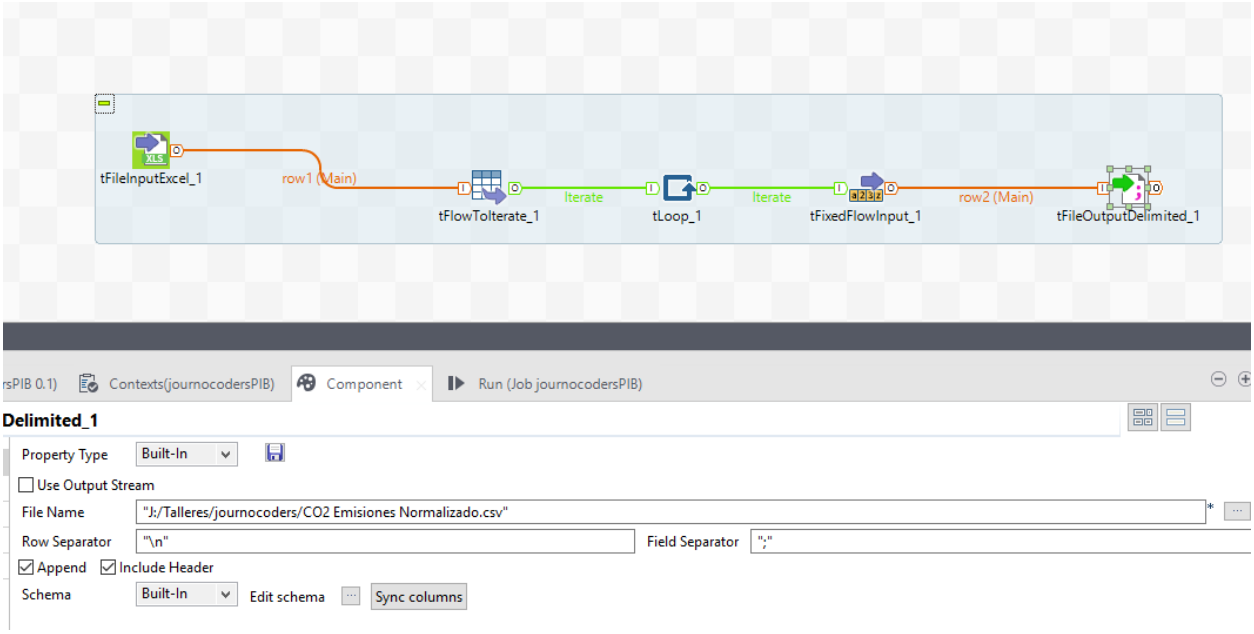


Los valores vienen de los 2 componentes anteriores, son autocompletados, basta con digitar “TFlow” + [TAB] y nos salen los valores actuales que maneja este componente

Para el valor correspondiente al año vamos a utilizar el valor actual del Loop para obtener el valor de la columna deseado.

5) Y Finalmente cargar de los datos

Ahora solo es cuestión de escoger el formato de salida de los datos ya transformados, por ejemplo si queremos generar un csv hay un componente para ésto: tFileOutputDelimited, al cual le enviamos el flujo de datos principal (Main), le definimos la ruta del archivo a generar y los demás parámetros de configuración según muestra la imagen:



Y ya estamos listos para poner a correr el Job, lo hacemos desde el tab “Run” donde hay un botón de “Run”