Almacenes de datos:

importancia de la estandarización de las direcciones para las empresas

Por Lic. Liudmila Padrón Torres
Especialista en Informática, Grupo de Gestión y Administración, Filial TISW, Gerencia Territorial Villa Clara, ETECSA
lumy@vcl.etecsa.cu

Introducción

Desde un inicio, las bases de datos se convirtieron en herramienta fundamental para el control y manejo de las operaciones comerciales. Fue así como en pocos años en las grandes empresas y negocios existía un número considerable de información almacenada en diferentes fuentes de datos que fueron alcanzando un tamaño considerablemente grande.

Con el gran cúmulo de información, los directivos de tales empresas y negocios se dieron cuenta de que esta podría tener un fin útil, al estar la mayoría de sus operaciones comerciales reflejada durante los llamados ciclos de negocios propios del mercado.

A su vez, los mercados empresariales han experimentado una transformación radical. Las empresas demandan mayor rapidez y eficiencia en la entrega de productos, y mejora en todos los servicios existentes, por lo que se hace imprescindible encontrar formas más eficaces de distribuir los productos, más facilidades para hacer estudios de mercado basados en la información de las operaciones comerciales de las empresas y de sus clientes y, en definitiva, mayor rapidez en la toma de decisiones.

Por lo tanto, se pensó que sería ideal unificar las diferentes fuentes de información de las que se disponían en un único lugar al que sólo se le incorporaría información relevante, sobre la base de una estructura organizada, integrada, lógica, dinámica y de fácil explotación. La respuesta a esto fueron los Almacenes de Datos o *Data Warehouse* (DW).

Sin embargo, para hacer un uso eficiente de la información histórica almacenada en un DW que ayudara a la toma de decisiones, era vital garantizar que estos datos fueran fáciles de obtener, estandarizados y confiables.

Aún así, el problema de la limpieza de datos es poco tratado o evitado por muchas empresas, que no consideran adecuadamente el impacto que se producirá en el negocio al tener información deficiente almacenada.

Almacenes de datos: conceptos básicos

Un *Data Warehouse* es un almacén de información temática orientado a cubrir las necesidades de aplicaciones de los Sistemas de Soporte de Decisiones (DSS) y de la Información de Ejecutivos (EIS), que permite acceder a la información corporativa para la gestión, el control y apoyo a la toma de decisiones [8].

Dicha información es construida a partir de bases de datos que registran las transacciones de los negocios de las organizaciones —bases de datos operacionales— y su importancia reside en algunos elementos como los siguientes:

- Contribuye a la toma de decisiones tácticas y estratégicas al proporcionar un sentido automatizado para identificar información clave desde volúmenes de datos generados por procesos tradicionales o elementos de software.
- Posibilita medir las acciones y los resultados de una forma mejor.
- Los procesos empresariales pueden ser optimizados. El tiempo perdido en la espera de información que finalmente es incorrecta o no encontrada, es eliminado.
- Permite a los usuarios priorizar decisiones y acciones, por ejemplo, a qué segmentos de clientes deben ir dirigidas las acciones de marketing.

En general, un DW es un conjunto de datos con las siguientes características:

Temático

Los datos están almacenados por materias o temas —clientes, campañas, productos—. Se organizan desde la perspectiva del usuario final, mientras que en las bases de datos operacionales se organizan desde la perspectiva de la aplicación, con vistas a lograr una mayor eficiencia en el acceso a los datos.

Integrado

Todos los datos almacenados en el DW están integrados. Las bases de datos operacionales orientadas hacia las aplicaciones fueron creadas sin tener en cuenta su integración, por lo tanto, un mismo tipo de dato puede ser expresado de forma diferente en dos bases de datos operacionales distintas, por ejemplo, para representar el sexo: Femenino y Masculino ó F y M.

No volátil

Únicamente hay dos tipos de operaciones en el DW: la carga de los datos procedentes de los entornos operacionales —carga inicial y carga periódica— y la consulta de los mismos. La actualización de datos no forma parte de la operatividad normal de un DW.

