IIC2413

AYUDANTÍA 3

LIMPIEZA DE DATOS

¿QUÉ VEREMOS?

Teoría carga y liempiza de datos

PHP para el procesamiento

Limpieza en la practica

¿CÓMO SE HACE?

1. Analisis de los datos

   Analizar los datos para ver la detección de errores.

2. Limpieza de Datos.

   Leer los datos desde los archivos y limpiar los datos con un programa de PHP. Entregar los datos en nuevos archivos.

¿LIMPIEZA DE DATOS?

ERRORES EN LOS DATOS

01

No estandarizados

02

Tipos de dato incorrecto

03

Datos nulos (cuando no se permite)

IMPORTANTE

No pueden alterar el archivo base, en caso de corregir un dato tiene que ser mediante el programa php.

1. ESTANDARIZACIÓN DE FORMATO

- Datos que represetan lo mismo pero en diferente forma. Deben ser estandarizados. Ejemplo: DD/MM/YY, MM/DD/YYYY

- Datos con errores menores. Son fáciles y rápidos de solucionar

1. TIPO DE DATOS INCORRECTOS

- El tipo de dato recibido no corresponde al tipo de dato definido

Ejemplo: Se espera un int (500) y se recibe un float (500.0)

- El dato no representa lo pedido, se coloca donde corresponde y se elimina donde no debe

Ejemplo Se espera un apellido (Parra) y se recibe un correo (parra@uc.cl).

¿QUE HACER ANTE UN ERROR?

Principalmente nos vamos a enfocar en 3 opciones para corregir un dato.

1. Correción de datos
2. Asignación de nulo
3. Eliminación

Siempre tenemos que buscar la pérdida mínima de información

2. ASIGNACIÓN DE NULO

Se utiliza en casos que el dato no puede ser corregido, pero no es fundamental para la tupla

Esto se realizan en casos que el valor por default no pueda ser nulo, por lo que se asigna valor decidido por ustedes.
Ejemplo: numero de telefono 5427 se reemplaza por 000000000.

3. ELIMINACIÓN DE DATO

- Esta medida debe aplicarse solamente como último recurso
- ¿Por qué? Siempre debemos piorizar la pérdida mínima de datos
- Consdiderar esta opción solo en casos donde los datos hayan perdido casi todo su valor
- Si se llega a optar por esta opción se debe de eliminar en cascada

ACTIVIDAD: ANALISIS LIMPIEZA

1.
Errores en identificadores de aviones: En una base de datos de vuelos, los códigos de avión presentan inconsistencias: algunos tienen números faltantes (B EING- 47), la marca que existe es (BOEING-747). Hint: Se necesita información adicional.

2.
Uso de diferentes unidades de medida: En un base de datos de productos, algunos pesos están en kilogramos (kg), otros en gramos (g) y algunos en libras (lb). Hint: Hacer la conversión.

ACTIVIDAD: ANALISIS LIMPIEZA

3.
Inconsistencia en nombres de categorías: En una base de datos de clientes, el campo "País" tiene valores como USA, United States y EEUU para referirse a lo mismo. Hint: estandarizar un valor.

4.
Errores en identificadores de vehículos: En una base de datos de automóviles, algunos registros presentan números de chasis escritos con distintos formatos (1HGCM82633A123456, 1HG CM8-2633A 123456 o 1HGCM8263A12345 con un dígito faltante), lo que impide validar el identificador correctamente. Hint: análisis de caso por caso

ACTIVIDAD: ANALISIS LIEMPIEZA

5.
Errores en números de teléfono: En una base de datos de contactos, algunos números de teléfono celular tienen menos de 9 dígitos (12345678) o más de lo esperado (9123456789)

6.
Formato de números incorrecto: En una base de datos financiera, algunos valores monetarios están escritos con coma como separador decimal (1,500.75), mientras que otros usan punto (1500,75), siendo este último el formato correcto.

ACTIVIDAD: ANALISIS LIMPIEZA

7.

Atributo opcional con datos erróneos: En una base de datos de reseñas de productos, algunos registros tienen valores incorrectos en la columna "Calificación", donde se encuentran valores fuera del rango permitido como 8, dado que el formato es de 1 a 5 y la calificación no es un dato obligatorio para el resto de la información. Hint: Anular el dato, datos fuera de rango igual a 0.

CIERRE

1. Se arregla lo que se puede arreglar, muchas veces obteniendo información adicional. Ejemplo: RUN sin dígito verificador.
2. Si no se puede arreglar:

- Dejarlo en NULL.
- En caso de campo obligatorio (no NULL): valor que “signifique" nulo en nuestro contexto. Ejemplo: teléfono: 999999999.
- Si no es crítico: borrarlo.
- Si es crítico: borrar la tupla (ON DELETE CASCADE).

¡MUCHAS GRACIAS!

IIC2413

AYUDANTÍA 3

LIMPIEZA DE DATOS