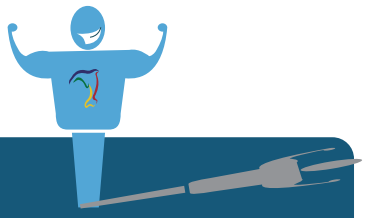
autentia

Avenida de Castilla,1 - Edificio Best Point - Oficina 21B
28830 San Fernando de Henares (Madrid)
tel./fax: +34 91 675 33 06
info@autentia.com - **www.autentia.com**

¿**Qué ofrece** Autentia Real Business Solutions S.L?

Somos su empresa de **Soporte a Desarrollo Informático**.
Ese apoyo que siempre quiso tener...

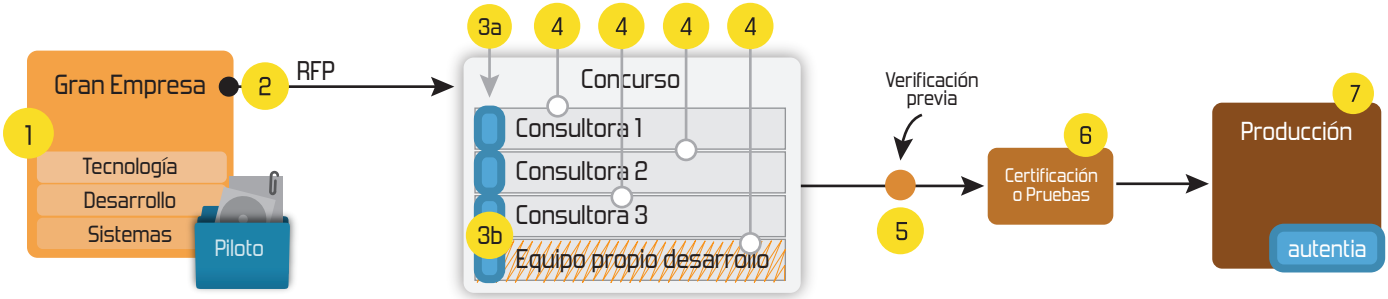
1. Desarrollo de componentes y proyectos a medida



2. Auditoría de código y recomendaciones de mejora

3. Arranque de proyectos basados en nuevas tecnologías

1. Definición de frameworks corporativos.
2. Transferencia de conocimiento de nuevas arquitecturas.
3. Soporte al arranque de proyectos.
4. Auditoría preventiva periódica de calidad.
5. Revisión previa a la certificación de proyectos.
6. Extensión de capacidad de equipos de calidad.
7. Identificación de problemas en producción.



4. Cursos de formación (impartidos por desarrolladores en activo)

Spring MVC, JSF-PrimeFaces /RichFaces, HTML5, CSS3, JavaScript-jQuery

Gestor portales (Liferay)
Gestor de contenidos (Alfresco)
Aplicaciones híbridas

Control de autenticación y acceso (Spring Security)
UDDI
Web Services
Rest Services
Social SSO
SSO (Cas)

JPA-Hibernate, MyBatis
Motor de búsqueda empresarial (Solr)
ETL (Talend)

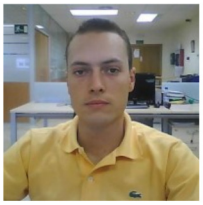
Dirección de Proyectos Informáticos.
Metodologías ágiles
Patrones de diseño
TDD

Tareas programadas (Quartz)
Gestor documental (Alfresco)
Inversión de control (Spring)

BPM (jBPM o Bonita)
Generación de informes (JasperReport)
ESB (Open ESB)

Compartimos nuestro conocimiento en:
www.adictosaltrabajo.com

Para más información visítenos en:
www.autentia.com



Daniel Casanova Frutos

Consultor tecnológico de desarrollo de proyectos informáticos.

Ingeniero Técnico En Informática De Sistemas por la Universidad Alfonso X El Sabio.

Puedes encontrarme en Autentia: Ofrecemos servicios de soporte a desarrollo, factoría y formación

Somos expertos en Java/J2EE

Ver todos los tutoriales del autor

Fecha de publicación del tutorial: 2009-02-26

Tutorial visitado 6 veces Descargar en PDF

Lectura y tratamiento de ficheros Excel con Talend (I): nociones básicas.

0. Índice de contenidos.

1. Introducción

Este tutorial pretende mostrar de manera sencilla la extracción, manipulación e inserción de datos de ciertos tipos de ficheros mediante la herramienta **Talend**.

Talend, como ya se ha enseñado en tutoriales publicados con anterioridad en http://www.adictosaltrabajo.com/, es una herramienta de diseño **ETL (Extract, Transform, Load)**, es decir, su finalidad es extraer datos, transformación y carga de los mismos, a partir de distintas fuentes de datos como pueden ser ficheros, conexiones a distintas bases de datos y un sin fin de posibilidades.

En este tutorial nos centramos en el tratamiento de información sobre ficheros del tipo **XLS** (tipo de ficheros excel).

En un primer paso del tutorial veremos de manera rápida donde podemos obtener la herramienta Talend, así como una rápida instalación de ésta.

A continuación analizaremos como extraer y manipular datos de ficheros del tipo XLS.