Histórico

El tiempo debe estar presente en todos los registros contenidos en un DW. Las bases de datos operacionales contienen los valores actuales de los datos, mientras que los DW contienen información actual y resúmenes de esta en el tiempo.

Arquitectura

Los bloques funcionales que se corresponden con un sistema de información completo que utiliza un DW se muestran en la figura 1.

Nivel operacional

Contiene datos primitivos (operacionales) que son actualizados permanentemente, usados por los sistemas operacionales tradicionales que realizan operaciones transaccionales.

Almacén de datos o DW

Contiene datos primitivos correspondientes a sucesivas cargas del DW y algunos datos derivados. Los datos derivados son generados a partir de los datos primitivos al aplicarles algún tipo de procesamiento (resúmenes).

Nivel departamental —*Data Mart*—

Contiene casi exclusivamente datos derivados. Cada departamento de la empresa determina su nivel departamental con información de interés para él. Va a ser el blanco de salida sobre el cual los datos en el almacén son organizados y almacenados para las consultas directas por los usuarios finales, los desarrolladores de reportes y otras aplicaciones.

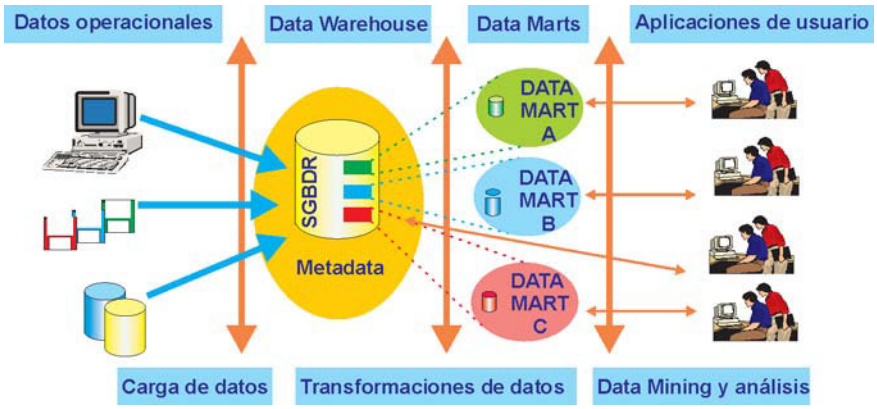


Figura 1 Arquitectura de un *Data Warehouse* [4]

Nivel individual

Contiene pocos datos, el resultado de aplicar heurísticas, los procesos estadísticos, etc., a los datos contenidos en el nivel anterior. El nivel individual es el objetivo final de un DW. Desde este nivel accederá el usuario final y se podrán plantear diferentes hipótesis, así como navegar a través de los datos contenidos en el DW.

Detrás de la arquitectura del DW existe un conjunto básico de procesos, entre los cuales pueden mencionarse algunos elementales como:

El **proceso de extracción** que consiste en estudiar y entender los datos fuente, tomando aquellos que son de utilidad para el almacén.

El **proceso de transformación** de los datos hacia una forma presentable y de valor para los usuarios.

Al terminar el proceso de transformación, se realiza la **carga de los datos** en el DW y seguidamente se efectúan los **controles de calidad** para asegurar que sea correcta.

De estos procesos, es importante para las empresas ponerle atención a la transformación de datos donde se incluyen operaciones de corrección de errores, resolución de problemas de dominio, borrado de campos que no son de interés, generación de claves, agregación de información, etc.

La transformación de datos es necesaria porque no siempre los datos están en la forma más adecuada para poder aplicar los métodos que hacen falta para la tarea que se ha de llevar a cabo y el modelo que se quiere obtener [7].

Esta fase, aunque parezca sencilla, conlleva aproximadamente el 70 % del esfuerzo en los proyectos de DW [7].

Limpieza de datos

"Every meaningful data warehouse application needs good data" [6]

Un problema en DW que es universalmente reconocido, pero la mayoría de las veces ignorado, es la limpieza de datos de un almacén, lo que conlleva en

muchas ocasiones, no considerar adecuadamente el impacto negativo para el negocio, de tener almacenada información deficiente.

