Transcripciones en castellano y en catalán.

Pasos comunes de transformación

[Información contextual: tutorial sobre cómo realizar la funcionalidad objeto de este vídeo, con una voz en off que acompaña las imágenes del videotutorial.]

Hola, mi nombre es Juan Vidal y soy profesor colaborador de la UOC en el área de Informática, Telecomunicaciones y Multimedia. En este vídeo vamos a continuar trabajando con Pentaho Data Integration y vamos a ver diferentes pasos comunes de transformación.

Como siempre, arrancamos Spoon para acceder a la herramienta y vamos a partir de una transformación, compuesta por varios pasos en los cuales vamos a ir analizando diferentes pasos, que son pasos comunes de transformación y que nos van a permitir elaborar y transformar la información de origen.

Esta es nuestra transformación, que como veis se compone de varios pasos. Podemos ejecutar esta transformación para poder ver un poco qué información está generando y vamos a ir analizando los diferentes pasos que componen esta transformación. El objetivo de este vídeo va a ser conocer una serie de pasos comunes que nos van a ser muy útiles para generar transformaciones.

El primer paso de esa transformación es una lectura que ya vimos en vídeos anteriores, de la que podemos ver el resultado accediendo a la solapa de "*Preview data*". Aquí tenemos una información, que es la información que leemos tal cual, en bruto, que vamos a ir elaborando en los diferentes pasos. El primer paso que vamos a ver es un paso de normalización de filas.

Este paso lo encontramos en la ventana de la izquierda, dentro de la carpeta de "Transformación". Encontramos por aquí este paso, este paso de normalización. Como vimos en vídeos anteriores, incluirlo dentro del área de trabajo es simplemente arrastrarlo y tendríamos, posteriormente, que unirlo con el paso que corresponda. Este paso de normalización nos va a permitir pasar información que tenemos ahora mismo en columnas, pasarla a filas, de forma que los diferentes valores, estos diferentes campos que tenemos, van a convertirse en una fila con diferentes valores y el valor o la medida de cada campo va a aparecer en una columna adicional.

Vamos a ver cómo parametrizar este paso. Le indicamos cuál va a ser la variable nueva que vamos a generar. Los valores que va a tomar esta variable son los diferentes valores de las columnas. Y vamos a crear un nuevo valor que contiene las diferentes medidas que tiene cada columna. Vamos a verlo sobre los datos, que es donde mejor vamos a poder comprobar.

La ejecución del segundo paso para poder ver los datos de resultado. Simplemente, nos posicionamos en él y vemos los datos resultado. Aquí vemos que este paso nos ha convertido lo que antes eran columnas en diferentes valores de la fila y ha generado una nueva columna con el valor que tenía cada columna. Según le hemos indicado en la parametrización del paso que indicamos, la nueva variable en generar los diferentes

valores que toma y en qué campo incluimos el valor de cada columna. De forma que hemos transformado la información que teníamos en bruto, de columnas a filas.

Esto es un primer paso de normalización de filas que nos permite empezar a elaborar la información. El segundo paso, que es un paso común, es un paso que trabaja con cadenas de caracteres, pudiendo obtener una subcadena del mismo. Podemos encontrar esta transformación en la carpeta de transformaciones, que tendríamos aquí. Sería este paso. En este paso vamos a ver cómo está parametrizado, siempre entramos por ahí, y vemos cómo se parametriza. Y aquí encontramos que se realiza una operación, digamos de obtención de una subcadena, que se indica que se obtenga desde la posición 8 hasta la posición 12 y lo que obtiene lo incluye en este campo. Aquí está generando un campo año partiendo del contenido del campo, fichero del cual desde la posición 8 hasta la 11 obtendría el ancho.

Si nos fijamos en la ventana de resultados, el campo “Fichero”, efectivamente, de la posición 8 a la posición 12 contiene el año. Eso lo obtenemos con este paso. Podemos ver los resultados de este paso, en el cual ha añadido un campo año, lo vemos a la derecha, proveniente de la información del campo “Fichero”. Bien, el flujo continúa. El siguiente paso, tenemos un paso de mapeo de valores, un paso también bastante común que encontramos también dentro de la carpeta “Transformación”.

