La calidad de los datos: Su importancia para la gestión empresarial

Jobany José Heredia Rico[1]

José Alberto Vilalta Alonso[2]

Artículo de reflexión
no derivado de investigación

Fecha de recepción: 05-05-09 Fecha de aceptación: 05-06-09

Abstract

In this paper it is commented the importance of using quality data within an enterprise for its appropriate management. The impact of poor data quality, beyond decision-making processes, is reflected on operative processes and frequently on costumers. Reference is made to some ways of measuring data quality level and the necessity of taking initiatives, which consider current trends, regarding the problems presented in this field.

Key Words

Data, data quality, impact, improvement.

Resumen

En este trabajo se aborda la importancia que tiene utilizar datos con calidad en una empresa para la correcta gestión de la misma. El impacto de la pobre calidad de los datos, más allá

1. Ingeniero Industrial. Profesor del Departamento de Ingeniería Industrial de la facultad de Ingeniería Industrial del Instituto Superior Politécnico José A. Echeverría. Investiga en temas relacionados con la Calidad y la aplicación de técnicas estadísticas para la gestión.
2. Ingeniero Industrial. Máster en Aseguramiento de la Calidad. Doctor en Ciencias Técnicas. Profesor y Jefe del Departamento de Ingeniería Industrial de la facultad de Ingeniería Industrial del Instituto Superior Politécnico José A. Echeverría. Investiga en temas relacionados con la calidad y la aplicación de técnicas estadísticas para la gestión.

de los procesos de toma de decisiones, se refleja en los procesos operativos de la empresa y, muy frecuentemente, en los clientes. Se hace referencia a algunas vías de medición del nivel de calidad de los datos y a la necesidad de acometer iniciativas que tengan en cuenta las tendencias actuales para resolver los problemas que se presentan en este campo.

Palabras clave

Datos, calidad de los datos, impacto, mejora.

Introducción

Como consecuencia de los avances que hoy día existen en las tecnologías de captura y almacenamiento de datos e información, las empresas se enfrentan a un crecimiento exponencial en cuanto a la cantidad y diversidad de datos a gestionar. Esto significa que no sólo aumentan los volúmenes de datos sino también los elementos a los cuales se les asocian datos e información (Loshin, 2001; Pipino y otros, 2002; Ponjuán, 2006).

Por esta razón, la pobre calidad de los datos es un factor que afecta cada vez más el desempeño de la empresa, ya que deteriora de alguna forma las relaciones que se mantienen con los elementos a los que están asociados los datos (proveedores, clientes internos y externos, empleados) (Loshin, 2001; Redman, 2001).

Este artículo tiene como objetivo alertar a los directivos de empresas respecto a las dificultades en la gestión organizacional derivadas de una inadecuada calidad de los datos.

Importancia de los datos

Los datos son "un término general para denotar alguno o todos los hechos, letras, símbolos y números referidos a, o que describen un objeto, idea, situación, condición u otro factor" (Maynard, 1982) y constituyen un elemento fundamental para la toma de decisiones objetivas a todos los niveles de una organización (Javed y Hussain, 2003; Levy, 2004; Naveh y Halevy, 2000). Es más, para una organización moderna, los datos constituyen uno de sus recursos estratégicos (Olson, 2002; Tayi y Ballou, 1998).

En la actualidad existe un gran interés organizacional por lograr lo que se ha denominado "gestión del conocimiento". Esto implica, primeramente, tomar los datos generados en los procesos empresariales y convertirlos en información al agregarles valor mediante procesos de agrupación, clasificación, etc; para posteriormente convertir esta información en conocimiento, a través de procesos de separación, evaluación, comparación, etc. (Ponjuán, 2006). Por lo tanto, sin la existencia de datos no se llegaría nunca a obtener conocimiento.

Por otra parte, el uso de datos como base para la toma de decisiones ha sido una práctica ampliamente recomendada, en contraposición al hecho de desarrollar la toma de decisiones con base en la intuición. De hecho, uno de los principios de gestión de la calidad es el enfoque basado en hechos para la toma de decisión, el cual plantea que las decisiones eficaces se basan en el análisis de los datos y la información (NC ISO 9000: 2000).

Sin embargo, no basta con la existencia de datos ni con la voluntad de basar las decisiones que se tomen, en ellos (Gil-Aluja, 2000; Javed y Hussain, 2003), se requiere que éstos tengan la calidad adecuada. Es decir que, cuando con base en los datos se favorezca una decisión sobre otra, se tenga la certeza de que los datos estén libres de errores y que, además, posean atributos relevantes (Redman 2001).

