Universidad Central "Martha Abreu" de Las Villas

Facultad de Matemática, Física y Computación

Sistema para la Limpieza de Datos

Tesis presentada en opción al Título Académico de Master en Computación Aplicada

Autor: **Ing. Edgar Alejandro Martínez Aguiar**
Tutor: **Dr. Ramiro Pérez Vázquez**

Guadalajara, Jalisco, México
2003

INDICE

RESUMEN

En la actualidad la limpieza de datos es un conjunto de técnicas muy especializadas y que podría llegar a considerarse toda una ciencia.

La limpieza de datos es mucho más que actualizar un registro con datos "correctos". La Limpieza de Datos especializada incluye "Descomposición" y "Reensamblamiento" de Datos.

Se podría separar esta metodología en 6 principios básicos:

- Elementarizar
- Estandarizar
- Verificar
- Relacionar
- Domiciliar
- Documentar.

Este trabajo pretende tomar un algoritmo especializado en "Limpieza de datos" y mejorarlo, para finalmente aplicar los conocimientos adquiridos desarrollando una herramienta especializada en "Limpieza de Datos".

ABSTRACT

Actually, the Data Cleaning is a group of speciality techniques and could be consider a Science.

The Data Cleaning is much more than simply updating a record with good data. Serius Data Cleaning involves decomposing and reassembling the data.

Can break down the cleaning into six steps:

- Elementizing
- Standardizing
- Verifying
- Matching
- Householding
- Documenting

This work pretends enhance an speciality "Data Cleaning" algorithm, and finally apply the knowledge acquired developing a speciality Data Cleansing tool.

INTRODUCCIÓN

El almacenamiento de la información siempre ha sido y será un gran reto de la informática desde los inicios de los Sistemas Computacionales.

Originalmente se pensó en una forma sencilla y natural de almacenar toda esta información. Los datos se empezaron a guardar de una manera ordenada y en cierta forma catalogada. Esto no fue suficiente, ya que mientras la información seguía llegando, el almacenamiento se hacía cada vez más complejo.

Poco después de la gran revolución que causaron las Bases de Datos, los Gestores de Bases de Datos, y todas las aplicaciones relacionadas para almacenar la información de una manera segura los empresarios principalmente cada vez más se preguntaban cómo explotar toda esta información.

Inicialmente se crearon sistemas "a la medida" que fueron cubriendo esta necesidad de "tratar" la información. Poco a poco se crearon técnicas como las bases de datos, aparecieron diferentes modelos de datos capaces de representar la información, se desarrollaron gestores potentes que ayudaban a la creación de Sistemas de Bases de Datos.

Con la complejidad de los grandes Sistemas de Información como ERP´s, MRP´s, etc. se tuvo la necesidad de llevar un ordenamiento especializado de la información, se buscó la forma de tener un control del correcto almacenamiento de estos datos, gracias a esta necesidad se implementaron filtros, esquemas, políticas y hasta restricciones para que toda la información retenida en la Base de Datos contara con una especie de "Logística Informática".

Esto fue posible hasta cierto punto, ya que con el surgimiento de herramientas y técnicas especializadas para el tratamiento y explotación de la información se llegó a la conclusión de que se tenía un gran problema.

Después de gastar miles de dólares en la elaboración de sistemas a la medida, Sistemas Gestores de Bases de Datos, consultoría especializada para el diseño y elaboración de una

Base de Datos y toda una gran infraestructura de hardware se llegó a una triste pero acertada conclusión: los datos de ese gran Almacén de Datos no eran confiables.

La razón, es muy simple: si tenemos los datos de miles de productos, precios, personal, salarios, etc. y queremos extraer información con el fin de hacer un análisis exhaustivo para un estudio de mercado o algún otro análisis de información nos damos cuenta de que nuestros datos NO son correctos por diversas razones:

- Nuestros datos sufrieron un error humano. Es decir, no se capturaron correctamente.
- Los datos (algunos campos) no fueron capturados generando campos vacíos.
- Nuestra información está parcial o totalmente duplicada en uno o todos los registros.
- Los datos al ser importados pueden no almacenarse correctamente por diversas razones. No todos los gestores de bases de datos cubren este tipo de fallas.

Es por esta y diversas razones las que nos pueden llevar a encontrar alguna metodología que nos auxilie en corregir este tipo de errores lógicos o humanos. Estas técnicas son prácticamente "nuevas" y son conocidas como "Limpieza de Datos" ("Data Cleansing").

Con los grandes volúmenes de información que se manejan en la actualidad y las cantidades tan elevadas en el costo de cualquier proyecto de Tecnología de Información (como por ejemplo la creación de un Almacén de Datos) es necesario tener la certeza de que esa información que va a ser explotada tenga un porcentaje de confiabilidad de un 100 %.

De nada nos serviría tener una gran base de datos con toda la información que pudiéramos necesitar (como estados financieros, estados de resultados, proyecciones, etc.) si cuando tratamos de consultar algo de esa información resulta ser que la misma NO es tan confiable como pensamos.

Igualmente podríamos contratar al mejor consultor de DataWareHouse (Ralph Kimball) para que nos elaborara el proyecto más costoso jamás creado con las mejores herramientas de Inteligencia de Negocios y/o toma de decisiones y de nada nos serviría si nuestros datos NO son confiables.

Lo que tenemos que tratar como primer paso es contratar los servicios de un consultor especializado en Limpieza de Datos para que a través de técnicas especializadas en el tema realice una "purga" y estandarización de TODOS nuestros datos y mediante esto tener la certeza de que nuestros datos son 100 % confiables y gracias a esto aplicar cualquier metodología o herramienta propia de la Inteligencia de Negocios.

Las empresas pierden diariamente cientos de miles de dólares por realizar llamadas telefónicas incorrectas, poseer datos erróneos, enviar correspondencia a direcciones equivocadas u ofrecer un servicio a un cliente no calificado. El tema puede ser grave: desde un aumento de los costos operativos del 40%, hasta un considerable daño de la imagen corporativa.

La reingeniería de datos y la provisión de soluciones para las áreas de Marketing y Ventas pueden llegar a generarle a una empresa con una cartera de 2 millones de clientes - el promedio de los grandes bancos y tarjetas de crédito, por ejemplo -, un ahorro mensual de 180.000 dólares. Esto, entre otras cosas, es consecuencia de que no se envíe folletería, cartas o resúmenes de cuentas a clientes equivocados. Con una lista de datos ordenada y sin superposiciones ni duplicaciones, las compañías pueden añadir información en forma eficiente con la garantía de que no habrá errores, además de reasegurarse de que los envíos masivos tendrán el destino correcto, y que habrá una mayor personalización con el cliente.

Aunque la mayoría de las empresas saben que necesitan invertir dinero en este tipo de servicios, pocas lo hacen por la complejidad a la que suponen deben enfrentarse, cuando en realidad no sólo es sencillo sino que también muy redituable".

Esto se puede resumir en la frase siguiente: "Conseguir un nuevo cliente cuesta seis veces más que mantener uno que ya tenemos".

Objetivos:

Objetivo General

Crear una primera versión de una herramienta de limpieza de datos que incluya la sustitución de valores nulos o vacíos, reemplazo de cadenas y homogeneización de datos.

Objetivos específicos

- Realizar una sistematización de los conceptos y metodologías relacionados con la limpieza de datos de manera que sirva de material de estudio sobre el tema.
- Determinar los elementos más importantes que se deben tener en cuenta en la limpieza de datos.
- Realizar el análisis, diseño de una herramienta que ayude a la limpieza de datos e implementar una primera versión de la herramienta de limpieza.

FORMAS DE PRESENTACIÓN DE RESULTADOS

La principal forma de presentación será la tesis escrita y digital, además de un programa (herramienta) para la limpieza de datos de los Gestores de Bases de Datos más utilizados.

Este trabajo se enfocará principalmente en la elaboración de la herramienta antes mencionada, pero sin olvidar la importancia de los conceptos para seguir una eficiente "Limpieza de Datos".

Es por ello que tanto el análisis de los conceptos como la herramienta propia tendrán igual prioridad e igual divulgación dentro del ámbito informático en las posibilidades que los medios lo permitan.

Para lograr estas metas se pretende elaborar diversos artículos para revistas científicas, revistas electrónicas, ponencias, programas de radio además de la elaboración de la misma Tesis.

Por lo tanto, los preguntas a responder serían:

COMO ?

- Evitar duplicidad en los datos.
- Corregir errores de captura u ortográficos.
- Lograr una reducción del tamaño físico de las bases de datos.
- Lograr una optimización en la respuesta de los SGBD.
- Evitar alias en los registros.
- Homogeneizar información de fuentes dispares.