2. Entorno.

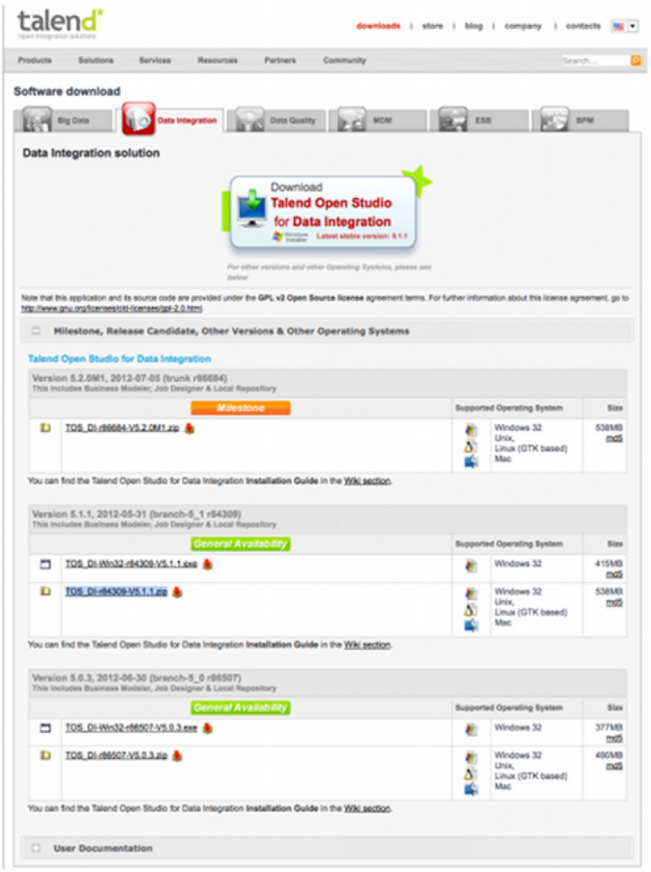
El tutorial está escrito usando el siguiente entorno:

- Hardware: Portátil MacBook Pro 15' (2.2 GHz Intel Core i7 Duo, 8GB DDR3 SDRAM).
- Sistema Operativo: Mac OS X Lion 10.7.3 (11D50d)
- Talend Open Studio V5.1.1 r84309

3. Instalación de la herramienta Talend.

Veamos en unos sencillos pasos como instalar la herramienta en un entorno MAC. Instalaremos la versión 5.1.1, versión utilizada durante el desarrollo del tutorial.

- 1. Lo primero de todo es descargar la aplicación, para lo que acudiremos a la página http://www.talend.com/
- 2. Pinchar en el enlace superior 'downloads'
- 3. Abrir pestaña 'Milestones, etc' y descargar el fichero ZIP **'TOS_DI-r84309-V5.1.1.zip'**



Últimos Tutoriales del Autor

» JMeter. Uso de funciones.

» JMeter. Gestión de usuarios

» JMeter y JSF. Extracción del parámetro ViewState

» Instalación de Glassfish 2.1

» Blender y JMonkeyEngine. Exportación de archivos Blender y uso de los mismos en JMonkeyEngine

- 4. Descomprimir y ejecutar 'TOS_DI-macosx-cocoa'

Tras seguir unos sencillos pasos la aplicación TALEND quedará instalada. Versiones anteriores de Talend no se distribuyen en su web. En caso de querer una versión anterior, se ha de perdir por mail a uno de los administradores de la misma.

4. Tratamiento de ficheros Excel (XLS) mediante la herramienta Talend.

En este apartado veremos como recuperar y manipular mediante la herramienta Talend un fichero con formato Excel, de manera más concreta con extensión xls, que como ya sabemos es el estándar de ficheros Excel en versiones anteriores o iguales a Excel 2003. Como veremos, los componentes utilizados de la herramienta Talend, también sirven para el tratamiento de ficheros Excel con versiones posteriores.

Realizaremos un ejemplo que analice el fichero xls y muestre por pantalla el contenido del mismo.

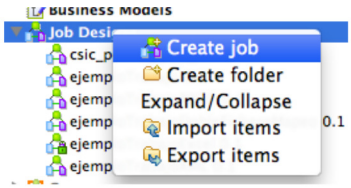
Para realizar la lectura y mapeo del fichero XLS utilizamos uno de los componentes de ficheros de entrada (File/Input) denominado **tFileInputExcel.**
Dicho componente básicamente lee las celdas del fichero xsl, fila por fila.

Partimos de un ejemplo sencillo xls 'Libros.xls' con el siguiente contenido:

	A	B
1	**TITULO**	**Resumen**
2	Titulo Libro1	Resumen del libro 1
3	Titulo Libro2	Resumen del libro 2

Como podemos observar el contenido del fichero son dos columnas con dos filas cada una, representando cada una el título y resumen de un libro.

Para comenzar con el ejemplo creamos un nuevo **Job** (concepto de un trabajo a realizar en la herramienta Talend). En la parte izquierda de la pantalla:

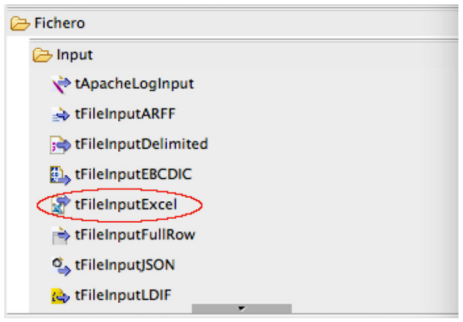


Y le damos un nombre, por ejemplo **'Job1'.**

Ya tenemos listo nuestro entorno de trabajo en Talend para comenzar a diseñar.