Calidad de los datos. Definición

El término calidad, en relación con los datos, toma sentido por el hecho de que los datos al igual que los productos y servicios, deben adecuarse al uso que se les pretende dar. El término preciso para el uso en este caso implica que dentro de cualquier contexto operacional, el dato que va a ser utilizado satisfaga las expectativas de los usuarios de los datos. Dichas expectativas se satisfacen en gran medida si los datos son útiles para lo que estos los necesitan, son fáciles de entender e interpretar, y además son correctos (Loshin, 2001; Redman, 2001).

Para garantizar estos aspectos se debe hacer, en principio, un diseño apropiado de la base, tabla o lista de datos, con el fin de definir correctamente los atributos o tipos de datos en la misma; y posteriormente realizar un adecuado diseño de los procesos de producción de datos, garantizando que los datos lleguen a la base o tabla de datos, libres de defecto y con las demás características deseadas (Lee y Strong, 2003; Redman, 2001).

De estas definiciones se puede deducir que la calidad de los datos es un concepto relativo. Por ejemplo, los datos que un consumidor puede considerar como de calidad aceptable, son de calidad inaceptable para otro consumidor con requisitos más rigurosos de uso o con otros usos previstos. Por lo tanto, al variar las expectativas de los usuarios respecto a los datos, tambien varían las características que deben tener los mismos para ser considerados como adecuados. Estas características o cualidades que deben poseer los datos para ser considerados como adecuados se denominan dimensiones de calidad de los datos (Abate, 1998; Pipino y otros, 2002).

Esto quiere decir que la calidad de los datos está asociada a un conjunto de dimensiones o atributos que son los que la definen. Un objetivo fundamental de la definición de las dimensiones es poder establecer un lenguaje común y también focalizar los problemas de calidad de los datos y las oportunidades de mejora (Javed y Hussain, 2003; Naveh y Halevy, 2000). Entre las dimensiones más importantes, pues son las más utilizadas y referenciadas están la exactitud, la integridad, la consistencia y la coherencia (Cong *et al*., 2007; Levy, 2004; Olson, 2002; Redman, 2001; Strong, Lee y Wang, 1997), es conveniente señalar que éstas deben ser definidas teniendo en cuenta las características propias de cada sector (Gendron y D'Onofrio, 2001).

Implicaciones de la mala calidad de los datos

La mala calidad de los datos afecta de diversas maneras la gestión empresarial. Obviamente una afectación primaria de la mala calidad de los datos, la constituye su efecto sobre la toma de decisiones. Sucederá entonces que datos de mala calidad implicarán procesos de toma de

decisiones inefectivos y, en última instancia, ineficientes. La ineficacia se materializa en el hecho de que cuando los datos son erróneos, implican decisiones erróneas en las diversas alternativas sobre las que se decide. Considerando que las decisiones que se toman en la empresa están relacionadas con una multitud de elementos (clientes, proveedores, productos, procedimientos de trabajo, etc), se deduce a su vez la afectación que produce una inadecuada toma de decisiones. Por otra parte la ineficiencia se debe al hecho de que muchas veces se logra tomar las decisiones con datos correctos, pero con un costo adicional en tiempo debido a la demora o falta de puntualidad de los datos. (Redman, 2001; Redman, 2004).

Otro resultado de la inadecuada calidad de los datos que resulta muy costoso es el efecto sobre los clientes de la empresa. Este se puede materializar en la insatisfacción de los clientes debido a nombres incorrectos, facturas con cantidades erróneas, envío de productos o cantidades equivocadas, etc. Además pudiera materializarse en costos que se generen en los clientes, por ejemplo, el tiempo que dedique el cliente a solucionar el problema creado por el error. Estas dificultades probablemente provocarán la pérdida del cliente, e incluso de otros clientes potenciales (Klein, 1998; Redman, 2001; Redman, 2004).

Los costos en tiempo y demás recursos que las empresas dedican a la detección y corrección de errores en los datos es otra secuela de la baja calidad de los datos. En algunas empresas de producción, gran parte del personal administrativo e incluso parte del personal directamente relacionado con la producción, dedica un porcentaje no despreciable de su tiempo de trabajo a la corrección de errores en los datos, aspecto particularmente negativo en el caso de los trabajadores vinculados a la producción (Heredia y Vilalta, 2008). En otras ocasiones el departamento de ventas o de servicio al cliente de la empresa incurre en diversos costos al tener que realizar continuamente correcciones de las direcciones, pedidos y facturas de los clientes (Klein, 1998; Redman, 2001; Redman, 2004).