CAPITULO 1: ESTUDIO GENERAL SOBRE LIMPIEZA DE DATOS

1.1 Concepto de limpieza de datos

La limpieza de datos tiene varias definiciones desde el punto de vista de diferentes autores, cada autor define la limpieza de datos según el uso que le da a este complejo proceso y a la complejidad a la que se enfrenta al realizar este proceso.

Trataremos de dar el punto de vista de diferentes autores y formar una opinión al respecto sobre como definir la Limpieza de Datos:

Elymir Urdaneta en [URD01] ofrece la siguiente definición: Limpieza de Datos: proceso de asegurar que todos los valores en un conjunto de datos sean consistentes y correctamente registrados.

A pesar de que la "Limpieza de Datos" puede tomar muchas formas, muchos de los más importantes ejemplos de "Limpieza de Datos" se tienen de la necesidad de contar con buenas descripciones de cosas tangibles como los clientes, productos, procedimientos y diagnósticos. El mercado actual y la tecnología actual para la limpieza de datos están fuertemente enfocados en las listas de clientes, o sea la limpieza se realiza para los datos asociados a estos clientes: nombres, direcciones, teléfonos, sexo, etc.

En [KIM96] Kimbal señala que la limpieza de datos es mucho más que simplemente actualizar un registro con datos correctos. La "Limpieza de Datos" seria involucra descomposición y reensamblamiento de datos. Plantea la descomposición de la limpieza en seis pasos:

- Elementarizar
- Estandarizar
- Verificar
- Relacionar

- Asignar Origen de Datos
- Documentar

Por su parte Helena Galhardas en [GAL01] señala que el problema de la limpieza de datos, que consiste en remover inconsistencias y errores desde los conjuntos originales de datos, es bien conocida en el área de "Sistemas de Soporte de Decisiones y DataWareHouse".

La calidad de los datos surge cuando se eliminan las anomalías en una sencilla fuente de datos, o integrando datos provenientes de múltiples fuentes en un simple repositorio de datos. La información manipulada podría necesitar experimentar un proceso de formateo y normalización que resultará en datos estructurados y presentado acordando a los requerimientos de la aplicación.

La misma autora en [GAL01A] plantea cuales a su juicio son las principales causas de las anomalías en los datos:

- La ausencia de Llaves Universales sobre diferentes Bases de Datos que es conocido como el Problema de Identidad del Objeto.
- El uso de diferentes formatos de datos.
- La existencia de errores en la entrada de datos.
- La presencia de inconsistencias en los datos.

Y señala que tratar con estos problemas es globalmente llamado el proceso de "Limpieza de Datos".

En [LOS01] se afirma que la Limpieza de Datos se da por que los clientes se cambian de dirección, se casan o divorcian, y cambian sus nombres, largas bases de datos de clientes frecuentemente pierden entropía en términos de exactitud. Además, cuando los humanos son responsables de la entrada de datos, los errores son posibles. Estos escenarios pueden conducir a la existencia de registros duplicados para muchas entidades individuales o corporativas. Como en una tabla relacionada (imaginaria que contiene datos extraídos de

"U.S. Securities and Exchange Commission's Edgar database"), que muestra que muchos de esos duplicados no son siempre fácilmente reconocidos automáticamente. Ligando los registros duplicados es una forma de limpiar la Base de Datos y eliminar esos duplicados. Muchas veces hay varias formas de representar el mismo concepto.

Este mismo autor plantea que otro proceso llamado "householding", el cual envuelve la consolidación de un conjunto de unidades individuales en simples agrupamientos basados en los atributos de los registros. Un sencillo ejemplo es aquel en el que consolidamos el registro del Esposo y de la Esposa en un origen de datos. [LOS01]

A partir de todos estos elementos proponemos la siguiente definición de "Limpieza de Datos":

> La Limpieza de Datos es una técnica que involucra varios procesos para generar consistencia y credibilidad en los datos que contienen las bases de datos, además la complejidad de estas técnicas son directamente proporcionales al tamaño de la Base de Datos a "limpiar".

1.2 Necesidad de la limpieza de datos

La limpieza de datos se hace necesaria cuando queremos hacer una explotación de los datos almacenados en una base de datos operacional o cuando verificamos que hay inconsistencias en los datos que queremos explotar. También se hace necesaria cuando nos vemos en la necesidad de extraer datos de diferentes fuentes de datos (como pueden ser diferentes Bases de Datos) y encontramos que los tipos de datos entre las fuentes de datos son diferentes, esto puede generar una seria inconsistencia de datos que hace inexplotable los datos extraídos de estos sistemas.

Enseguida se citan las opiniones de varios autores acerca de cuando realizar una limpieza de datos profesional:

En [DAT98] se afirma que hoy en día los negocios tratan de entender su base de consumo. Para ayudarlos con esta misión, las organizaciones ponen información de contacto, patrones de compra de los clientes y otros datos orientados a los clientes en los

Sistemas de Soporte a Decisiones. Sin embargo, ¿qué pasaría cuando un gerente encuentra información acerca de Janet Smith, Jan Smith, and J. Smith?. ¿Son todas estas entradas de datos independientes?. ¿Es un registro duplicado?. ¿Qué pasa si todos ellos son la misma persona? ¿Será posible determinar tales cosas o encontrar la forma para mejores decisiones basadas en datos puros, o datos "limpios".

Orli R en [ORL96] se pregunta ¿por qué herramientas?. Señala que las herramientas pueden ayudar a lograr la eficiencia técnica, y pueden realizar tareas de calidad de datos que de otra forma sería impráctico. En los procesos de migración de datos es importante utilizar este tipo de herramientas para lograr un correcto manejo de datos.

Kimball en [KIM96] plantea que una de las razones para la ausencia de herramientas para comprobar la integridad de datos está en los equipos (de desarrollo) que no consideran adecuadamente el impacto en el negocio de los datos erróneos. Señala con mucha fuerza que la robustez de las aplicaciones de Almacenes de Datos depende de contar con buenos datos.

En el trabajo [GAL01A] se plantea que asegurar la calidad de los datos en los sistemas de información es crucial para el soporte a decisiones y las aplicaciones orientadas a negocios (correo electrónico, detección de fraudes, análisis de riesgos, etc.).

Se considera que la calidad de datos tiene sus orígenes en 3 diferentes contextos:

a) Cuando se busca corregir anomalías en un sencillo origen de datos (Eliminación duplicada en un archivo).

b) Cuando los datos pobremente estructurados o no estructurados son migrados en datos estructurados.

c) Cuando se quiere integrar datos provenientes de múltiples fuentes en una simple fuente de datos nueva (Construcción de un Almacén de Datos).

La principal meta de la Limpieza de Datos es eliminar datos anómalos en cada una de estas situaciones.

Roger Frauca en [FRA01] coincide con otros autores en que el primer gran problema que se enfrenta en la carga de un Almacén de Datos es la diversidad de fuentes posibles que suelen abastecerlo de datos de origen, así no es nada extraño que un Almacén de Datos tenga que adquirir cierta información de una parte de un AS/400 con su respectivo DB2, otra parte de una máquina Unix con Oracle por ejemplo y otra parte de una máquina con NT y SQL Server. Tan sólo para recopilar las tablas de cada base de datos habría que resolver un buen número de problemas, tanto de compatibilidad de datos, como de programación para acceder a ellos, entre otros muchos. Una posible solución sería homogeneizar las fuentes heterogéneas y la adquisición de los datos, independizándolas de su origen. Así, a la hora de tratar los datos se verían igual los datos de Oracle, Informix, DB2, SQL Server, etc. [FRA01]

Por ejemplo, comúnmente los sistemas de procesamiento de transacciones no tienen control sobre los nombres de los clientes. El resultado puede dar como resultado que un sencillo cliente pueda ser identificado como "Acme Dist.", "Acme Distributing Company", y "Acme Distribution". En un simple ejemplo como este se puede ver que este tipo de problemas origina gran dificultad para el desempeño de usuario e igualmente en la selección de la consulta de un nombre de cliente.

En otros casos, los números de identificación de productos pueden ser traspuestos, es decir, varios códigos para identificar a un mismo producto. Los clientes más antiguos podrían resistirse a cambios en los códigos de sus productos y entonces pudiera pasar que el mismo producto aparezca con códigos diferentes.