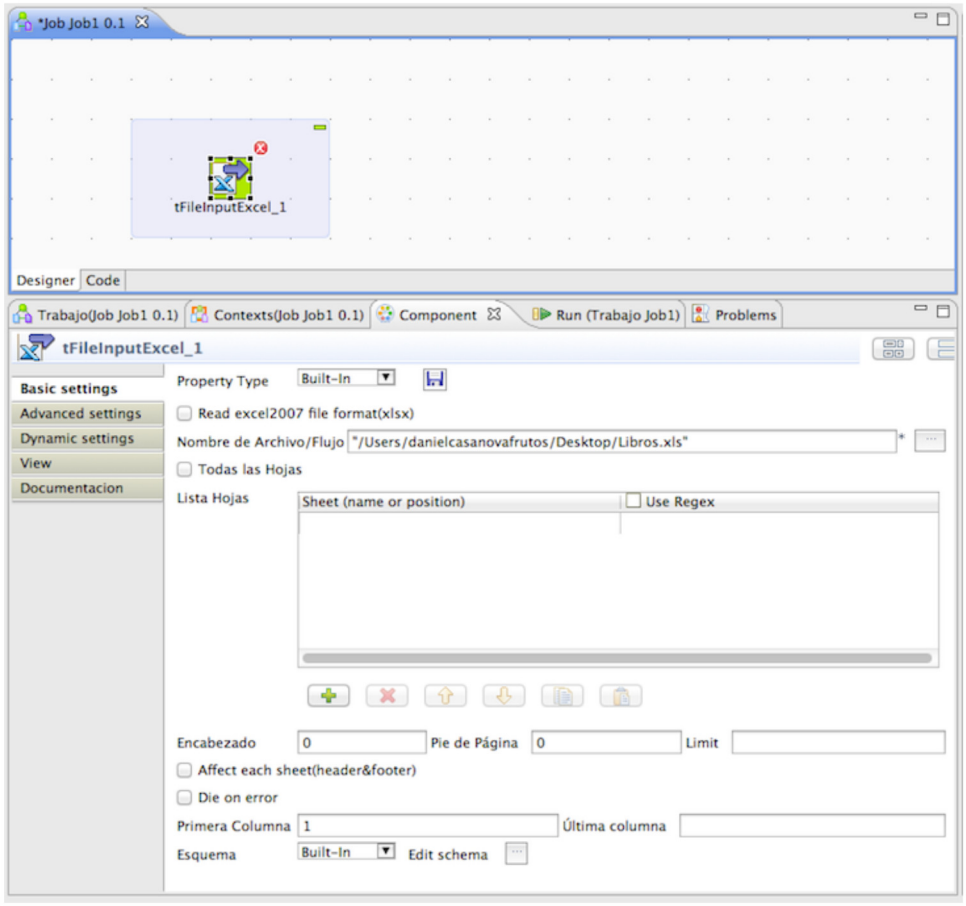
En la parte derecha de la pantalla podemos ver una **paleta** con todos los componenentes que Talend nos ofrece, agrupados por grupos de componentes, representando conceptualmente la funcionalidad de los mismos. En nuestro ejemplo nos interesan aquellos bajo la agrupación '*Fichero*', es decir, componentes preparados para realizar acciones sobre ficheros.

Seleccionamos en la paleta de componentes el componente **'tFileInputExcel'** y lo arrastramos a la zona donde realizaremos el diseño del sistema, también llamada diagrama de trabajo.



Como su nombre indica dicho componente se utiliza para la lectura de un flujo de de datos de un determinado fichero excel.

Al realizar un doble click sobre el componente accedemos a su confguración:



Como podemos observar las preferencias básicas de este componente son:

- **Property type:** Permite seleccionar si las propiedades del componente se crean de manera dinámica, o a través de metadatos ya definidos con anterioridad al trabajo
- **Read Excel 2007 file format (xlsx):** Marcar en caso de que el fichero a analizar pertenezca a una versión de Excel de 2007 o superior
- **Nombre del archivo/Flujo:** Ruta física donde se encuentra el fichero xls
- **Todas las Hojas:** Indica si se deben analizar todas las hojas del fichero xls (nuestro ejemplo del tutorial sólo tiene una hoja)
- **Lista Hojas:** Permite mapear que hojas y cuáles no se analizan
- **Encabezado:** Permite indicar qué fila del fichero Excel representa el encabezado del mismo
- **Pie de Página:** Permite indicar qué fila del fichero Excel representa el pie de página del mismo
- **Limit:** Limite de resultados del encabezado o pie
- **Affect each sheet:** Permite elegir si la definición de los límites y filas del encabezado y pie afecta a todas las páginas del fichero xls

- **Die on error:** En caso de producirse algún tipo de error de ejecución del componente, no continuar con el trabajo
- **Primera Columna:** Número de la primera columna del fichero xls. Permite escoger cualquier columna como la primera
- **Última Columna:** Número de la última columna del fichero xls. Permite escoger cualquier columna como la última
- **Edit schema:** Permite editar el esquema o flujo de salida (schema) del componente en caso de que sea del tipo 'Built-In', es decir, definición del esquema de salida en el propio componente, o seleccionar el esquema de un repositorio de esquemas creados con anterioridad mediante la opción 'Repository'

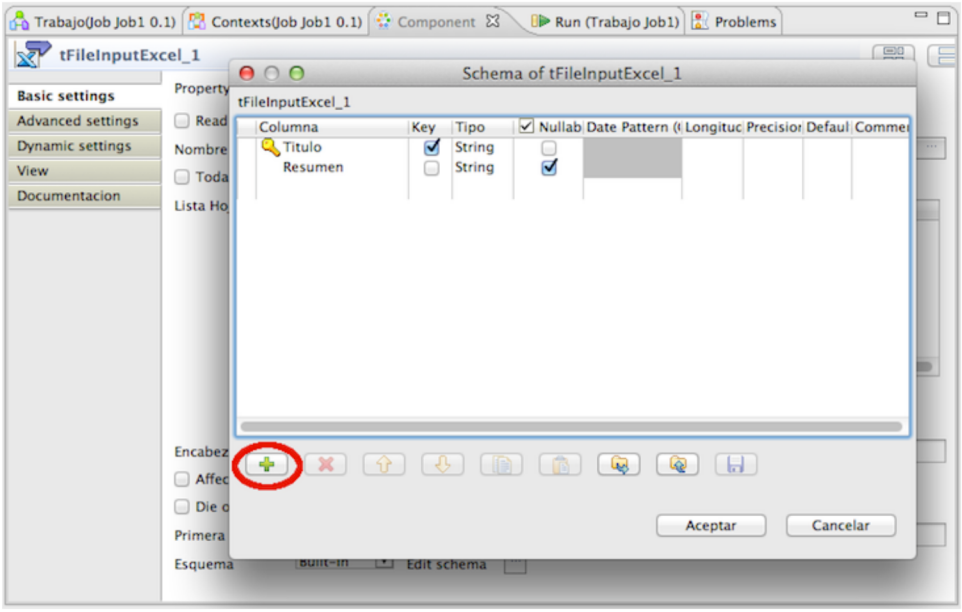
Todo componente de Talend tiene unas preferencias avanzadas. Podemos verlas pulsando sobre 'Advanced settings'. Las preferencias avanzadas del componente tFileInputExcel en cuestión son:

- **Separador avanzado (para números):** Checkeando esta opción, nos permite definir los separadores para delimitar los números decimales y aquellos donde se diferencien los millares
- **Trim all columns:** Elimina del flujo de salida del componente los datos de todas las columnas de nuestro fichero
- **Check column to trim:** Nos permite selecionar que columnas de nuestro fichero serán excluidas del flujo de salida del componente
- **Codificación:** Tipo de codificación de los datos del fichero Excel
- **Read real values for numbers:** Leer la parte real de los números
- **Stop to read on empty row:** Se para la lectura del fichero cuando se encuentre una fila sin datos en el mismo
- **Ignore the warnings:** Indica que no vuelque información del tipo 'warning' a ningún log ni componente paralelo
- **tStatCatcher Statistics:** Volcar información en el log que se registra en un componente especifico para ello llamado tStatCatcher

Después de indicar cuál es el fichero fuente de la información mediante la preferencia **'Nombre del Archivo/Flujo'**, e indicar mediante la preferencia **'Todas las hojas'** que lea todas las hojas del fichero xsl, nos disponemos a editar el esquema de salida del mismo. Talend al realizar la lectura del fichero xls no genera un esquema de salida por defecto, sino que nos obliga a indicarle por nosotros mismos el esquema de salida, es decir, que columnas y filas, y qué información, queremos en la lectura del fichero. Para ello pulsamos el botón **'Edit schema'** de la parte inferior:

A continuación definimos el **esquema de salida**.

En un primer ejemplo creamos un esquema de salida que represente un objeto o flujo de datos, con ambas columnas del fichero xls y los datos de la mismas. Para ello vamos añadiendo columnas mediante el botón '**+**', y definiendo el nombre de la columa de salida, el tipo de salida, si será o no clave primaria mediante el ckeck 'key', y si dicha columna puede tener valores nulos o no mediante la opción 'Nullable':

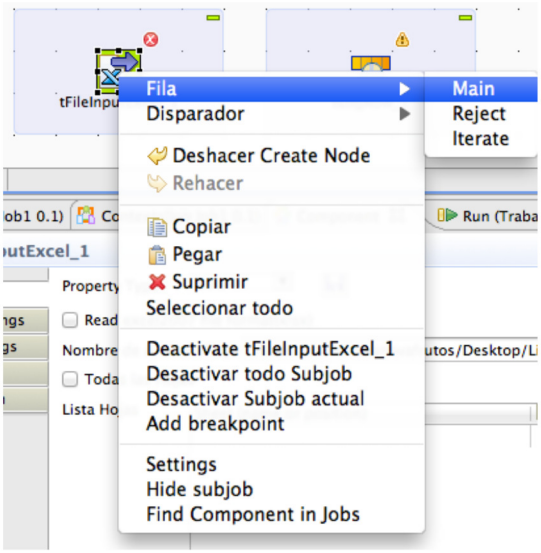


Como podemos ver definimos dos columnas de salida. Una primera con el nombre **'Titulo'**, la cual es key, es decir, es la clave primaria del esquema de salida, y además no puede tener valores nulos. Y una segunda columna **'Resumen'** que puede tener valores nulos.

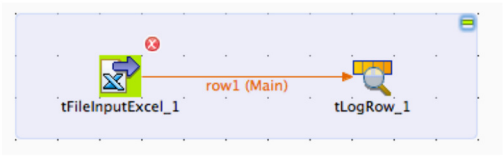
A continuación utilizamos en nuestro trabajo un componente de log que nos trace la salida del esquema por consola, de esta manera vemos si el resultado del flujo de salida es el que esperabamos. Dicho componente recibe el nombre de **'tLogRow'**. Seleccionamos el componente en la paleta de componentes y lo arrastramos a nuestro diagrama de trabajo.

A continuación propagamos el flujo de salida del componente **tFileInputExcel** hacía el componente **tLogRow**. Esto es una acción muy típica en la herramienta de Talend, que no explicamos en profundidad en este tutorial ya que no es el objetivo del mismo, pero básicamente consiste en propagar el esquema de salida con la información correspondiente al siguiente componente, de manera que el siguiente componente tendrá como esquema de trabajo dicho flujo.

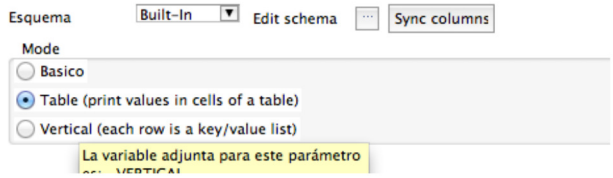
Para ello realizamos un click derecho sobre el componente tFileInputExcel, seleccionando 'Fila/Main' (basicamente creamos un flujo de salida principal)



A continuación pinchamos el flujo de salida, representado por una línea, sobre el componente **tLogRow**. Esto nos indica con una flecha roja el flujo de salida, así como la dirección del flujo:

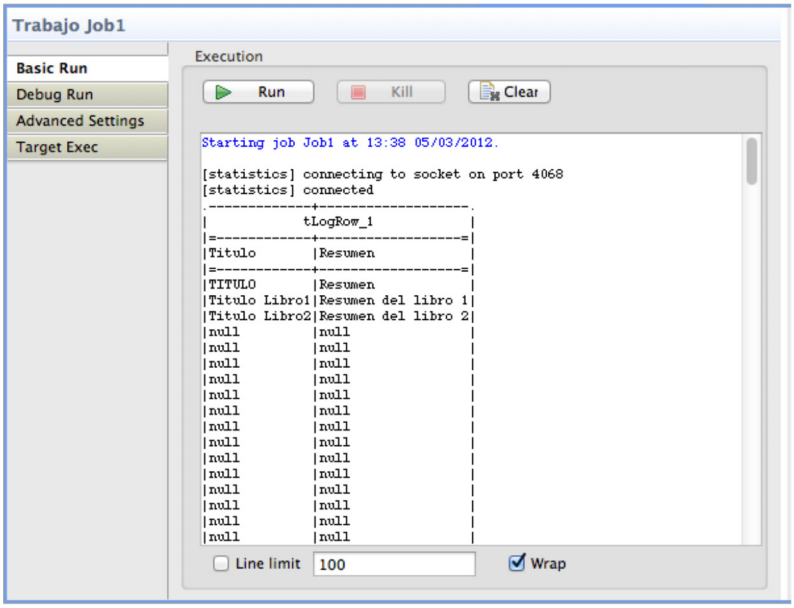


En el siguiente paso configuramos el componente tLogRow para que mueste los resultados en forma de tabla en pantalla, que es una manera más legible de ver el flujo de salida:



