Sesión 3b: 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, procesan 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:

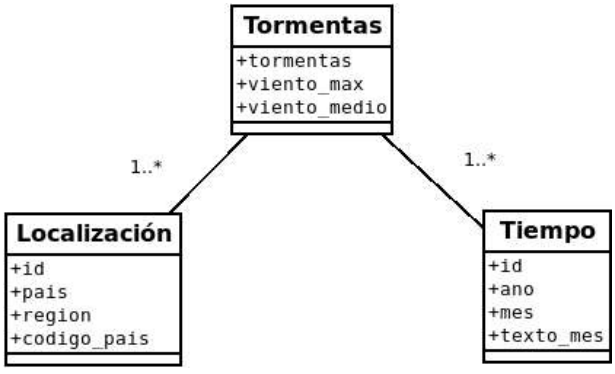
- Pasos de transformaciones:
  https://help.hitachivantara.com/Documentation/Pentaho/Data_Integration_and_Analytics/9.5/Products/Transformation_step_reference
- Tareas de trabajos:
  https://help.hitachivantara.com/Documentation/Pentaho/Data_Integration_and_Analytics/9.5/Products/Job_entry_reference

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.

Ejemplo de creación de un trabajo con varias transformaciones

Vamos a desarrollar un ejemplo de creación de un trabajo empleando el dataset de tormentas modificado para la ocasión. En esta ocasión la información se nos proporcionará a través de 2 fuentes de datos externas. Una de las fuentes nos proporcionará XMLs con información sobre las tormentas registradas y la otra nos proporcionará los códigos de los países en los que se produjo la tormenta en formato CSV. Deberemos acceder a las dos fuentes y transformar los datos para poder acomodarlos en nuestro esquema en estrella.

Lo primero que vamos a hacer es crear un esquema en estrella basado en el siguiente modelo de datos multidimensional:



Creamos la BD *dwtormentas*.

Para generar las tablas asociadas:

```
CREATE TABLE geo (
        id serial primary key,
        pais varchar(50),
        region varchar(50),
        codigo_pais varchar(2)
);


CREATE TABLE tiempo (
        id serial primary key,
        ano int4,
        mes int4,
        texto_mes varchar(25)
);



CREATE TABLE tormentas (
        tiempo int references tiempo(id),
        geo int references geo(id),
        viento_max int4,
        viento_medio float4,
        tormentas int4,
        PRIMARY KEY (tiempo, geo)
);
```

A continuación crearemos una carpeta *"etl*" dentro de la carpeta "*IN*" en la */home*/alumnogreibd. Dentro de la carpeta "*etl*" crearemos otras dos subcarpetas: "*entrada*" y "*trabajo*". En la primera pondremos los archivos .xml y .csv necesarios para hacer la práctica y en la segunda guardaremos los ficheros generados por PDI. Esta estructura NO es necesaria. La crearemos simplemente por llevar un orden.

En este punto iniciaremos la interfaz de PDI (spoon): "*./pentaho/data-integration/spoon.sh*"

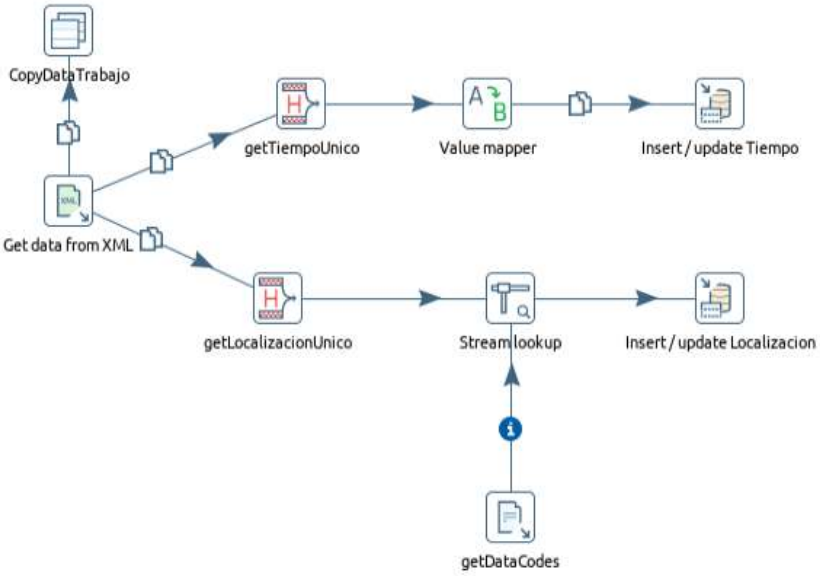
Crearemos un trabajo nuevo y las transformaciones necesarias para llevar a cabo el proceso de ETL relacionado con las tormentas.

Aunque lo trabajaremos en clase, a continuación os muestro los esquemas gráficos resultantes:

Transformación para crear las dimensiones

En esta transformación se realizarán las diferentes operaciones para obtener, preparar y cargar los datos asociados a las dimensiones en las tablas asociadas. Pasos:

1. Primero se extrae el contenido de los .xml. El resultado se copia para futuras transformaciones**.**
2. Se obtienen los valores únicos de tiempo (combinaciones únicas de año y mes).
3. Se crean los valores de texto (nombres de los meses) empleando un "paso" para mapear.
4. Se guardan los valores en la tabla.
5. Se obtienen los valores únicos de tiempo (combinaciones únicas de país y región).
6. Se asigna el código de país empleando la información del .csv empleando un "paso" stream lookup.
7. Se guardan los valores en una tabla.



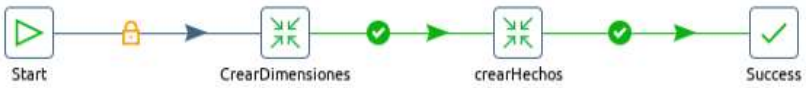
Transformación para crear la tabla de hechos

En esta transformación se preparán los datos para poder introducir en la tabla de hechos. Los pasos asociados son los siguientes:

- Se recuperan los datos almacenados de las transformaciones anteriores.
- Se recupera el identificador (id, clave primaria) de la tabla de tiempos (identificador asignado a la combinación año-mes que usaremos como clave foránea).
- Se recupera el identificador (id, clave primaria) de la tabla geo (identificador asignado a la combinación país-región que usaremos como clave foránea).
- Almacenamos el stream en la tabla correspondiente.

Trabajo

Flujo de transformaciones asociadas a nuestra ETL**.**



Última modificación: mércores, 11 de outubro de 2023, 18:07