Almacenes y Minería de Datos

Facultad de Ciencias UNAM

Dra. Amparo López Gaona < *alg@ciencias.unam.mx* >
M.I. Gerardo Avilés Rosas < *gar@ciencias.unam.mx* >

Tarea 2	**Extracción, transformación y carga**
Fecha de entrega	**28 de febrero de 2018**

1. Extracción, transformación y carga

Esta tarea tiene por objetivo que revises los aspectos básicos del proceso **ETL**. En este caso, se proporcionan tres conjuntos de datos separados y deberás obtener una **sola fuente** consolidada.

Los conjuntos de datos tienen varias diferencias, las cuales impiden que los registros se comparen directamente entre sí. El proceso que tendrás que realizar, será: ***extraer los datos de las tres fuentes que se proporcionan, transformar los datos y cargar los mismos en una nueva fuente, la cual contenga los datos combinados*** (*81 registros*).

Las características de las fuentes de datos son las siguientes:

- **Fuente 1**, la cual contiene **29 registros**.
- **Fuente 2**, la cual contiene **30 registros**.
- **Fuente 3**, la cual contiene **22 registros**.

Para este proceso será importante que **documentes** las fuentes de datos indicando, por ejemplo: ***nombre del atributo, tipo de dato, valores válidos, si debiera o no tener alguna regla de integridad***. De igual forma, deberás de generar una tabla (se recomienda hacer una pequeña **integración de esquemas**) en donde, después de realizar un análisis de los atributos, puedas indicar las **discrepancias** que se presentan en cada caso. El siguiente paso, será realizar una **reconciliación de los datos**, para esta parte, te puedes apoyar en herramientas como *Open Refine*, pero deberás indicar en una tabla, las transformaciones que tuviste que realizar (solo se muestran algunos campos a manera de ejemplo):

Atributo	Fuente	Transformación empleada
OrdenID	Fuente 1	
	Fuente 2	
	Fuente 3	
Nombre completo	Fuente 1	
	Fuente 2	
	Fuente 3	
Ciudad	Fuente 1	
	Fuente 2	
	Fuente 3	
Estado/Provincia	Fuente 1	
	Fuente 2	
	Fuente 3	
…	Fuente 1	
	Fuente 2	
	Fuente 3	

Puedes usar cualquier transformación que quieras, pero cuando termines, los datos deben ser formateados consistentemente en todo el conjunto de datos.

2. Limpieza de datos

Se proporciona un conjunto de datos pertenecientes al **MoMA (Museo de Arte Moderno)**. Deberás realizar un análisis de los atributos que lo componen y efectuar algunas tareas de limpieza. Utiliza para este proceso ***Open Refine*** (*apóyate en el material de clase*). Realiza una **tabla** donde indiques ***los atributos que seleccionaste, las inconsistencias que presentaban, la tarea que realizaste para resolver dichos problemas y las decisiones que tomaste***. Deberás entregar adicionalmente el conjunto de datos ya sin errores.

Tu tarea deberá estar correctamente documentada y la entrega es de acuerdo a los criterios para entrega de tareas (descritos en la página del curso). Deberás de enviar un comprimido (**.zip o .tgz**) a la dirección de correo electrónico (a más tardar a las **23:59 horas** del día indicado) al correo:

gar@ciencias.unam.mx

que contenga lo siguiente:

- Todos los aspectos que realizaste en la tarea ETL (indicados en el documento).
- La fuente integrada, en formato **CSV**.
- El análisis que realizaste para efectuar el proceso de limpieza
- El dataset sin errores, en formato **CSV**.

Nota:

Cualquier duda o comentario que pudiera surgirte al hacer tu tarea, recuerda que cuentas con la **Lista de Correo** del grupo: **dwhYdm@ciencias.unam.mx** en donde seguramente encontrarás las respuestas que necesites.

¡SUERTE!