El útimo paso es ejecutar nuestro trabajo mediante el botón **'Play'** de la barra de herramientas superior

Veremos en la salida de consola que el resultado del flujo de salida es el esperado:

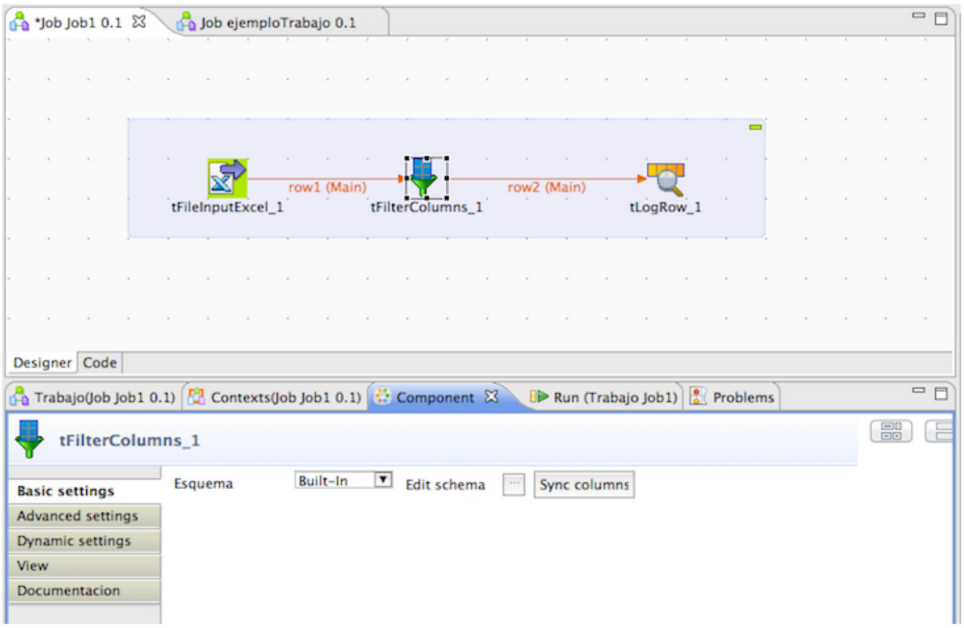


Básicamente podemos observar que tenemos una primera columna en el flujo de salida con el nombre **'Titulo'** y el contenido de la columna titulo del Excel (incluyendo el valor de la primera fila), y una segunda columna con el nombre **'Resumen'** y el contenido de la columna Resumen del fichero xls.

Veamos a continuación dos componentes básicos para aplicar condiciones a las columnas y filas de un flujo de datos. Dichos componentes son **tFilterColumns**, para filtrar columnas, y **tSortRow** para filtrar filas.

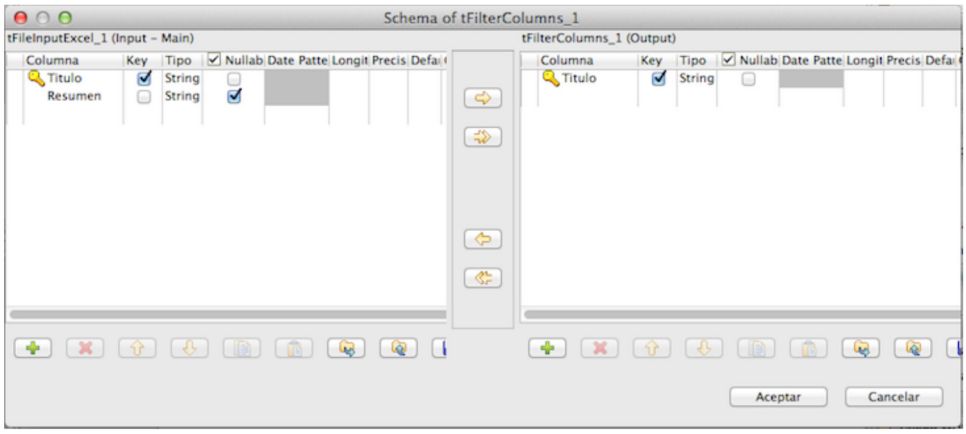
El componente tFilterColumns permite filtrar las columnas del flujo que recibe como entrada. Para añadir dicho componente lo arrastramos de la paleta a nuestro diagrama y eliminando el flujo que salía del componente tFileInputExcel (mediante click derecho sobre la flecha de flujo y suprimir), creamos un nuevo flujo sobre el componente tFileInputExcel y lo redirigimos hacia

tFilterColumns. A continuación creamos un flujo de salida desde el componente tFilterColumns y lo redirigimos hacia el componente de log. En resumen hemos puesto entre medias del fichero y el log, el componente de filtrado:



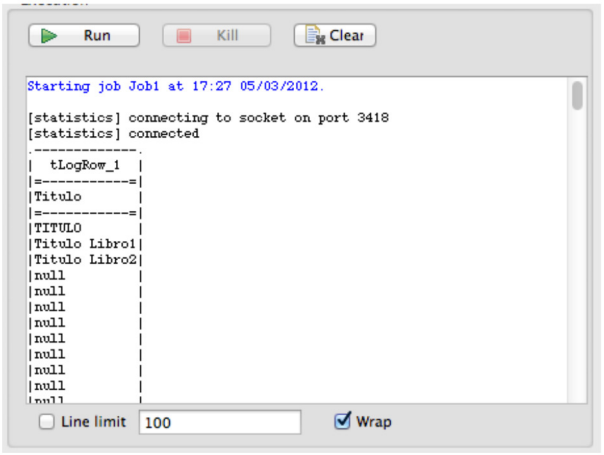
Muchas veces cuando eliminamos flujos de un componente y lo redirigimos hacia otro, Talend nos pregunta *'¿do yoy want to get the schema of the target component?'*, esto quiere decir si queremos propagar el esquema del componente origen al siguiente componente, de manera que no haya que redifinir el esquema en el componente destino.