Esa condición de "suciedad" en los datos descritos anteriormente da como resultado la necesidad de hacer una limpieza de datos existentes de los sistemas fuentes, que es tan importante como el mejor esfuerzo de programación para implementar campos de edición y controles en esos sistemas de operación para prevenir la "suciedad" de los datos desde que son capturados en el primer módulo. A menos que el equipo del Almacén de Datos esté preparado para tratar con esto rápida y eficientemente, hay un alto riesgo de que el proyecto se retrase o se tenga una pérdida de credibilidad por parte del Almacén. [DWI01]

1.3 Problemas fundamentales al limpiar Bases de Datos

Uno de los principales problemas al realizar la limpieza de datos es la **homogeneización** de los datos de diferentes fuentes de datos de las que se extrae la información, es decir, buscar una forma de igualar y estandarizar los datos es algo realmente complejo que como se mencionó anteriormente, se vuelve más complejo cuando se tienen más registros a igualar.

Un ejemplo clásico sería el siguiente:

Imagine que en una base de datos de una agencia de viajes se capturan los datos del cliente, entre ellos el lugar de origen de donde viajara y el lugar del destino. Entre los datos a capturar se encuentra la ciudad y los usuarios la capturan con las siguientes expresiones:

- Guadalajara
- GDL
- Guad
- 03
- Gdl

En este caso tendríamos que ingeniárnosla para igualar todas las entradas de datos que se refieren a la ciudad de Guadalajara.

Otro caso muy común es aquel en el cuál los **campos** de los registros aparecen **vacíos**, o **nulos**. Este es otro grave problema que se enfrenta en la limpieza de datos, ya que en la construcción de un Almacén de Datos los campos con ceros o vacíos al momento de ser explotados por un Sistema de Toma de Decisiones son contabilizados y graficados afectando los promedios, pronósticos y en gran medida afectando la confiabilidad de un sistema que fue diseñado para la toma de decisiones a nivel gerencial o directivo.

Un ejemplo sería el siguiente:

Imagine que tiene una base de datos con 100 registros, de los cuales 20 tienen una columna con campos vacíos que no se capturaron. Al momento de promediar la columna

citada (suponiendo que se trata de la columna donde se registra el total de la venta hecha a un cliente) nos daría un resultado de "X" cantidad que sería la suma de los totales de la columna divididas entre 100, no entre 80 como sería la forma más razonable de hacerlo, ya que el gestor de Bases de Datos toma en cuenta todos los registros con que cuenta la tabla para realizar el cálculo solicitado, en este caso el Promedio del total de los registros.

El problema de **llaves únicas** entre diferentes fuentes de datos es algo que se da frecuentemente cuando se extraen datos de diferentes bases de datos y que el producto sigue siendo el mismo. Es decir, imaginemos que se tiene una tienda de autoservicio que agrupa a diferentes estados de la república mexicana, cada una de las tiendas maneja un refresco de cola con una presentación de 500 mililitros. El problema se da cuando se desea tener una estadística de cuantas botellas se vendieron en el mes de marzo y cada una de las tiendas maneja un código diferente para el mismo producto:

- Coca Cola 500 ml. (Código CL-0012500). Michoacán
- Coca Cola 500 ml. (Código CC-5000012). Guanajuato
- Coca Cola 500 ml. (Código CO-0050012). Monterrey

En este ejemplo sería un gran reto igualar los códigos para llegar a tener esa estadística o gráfica de las ventas mensuales de este producto, si tenemos en cuenta que solo contamos con el código, la cantidad vendida y el importe de cada uno de los productos de las 3 sucursales.

1.4 Técnicas para limpiar

Algunas de las técnicas para limpiar datos se toman de las diferentes necesidades que presenta cada una de las bases de datos con las que se desea trabajar, es decir, hasta el momento no existe una técnica "Universal" para la limpieza de cualquier tipo de Base de Datos.

Algunas de las técnicas que se plantea en la literatura referente a este problema son las siguientes:

1.4.1 Algoritmo Soundex.

El algoritmo Soundex es utilizado para codificar palabras a partir de su sonido y realizar búsquedas de manera que no se igualen cadenas de caracteres, como habitualmente sucede, si no se busque por el sonido.

Algunas características de este algoritmo se explican a continuación

- Los códigos del algoritmo soundex empiezan con la primera letra de la palabra seguida de un código de 3 dígitos que representan las 3 primeras consonantes. Los ceros pueden ser agregados a los códigos cuando no cuente con las letras suficientes para ser modificado.

- La guía para el codificado soundex es la siguiente (Consonantes que suenen igual tienen el mismo código).
  1 - B,P,F,V
  2 - C,S,G,J,K,Q,X,Z
  3 - D,T
  4 - L
  5 - M,N
  6 - R

- Las letras A,E,I,O,U,Y,H, y W no son codificadas.

- Nombres con letras adyacentes tienen un número equivalente para ser codificadas como una letra con un solo número.

- Los prefijos de los apellidos como La, De y Van no son usados generalmente en el código Soundex. Mc, Mac y O generalmente no son considerados prefijos por la codificación soundex.

Para calcular un código Soundex se utiliza el siguiente algoritmo:

- Eliminar los espacios, puntuación, acentos y otras marcas

- Eliminar cualquiera de los siguientes caracteres A, E, I, O, U, H, W, Y

- Eliminar la segunda letra de los caracteres duplicados
- Eliminar la segunda letra de los caracteres adyacentes con el mismo número soundex.
- Convertir los caracteres en las posiciones 2 a 4 en un número
    1. B, P, F, V -> 1
    2. C, S, K, G, J, Q, X, Z -> 2
    3. D, T -> 3
    4. L -> 4
    5. M, N -> 5
    6. R -> 6
- Finalmente en cualquier posición no usada agregar ceros ejemplo: Lee es L00, Bailey es B400. Siempre será una letra seguida de 3 números.

Limitaciones del Algoritmo Soundex

- Los nombres que suenan igual NO siempre tienen el mismo código soundex. Por ejemplo, Lee (L000) y Leigh (L200) tienen idéntica pronunciación, pero tienen diferentes códigos soundex por que la letra g (que no se pronuncia) cuenta con un código soundex.
- Los nombres que suenen igual, pero que inicien con una letra diferente siempre tendrán un código soundex diferente. Así que, nombres como Carr (C600) y Karr (K600) tienen códigos distintos.
- El algoritmo Soundex está basado en la pronunciación del inglés, así que los nombres Europeos no podrían ser codificados correctamente. Por ejemplo, algunos apellidos Franceses con letras que no se pronuncian al final no se codificarían de acuerdo a la pronunciación. Esto se aplica para un nombre francés como Beaux –donde la X no se pronuncia. Algunas veces el apellido

es mal pronunciado. Beau (B000) y su pronunciación idéntica tienen código diferente Beaux (B200). Sin embargo solo se expone un ejemplo francés, que podría aplicarse a cualquier nombre que no utilice pronunciación en inglés. Hemos encontrado en la literatura el algoritmo aplicado al idioma alemán donde el agrupamiento de las letras es otro.

- Algunas veces los nombres que no suenan igual tienen el mismo código soundex. Cuando se busca el apellido Powers (P620), se tiene que esquivar el nombre Pierce, Price, Perez y Park que tienen también el mismo código soundex. Incluso Power (P600), es una manera común de pronunciar Powers hace 100 años, tiene un código soundex diferente.

- Un apellido con más de una letra, o un apellido que comúnmente viene después del nombre, como los nativos americanos y apellidos chinos, podrían ser codificados bajo el último nombre que aparezca (el apellido), aunque éste último no sea el apellido actual. En el caso de apellidos mundiales, solamente la última palabra será tomada en cuenta para ser codificada.

Usos para el Código Soundex.

El código Soundex ha sido extensamente en las informaciones de los censos de los EEUU. En este contexto se utiliza habitualmente el código Soundex como un índice para el apellido. Se utilizan además variaciones de este código de manera que se mantengan disponible la mayor cantidad posible de variantes de pronunciación de un mismo nombre o apellido.

1.4.2 Sustitución de los valores nulos o vacíos.

Como se comentó en el epígrafe 1.3 unos de los problemas de la limpieza de datos es la sustitución de valores nulos o vacíos que aparecen en diferentes campos de los registros de una base de datos.

El problema fundamental en esta sustitución es que la aparición de los valores nulos en una base de datos puede venir dada por dos razones fundamentales: datos omitidos o valores inaplicables para dichos registros. En el primer caso la sustitución del nulo es posible, en el segundo no se pudiera realizar, porque falsearía la información. Este es entonces el primer problema a resolver: ¿qué valores pueden ser sustituidos?. A esta interrogante no se han encontrado respuestas satisfactorias en la literatura consultada.

Otro problema interesante a considerar es que en muchos sistemas de bases de datos no se utilizan el valor NULO para indicar información ausente, sino que sencillamente se deja en blanco, algunos gestores sustituyen este blanco por CERO (si estamos en presencia de un campo numérico) o por CADENA VACÍA (en caso de datos alfabéticos). En el primer caso, por ejemplo, el problema sería ¿un CERO indica ausencia de información o el valor cero por sí?