El mapeo de valores realiza un mapeo entre una columna origen y una columna destino, en base a lo que le indiquemos en el paso. Vamos a editar y vamos viendo cómo está realizado este mapeo. El mapeo se realiza partiendo del campo “Año” y obtiene un campo “*Target*” que tiene los valores que se indican a la derecha. Creamos el campo “Fecha” partiendo del campo año.

En este caso, por ejemplo, está obteniendo siempre el último día de cada año. Necesita, para el procesado posterior, conocer el último día de cada año. Entonces, hace un mapeo directamente entre un valor origen y un valor destino. Es un paso bastante común, simplemente, desde una columna origen a una columna destino le indicamos un mapeo de valores, que va aplicando.

Podemos ver el resultado de aplicar este paso en la generación de una nueva columna que, desde el campo “Año” crea un campo “Fecha”, en base al mapeo de valores que se le ha indicado. Siguiendo con el flujo, encontramos otro paso también bastante común, que nos va a permitir trabajar con las columnas del *“Dataset”,* haciendo diferentes operaciones. También la tenemos disponible dentro de la carpeta de “Transformación”.

“*Select Values*”. Vamos a ver cómo hemos parametrizado esta tarea, este paso. Vemos que aquí encontramos diferentes solapas. La primera solapa la indicaríamos con qué campos del flujo nos queremos quedar. Todo campo que no esté indicado en este listado se va a eliminar. Sólo se mantienen estos. Podríamos optar, también, por indicar aquellos que queremos eliminar o, también, tenemos en la solapa de metadatos la posibilidad de hacer renombrados o cambios de tipo en los campos. De forma que es un paso muy común para temas de selección de campos, eliminación de campos o modificación de los tipos o renombrado de campos.

En este caso, hemos utilizado la funcionalidad de “Selección de campos”, de forma que el resultado de la ejecución de este paso nos limita ya a una serie de campos en el *“Dataset”*. De forma que hemos visto como de una información en bruto, diferentes

pasos de transformación llegan a un resultado, que veremos en posteriores vídeos cómo se vuelca finalmente a una base de datos. Todos estos pasos: selección de campos, mapeo de valores, operaciones de subcadena, normalización de filas… Son pasos comunes muy utilizados en los procesos ETL con Pentaho Data Integration y que nos van a permitir elaborar y manipular la información. Esperamos que el vídeo haya sido de utilidad. Muchas gracias por vuestra atención. Un saludo.

Passos comuns de transformació

[Informació contextual: tutorial sobre com fer la funcionalitat objecte d'aquest vídeo, amb una veu en off que acompanya les imatges del videotutorial.]

Hola, em dic Juan Vidal i soc professor col·laborador de la UOC a l'àrea d'Informàtica, Telecomunicacions i Multimèdia. En aquest vídeo continuarem treballant amb Pentaho Data Integration i veurem diferents passos comuns de transformació.

Com sempre, arrenquem Spoon per accedir a l'eina i comencem amb una transformació, composta per diversos passos, en els quals anirem analitzant diferents passos, que són passos comuns de transformació i que ens permetran elaborar i transformar la informació d'origen.

Aquesta és la nostra transformació que, com veieu, es compon de diversos passos. Podem executar aquesta transformació per poder veure una mica quina informació està generant i anirem analitzant els diferents passos que componen aquesta transformació. L'objectiu d'aquest vídeo serà conèixer una sèrie de passos comuns que ens seran molt útils per generar transformacions.

El primer pas d'aquesta transformació és una lectura, que ja vam veure en vídeos anteriors, de la qual en podem veure el resultat accedint a la solapa de "*Preview data*". Aquí tenim una informació, que és la informació que llegim tal qual, en brut, que anirem elaborant en els diferents passos. El primer pas que veurem és un pas de normalització de files.

Aquest pas el trobem a la finestra de l'esquerra, dins de la carpeta de "Transformació". Trobem per aquí aquest pas, aquest pas de normalització. Com vam veure en vídeos anteriors, incloure'l dins de l'àrea de treball és simplement arrossegar-lo i hauríem d'unir-lo posteriorment amb el pas que correspongui. Aquest pas de normalització ens permetrà passar informació que tenim ara mateix en columnes, passar-la a files, de manera que els diferents valors, aquests diferents camps que tenim, es convertiran en una fila amb diferents valors i el valor o la mesura de cada camp apareixerà en una columna addicional.