Como vemos este componente es muy sencillo y básicamente editamos su esquema (opción 'Edit schema') y elegimos que columnas y cuales no irán en el flujo de salida. En este caso por ejemplo eliminamos la columna de resumen:



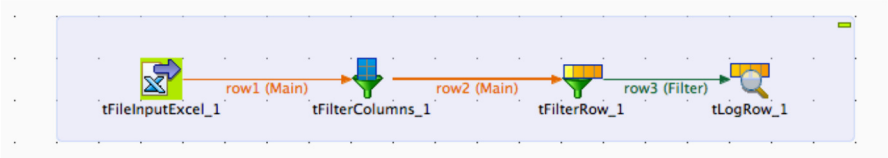
Podemos ver en la parte izquierda el flujo que corresponde a la entrada, y en la derecha el flujo de salida. Podemos elegir que pase o no una columna al flujo de salida, o bien que pase a ser de otro Tipo, por ejemplo conversión de String a Int...

Ejecutamos el trabajo mediante el botón play como hemos visto con anterioridad y vemos como el flujo de salida consta sólo de la columna Titulo con sus valores correspondientes:

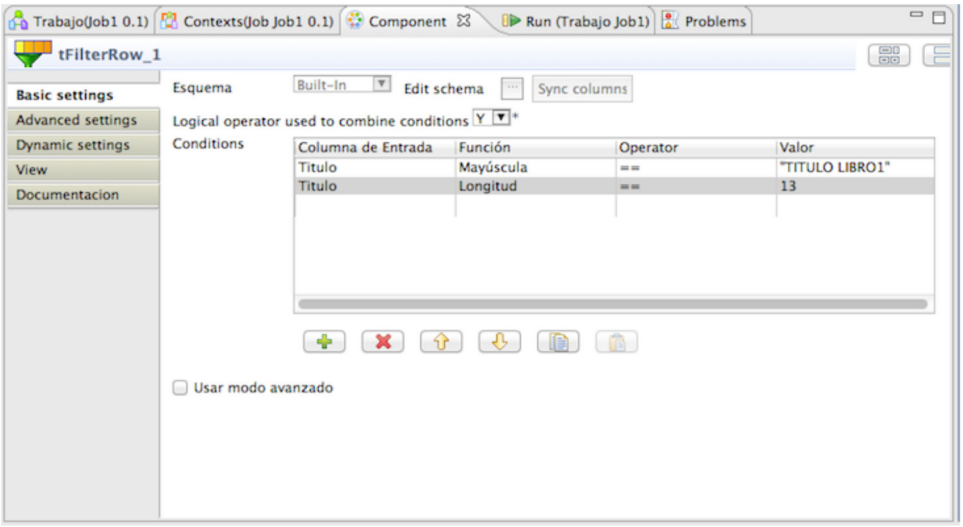


A continuación introducimos el componente **tFilterRow** que permite añadir condiciones para filtrar las filas del flujo resultante.

Para ello arrastramos el componente desde la paleta entre medias como hicimos con el anterior componente de filtrado, resultando:



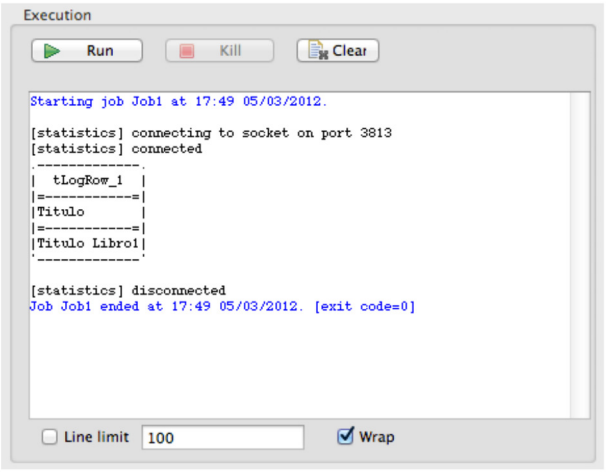
Configuramos el componente tFilterRow de la siguiente manera:



Como podemos ver las opciones básicas de configuración son:

- **Logical operator used to combine conditions:** Puede tomar los valores Y/O, y son los valores lógicos que utiliza para combinar las condiciones. El valor 'Y' indica que deben cumplirse todas las condiciones para que el valor pase al flujo de salida, y el valor 'O' indica que con cumplir una de las condiciones, el valor se añade al flujo de salida.
- **Conditions:** Es la tabla donde definimos las condiciones que deben cumplir los datos de una columna para dejarlos pasar al flujo de salida. Como podemos observar hemos añadido la condición de que el valor en mayúsculas sea igual que el valor 'TITULO LIBRO1', y otra condición que indica que la longitud del dato debe ser igual a 13.

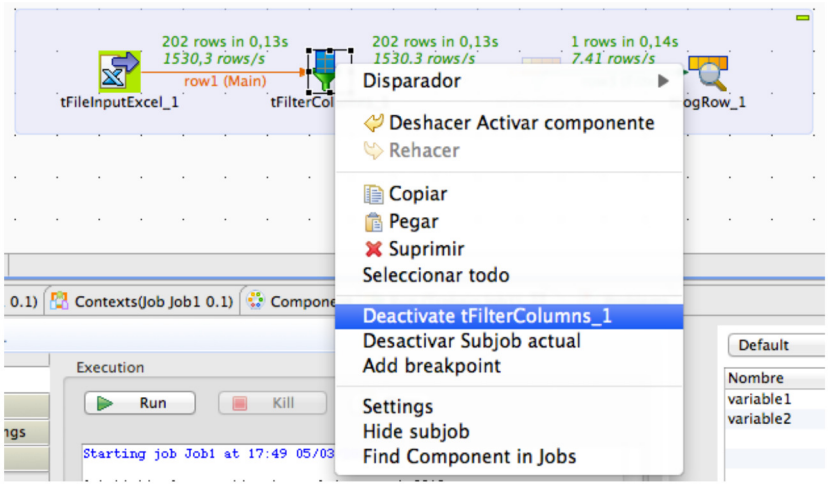
Al ejecutar el trabajo bajo estas condiciones obtenemos el resultado:



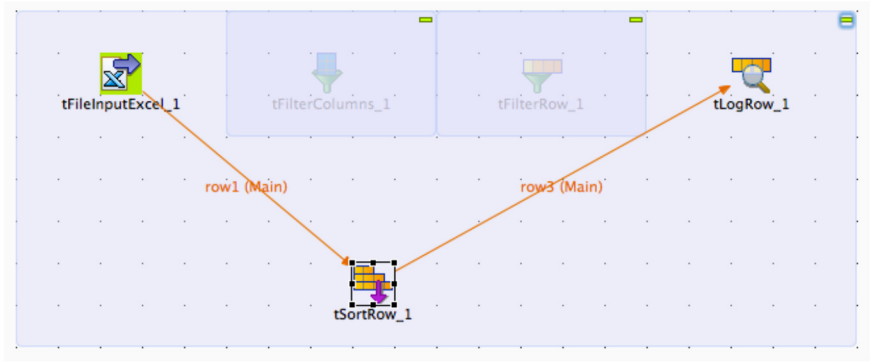
Como podemos observar en el flujo de salida se encuentra el único resultado de la columna Titulo que cumple ambas condiciones.

Por último vamos a ver un componente de ordenación de resultados muy útil cuando parseamos ficheros xls. Dicho componente es **'tSortRow'** y como su nombre indica sirve para ordenar resultados en el flujo de salida.

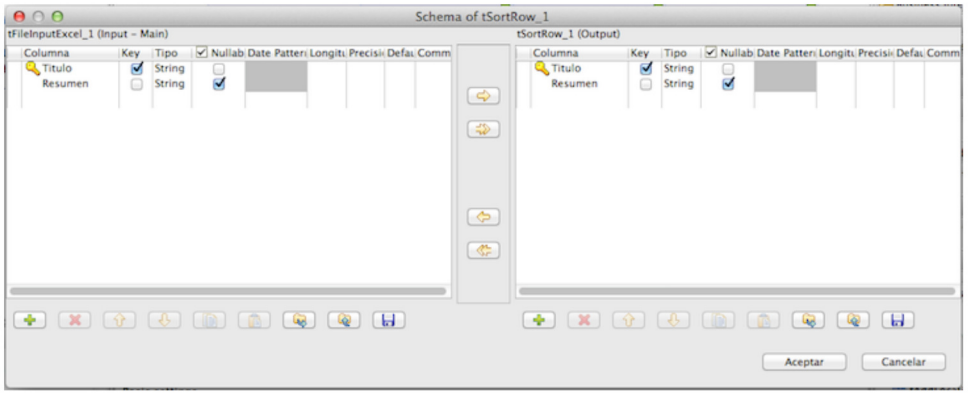
Lo que haremos será eliminar los flujos creados, desactivar los componentes de filtrado para que no influyan en el trabajo y colocar el componente tSortRow entre la salida al log y el componente de entrada del fichero xls. Para desactivar un componente basta con hacer click derecho sobre el mismo y marcar la opción 'Desactivate':



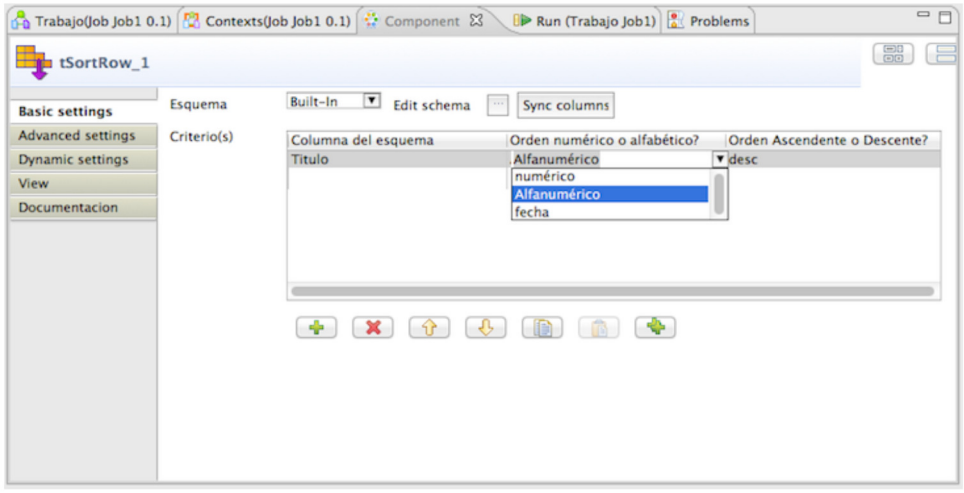
Una vez colocado el componente de ordenación resulta el diagrama:



Editamos el esquema del componente tSortRow.
Como vemos realizamos un simple mapeo de la información hacía el flujo de salida. Como cualquier esquema permite añadir o eliminar columnas hacía el flujo de salida.



Lo interesante de este componente se encuentra en las preferencia 'Criterio(s)':

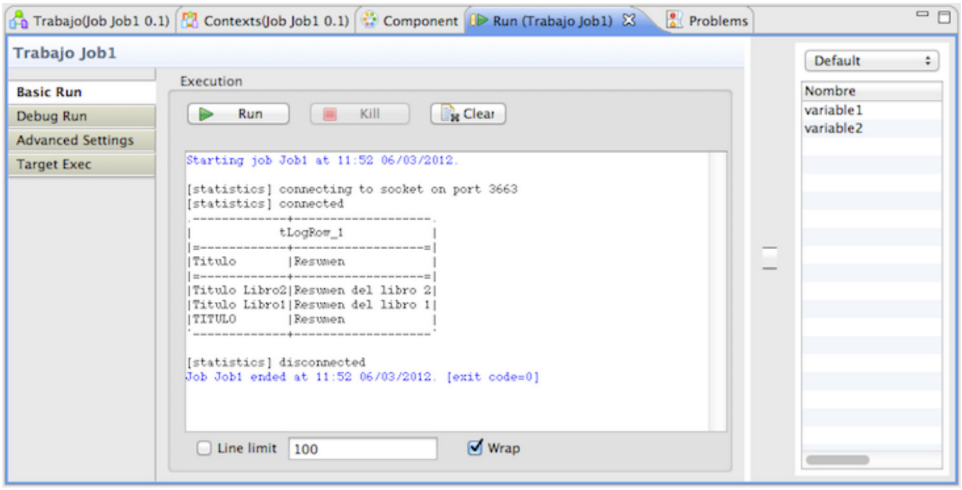


Este preferencia sirve para establecer criterios de ordenación sobre una o varias columnas del flujo de salida.