Según Dorian Pyle [PYL99] la sustitución de los valores nulos puede ser realizada siguiendo varias técnicas.

En el caso de valores numéricos se puede lograr que se mantengan las medidas de tendencia central: media, moda, mediana, varianza, desviación típica. En todos los casos se busca la medida de tendencia central que se trate en el conjunto de datos que se tiene sin considerar los valores nulos o vacíos y luego se sustituyen estos por valores tales que hagan que estas medidas de tendencia central no cambien.

También Pyle propone que se pueden obtener funciones de regresión lineal para a partir de ellas predecir el valor a sustituir en los campos con valores ausentes.

En el caso de valores no numéricos estas técnicas ya no son aplicables totalmente (solo el caso de la moda) entonces en [PYL99] se proponen utilizar técnicas de inteligencia artificial, en particular se propone el uso de Redes Neuronales Artificiales y Sistemas Basados en Casos.

1.4.3 Metodología para la limpieza de datos.

La técnica que propone Helena Galhardars en [GAL01], es dividir los datos de manera lógica y física. En la parte lógica se encuentran las llaves y las técnicas de normalización (ya que desde el análisis se podría considerar que se inicia la labor de limpieza de datos). En cambio en la parte física tratan de separar los datos por medio de sentencias SQL realizando agrupaciones de datos con el fin de clasificarlos de manera efectiva para concretizar una consistencia entre los mismos.

La técnica propuesta por Ralph Kimball en [KIM96] es con la que estoy de acuerdo totalmente, ya que involucra varios casos expuestos anteriormente al considerar una limpieza de datos. Esto combinado con el algoritmo Soundex podría resultar en una técnica interesante que podría cubrir diferentes problemas planteados por los autores antes mencionados y que se presentan con más frecuencia en las Bases de Datos. La técnica propuesta por Ralph Kimball es la siguiente:

La limpieza de datos Profesional consiste en la descomposición y reensamblamiento de los datos. Se podría dividir la limpieza de datos es 6 pasos:

- Elementarizar
- Estandarizar
- Verificar
- Ligar o relacionar
- Asignar origen de datos
- Documentar.

Para ilustrar estos 6 pasos, consideraremos las siguientes direcciones ficticias:

Ralph B and Julianne Kimball Trustees for Kimball Fred C
Ste. 116
13150 Hiway 9

Box 1234 Boulder Crk
Colo 95006

Tal vez esta dirección fue ingresada en 5 campos llamados Direccion1, … , Dirección5. Eso no es muy confiable al momento de ordenar las partes de la dirección. Una parte crítica de esta dirección se ingresó incorrectamente. Tenemos que encontrar el error en un minuto.

El primer paso en la limpieza de la dirección es elementarizar: Los sistemas para la limpieza de datos utilizan esta jerga para pronunciarlo correctamente. Elementarizar la dirección produce un resultado como el siguiente:

Addressee First Name(1): Ralph
Addressee Middle Initial(1): B
Addressee Last Name(1): Kimball
Addressee First Name(2): Julianne
Addressee Last Name(2): Kimball
Addressee Relationship: Trustees for
Relationship Person First Name: Fred
Relationship Person Middle Name: C
Relationship Person Last Name: Kimball
Street Address Number: 13150
Street Name: Hiway 9
Suite Number: 116
Post Office Box Number: 1234
City: Boulder Crk
State: Colo
Five Digit Zip: 95006

Esta lista de elementos estándar es dependiente de lo que encontremos al momento de analizar los datos. Ralph y Julianne pueden ser los administradores de una organización individual, en ese caso algunos de los tipos de los elementos serán diferentes.

El segundo paso es estandarizar los elementos. Los últimos 4 elementos se podrían pasar a una forma más estándar. Encontramos "Ste" por que lo reconocemos como "suite" Sospechamos de "Hiway 9" cuando deberíamos leer "Highway 9." Haremos un cambio provisional, y en el paso de verificación debemos asegurarnos que la calle actual es llamada "Highway 9." Así que cambiaremos "Boulder Crk" a "Boulder Creek" y "Colo" a "Colorado".

El tercer paso es verificar la consistencia del estandarizado de los elementos. En otras palabras, ¿hay errores en el contenido? Aunque no pueda parecer obvio, hay claramente un error en la dirección. Boulder Creek en el código Postal 95006 está en California, no en Colorado. Por que 2 de las 3 piezas del código postal indican el estado que apunta a California, cambiar el estado a California. Probablemente se considerarían los campos similares para futuras verificaciones. Por ejemplo, si se tiene otra dirección para Ralph Kimball o Julianne Kimball, se verificaría el estado. Esto sería más importante si se descubre una ciudad llamada Boulder Creek en Colorado.

Ahora se ha elementarizado, estandarizado y verificado la dirección, estamos en buena posición para seguir con el paso 4 y 5: ligar o relacionar y asignar origen de datos.

Ligar o relacionar consiste en buscar cualquiera de los nombres como Ralph Kimball o Julianne Kimball en otros registros de clientes asegurándonos de que todos los elementos de todas las direcciones son idénticas. Nótese que el problema de legitimar cambios de direcciones para asignar elementos separados de dirección "anterior" y "actual" se está tomando en este momento.

Asignar origen de datos consiste en reconocer que Ralph y Julianne constituyen un mismo origen por que cuentan con la misma dirección, así mismo se debe de tener cuidado de excluir personas que vivan en diferente apartamentos en un mismo edificio. Se podría tener información en otro origen de datos interno o externo que indique que Ralph y Julianne están casados.

El sexto paso de la limpieza de datos consiste en documentar los resultados de elementarizar, estandarizar, verificar, ligar, y asignar origen en metadatos.

Esto ayudará en subsiguientes episodios de limpieza de datos para reconocer de una manera más eficiente direcciones y las aplicaciones de usuarios finales podrán ser más fáciles de analizar y más fácil de comprender la base de datos de clientes.

Los 6 pasos para la limpieza de datos requieren de software sofisticado y un experto con los conocimientos adecuados. El experto debe de conocer algoritmos de búsqueda, algoritmos de búsqueda de direcciones, y una larga lista de millones de entradas en tablas que proveerán sinónimos para las partes y las direcciones. En otras palabras una limpieza de datos seria es un robusto sistema de software.

1. 5 Herramientas de limpieza de datos

Actualmente los diferentes sistemas para la limpieza de datos que existen en el mercado son muy costosos y realmente no ofrecen una solución "completa" a varios problemas que se exponen en este documento.

Se analizaron varias herramientas de diferentes "marcas" y/o fabricantes, los resultados se explican a continuación:

En general las diferentes herramientas para la limpieza de datos se especializan en un rasgo característico o en un enfoque específico del problema.

Un ejemplo claro es el software del fabricante TRILLIUM llamado TRILLIUM SOFTWARE que única y exclusivamente se encarga de eliminar errores en los códigos postales tomando como base la "colonia" en la que se encuentran.

Otras herramientas tienen la limitante de que únicamente funcionan bajo ciertos sistemas operativos, como la herramienta Dataflux del fabricante DATAFLUX que únicamente puede funcionar bajo el sistema operativo Windows NT 4.0. Su funcionalidad está dada a la limpieza de nombres, direcciones, y códigos postales, todo esto dirigido a datos de Estados Unidos.

Otro software llamado “Melissa Data” es especialista en eliminar errores en campos como Nombre, Dirección, Código Postal, Teléfono. A pesar de que es muy sencillo de usar, resulta algo ineficiente si queremos realizar una limpieza pensando en un “Dataware House”, ya que un almacén de datos de una inmensa capacidad podría resultar muy complejo de “optimizar” ya que tendríamos que utilizar varias herramientas a la vez y tal vez el utilizar varias herramientas de “Limpieza de Datos” podría resultar contraproducente al momento de verificar la integridad de los datos.

Es decir, el usar más de una herramienta a la vez, podría dejar serias inconsistencias en nuestro almacén de datos que se pueden ver reflejados en los reportes de Inteligencia de Negocios y por lo tanto, nuestros datos terminarían por ser muy poco confiables, muy a pesar de que usamos varias herramientas de “Limpieza de Datos”.

Por otro lado la compañía ARKIDATA ofrece una solución de Limpieza de Datos basada en un motor “Unico” y exclusivo propiedad de la compañía. Por supuesto, no mencionan a detalle el funcionamiento de dicha herramienta, pero prometen solucionar de un 90% a un 98 % los problemas en los datos. Es decir, dejan hasta un 10 % (Máximo) de errores en las Bases de Datos una vez utilizada la herramienta.

