Sesión 3: Implementación de procesos ETL con Pentaho Data Integration

3.1 Pentaho Data Integration (PDI)

En la siguiente página tenéis información sobre PDI:
https://help.hitachivantara.com/Documentation/Pentaho/9.1/Products/Pentaho_Data_Integration.

Esta herramienta permite diseñar procesos de Extracción, Transformación y Carga (ETL), y planificar su ejecución.

Los conceptos fundamentales relacionados con la definición de los procesos los tenéis en esta página:
https://help.hitachivantara.com/Documentation/Pentaho/9.1/Products/Data_Integration_perspective_in_the_PDI_client.

- **Transformación** (Transformation): Flujo de pasos (step) de procesamiento de datos, unidos por saltos (hop), que se ejecutan en un pipeline en el que cada paso se inicia en paralelo.
- **Trabajo** (Job): Flujo de tareas (Normalmente transformaciones), en las que una tarea no se inicia hasta que se completa la que le precede en el flujo.
- **Paso** (Step): Son los bloques con los que se generan las transformaciones. En general, procesa un stream de filas de entrada para producir un stream de filas de salida (por eso tiene sentido que todos los pasos se inicien en paralelo).
    - Existen pasos de muchos tipos distintos, incluyendo pasos para entrada y salida de datos a distintos tipos de fuentes y destinos de datos, y pasos que permiten leer el stream de la tarea precedente en el trabajo y escribir el stream en la tarea o tareas siguientes.
- **Salto** (Hop): Son enlaces entre pasos. Los saltos pueden habilitarse o deshabilitarse (para evitar que una determinada parte de un trabajo o transformación se ejecute).
    - No se permiten ciclos en las transformaciones, pero si en los trabajos.
    - No se pueden mezclar streams de entrada que no son compatibles (mismo esquema). A no ser que el paso esté diseñado específicamente para mezclar (como el caso de los JOIN).
    - Cuando sacamos dos saltos de salida de un trabajo podemos especificar si la salida se copia en los dos o si se distribuye entre los dos.

En las siguientes URLs podéis acceder a una referencia de los tipos de pasos y tareas disponibles:

- Pasos de transformaciones:
  https://help.hitachivantara.com/Documentation/Pentaho/Data_Integration_and_Analytics/9.5/Products/Transformation_step_reference
- Tareas de trabajos:
  https://help.hitachivantara.com/Documentation/Pentaho/Data_Integration_and_Analytics/9.5/Products/Job_entry_reference

Nota: La versión instalada en la máquina virtual es la 9.1, es decir, pueden existir pequeñas diferencias respecto a los documentos enlazados pero la documentación de la versión 9.1 tiene problemas de mantenimiento.

3.2 Ejemplo de creación de un trabajo con una transformación.

Vamos a intentar ahora crear la primera de las transformaciones ("CargarPeliculas") descritas en documento del proyecto de Producción de Cine.

Primero creamos una carpeta "etl" dentro de la carpeta "IN". Aquí vamos a colocar los datos de entrada y también vamos a guardar los archivos generados por PDI en los que se definen las transformaciones y trabajos.

Dentro de "etl" creamos dos carpetas, una "entrada" y otra "trabajo". En la carpeta de entrada vamos a colocar un par de archivos .xml y otro par de archivos .csv, correspondientes a dos meses consecutivos.

Vamos a crea la tabla en la que almacenaremos, en el área de ensayo de la etl, los datos resultantes de esta transformación.

```
drop schema if exists etl cascade;
create schema etl;

create table etl.pelicula_productora (
  pelicula int not null,
  fecha_emision date not null,
  id_productora int not null,
  text_id_director varchar,
  text_id_productor varchar,
  primary key (pelicula, id_productora)
);

create index on etl.pelicula_productora(pelicula);
```

Iniciamos PDI "./pentaho/data-integration/spoon.sh"

Creamos un trabajo nuevo, que guardaremos en la carpeta "trabajo" creada arriba. Y creamos una transformación nueva que también almacenaremos en esa carpeta. Ya podemos definir el flujo tanto del trabajo como de la transformación en PDI.

ENTREGA INDIVIDUAL 3: Completa el trabajo añadiendo las transformaciones necesarias según lo descrito en el documento del proyecto. Entrega los archivos generados para el trabajo y las transformaciones a través del campus virtual.

Última modificación: luns, 9 de outubro de 2023, 21:03