Un estudio extranjero realizado en el año 2005 plantea que [9]:

- "El 25 % de nuestros datos son defectuosos, y un 48 % de las empresas no invierten esfuerzos y dinero suficiente en la depuración y el mantenimiento de sus bases de datos".
- "El éxito o fracaso en un proyecto *Costumers Relations Management* (*CRM*), DW, *e-Business o Enterprise Resource Planning* (ERP), depende en gran parte de la calidad de datos e información interna".
- "El 75 % de 600 empresas confesaron tener problemas internos serios por problemas de calidad de datos".

Y prosigue la misma referencia [9]: "Así, surge la pregunta: Si tan solicitadas son las bases de datos, si todos estamos de acuerdo en que la información de nuestros clientes y nuestros mercados es nuestro activo más importante, si la calidad de la información es el primer paso en cualquier proyecto de implantación de tecnología de información, entonces, ¿cómo puede convertirse en un problema?".

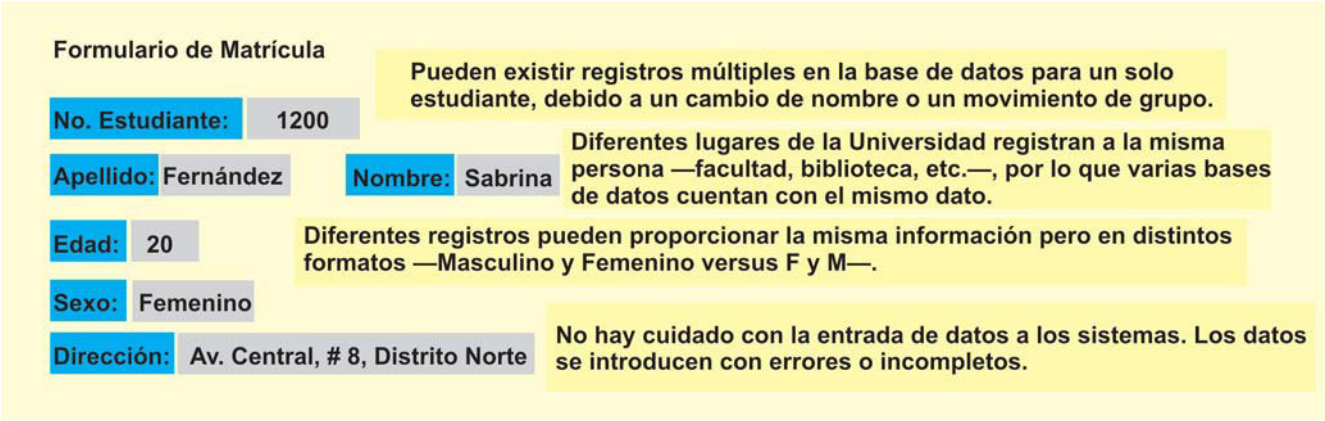
Generalmente esto se debe a que las empresas no cuentan con aplicaciones únicas para cada parte de la operativa del negocio, sino que pueden tener replicaciones y distintos sistemas para atender un mismo conjunto de operaciones, y en esos casos es probable que las bases de datos de los sistemas operacionales contengan datos duplicados, a veces erróneos, superfluos o incompletos. Además la información es dinámica y sometida a constantes cambios; y se suman también, los posibles errores a la hora de la entrada de datos a los sistemas de datos operacionales. Estas son algunas de las cuestiones que contribuyen a la suciedad de datos (Figura 2).

La limpieza de datos se encuentra dentro del proceso de transformación de datos. Es mucho más que la simple actualización de registros con datos buenos, involucra descomposición y reensamblaje de datos. La limpieza de datos se puede dividir en seis pasos: separar en elementos, estandarizar, verificar, machear, agrupar y documentar [6].