También promete utilizar las mismas reglas de negocio que se utilizan en la compañía que planee contratarlos, es decir, no tiene que preocuparse en caso de que los orígenes de datos den lugar a datos más complejos en una base de datos o que las llamadas “Reglas de Negocio” con las que actualmente su negocio y gracias a esas reglas ha resultado tan redituable. Esto es importante, por que quiere decir que podemos aplicar esas mismas reglas desde la misma limpieza de datos y no como en algunos casos, en los que las reglas de negocio se aplican como último paso al momento de realizar una exitosa extracción, transformación y limpieza de datos.

Esto también nos dice que la herramienta de la compañía ARKIDATA es también una herramienta totalmente configurable y adaptable a las necesidades de cualquier compañía, aunque lamentablemente, sigue siendo una herramienta que todavía se

preocupa más por los detalles más comunes en una Base de Datos operacional como el nombre del cliente, dirección, teléfono, código postal, etc.

La herramienta que ofrece la compañía CENTRUS promete eliminar todos los errores basados en los datos erróneos que suelen producirse en la dirección del cliente, calles, códigos postales, etc. Promete reducir en hasta un 40% los costos del servicio postal (Muy atractivo si se trata de un Banco o un negocio similar), ya que al eliminar los errores con la limpieza de datos, automáticamente se eliminarán registros duplicados, envío equivocado de cartas o estados de cuenta, etc. Que al momento de realizar un envío masivo de cartas utilizando el servicio postal podemos generar un gran ahorro de recursos gracias a esta herramienta.

Basado en los datos proporcionados por la compañía CENTRUS tenemos un ejemplo muy sencillo como lo es el siguiente:

Imagine que se tiene una cartera de clientes de un banco, esta cartera de clientes está compuesta por 1’000,000 de clientes. A ese millón se le tiene que enviar su estado de cuenta cada mes. Si el costo unitario por el envío de cada estado de cuenta es de 2 pesos tenemos 1’000,000 * 2 = 2’000,000 de pesos. Si tomamos como base que de esos 2 millones un 40 % de los domicilios y o destinatarios no existe o contiene algún error en sus datos, ya sea el domicilio, el código postal, la zona, la colonia, etc. Entonces estamos hablando que de ese 1’000,000 de sobres, estamos enviando directamente a la basura a 400,000 sobres, es decir, 400,000 * 2 = 800,000 pesos, es decir, que con la herramienta adecuada, podemos lograr un ahorro anual aproximado de 9’600,000 pesos y esto es un ahorro muy significativo en los costos de cualquier empresa y por lo tanto, una herramienta de limpieza de datos puede convertirse en una herramienta muy valiosa para cualquier mercado.

CAPITULO 2: SOLUCIONES ADOPTADAS

Después de un exhaustivo análisis de las escasas herramientas que se encuentran actualmente en el mercado se llegó a la conclusión de que ninguna herramienta cumple con los requisitos básicos para ser considerada una herramienta "Universal" que pueda utilizarse en cualquier Sistema Gestor de Bases de Datos.

Tratar de construir una herramienta "universal" de limpieza de datos es realmente una tarea compleja, ya que si tomamos en cuenta la forma de organizar el catálogo (los metadatos) de cada SGBD tendremos un reto realmente complejo a realizar. La complejidad de tal herramienta realmente reside en que la diferencia entre la forma de organización de los catálogos hace que la homogeneización puede convertirse en una tarea titánica. Es por ésta razón que por cuestiones de tiempo se decidió realizar una pequeña parte de esta herramienta a fin de ilustrar el proceso de la limpieza de datos, así como plasmar las bases teóricas de las cuáles se habla en este documento.

En un principio y haciendo un análisis de las necesidades reales de cualquier empresa se llegó a la conclusión de que se debía elaborar una herramienta totalmente programable y que se adapte a las necesidades básicas y complejas al momento de enfrentarse a una profunda y profesional limpieza de datos.

2. 1 Propuesta del Soundex

Iniciamos con el análisis de los principales algoritmos utilizados para la limpieza de datos y como se explicó anteriormente, el algoritmo "Soundex" es uno de los algoritmos más utilizados al momento de hablar de "Limpieza de Datos" pero como también ya se explicó, es un algoritmo muy limitado al momento de utilizarlo, y el principal factor que lo limita (después de las limitadas reglas que posee) es que es un algoritmo totalmente "Americanizado", es decir, puede funcionar hasta en un 90 % de efectividad, pero claro,

con datos Estadounidenses, ya que al utilizarlo con datos de cualquier otro país el porcentaje de efectividad disminuye notablemente.

Por tal motivo se pensó en adaptar el algoritmo "Soundex" de una manera en que se pudiera utilizar internacionalmente, pero otra muy grande limitante es las reglas gramaticales de cada lenguaje y de cada país, es decir, no se tiene la misma gramática y sintáctica en España que en México o que en Venezuela, a pesar de que en los 3 países se habla español y por lo tanto podríamos pensar que una herramienta "Latina" para la limpieza de datos podría sernos muy útil en el idioma español.

Se propone utilizar el siguiente algoritmo, el cual está en fase de prueba y es experimental, no es un algoritmo definitivo y por lo tanto, 100% confiable al momento de utilizarlo (aunque realmente el algoritmo soundex original tampoco lo es).

Las reglas para iniciar la codificación son las siguientes:

1. Los códigos del algoritmo H-Soundex empiezan con la primera letra de la palabra, seguida de la primera consonante que se encuentre.
2. Todas las vocales serán tomadas con valor 0 además de las letras "Y" y "H".
3. Utiliza los siguientes grupos para codificar el resto de las consonantes:
   0 – Vocales, Y, H
   1 – B, P, V
   2 – C, S, Z, K, X, Q
   3 – J, G
   4 – D, T
   5 – L, LL
   6 – M, N, Ñ
   7 – R
   8 – F
4. Nombre con letras adyacentes que pertenecen al mismo grupo son codificadas con un solo número.
5. Se completa, si es necesario, con 0 hasta tener 6 caracteres

Ejemplo:

SOUNDEX		H-SOUNDEX	
Aguiar	A260	Aguiar	AG0700
Aguilar	A246	Aguilar	AG0506

Una vez codificada la palabra, se puede tomar como base los primeros5 caracteres del código con el fin de tomar todas las palabras similares o que tengan un sonido similar o el código completo en el caso que se desee más precisión.

Aquí vemos otro ejemplo:

SOUNDEX	H-SOUNDEX
Nuno (N500)	Nuno (NU600)
Nuño (N000)	Nuño (NU600)

2. 2 Conectividad con diferentes gestores.

Otro gran problema con los que comúnmente se encuentran las herramientas especializadas en limpieza de datos es que tienen diversas limitantes en cuanto a los sistemas operativos bajo los cuáles pueden trabajar, igualmente cuentan con bastantes limitantes al momento de trabajar con Sistemas Gestores de Bases de Datos.

En las herramientas actuales de programación se utilizan dos técnicas básicas de conexión a los SGBD: utilizar ODBC (Open Data Base Conectivity) y ADO (Activex Data Object).

Los controles ODBC se tienen que configurar con las referencias estándar que ofrece Windows (ODBC 32) dentro del panel de control. En el entorno de programación se

llama a este origen de datos para lograr la conexión con el SGBD y el catálogo (Base de Datos). El usuario de la herramienta necesitaría realizar esta conexión para su origen de datos y luego indicarle a la herramienta la conexión realizada.

En cambio, con el control ADO utilizamos las cadenas de conexión directamente en el entorno de programación. Es decir, si utilizamos la facilidad de conexión de ADO dentro del entorno de programación, podemos conectarnos con casi cualquier SGBD deseado, siempre y cuando tengamos las bibliotecas referenciadas en el entorno de programación referido.

Analizando la forma de trabajar de las diversas herramientas de Limpieza de Datos, se logró observar que muy pocos tocan el tema de el ordenamiento y clasificación de los diversos datos provenientes de diversas fuentes. Es decir, datos que tienen un origen en diversos SGBD o en datos generados por entidades externas. La verdadera tarea compleja de estos datos consiste en identificar que estamos hablando de un mismo producto o registro, muy a pesar de que en la Base de Datos está identificado de muy diversas formas haciéndonos creer que son registros totalmente diferentes, y esto puede reflejar graves errores o grandes inconsistencias en los datos que nos pueden hacer dudar de los datos que se tienen una vez realizada la extracción de los datos de esas diversas fuentes tan dispares.

