Facultad de Ciencias UNAM
Almacenes y Minería de Datos 2018-2

Práctica III: Integración de Datos

Profesor: Amparo López Gaona
Ayudante de laboratorio: José Luis Vázquez Lázaro

24 de febrero de 2018

1. Objetivo

Al finalizar esta práctica el alumno será capaz de:

- Entender las inconsistencias entre los esquemas, dominios, definiciones y abreviaturas.
- Conocer una metodología para atacar el problema de integración de orígenes de datos.

2. Marco teórico

La Integración de Datos la podemos definir como el proceso de combinar datos que residen en diferentes fuentes y permitirle al usuario final tener una vista unificada de todos sus datos. La habilidad de transformar datos de fuentes heterogéneas para su integración es un plan de acción que se ha convertido en un reto y en una ventaja competitiva para compañías que requieran aplicaciones de Inteligencia de Negocio. La integración de datos es un elemento fundamental y crítico en la construcción de los Almacenes de Datos.

3. Instrucciones

- Descargar el archivo `BD_Autos1.backup` (base de datos correspondiente al caso de uso) y restaurarla en el Sistema Manejador de Bases de Datos PostgreSQL.

- Descargar el archivo `BD_Autos1_ModeloER.png` (modelo E-R correspondiente a la base de datos anterior)
- Descargar el archivo `BD_Autos2.cvs` (pseudo base de datos correspondiente al caso de uso).

4. Caso de uso

Varias de las sucursales de la Agencia "Continental" no utilizan la BD creada para su proceso de negocio, en su lugar utilizan la aplicación de hojas de cálculo Excel para registrar todas las transacciones que realizan. Esto ha provocado el descontentos de los directivos de la empresa, quienes exigen una solución pronta y eficaz. Por lo que la empresa ha decidido contratar de nueva cuenta a estudiantes de la carrera en Ciencias de la Computación (específicamente a los estudiantes de Almacenes y Minería de Datos 2018-2) para que unifiquen toda esta información.

5. Actividades

1. Analizar los diversos orígenes de datos de acuerdo al listado de puntos que se describen en la siguiente sección y construir un documento que describa los puntos del listado para cada conjunto de datos.
2. Considerando los puntos descritos en su documento de análisis, reconciliar las definiciones, nomenclaturas, tipo de datos y nivel de grano; y diseñar una nueva base de datos que sirva como receptáculo de todos los datos.

6. ¿Qué debo considerar para integrar fuentes de datos?

Cuando se tengan disponibles los diversos orígenes de datos, deben ser integrados en un solo punto, para ello debe tener en cuenta lo siguiente:

- ¿Qué entidades posee el conjunto de datos? y ¿cómo se relacionan?
- ¿Cómo se define cada entidad?
- ¿Cómo se define cada atributo? y ¿cuál es su dominio de datos (valores válidos)?
- ¿Cuál es la representación de valores ausentes?

- Mantener el nombre de las relaciones y sus atributos
- Identificar los tipos de datos de cada atributo.
- Identificar llaves primarias.
- Identificar si existen atributos que:
  - Son parte de un mismo atributo y se fraccionaron.
  - Están fraccionados y deben unificarse.
- Información de la precedencia de los datos.

7. Entregables

Deberás enviar un archivo `.zip`, con nombre `<número de cuenta>_practica03`, que contenga lo siguiente:

- Un archivo `.pdf` que contenga el análisis de los diversos orígenes de datos, así como la descripción de la unificación.
- El esquema de la BD Unificadora, integrado por el Modelo E-R (archivo `.png`) y el Modelo relacional (archivo `.png`).
- Un archivo `README.txt` que contenga tu nombre completo, tu número de cuenta y tu correo.

a la direccion de correo `luis_lazaro@ciencias.unam.mx` con el asunto `[A&MD2018-2]Practica03`.