Permite seleccionar un orden númerico, alfanumérico o de fecha, y si el orden debe ser ascendente o descendente.

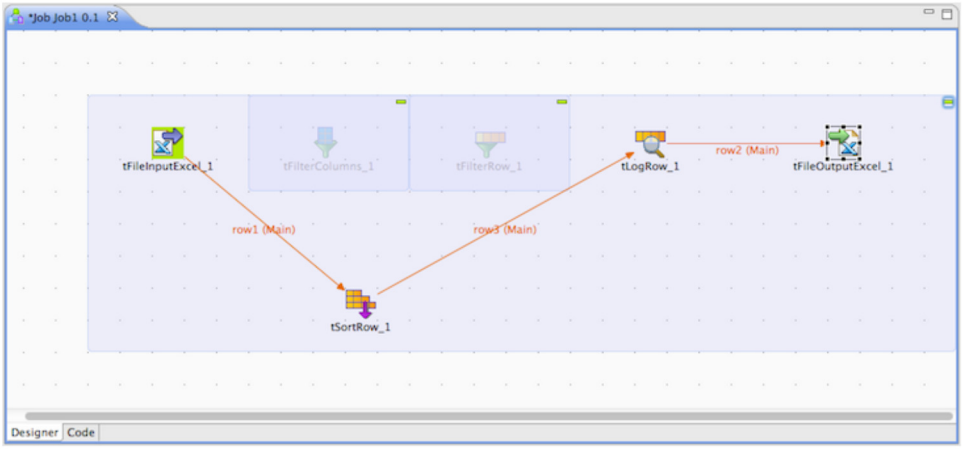
En este caso por ejemplo aplicamos una ordenación alfanumérica descendente sobre la columna Resumen. Al ejecutar el trabajo

podemos observar la correcta ordenación:

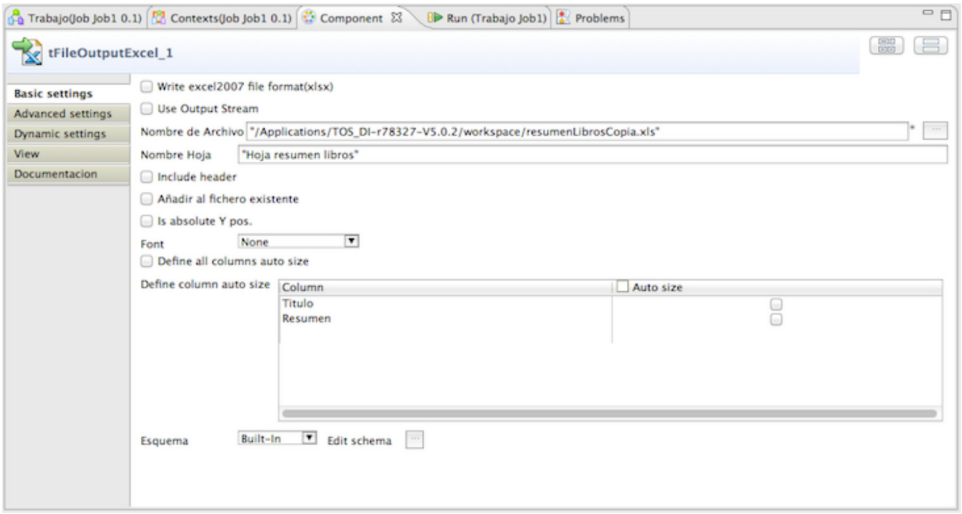


Una vez hemos extraido y manipulado la información proveniente del fichero Excel, nos es de interés volcar o cargar dicha información en otro fichero Excel, cualquier otro tipo de fichero, o incluso en una base de datos.

Veremos como cargar dicha información en otro fichero Excel mediante el componente **tFileOutpuExcel**.
Colocamos dicho componente en nuestra zona de trabajo y la salida de flujo del componente tLogRow1 la apuntamos hacia el componente de salida a fichero Excel, resultando:



Las propiedades básicas de dicho componente son:



Como podemos ver las propiedades básicamente nos permiten definir el nombre y ruta física del fichero de Excel de salida, así como el nombre y tamaño de las columnas del fichero destinto mediante la opción **'Define colum auto size'**. Es aconsejable marcar la opción 'Auto size' en todas las columnas si queremos asegurarnos que las columnas del fichero Excel mantienen el tamaño adecuado para los datos.

Ejecutamos nuestro trabajo y podemos observar como el fichero Excel se crea de manera correcta y en la ruta adecuada.

5. Referencias.

- http://www.talendforge.org/components/

6. Conclusiones.

Como podemos ver la herramienta **TALEND** nos permite de manera sencilla tratar la información que contienen ficheros Excel. Es una herramienta con versión 'GPL v2 Open Source license' lo que lo hace aún más atractiva. Por otro lado es una herramienta muy intuitiva por la representación de los componentes como cajitas que vamos colocando y van realizando las funciones adecuadas. Sin duda, una herramienta muy aconsejable para realizar ETL.

Un saludo.

Daniel Casanova

dcasanova@autentia.com

A continuación puedes evaluarlo:

Regístrate para evaluarlo

Por favor, vota +1 o compártelo si te pareció interesante

Share | 0

Anímate y coméntanos lo que pienses sobre este **TUTORIAL:**

» **Registrate** y accede a esta y otras ventajas «