ETL de datos mediante Pentaho

Contexto y objetivo general

En esta actividad vais a emplear Pentaho Spoon para limpiar un set de datos, que tenemos en un csv, e introducirlo en un sistema transaccional, el propio MySQl que viene instalado en la máquina virtual.

Para el ejercicio, vamos a suponer que una empresa dispone de un fichero CSV con información de clientes, este fichero, procedente de un sistema externo de origen americano, no nos da las fechas en el formato que queremos. Así que antes de poder cargar estos datos en su sistema transaccional corporativo, es necesario validar, limpiar y normalizar la información mediante un proceso ETL. Además, el fichero presenta errores intencionados de calidad de datos, que deberán ser detectados y tratados correctamente.

Los datos que contiene cada fila del fichero son: Nombre del cliente, Apellidos del cliente, Fecha de registro en formato americano (MM/dd/yyyy), Número total de pedidos realizados por el cliente en el último mes y Ciudad de residencia del cliente.

En el MySQL que tenéis instalado en la máquina virtual va a ser necesario crear una base de datos, a la que llamaremos OLTP (OnLine Transaction Processing), y que tendrá una tabla llamada Clientes, cuyos campos deben coincidir con los del fichero.

Tarea a realizar

Los pasos a realizar para esta tarea son los siguientes:

1. Preparar el entorno en la base de datos
2. Importar los datos a Pentaho
3. Detección de errores de calidad de datos, almacenando en un fichero de rechazos por ejemplo clientes_rechazados.csv, debéis detectar y rechazar los registros duplicados e incompletos
4. Convertir el campo Fecha de Registro de formato americano a español
5. Normalizar los datos convirtiendo a mayúsculas los siguientes campos: Nombre, Apellidos y Ciudad
6. Generar un fichero de salida llamado, por ejemplo clientes_limpios.csv
7. Introducir los datos en la tabla Clientes de MySQL

Entregable

En esta tarea cada uno de vosotros deberá entregar:

- El fichero clientes_limpios.csv
- El fichero clientes_rechazados.csv
- Un documento con una breve explicación (1-2 páginas) indicando:
    - Qué pasos de Pentaho se han utilizado con una captura del flujo
    - Qué tipos de errores se han detectado en los datos
    - Cómo se ha realizado la conversión de fechas
    - Qué posibilidades de uso no exploradas ves a Pentaho

Calificación

Aspecto evaluado	**Peso**
Lectura correcta del CSV	10%
Detección de errores de datos	25%
Conversión correcta del formato de fecha	25%
Normalización de datos	10%
Creación de tabla MySQL	20%
Claridad y orden del proceso ETL	10%