Bases de datos estructuradas
Práctica VII
Ing. Fernando Arreola

Objetivo

- Elaborar un flujo de procesamiento en la herramienta *pentaho* que permita extraer información de distintos fuentes, aplicar una serie de transformaciones y concentrar en un destino.

Instrucciones

- La presente práctica se realizará en los equipos previamente organizados.
- Se deberá entregar un documento pdf donde muestre los steps empleados para cada parte del ejercicio, agregando una captura de la configuración de los mismos, captura de muestra de información (preview) de esa parte del ejercicio y explicación breve de su propuesta de solución, lo que implica que el documento estará dividido en 3 secciones, una para extracción, otra para transformación y otra para carga.
- Adicional, tambien se deberá entregar el archivo(s) correspondiente a su flujo de procesamiento.
- Organizarse adecuadamente, una parte del equipo puede trabajar una fuente, una parte la otra fuente, ya que no hay dependencia entre ellas y el destino es remoto.
- En el documento debera agregar conclusiones individuales, mencionando los errores, dudas, aciertos, etc. que se enfrentaron en el desarrollo de la práctica.
- La hora de entrega es máximo a las 14 horas por medio de mensaje directo en la plataforma slack.

Ejercicio

Diseño e implementación de un flujo *ETL*

Extracción:

Se cuenta con dos fuentes de información:

- Anexo a los documentos de la práctica se encuentre el archivo *vuelos.csv*, el cual deberá almacenar dentro de una ubicación dentro de la pc donde van a trabajar.
- El resto de información vive en una tabla dentro de una base de datos remota. Los datos de conexión son los siguientes:
  - ip: 132.248.59.32
  - puerto: 4321
  - usuario: lcd20251
  - contraseña: pr*ctIc47$
  - base de datos: practica7
  - tabla: aerolineasaeropuertos

Agregue los steps necesarios para incorporar ambas fuentes de información a su flujo de procesamiento.

Transformación:

Una vez incorporadas las fuentes de información realizar las siguientes transformaciones. Para el archivo vuelos:

- Reemplace el valor numérico del atributo *month* por el mes (en inglés) correspondiente
- Reemplace el valor numérico del atributo *day_of_week* por el día (en inglés) correspondiente
- La fecha se encuentra separada en 3 atributos, junte los atributos en uno solo, con el formato dia-mes-año

Para la información extraída de la tabla aerolineasaeropuertos:

- Los atributos *iata_code2, airline e id* son parte de las aerolíneas, los restantes son parte de los aeropuertos, por lo que debe separar los datos.
- Remover el atributo *id*
- Renombrar el atributo *iata_code2* por *code_airline*
- Del nombre de las aerolíneas sustituir la palabra Inc por Incorporation
- Filtrar aquellos aeropuertos que tengan una latitud con valor de 0, no deben formar parte del flujo
- Remover de los nombres del aeropuerto su nombre popular (aquel que se encuentre entre paréntesis), ejemplo: Northeast Florida Regional Airport (St. Augustine Airport) debe quedar sólo como Florida Regional Airport.

Una vez realizadas las transformaciones indicadas, la información está lista para ser entregada en un destino, el cual puede ser una base de datos, archivos de texto, rutas remotas, un data warehouse, etc.

Carga:

Para concluir el ciclo de procesamiento, deberá cargar su información en la siguiente base de datos destino:

- ip: 132.248.59.32
- puerto: 4321
- usuario: equiposp7
- contraseña: bd3qu1p*s+
- base de datos: load_results_p7

Para ello deberá crear las tablas indicadas en el script *tablasp7.sql*, agregando los datos de su equipo como parte del nombre, ejemplo: *aerolineas_equipo1*

NOTA: Puede que deba hacer modificaciones al script tomando como base las transformaciones que hizo a sus datos, ya que pueden surgir errores o inconsistencias en su información a la hora de hacer la carga.