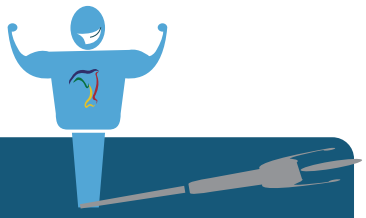
autentia

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...

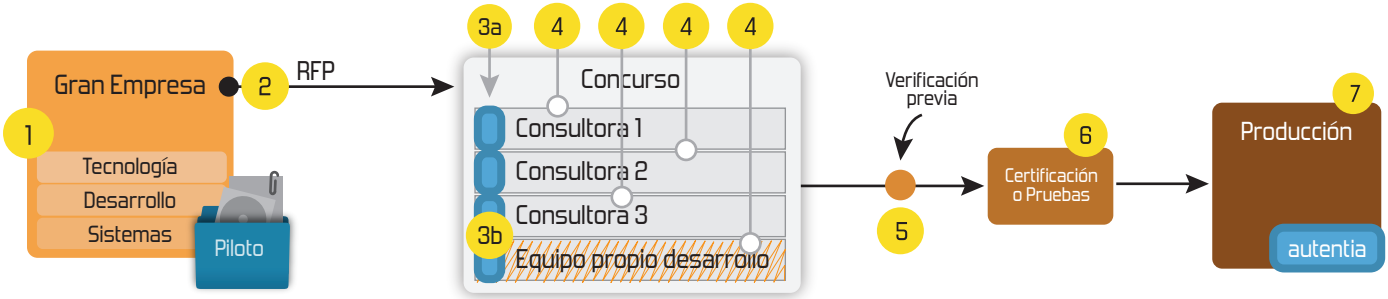
1. Desarrollo de componentes y proyectos a medida



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 conociemiento en:
www.adictosaltrabajo.com

Para más información visítenos en:
www.autentia.com

Jose Manuel Sánchez Suárez

Consultor tecnológico de desarrollo de proyectos informáticos.

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

Lectura y tratamiento de ficheros Excel con Talend (II): filtros y splits

0. Índice de contenidos.

1. Introducción

Siguiendo el hilo argumental del tutorial sobre indexación de documentos en Solr con el soporte de Talend, y tomando como referencia el tutorial sobre Lectura y tratamiento de ficheros Excel con Talend (I): nociones básicas, supongamos ahora que no solo disponemos de la información sobre el catálogo de libros en un formato estandar xml, si no que, además, existe información dispersa en documentos Excel, como, por ejemplo, las editoriales o el número de ejemplares en stock de los libros. Por lo que sea, esa información no existe en xml y se ha ido almacenando en ficheros Excel. Cierto es que cuando no se dispone del software adecuado las hojas de cálculo son el comodín perfecto, la de departamentos y unidades de negocio que basan su día a día en la elaboración, envío y análisis de hojas Excel.

Sobre las nociones básicas comentadas por Dani en el tutorial sobre Excel con Talend, en este tutorial vamos a ver como analizar, de una manera muy simple, el contenido de una hoja Excel, para darle, por ejemplo, una salida a una tabla de base de datos, aunque no lo veremos explícitamente en este tutorial, con lo que la salida podría ser cualquiera.

Dentro de ese análisis haremos un filtro y un split de una columna para tratar la información de una manera normalizada.

2. Entorno.

El tutorial está escrito usando el siguiente entorno:

- Hardware: Portátil MacBook Pro 15' (2.4 GHz Intel Core i7, 8GB DDR3 SDRAM).
- Sistema Operativo: Mac OS X Lion 10.7.4
- Talend Open Studio 5.1.1.

3. Fuente de datos.

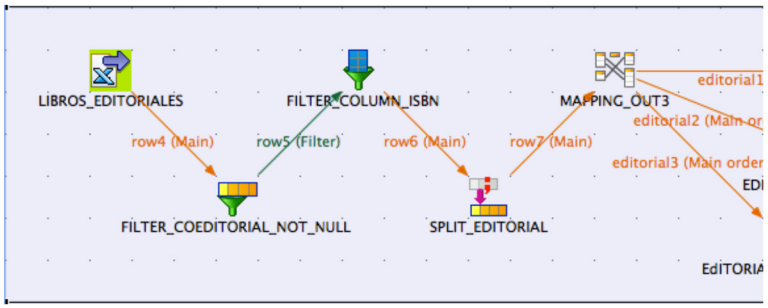
Supongamos que disponemos de la información estructurada en las siguientes columnas de una hoja Excel:

	A	B	C	D
1	ISBN	TITULO	COPIAS	EDITORIALES
2	9788483062982	Sobre la felicidad	225	Editorial Debate, Planeta, Starbook Editorial

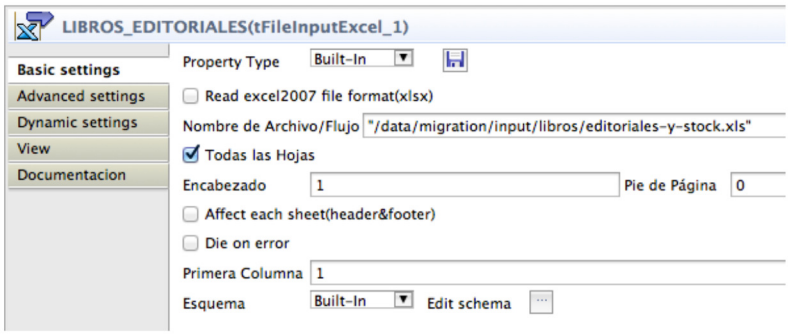
Disponemos información sobre el número de ejemplares en stock y las editoriales asociadas a un libro se encuentran en una única columna separadas por comas.

4. Diseño del Job de Talend.

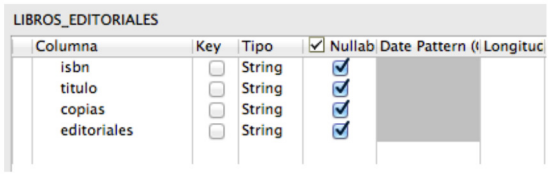
La parte de lectura del fichero Excel y normalización de su información podría tener, en la vista de diseño, un aspecto similar al siguiente:



La configuración de la lectura del Excel se basa en la asignación de un fichero de entrada y una serie de parámetros, como por ejemplo, que empiece a procesar el fichero a partir de una columna específica para escapar la típica fila de cabeceras de las columnas.

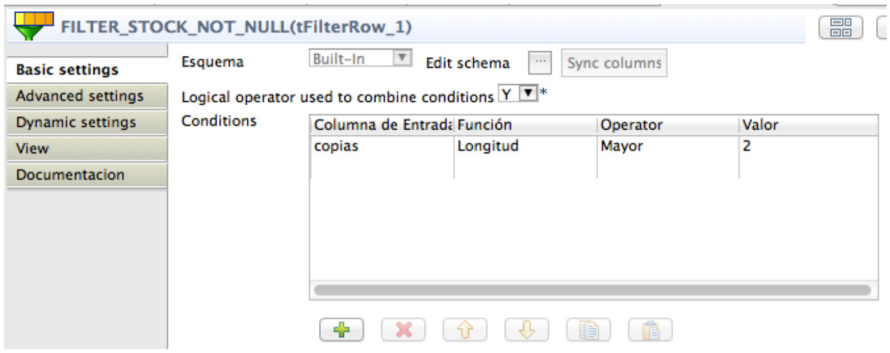


Pulsando sobre "Edit schema" podemos establecer, por orden, las columnas a leer del fichero asignando un nombre. Aunque no nos interese el contenido de una columna debemos mapearlo si se encuentra antes de otra interesante, esto es, no nos podemos saltar columnas.

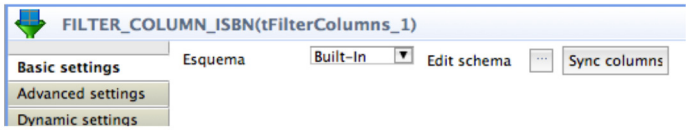


Con ello se realizará una lectura secuencial de las filas del fichero Excel almacenando el valor de las columnas para su tratamiento.

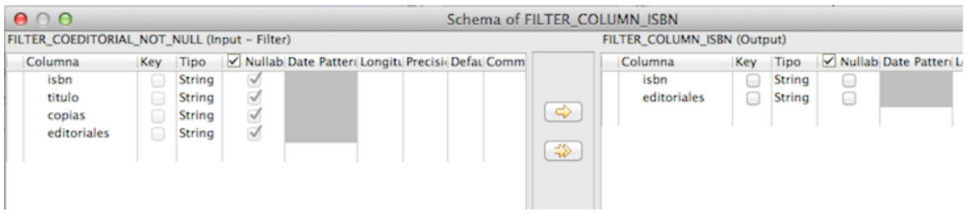
Lo siguiente podría ser introducir un filtro para que, en función del valor de una columna, se continúe con el tratamiento de la fila o no. En nuestro caso hemos establecido un filtro para que no procese los libros sin stock, usando una función de longitud de cadena.