Haciendo un estudio analítico de los principales SGBD, se decidió tomar el SGBD "Microsoft SQL Server 2000" como el SGBD con el cuál se interactuará y se realizarán todas las pruebas con la herramienta que se pretende crear a partir de estos principios básicos de la "Limpieza de Datos".

2.3 Propuesta de Herramienta de Limpieza

A partir del análisis realizado se darán los principios sobre los cuales se ha elaborado una herramienta para la Limpieza de Datos.

La herramienta que se pretende elaborar es una primera versión de un prototipo experimental, la herramienta que el mercado necesita es un proyecto ambicioso que puede lograrse con el tiempo suficiente y la adecuada dedicación, pero precisamente por cuestiones de tiempo, este documento y la herramienta a generar se tendrán que ver limitadas a solamente los principios más básicos y simples de una limpieza de datos que implica una extracción, transformación de datos más comúnmente utilizadas en la "Inteligencia de Negocios".

Las fases de creación de ésta herramienta tocarán todos y cada uno de los puntos básicos que este documento menciona como los principios fundamentales para lograr una correcta y efectiva "Limpieza de Datos".

Primeramente partiremos por analizar que el servidor válido, es decir, que el SGBD sea efectivamente "Microsoft SQL Server 2000", ya que como se comentó anteriormente la herramienta en ésta fase inicial únicamente pretende trabajar con el SGBD antes mencionado.

Enseguida, se determinará una Base de Datos que contenga lo menos una tabla, ya que por lo menos necesitaremos una tabla para mostrar todas y cada una de las fases de una "Limpieza de Datos" profesional y efectiva.

Al momento de elegir la tabla debemos tomar en cuenta que por lo menos cuente con un campo y por lo menos 1 registro para que la "Limpieza de Datos" se realice de una forma efectiva.

Una vez validados todos estos elementos, se procederá a realizar una limpieza básica a la tabla elegida, iniciando con los criterios más básicos para la limpieza de datos:

El problema de limpieza se abordará a partir de cuatro etapas:

1. Tratamiento de valores nulos y/o vacíos.
2. Tratamiento de registros duplicados.

3. Filtrar los datos semejantes.
4. Homogeneizar los datos.

En cada etapa la solución que se pretende dar a este problema consiste en analizar tabla por tabla de la Base de Datos que se desea limpiar y aplicar consultas para iniciar la labor de limpieza.

1. Antes que nada, debemos elegir el Servidor de Bases de Datos (Microsoft SQL Server).
2. Después tendremos que elegir una Base de Datos (Microsoft SQL Server)
3. Enseguida debemos elegir una tabla.
4. Una vez elegida la tabla la herramienta nos mostrará los campos vacíos y/o nulos y el tipo de datos de éstos.
5. Una vez mostrado el tipo de dato, la herramienta nos dará una sugerencia de reemplazo del campo dependiendo del tipo de dato.
6. Si no deseamos reemplazar el campo vacío y/o nulo con el dato sugerido por la herramienta, tendremos la opción de ingresar el dato se desee.

El proceso que hemos descrito es aplicado en cada una de las etapas.

La primera etapa cubrirá únicamente a los campos nulos y vacíos, la segunda etapa cubrirá los registros iguales.

Una tercera etapa filtrará registros parecidos o con alguna similitud con el fin de Homogeneizarlos y actualizarlos y/o eliminarlos dependiendo el caso.

La cuarta etapa cubrirá la homogeneización de datos totalmente diferentes y que probablemente provengan de diversas fuentes de datos y posiblemente hasta con diferentes tipos de datos entre si.

Con estos pasos a seguir estamos siguiendo un patrón como el que se describe en el capítulo 1 que nos dice la base de la cual parte la “Limpieza de Datos”, es decir:

- Elementarizar
- Estandarizar
- Verificar
- Ligar o relacionar
- Asignar origen de datos
- Documentar.

CAPITULO 3: HERRAMIENTA DE LIMPIEZA

A partir de los elementos estudiados anteriormente se diseñó una herramienta de Limpieza de Datos, que incluya algunos de los aspectos vistos.

La misma se realizó en Visual Basic 6, la conexión con las bases de datos se logra a través de la tecnología ADO.

Los componentes utilizados para la construcción de la misma son los que brinda el lenguaje.

A continuación se explicará de una manera breve la forma de utilizar la herramienta (en fase de prueba todavía) con los fundamentos de la "Limpieza de Datos" mencionados anteriormente.

Los parámetros mínimos para el buen funcionamiento de la herramienta son:

- Nombre del Servidor
- Usuario
- Contraseña
- Base de Datos

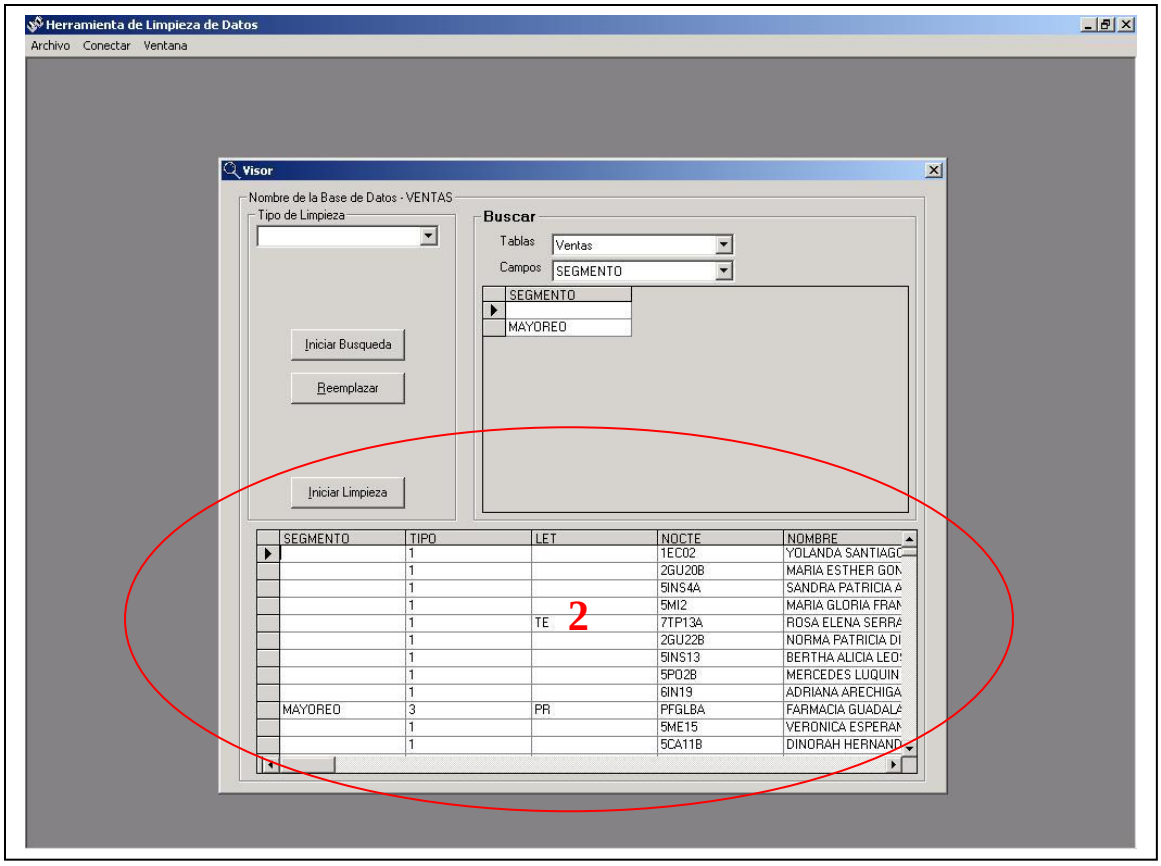
Esto con el fin de que el programa muestre al iniciar las tablas con las que cuenta, así como los campos de la tabla por defecto.

