
**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 Kettle/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 el backup de una Base de Datos de PostgreSQL, denominado *01-02-cursadas1_2003.backup*, con los registros de cursada de los estudiantes de la Universidad durante el 1er Cuatrimestre 2003:

- **Cursadas:**
  - Legajo,
  - Plan de estudios,
  - Cuatrimestre,
  - Año,
  - Materia,
  - Condición,
  - Calificación.

Por otro lado, se posee un archivo de texto plano, denominado *01-02-planes.txt* con los códigos de Plan existentes y la Carrera a la cual corresponden:

- **Planes y Carreras:**
  - Código de Plan,
  - Nombre Carrera.

Para la generación de un cubo con información gerencial de menor grado de granularidad, se solicita que genere una nueva DB con una tabla denominada **rendimiento_estudiantes** con los siguientes atributos:

- Legajo,
- Código de Carrera (donde el código de la Carrera es la parte entera de la división por 100 del código de Plan),
- Nombre Carrera,
- Cantidad de Cursadas,
- Cantidad de Aprobadas,
- Promedio.

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

3. A continuación, defina conceptualmente las estructuras ETL de Kettle:
   a. Transformation,

b. Job,
c. Step,
d. Hop.

4. Ahora, resuelva la **consigna 1)** con la herramienta PDI/Kettle 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/Kettle.
5. 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.
6. Guarde los archivos resultantes de las actividades prácticas en una carpeta denominada **tp1-<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:

Getting Started with PDI

https://help.pentaho.com/Documentation/7.0/0J0/0C0/020

[Accedido el 22/08/2017]