Bases de Datos Masivas (11088)
Departamento de Ciencias Básicas

TRABAJO PRÁCTICO I: Definición de Procesos ETL

Introducción:

Para la definición de procesos ETL, en primer término se solicitará al estudiante que utilice las herramientas que utiliza actualmente para luego utilizar el software PDI de la suite Pentaho.

El objetivo de esta metodología es que el estudiante incorpore los conocimientos acerca del proceso y luego conozca las herramientas especializadas para procesos ETL. Para todas las consignas, el equipo docente proveerá las fuentes de información necesarias.

Consignas:

1. Se cuenta con el dataset *Medios* que cuenta con 7000 medios nacionales. Se desea normalizar esta información en una Base de Datos transaccional teniendo en cuenta que cada medio posee atributos correspondientes a su nombre, ubicación, tipo de medio y especialidad. Migre la información del archivo a una Base de Datos PostgreSQL con la siguiente estructura:
    a. Medios(id, nombre, id_especialidad, id_tipo_medio, dirección, id_ciudad),
    b. Especialidades(id, descripción),
    c. Tipos_medio(id, descripción),
    d. Ciudades(id, nombre, id_provincia).
    e. Provincias(id, nombre).

    Explique someramente la metodología utilizada y estime el tiempo que le demandó la actividad.

2. Se cuenta con los orígenes de datos *etl_cursadas, etl_estudiantes* y *planes* con información de los estudiantes de la Universidad y sus cursadas durante el 1er Cuatrimestre 2003. Se solicita que genere una nueva DB con la siguiente estructura:

a. Rendimiento_Academico(id_estudiante, id_plan, id_sede, id_ciudad, id_sexo, id_cohorte, cantidad_cursadas, cantidad_aprobadas, promedio),
b. Planes(id_plan, codigo_plan, código_carrera, nombre_carrera),
c. Ciudades(id_ciudad, código_postal, nombre_ciudad, provincia),
d. Sedes(id_sede, sede),
e. Sexo(id_sexo, sexo),
f. Cohortes(id_cohortes, cohorte).

Utilice el software PDI y estime el tiempo que le demandó la actividad.

3. Ahora, resuelva la **consigna 1)** con la herramienta PDI de la suite Pentaho, a través de las transformations y Jobs necesarias para llevar adelante la solución. Tome el tiempo que demora en resolver este ejercicio con PDI.
4. Cree un Job que verifique todos los días a las 14 hs si existe el archivo *01-01-medios.csv,* trabajado en el punto 1), en un directorio determinado y en caso afirmativo ejecute el Job para actualizar la DB generada antes.
5. Guarde los archivos resultantes de las actividades prácticas en una carpeta denominada **tp01-<legajo>** que a su vez tenga un directorio por cada uno de los puntos de este trabajo, comprima la carpeta y envíelo al equipo docente.

Referencias sugeridas:

Guía de LABORATORIO: Definición de Procesos ETL con Pentaho Data Integration (Pentaho):

https://github.com/jumafernandez/BDM/blob/master/Guias/PDI.md

Getting Started with PDI:

https://help.pentaho.com/Documentation/8.3/Products/Pentaho_Data_Integration