# Parcial 1 - Gestión de Datos

Formato de entrega: UN ÚNICO archivo .ZIP por pareja por medio de la plataforma virtual.

**Ejercicio 1: Directorios Variables y extracción SQLServer**

*"Server=tcp:186.154.145.238;Port=2433;Database=ExampleGDD;Uid=gdd;Pwd=gdd2022* Encrypt=yes;TrustServerCertificate=no;Connection Timeout=30;"*

El objetivo del ejercicio es construir un ETL (se debe llamar *ej1.ktr*) que extraiga los datos de las especies ([bi_Species]) y sus nombres comunes ([bi_SpeciesCommonNames]), también se debe seleccionar el investigador sobre la especie que están en ([bi_SpeciesInvestigators]) y ([bi_Investigators]). Estos datos se encuentran en el servidor SQL Server mencionado anteriormente, finalmente almacene los datos de las siguientes especies en un archivo CSV:

[SpeciesID]:

| | | | |
|---|---|---|---|
| 271383 | 205667 | 279652 | 285557 |
| 208045 | 282209 | 272325 | 202319 |
| 241523 | 285318 | 261205 | 244210 |
| 251105 | 261205 | 275717 | 202414 |
| 272659 | 281039 | 279842 | 244166 |

**Entregable:**

- El archivo 1 generado debe contener los campos de id de la especie, nombre científico de la especie, nombre común de la especie, su respectivo reino y el investigador.
- El archivo 2 generado debe contener los campos de el nombre del investigador y el reino de la especie con un conteo de cuantas especies él ha investigado por reino.
- Debe contener encabezado con los nombres de columnas
- El separador debe ser el punto y coma ";"

Para la extracción del servidor SQL se debe agregar el driver sqljdbc42.jar en la carpeta lib de Pentaho.

En clase las rutas de los archivos que utilizamos eran estáticas. Es decir, al seleccionar de donde leer o donde guardar el archivo, la ruta quedaba de esta manera:

> *Ejemplo: C:\usuario\estudiante\gestionDeDatos\codigosPostalesMedellin.csv*
>
> *En la práctica esto no es recomendable ya que no sabemos en cual computador se va a ejecutar el ETL y la ruta "C:\usuario\estudiante\gestionDeDatos\" podría no existir.*

Para solucionar esto la ruta se puede construir variable, es decir que dependa de un valor fijado anteriormente. Para el ejercicio la ruta debe ser la misma carpeta en la cual está el archivo de Transformación (.ktr). Esto se logra introduciendo la variable interna *Internal.Transformation.Filename.Directory*[1][2] al inicio de la ruta. Para el

[1] Más información sobre el uso de variables en Kettle en https://wiki.pentaho.com/display/EAIes/.07+Variables

[2] Los nombres de las variables aparecen automáticamente al presionar CRTL+Espacio

ejemplo de clase con codigosPostalesMedellin.csv, en el campo 'Filename' de las propiedades del CSV Input se cambia:

*"C:\usuario\estudiante\gestionDeDatos\codigosPostalesMedellin.csv"*

*"${Internal.Transformation.Filename.Directory}/codigosPostalesMedellin.csv"*

El cambio para los Output de tipos Text file Output es similar.

El ETL debe generar el archivo de salida con las características enunciadas al inicio del ejercicio y ninguna de las rutas de salida en el ETL deben ser absolutas.

**Ejercicio 2: Crear un TSV con un campo calculado**

Para el ejercicio 2 se debe crear un ETL con el nombre *ej2.ktr* que transforme el archivo *iris.data.csv* en un archivo TSV (tab-separated values)[1] llamado ***iris.data.tsv*** con las siguientes características:

- *Debe contener encabezado con los nombres de columnas [ sepalLength , sepalWidth , petalLength , petalWidth , sepalProd , petalProd , class ]*
- *El separador debe ser un TAB*
- *Los campos de texto no se deben encerrar entre comillas*

Para definir TAB como separador del archivo de texto se utiliza el botón "Insert Tab" al lado derecho del campo "Delimiter" el cual crea un tab transparente dentro del separador.

Para crear los campos sepalProd y petalProd se utiliza la transformación básica[2] "Transform/Calculator"( https://wiki.pentaho.com/display/EAIes/.07+Variables ) con la cual se crea nuevos campos haciendo operaciones matemáticas sobre los que ya existen. La fórmula para estos campos es un producto de esta manera:

- sepalProd= sepalLength* sepalWidth
- petalProd= petalLength* petalWidth

El ETL debe generar el archivo TSV con TODAS las columnas y además debe cumplir la misma característica de Directorios Variables del Ejercicio 1.

**Entregable:**

UN UNICO archivo .zip que contenga los siguientes archivos:

- iris.data.csv
- ej1.ktr
- ej2.ktr
- species.csv
- iris.data.tsv

***Nota:***

*Para entregar los dos puntos con sus evidencias deben también adjuntar un PDF que contenga la explicación de la estrategia de desarrollo del parcial y las decisiones que tomaron técnicamente.*

[1] https://en.wikipedia.org/wiki/Tab-separated_values

[2] La transformación se agrega como un paso intermedio entre el Input y el Output