Estrategias para Evaluar base de datos con datos duplicados y Mejora de los datos

Autores: Johao Hernandez, Jeiner Cantillo, Jader Gonzalez

IES INFOTEP CIENAGA,

Abstract

Este informe se enfoca en el desafío de los datos duplicados en bases de datos, un problema común que afecta la precisión y eficiencia del análisis de datos. Estamos investigando y documentando estrategias efectivas para evaluar la calidad de las bases de datos, identificar y eliminar datos duplicados y mejorar la integridad general de los datos. Esto incluye la exploración de diversas técnicas y herramientas, desde métodos manuales hasta soluciones automatizadas .

palabras clabes: Deduplicación,Calidad de datos,Integridad de datos y Bases de datos

1 Introduction

Este informe se centra en el desarrollo de estrategias efectivas para la evaluación y gestión de datos, con énfasis en la identificación y eliminación de duplicados. Abordaremos la selección y de los datos, así como la implementación de técnicas para garantizar la integridad y calidad de la información. El objetivo es establecer un marco de trabajo que permita mejorar la precisión y eficiencia en el manejo de datos, minimizando errores y optimizando los procesos de análisis.

2 Objetivos del Informe

Este informe tiene como objetivo proporcionar una guía de estrategias y de las mejores prácticas para la identificar y prevenir los datos duplicados en bases de datos de cualquier tipo, permitiendo a los usuarios implementar soluciones efectivas para mantener la integridad y calidad de sus datos.

3 Descripcion de la actividad

Esta actividad se centra en abordar este problema mediante la implementación de estrategias efectivas para evaluar una base de datos en busca de datos duplicados. Además, se enfoca en la mejora de los datos, lo que implica no solo identificar y eliminar duplicados, sino también garantizar la integridad y precisión de la información.

3.1 Estrategias para evitar la insercion de datos duplicados:

Para prevenir la inserción de datos duplicados en una base de datos, se pueden aplicar técnicas determinísticas antes de almacenar los datos. Estas técnicas permiten identificar posibles duplicaciones en tiempo real y asegurar la integridad de la información almacenada.

Funciones Determinísticas: Se entrena un modelo de funciones como "duplicated()" (para identificar si los datos son duplicados) y "drop-duplicates()"(para eliminar esos datos duplicados) Implementar estas soluciones permite una mejor organización de la información, reduciendo errores y facilitando la toma de decisiones basada en datos confiables.

Normalización y Limpieza de Datos: Convertir texto a minúsculas, eliminar caracteres especiales y espacios redundantes para facilitar la detección de duplicados.

Validación en Tiempo Real: Implementar verificaciones antes de insertar nuevos datos mediante consultas SQL (SELECT EXISTS o ON DUPLICATE KEY UPDATE).

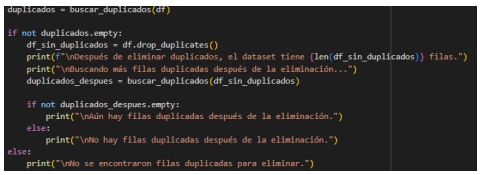

Figure 1: en esta imajen podemos observar se implementa las funciones duplicado() y drop-duplicates()

3.2 Mejora de Datos en Database con datos duplicados :

Los datos duplicados en una base de datos pueden generar inconsistencias, errores y un rendimiento deficiente. Por lo tanto, es crucial implementar estrategias efectivas para identificar, eliminar y prevenir la duplicidad de datos. Este informe presenta métodos y consideraciones generales aplicables a la mayoría de las bases de datos.

Consultas SQL: El Poder de la Precisión :

SQL es una herramienta fundamental para manipular bases de datos relacionales. Al utilizar GROUP BY, podemos agrupar registros que comparten valores en columnas específicas. Esto nos permite contar cuantas veces aparece cada combinación única de valores. La cláusula HAVINGactúa como un filtro posterior a la agrupación, permitiéndonos seleccionar solo aquellos grupos que cumplen con ciertas condiciones, como aquellos que aparecen más de una vez, indicando duplicados.

COUNT(*): Es una función agregada que cuenta el número total de filas en un grupo, revelando la frecuencia de cada registro.

ROW NUMBER(): Esta función de ventana asigna un número secuencial a cada fila dentro de una partición de un conjunto de resultados. Esto es de gran ayuda para identificar el "número de orden" de cada registro, así poder diferenciar entre los registros duplicados.

Herramientas de Software: Simplificando el Proceso : Los DBMS modernos, como MySQL, PostgreSQL y Microsoft SQL Server, incluyen funcionalidades para detectar y gestionar duplicados. Estas herramientas suelen ofrecer interfaces gráficas y asistentes que simplifican el proceso de identificación y eliminación. Hojas de cálculo como Excel son valiosas para conjuntos de datos más pequeños. Sus funciones de filtrado y formato condicional permiten resaltar y eliminar duplicados con facilidad.

Eliminación de Duplicados: Estrategias Efectivas

Eliminación Directa: Precaución y Control: La sentencia DELETEen SQL permite eliminar registros específicos de una tabla. Es crucial utilizarla con precisión, ya que la eliminación incorrecta puede provocar la pérdida de datos valiosos. Las copias de seguridad son indispensables antes de realizar cualquier eliminación masiva. Esto garantiza que puedas restaurar la base de datos en caso de errores.

Desduplicación Basada en Reglas: Decidiendo Qué Conservar : En muchos casos, los duplicados no son idénticos. Pueden tener pequeñas diferencias en ciertos campos. En tales situaciones, es necesario definir reglas para determinar qué registro conservar.

Por ejemplo, si tienes registros duplicados de clientes, podrías decidir conservar el registro con la información de contacto más reciente o el registro con la información más completa. Normalización de Datos: Prevención a Largo Plazo : La normalización es un proceso de diseño de bases de datos que reduce la redundancia y mejora la integridad de los datos. Implica dividir los datos en tablas relacionadas y definir claves primarias y foráneas para establecer relaciones entre ellas. Una base de datos bien normalizada es menos propensa a la duplicación de datos.

4 Conclucion

La gestión de una base de datos no solo se trata de almacenar información, sino de garantizar que esta sea precisa y sin errores como duplicaciones. Evitar y eliminar datos redundantes no solo optimiza el rendimiento del sistema, sino que también mejora la confiabilidad de la información para la toma de decisiones. La implementación de estos métodos, como validación en tiempo real y uso de índices, nos permite mantener la integridad de los datos.