El programa se divide como sigue:

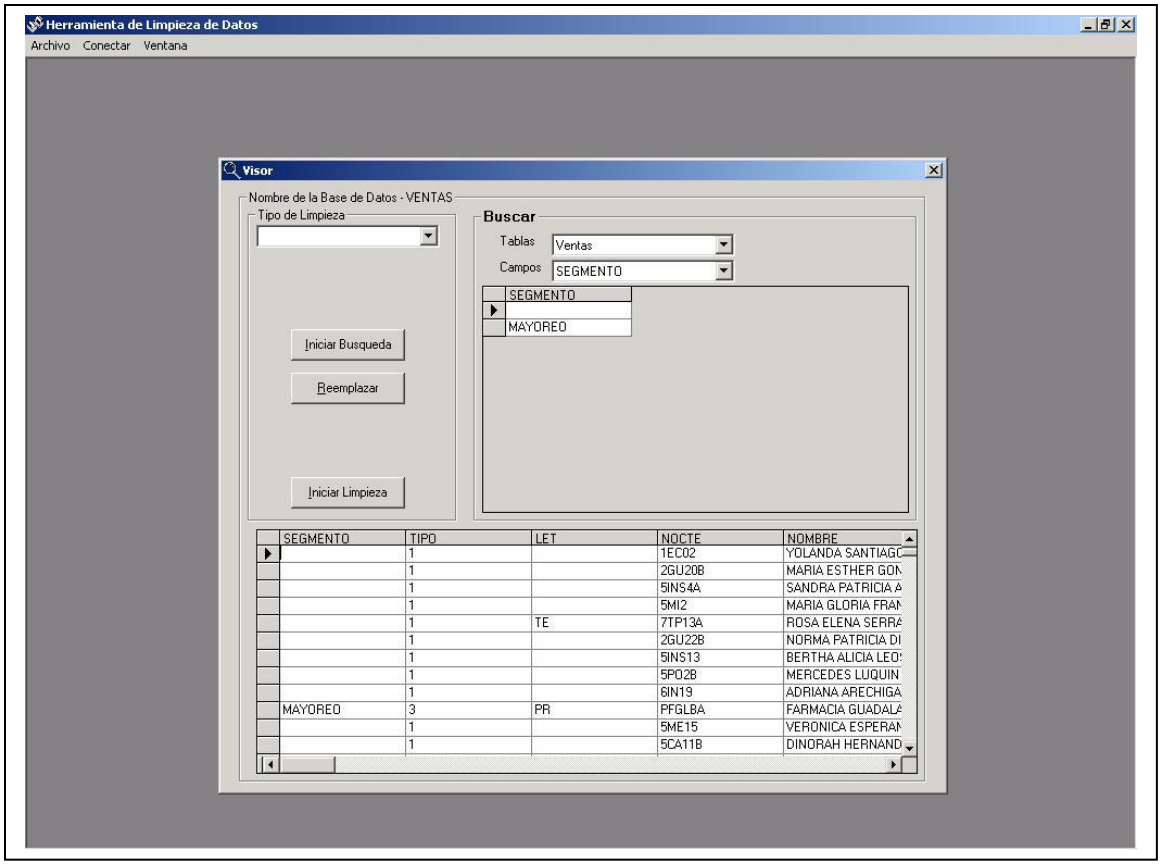
1. Primeramente en la parte superior de la ventana se tiene el nombre de la tabla y enseguida el nombre del campo que estamos analizando actualmente. En este

ejemplo se eligió como la tabla a analizar la llamada “Ventas” y el nombre del campo es “Segmento”.

Seguidamente aparecen todos los registros que son diferentes dentro de la tabla que se analiza. En esta ocasión solamente se muestran 2 registros diferentes, pero si se analizan otros, pueden aparecer tantos como datos diferentes se guarden en él los registros de ese campo.



2. La parte inferior de la ventana se muestra los registros que contiene la tabla y el campo que se muestra el punto 1, es decir, de la tabla “Ventas” y el campo “Segmento” además de todos los campos de que está compuesta la tabla. Es importante mencionar que la tabla de la parte inferior muestra todos los registros que contiene esta tabla, tal y como aparecen y se registraron en la tabla “Ventas”.



En la primera lista desplegable se elige si se trata de “Limpiar” registros nulos y/o registros vacíos, si se desea reemplazar registros con errores por otros registros adecuados o se si desea encontrar registros con campos similares, todo esto acorde con el campo que se muestra a la derecha (Sección 1). Es muy útil la vista que está dentro de la ventana en general, ya que se puede ver de manera “filtrada” los

registros el campo que se eligieron y en la parte inferior todos los registros de la tabla que se configuró.

Es necesario aclarar que la medida de similitud que se emplea en esta primera versión es la coincidencia del código Soundex (a partir de la propuesta realizada – HSoundex) de la cadena que se digita con el código Soundex de las cadenas que se encuentran en el campo seleccionado.

Esta vista da una idea primordial del tipo de limpieza que se quiere y realmente si "el" o "los" registros a limpiar son los adecuados y los que realmente se necesita limpiar.

4. Si se elige "Limpiar" registros nulos y/o vacíos, entonces se debe tomar en cuenta que la herramienta nos dará 2 tipos de sugerencias. Si el tipo de dato es carácter (o similar), entonces dará la sugerencia de ingresar en el campo la cadena "No capturado" y en caso de que el tipo de dato sea numérico (o similar) sugerirá ingresar el promedio de todos los datos que contiene el campo (la suma de el valor de todos los registros entre el numero de registros), esto con el fin de NO afectar el resultado promedio de el campo que estamos tratando de limpiar. Esto resulta particularmente importante en la "Inteligencia de Negocios" al momento de realizar cálculos y reflejarlos en un escenario de negocios, ya que por ejemplo, si se tienen 10 registros con valores de "10" cada uno, excepto uno de ellos que se olvidó capturar y que por tal motivo está vacío tenemos 11 registros, por lo tanto el promedio no es "10" como se pensaría, si no que el resultado es "9.0909", ya que tenemos un registro con valor "0" y esto afecta implícitamente el cálculo del promedio, ya que la operación se realiza como se explica a continuación:

Número de Registro	VALOR
1	10
2	10
3	10
4	10
5	10
6	10
7	10
8	10
9	10
10	10
PROMEDIO	**10**

Número de Registro	VALOR
1	10
2	10
3	10
4	10
5	10
6	10
7	10
8	10
9	10
10	10
11	
PROMEDIO	**9.0909**

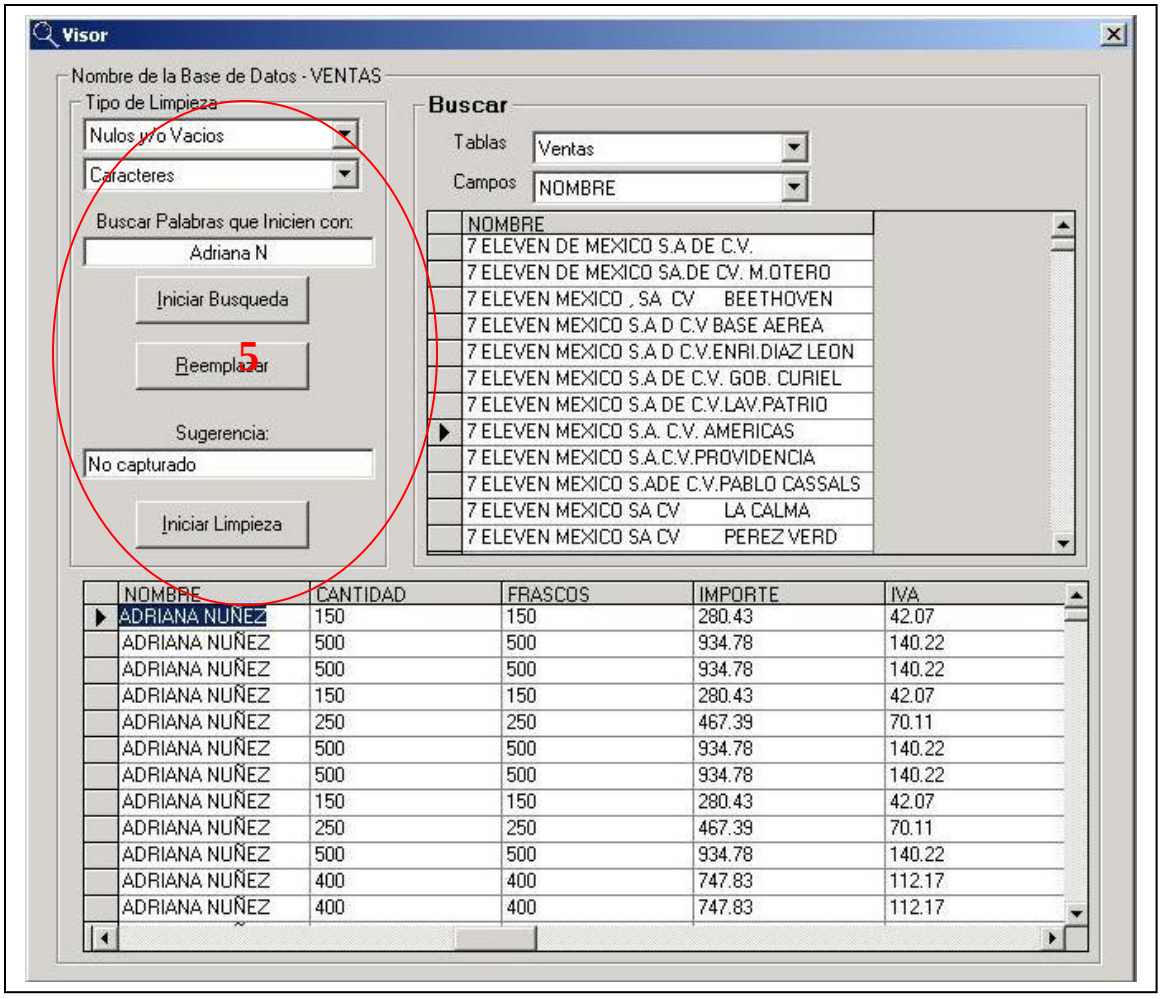
Este cálculo simplemente con un registro vacío resulta fatal para un escenario de "Inteligencia de Negocios" ya que afecta de manera dramática el resultado y la confiabilidad de los datos que pretendemos mostrar. Por esta razón la herramienta dará como sugerencia el promedio de esa columna (Sin contar los campos nulos, obviamente) con el fin de seguir manteniendo el valor promedio *real* de la columna.