Por ejemplo, si tenemos direcciones postales de clientes que queremos limpiar, lo primero sería separar este campo en sus elementos principales —calle, No., entre calles, código postal, etc.—. Lo segundo sería estandarizar los elementos, o sea lograr que estos queden de forma normalizada. Luego se verificaría si los elementos estandarizados contienen errores en su contenido, y ya estaríamos listos para machear —hacer parejas o correspondencias— y agrupar, que consiste en reconocer que algunas de las partes de la dirección constituyen una agrupación, por ejemplo, si se tienen dos direcciones iguales de diferentes clientes que están relacionados de alguna forma —son hermanos o están casados—, estos forman un grupo. Por último, se documentarían los resultados de los pasos anteriores en metadatos. Esto ayuda a que las siguientes limpiezas sean más capaces de reconocer direcciones y a que los usuarios finales de las aplicaciones puedan llevar a cabo mejor las operaciones de un DW.

Como se puede apreciar, sería bastante tedioso llevar manualmente este proceso, y para hacerlo automatizado se necesitaría de aplicaciones sofisticadas que contengan algoritmos de análisis gramatical (*parsing*) de direcciones, algoritmos de macheo, e inmensas tablas con gran cantidad de entradas que provea sinónimos para las diferentes partes de las direcciones.

Actualmente se dispone de numerosas herramientas que se ocupan de la limpieza de datos. Para trabajos intensos de este tipo, deben considerarse herramientas que se han desarrollado para esas tareas. Existen dos grandes competidores: Enterprise/Integrator de Apertus Technologies y la herramienta Integrity Data Reengineering de Vality [7, 3].



La empresa Enterprise/Integrator ofrece no solamente limpieza de datos, sino también extracción, transformación, carga de datos, repetición, sincronización y administración de la *metadata*. Es bastante caro —de $130000 a $250000—, pero se puede ahorrar dinero si se tiene en cuenta que elimina la necesidad de otras herramientas de gestión de DW [7].

La desventaja principal del enfoque que utiliza Enterprise/Integrator es que el usuario tiene que conocer, o ser capaz de deducir, las reglas del negocio y de la limpieza de datos.

Por su parte, la herramienta Integrity Data Reengineering de Vality tiene un enfoque diferente; ella analiza los datos caracter por caracter y automáticamente emergen los modelos y las reglas del negocio. Este enfoque tiende a dejar pocas excepciones para manejarse manualmente y el proceso tiende a consumir menos tiempo.

Integrity puede costar $250000, además de requerir también una herramienta como *Warehouse Manager* o *Passport* para las operaciones de extracción/carga. Sin embargo, pueden ser suficientes los utilitarios disponibles con las bases de datos. [7]

Si los datos que requieren limpieza son predominantemente nombres —incluyendo nombres de compañías— y direcciones, las compañías como *Harte-Hanks Communications* e *Innovative Systems* proveen no solamente herramientas de software, sino que actualizan periódicamente los archivos de datos para ayudar a combinar las variantes de los nombres de las compañías, detectar códigos postales que no corresponden a las direcciones proporcionadas y encontrar anomalías similares [7].

Las soluciones orientadas al nombre y la dirección pueden costar en cualquier parte desde $30000 a más de $200000, dependiendo del tamaño del DW en cuestión [7].

Otras herramientas de este tipo son:

- Trillium Software System, Trillium Software [4, 2]
- NADIS, MasterSoft International [3]
- Ultra Address Management, The Computing Group [3]
- PureName PureAddress, Carleton [3]

Importancia de la estandarizacion de direcciones para las empresas

Como se menciona anteriormente, la estandarización forma parte de los seis pasos necesarios para llevar a cabo la limpieza de datos. Esta consiste en separar la información en diferentes campos, así como unificar ciertos criterios para un mejor manejo y manipulación de los datos.

Tener datos estandarizados, consistentes y con calidad, resulta muy útil y a veces de vital importancia para las empresas que utilizan DW. Un ejemplo de ello son aquellas organizaciones cuyos datos referentes a sus clientes son de gran valor.

Una información almacenada y utilizada por varias organizaciones son los nombres y direcciones de sus clientes. El manejo de este tipo de datos no es tarea fácil. Más del 50 % de las compañías en Internet no pueden responder a las necesidades de todos sus clientes y no se pueden relacionar con ellos a causa de la falta de calidad en sus datos [5].

