DATA MINING

Práctica 4: Proceso de ETL, caso "precipitación pluvial en CDMX"

Objetivo: Desarrollar una herramienta ETL para procesar archivos de Excel para el **caso "Precipitación pluvial (PP), con la técnica de recolección para depósito húmedo (H))", durante el periodo "2000 al 2019"**

Procedimiento: Analizar la estructura de los archivos de Excel de origen, y determinar un proceso de integración de datos, que incluya tareas, reglas y/o validaciones. LA PRACTICA SE DESARROLLARÁ EN EQUIPOS DE DOS INTEGRANTES (no se aceptaran reportes individuales)

1. **Revise los metadatos** de la fuente, disponible en: *http://www.aire.cdmx.gob.mx/descargas/datos/excel/REDDAxls.pdf*
2. Descargue, ***de la red de depósito atmosférico*** *http://www.aire.cdmx.gob.mx/default.php?opc=%27aKBk%27* , los archivos con terminación: "****PPH.XLS**", para el periodo del **"2010 al 2019".***
3. **Definición de flujo de trabajo del ETL, es decir reglas, pasos y transformaciones, para extraer y procesar el:**
    a. *Año,*
    b. *Elemento de medición, por ejemplo "Precipitación pluvial (PP)"*
    c. *La semana de medición, mes y año*
    d. *La ubicación de la medición*
    e. *Valor de la medición, por ejemplo "Precipitación pluvial (PP)"*
    f. Defina las reglas necesarias para transformar los datos, por ejemplo, colocar el número de semana del año. Etc.
4. Desarrolla la estructura de la tabla de hechos principal (estrectura de datos big data o de datamining) y los respectivos catálogos, todos los que sean necesarios.
    a. Catálogo de estaciones de monitoreo
    b. Catálogo de elementos de medición (en este caso que solo contenga al elemento de medición en estudio junto con toda su información (e.g. unidad de medición)
    c. La tabla de hechos principal mínimo debe tener la siguiente estructura: **{elemento},{localizacion},{noDiaDeSemana },{fecha},{medicion}**
5. Explore los datos integrados, indique y documente:
    a. Cantidad de registros totales y por año
    b. Tendencia en el tiempo de la precipitación pluvial
    c. Indique los lugares con mayor precipitación durante todo el periodo de estudio
6. Documente el **pseudocódigo** que explique al menos
    a. Código fuente usado y capturas de pantallas que muestre la ejecución de los pasos principales:En el código se debe explicar los pasos para realizar el proceso de integración de datos, reglas y validaciones usadas
    b. Modelo de datos empleado para catálogos y tabla de hechos principal

c. **Documente** la exploración de lo datos realizados
d. **Reportar valores atipicos y valores vacios y su impacto, y otros errores o problemas al procesar los datos.**

7. **Desarrollé el documento con el diseño preliminar bajo el siguiente formato.**
   - **Autores**
   - **Introducción (explicación de la propuesta de solución)**
   - **Desarrollo (incluir los resultados del punto 1 al 6)**
   - **Conclusiones**
   - **Anexe el archivo zip con los archivos fuente y la base de datos integrados**