Número de Registro	VALOR
1	10
2	10
3	10
4	10
5	10
6	10
7	10
8	10
9	10
10	10
11	10
PROMEDIO	**10**

Basta con elegir "Iniciar Limpieza" para que los nulos y/o vacío sean reemplazados con los parámetros antes mencionados como se explicó en párrafos anteriores.

5. Si en cambio se tiene que elegir la opción de limpieza para caracteres entonces se puede elegir ya sea la limpieza para registros Nulos y/o vacíos (que se explicó anteriormente) o la opción de reemplazar registros por cualquier entrada de datos que nosotros deseemos.

   Esta opción funciona de una forma muy sencilla, simplemente hay que digitar en la casilla inferior la frase o palabra que sustituirá a los registros elegidos en la casilla superior y que se despliegan en el panel inferior de la ventana como se muestra a continuación:



Si se pide que busque palabras que inicien con “Adriana Nu …” el programa desplegará todos los registros del campo “Nombre” que inician con “Adriana Nu”, en este caso, todos los registros que contienen en el campo “Nombre” a “Adriana Nuñez” o nombres muy similares.

Si por lo menos un campo está escrito en los registros como “Adriana Nuñes” con una “S” al final y no con una “Z” como es correcto lo podemos solucionar de la siguiente forma:

En el campo superior se digita el nombre a buscar, es decir, “Adriana Nu…” y desplegará en la parte inferior todos los campos con el nombre “Adriana Nuñez”. Una vez desplegados y asegurándo que por lo menos un registro es incorrecto, es decir, Nuñez está escrito con “S” (Nuñes), entonces se debe digitar en la casilla

“Sugerencia” el nombre correcto con el que se desea reemplazar todos los errores posibles y/o encontrados en el campo que elegimos (Nombre), por lo tanto se digita “Adriana Nuñez”.

Ya con la cadena de corrección elegida, se elige “Reemplazar” y una vez efectuados los cambios por el programa, éste mostrará un mensaje informando que los cambios en los registros resultaron exitosos. Ver el punto 6.

CONCLUSIONES

Se diseñó e implementó una herramienta de limpieza de datos, la cual constituye una primera versión de una herramienta con una mayor funcionalidad. En esta primera versión se incluyeron tres problemáticas asociadas a la limpieza de datos: sustitución de valores nulos o vacíos, reemplazo de valores incorrectos y búsqueda de valores similares. Es necesario insistir que en todos los casos se utilizaron variantes sencillas de estos procesos. La herramienta se conecta a una Base de Datos SQL Server 2000.

Se realizó un estudio de los principales aspectos relacionados con la Limpieza de Datos, sistematizando algunos conceptos importantes que aparecen en diferentes fuentes; de tal manera que el presente documento puede servir como una referencia del tema. Se determinó el uso de una metodología a seguir en este proceso.

Se mejoró el algoritmo Soundex, planteando una versión para el idioma español. Esta versión es necesaria que se someta a mayores pruebas de comprobación.

RECOMENDACIONES

Se recomienda ampliar la conectividad de la herramienta programada de manera que se puedan limpiar datos de gestores diferentes. Esto no es difícil a partir de la utilización de la tecnología ADO.

Para el trabajo con campos nulos y/o vacíos, se recomienda agregar dentro de las recomendaciones al usuario las opciones como: moda, varianza, desviación estándar, además del ***promedio*** que ya está incluído dentro de la herramienta.

Se recomienda el estudio de otros criterios de semejanza distintos al Soundex para realizar búsqueda de información.

Se recomienda ampliar paulatinamente la herramienta de manera de incorporar nuevas posibilidades, por ejemplo homogeneización de direcciones, nombres, códigos postales, etc. Pero logrando una generalización tal que pueda ser utilizado en diferentes ámbitos, a partir del uso de diccionarios, catálogos, etc.

Bibliografía

Elmasri R., Navathe S. B., “Sistemas de Bases de datos”, 2da. Edición, Addison Wesley Iberoamericana.

Jonson James L., “Bases de datos: Modelos, lenguajes, diseño”, 1era. Edición, Oxford University Press México.

Silberschatz A. Korth H. y Sudarshan S., “Fundamentos de Bases de Datos”, 3era. Edición, Editorial McGraw-Hill

De Miguel Adoración y Piattini Mario, “Conceptos y diseño de bases de datos”, 2da. Edición, Editorial Ra-ma.

McFadden F. R., Hoffer J. A., Prescott M. B., “Modern Database Management”, Addison Wesley.

Gardarin G., “Bases de datos“, Editorial Paraninfo.

Date C. J., “Introducción a los sistemas de bases de datos”, 6ta. Edición, Addison Wesley.

Ullman, “Principles of Database System”, Editorial Computer Science Press.

González Alvarado Carlos, “Sistema de Bases de Datos”, 1era. Edición, Editorial Tecnológica de Costa Rica.

“Referencias De Computación”,

[http://www.monografias.com/trabajos/refercomp/refercomp.shtml]

Bibliografía referenciada.

Bibliografía Preeliminar

[Elmasri / Navathe, 1997] Ramez Elmasri, Shamkant B. Navathe, Sistemas de Bases de datos, 2da. Edición, Addison Wesley Longman, Págs. 1, 557.

[Kimball / Reeves / Ross / Thornthwaite , 1998] Ralph Kimball, Laura Reeves, Margy Ross, Warren Thornthwaite The Data Warehouse Lifecycle Toolkit, Wiley , Págs. 656, 661.

Consultas por INTERNET

[DataMirror, 2001]
Business Intelligence with DataMirror Transformation Server
White Paper
DataMirror Corporation
http://www.datamirror.com/Whitepaper.pdf 10/28/1998

[Loshin, 2001]
David Loshin
CTO of Knowledge Integrity Inc.
http://www.intelligententerprise.com/000209/docs/valad.htm 21/07/2000

[Kimball, 2001]
Ralph Kumball
Kimball University
http://www.ralphkimball.com/iydc.htm 01/09/2000

[Orli/Santos, 1996]
R. Orli / F. Santos
Based on a Public US Gov't document.
http://www.kismeta.com/deamt.htm 17/10/1996

REFERENCIAS BIBLIOGRÁFICAS

[DAT98] A DataMirror Corporation White Paper Octubre 1998. Pagina 6.Version 1B1027

[DWI01] Data WhareHouse Infocenter. "WHY DATA WAREHOUSE PROJECTS FAIL". http://www.dwinfocenter.org/whitepap.html 2001.

[FRA01] Frauca, Roger. Mejorando los procesos de depuración y carga en un proyecto Data WareHouse. 2001

[GAL01] Helena Galhardas, Daniela Florescu. "Declarative Data Cleaning: Languaje, Model, and Algorithms. November 2001. Page 1.

[GAL01A] Helena Galhardas, Daniela Florescu. "Declaratively Cleaning Your Data Using AJAX. November 2001. Page 1.

[KIM96] Ralph Kimball, "Dealing with Dirty Data" DBMS, September 1996. *The science of maintaining clean data in your warehouse, and why nobody talks about it.*

[LOS01] Loshin David. "Value Added Data: Merge Ahead". www.knowledge-integrity.com January 2001. Page 1.

[ORL96] Orli R., Santos, F. Data Extraction, Cleansing and Migration Tools to support Data Warehouse, Database Consolidation, and Systems Reengineering Projects**.** 1996. Pagina 1

[PYL99] Pyle Dorian, Data Preparation for Data Mining, Kaufman Press, 1999.

[URD01] www.monografias.com \ El DataMining. Elymir Urdaneta. universidadsr@cantv.net Caracas Venezuela 2001.