Para comunicarse efectivamente con sus clientes, por teléfono, por correo o por cualquier otra vía, una empresa debe mantener una lista de sus clientes extraordinariamente limpia. Esto no solo provoca que existan menos correos devueltos y más envíos precisos, sino que además, mejora la descripción y análisis de los clientes, que se traduce en un servicio más rápido y profesional.

Hay muchos ejemplos de aplicaciones basadas en la información del cliente que necesitan que sus datos, y principalmente sus direcciones tengan integridad, algunos de ellos son:

- Sistemas CRM —*Customer Relationship Management*, Gestión de las Relaciones con el Cliente—.
- E-Business —Negocios electrónicos—.
- Call Centers —Oficina o compañía centralizada que responde llamadas telefónicas de clientes o que hacen llamadas a clientes (*telemarketing*)—.
- Sistemas de Marketing.

Del mismo modo, se pueden mencionar algunas de las organizaciones que mayormente son beneficiadas por la limpieza de direcciones postales de sus clientes.

- Bancos y Finanzas
- Gobierno
- Salud
- Telecomunicaciones

Específicamente en la Gerencia Territorial de ETECSA, en Villa Clara, se vio la necesidad, por diferentes áreas, de tener las direcciones postales de forma estandarizada.

Por ejemplo:

- En el Centro de Facturación, se almacena toda la información relacionada con los servicios telefónicos y clientes de la Empresa, y dentro de ella, las direcciones postales, dato fundamental para la distribución de la factura telefónica.

El proceso de facturación tiene días planificados para la entrega de la factura, que no deben violarse. Del cumplimiento de estas fechas depende que el cliente pueda pagar su cuenta en el tiempo adecuado. Por esta razón, las direcciones asociadas al usuario deben ser consistentes. Si se añadiera la posibilidad de tenerlas estandarizadas y segmentadas en sus partes, las facturas podrían agruparse por algún elemento de la dirección —reparto, código postal—, lo que agilizaría y facilitaría el trabajo de los mensajeros.

- En la Filial de Red, ante cada nueva inversión necesita tener bien clara la densidad telefónica por áreas —cantidad de servicios que tiene en una zona determinada—, y para ello requiere saber la dirección correcta

de cada equipo terminal, factor que influye notablemente en el desarrollo de planes precisos y adecuados para las redes exteriores y la instalación de nuevos servicios.

- En la Filial de Clientes que se encarga, entre otras tareas, de instalar los nuevos servicios y de atender las reclamaciones e interrupciones. Para llevarla a cabo necesitan la localización exacta —es decir, la dirección postal— de los abonados para poder llegar hasta ellos en el menor tiempo posible.
- En el Grupo de Mercadotecnia, para desarrollar los planes de inversiones y presupuestos, es necesario disponer de la distribución de la densidad telefónica en los lugares en que se planea instalar los nuevos servicios en el año, debido a que depende, en gran medida, de la intensidad de uso —los ingresos promedio— de estos nuevos servicios y, por lo tanto, la cantidad de ingresos que será posible obtener por este concepto. Además, para la integración de las bases de datos que tiene el área de planta exterior en una base de información geográfica, tanto para la operación como para la realización de los análisis de mercado, es un requisito indispensable disponer de un catálogo de direcciones que tenga la calidad adecuada para que este proceso sea lo más fácil posible.
- El Grupo de Ventas realiza gestión de venta con los clientes de mayor poder de negociación, nivel adquisitivo y, en general, con aquellos que en alguna oportunidad han hecho compras en los puntos de ventas de ETECSA. Estos clientes son visitados en sus residencias en aras de lograr nuevas compras y se les propone nuevos servicios de la cartera de negocios. Un ejemplo de atención especializada lo constituye la postventa de insumos —película de recambio, papel térmico, baterías, cargadores— a aquellos usuarios que han invertido en equipos costosos como los Fax, y que, a veces, por desconocimiento o falta de tiempo adquieren estos insumos con otros proveedores, incluso a precios superiores. También este grupo hace promociones de algunos servicios específicos, como el de Tarifa Mixta, a clientes con alto histórico de Tráfico de Entrada Internacional, por representar clientes potenciales que pueden proporcionar grandes ingresos a la empresa. Con este fin se brinda atención personalizada llevando a cabo visitas a sus domicilios utilizando las direcciones postales asociadas a los datos de los clientes.
- Las Oficinas Comerciales realizan visitas a los clientes morosos o deudores de sus facturas telefónicas para que no se deje de ingresar por este concepto. Para poder cumplir con esta tarea las Oficinas Comerciales requieren de una lista las direcciones de abonados limpia y estandarizada.