Veurem com parametritzar aquest pas. Indiquem quina serà la variable nova que generarem. Els valors que pren aquesta variable són els diferents valors de les columnes. I crearem un nou valor que conté les diferents mides que té cada columna. Ho veurem sobre les dades, que és on millor podrem comprovar.

L'execució del segon pas per veure les dades de resultat. Simplement, ens hi posicionem i veiem les dades resultat. Aquí veiem que aquest pas ens ha convertit allò que abans eren columnes en diferents valors de la fila i ha generat una nova columna amb el valor que tenia cada columna. Segons hem indicat en la parametrització del pas que indiquem, la nova variable en generar els diferents valors que pren i en quin camp incloem el valor de cada columna. De manera que hem transformat la informació que teníem en brut, de columnes a files.

Això és un primer pas de normalització de files que ens permet començar a elaborar la informació. El segon pas, que és un pas comú, és un pas que treballa amb cadenes de caràcters, podent aconseguir-ne una subcadena. Podem trobar aquesta transformació a la carpeta de “Transformacions”, que tindríem aquí. Seria aquest pas. En aquest pas veurem com està parametritzat, sempre entrem per aquí, i veiem com es parametritza. I aquí trobem que es fa una operació, diguem d'obtenció d'una subcadena, que s'indica que s'assoleixi des de la posició 8 fins a la posició 12 i allò que aconsegueix ho inclou en aquest camp. Aquí esteu generant un camp “Any” partint del contingut del camp, fitxer del qual des de la posició 8 fins a la 11 aconseguiria l'amplada.

Si ens fixem a la finestra de resultats, el camp “Fitxer”, efectivament, de la posició 8 a la posició 12 conté l'any. Això ho assolim amb aquest pas. Podem veure els resultats d'aquest pas, al qual ha afegit un camp “Any”, el veiem a la dreta, provinent de la informació del camp “Fitxer”. Bé, el flux continua. El següent pas, tenim un pas de mapatge de valors, un pas també força comú que trobem també dins de la carpeta “Transformació”.

El mapatge de valors fa un mapatge entre una columna origen i una columna “Destinació”, a partir del que us indiquem en el pas. Editem i anem veient com està fet aquest mapatge. El mapatge es fa partint del camp “Any” i aconsegueix un camp “*Target*” que té els valors que s'indiquen a la dreta. Creem el camp “Data” partint del camp “Any”.

En aquest cas, per exemple, està obtenint sempre el darrer dia de cada any. Necessita, per al processament posterior, conèixer el darrer dia de cada any. Aleshores, fa un mapatge directament entre un valor origen i un valor “Destinació”. És un pas força comú, simplement, des d'una columna origen a una columna “Destinació” indiquem un mapatge de valors, que va aplicant.

Podem veure el resultat d'aplicar aquest pas a la generació d'una nova columna que, des del camp “Any”, crea un camp “Data”, a partir del mapatge de valors que se li ha indicat. Continuant amb el flux, trobem un altre pas també força comú, que ens permetrà treballar amb les columnes del *“Dataset*, fent diferents operacions. També la tenim disponible dins de la carpeta de “Transformació”.

“*Select Values*”. Veurem com hem parametritzat aquesta tasca, aquest pas. Veiem que aquí trobem diferents solapes. La primera solapa la indicaríem amb quins camps del flux ens volem quedar. Tot camp que no estigui indicat en aquest llistat s'eliminarà. Només es mantenen aquests. Podríem optar, també, per indicar aquells que volem eliminar o, també, tenim a la solapa de metadades la possibilitat de re-anomenar o fer canvis de tipus als camps. De manera que és un pas molt comú per a temes de selecció de camps, eliminació de camps o modificació dels tipus o renom de camps.

En aquest cas, hem utilitzat la funcionalitat de “Selecció de camps”, de manera que el resultat de l'execució d'aquest pas ens limita ja a una sèrie de camps al *“Dataset”*. De manera que hem vist com d'una informació en brut, diferents passos de transformació arriben a un resultat, que veurem en posteriors vídeos com es bolca finalment a una base de dades. Tots aquests passos: selecció de camps, mapatge de valors, operacions de subcadena, normalització de files… Són passos comuns molt emprats en els processos ETL amb Pentaho Data Integration i que ens permetran elaborar i manipular la informació.

Esperem que el vídeo hagi estat útil. Moltes gràcies per la vostra atenció.

Una salutació.