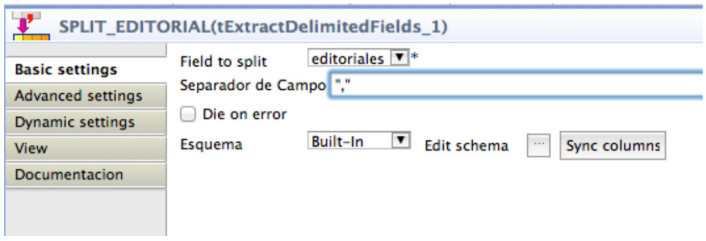
A continuación, filtramos las columnas para continuar el procesamiento solo de aquellas que nos interesan, el stock lo hemos usado como filtro, pero no nos interesa seguir procesándolo, de hecho, solo nos sigue interesando el isbn y las editoriales. Para ello, incluimos un tFilterColumns.



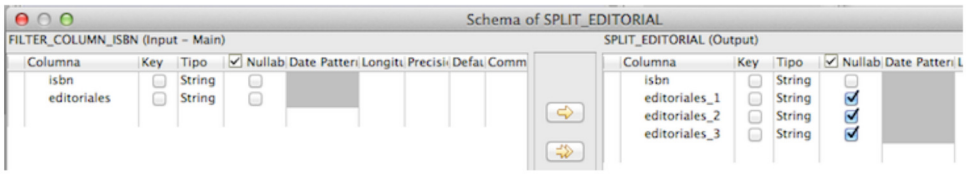
Y pulsando sobre "Edit schema", a la derecha nos quedamos con las filas que nos interesan:



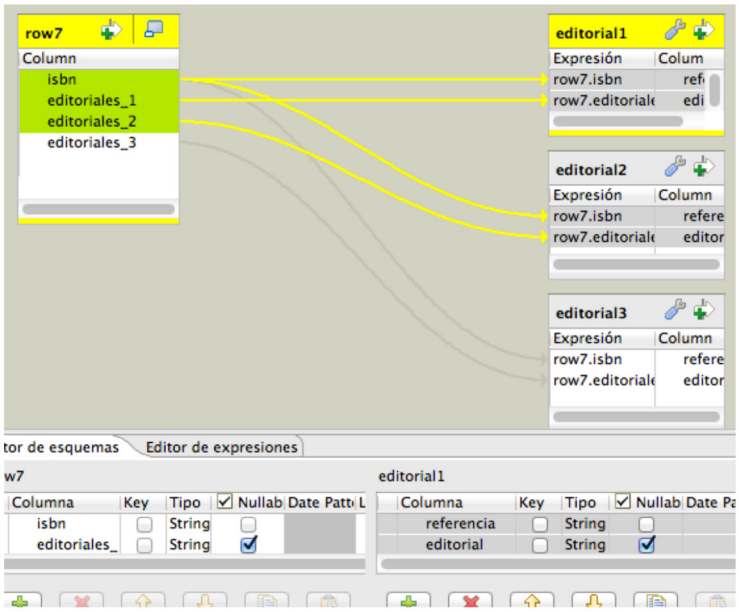
El siguiente paso es incluir un componente de tipo tExtractDelimitedFields que permite indicar una columna para realizar un split del valor de la misma, en función de un carácter de separación. En nuestro caso la columna es la de editoriales y el separador la coma.



Pulsando sobre "Edit schema" establecemos la salida del componente como sigue, de modo que, hasta tres editoriales, produciría un split del valor de la columna en los campos editoriales_1, editoriales_2 y editoriales_3. El número de campos de salida dependerá del número de ocurrencias máximo en cada caso.



En el siguiente componente, en nuestro caso un tMap, podríamos trabajar sobre esos campos de salida para dar un tratamiento individualizado a cada una de las ocurrencias en el split para las editoriales; recibiendo el isbn y cada una de ellas, nos serviría para añadir la relación entre libro y editorial en la tabla correspondiente.



5. Referencias.

- http://www.talend.com/resources/documentation.php
- Lectura y tratamiento de ficheros Excel con Talend (I): nociones básicas

6. Conclusiones.

Llegar a un Job de estas carcterísticas cuesta, la documentación sobre Talend es bastante "parca en palabras", aunque los foros sí son bastante activos; no es simple y requiere de muchas pruebas y componentes de debug, pero una vez dispones de una variedad de casos de uso y vas conociendo los componentes, vas reafirmándote en el juicio de que es infinitamente más productivo trabajar así que hacer la tarea de migración "a mano", tirando líneas de código.

Espero que os sirva de referencia.

Un saludo.

Jose

jmsanchez@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 «