Queda demostrado que para todas estas áreas es fundamental tener una lista de direcciones postales limpia y normalizada, además segmentada en sus elementos, para poder utilizarlas con mayor eficiencia.

Conclusiones

1.Los DW son el centro de atención de las grandes empresas actuales, ya que proporcionan una herramienta para hacer un mejor uso de la información y para el soporte al proceso de toma de decisiones gerenciales.

2.Existen numerosas causas que provocan suciedad en los registros de los sistemas operacionales, lo que trae como consecuencia que haya gran cantidad de datos almacenados en las empresas que carece de la calidad adecuada para ser utilizada de forma confiable.

3.El problema de la limpieza de datos es uno de los tres problemas fundamentales de los DW. Sin embargo, es poco tratado o evitado por muchas organizaciones, ya que no consideran adecuadamente el impacto negativo que puede ocasionar para el negocio el tener almacenada información deficiente.

4.Existen herramientas comerciales que se ocupan de la limpieza de datos.

5.Para las organizaciones actuales, la estandarización de las direcciones de sus listas de clientes es un punto fundamental a tener en cuenta, debido a que direcciones postales almacenadas que no tengan esta característica pueden provocar pérdida de credibilidad de las organizaciones, que a su vez, lleva a la pérdida de clientes como consecuencia de un servicio poco eficaz.

Bibliografía

[1] Casares, C. *Data Warehousing*. Disponible en: http://www.programacion.com/bbdd/tutorial/warehouse/15/#warehousing_desarrollo_confi. (Consulta: nov./2005)

[2] E. Corporation. *Calidad de Datos: Fundamento de la Empresa Exitosa*. Disponible en: http://www.eniaccorp.com/noticias2.htm] (Consulta: ene./2006).

[3] Galhardas, H. *Data Cleaning and Integration*. Disponible en: [http://web.tagus.ist.utl.pt/~helena.galhardas/cleaning.html]. (Consulta: oct./2005).

[4] H.H. Enterprise. *The Trillium Software System*. Disponible en: http://www.trilliumsoftware.com/site/content/products/tss/index.asp]. (Consulta: ene./2006).

[5] Hussain S.; Beg J. *Data Quality: A Problem and an Approach*. Disponible en: http://doc.advisor.com/doc/13060. (Consulta: oct./2005).

[6] Kimball, R. *Dealing with Dirty Data*. Disponible en: http://www.dbmsmag.com/9609d14.html. (Consuta: oct./2005)

[7] Llombart, Ó. A. *BI: Inteligencia aplicada al negocio*. Disponible en: http://www.icc.uji.es/asignatura/obtener. (Consultado: oct./2005).

[8] Martín, J.; Morrás, C.; García, M.L.; Tello, L.I.; Vivancos A.J. *Sistemas de soporte a la gestión del negocio*. Disponible en: http://www.tid.es/presencia/publicaciones/comsid/esp/articulos/vol812/soporte/soporte.html

[9] P. W. pwidlund@schober.es and A. G. d. S. asoto@schober.es, *Bases de Datos y Calidad de la Información*. Disponible en: [http://www.icemd.com/area-blogs/mes_actual.asp?id_seccion=46. (Consulta: oct./2005)

[10] ReallTech. *Data Warehousing*. Disponible en: [http://www.sqlmax.com/dataw1.asp]. (Consulta: ene./2006).

Nota editorial: se ha decidido hacer una excepción con las normas para citas, notas o referencias bibliográficas y la bibliografía de la revista. Se ha respetado la forma en que las ha utilizado la autora.