Otras consecuencias negativas de la inadecuación de los datos, quizás con una menor componente económica, son la insatisfacción en los empleados que produce el hecho de tener que corregir continuamente datos erróneos y el propio hecho de trabajar en un ambiente donde los datos que se procesan tienen frecuentemente problemas.

También es considerable el efecto de la mala calidad de los datos en el éxito de muchas de las nuevas aplicaciones informáticas de ayuda a la toma de decisiones, como los Almacenes de Datos (Caro *et al.* 2006), la Minería de Datos y los Sistemas de Gestión de Relaciones con los Clientes (Olson, 2002). Cuando en un proceso de producción de datos se obtienen malos resultados, no siempre es conveniente aplicar nuevas tecnologías, pues el efecto de la misma podría ser nulo e incluso negativo. Esta situación se debe al hecho, muy comentado en la literatura, que no es conveniente automatizar procesos ineficientes (Redman, 2001; Redman, 2004; http://web.mit.edu/tdqm).

No obstante todos los problemas hasta aquí comentados, en muchas organizaciones no se tiene en cuenta la calidad de los datos, situación que está motivada por diversos factores, como pueden ser las grandes cantidades de datos que se generan en una organización, lo cual dificulta el proceso de detección y corrección de errores (Javed y Hussain, 2003; Levy, 2004), el hecho de que los datos son, a diferencia de otros recursos, intangibles, lo que dificulta su medición. Además, la dificultad que existe al tratar de cuantificar las mejoras derivadas de un programa de mejora de la calidad de datos (Redman, 2001; Redman, 2004).

¿Cómo evaluar la calidad de los datos?

Una manera sencilla y práctica de evaluar la calidad de los datos es el cálculo de una tasa de error, para todos, o para los más importantes atributos dentro de una base de datos (Good Clinical Data Management Practices Guide, 2002; Heredia y Vilalta, 2008; Vilalta, 2008). Para esto es recomendable comparar un número de veces que sea estadísticamente adecuado, los datos entre la fuente original y la base, lista o tabla de datos[3] (Good Clinical Data Management Practices Guide, 2002).

Otra manera de medir la calidad de los datos es enfocándose en las dimensiones de calidad (Pipino y otros, 2002; Abate, 1998; Redman, 2004). Para esto se hace necesario, en principio, definir las dimensiones de calidad que sean importantes para el conjunto de datos en análisis, y después se deben establecer indicadores que permitan cuantificar o calificar el grado de adecuación del dato atendiendo a cada dimensión.

Algunas de las dimensiones que se definan, sobre todo las relacionadas con los valores de los datos (Redman, 2004), podrán ser medidas a partir del cálculo de un indicador que sea el resultado de comparar los datos entre la fuente original y la base, lista o tabla de datos. Sin embargo en el caso de otras dimensiones cuya medición directa sea más compleja (relevancia, puntualidad, accesibilidad, etc), una forma de evaluación sería la aplicación de encuestas al personal implicado en la producción y utilización de los datos, para obtener criterios cualitativos respecto a las dimensiones (Lee y Strong, 2003).

Otros indicadores del nivel de calidad de los datos menos relacionados con mediciones directas realizadas sobre la base de datos, serían las estimaciones que se puedan obtener del costo (en tiempo o dinero) dedicado a la detección y corrección de errores en los datos, las quejas y reclamaciones de los clientes de la empresa que estén asociadas a este factor, así como cualquier otro indicador que pueda ser reflejo de un inadecuado comportamiento empresarial a causa de la mala calidad de los datos (Redman, 2001; Redman, 2004).

Enfoques actuales para mejorar la calidad de los datos

En la actualidad la tecnología informática es ampliamente utilizada con el objetivo de mejorar el desempeño organizacional en cuanto a calidad de datos, por lo cual se han desarrollado una amplia gama de softwares.

En general, con estos softwares lo que se trata es de realizar un proceso denominado limpieza de datos (data cleaning) (López y Pérez, 2002; Loshin, 2001; Rahm y Hong, 2000). La limpieza de datos implica la exploración en el conjunto de datos, seguida de la validación y verificación del contenido mediante parámetros de validez lógicos lo que permite detectar los posibles problemas y trabajar en su corrección (López y Pérez, 2002). Dentro del proceso de limpieza de datos se desarrollan diversos pasos importantes, los cuales agregan valor en sí mismo al esfuerzo desarrollado por mejorar la calidad. Estos pasos son: el análisis, corrección y estandarización de los datos; la comparación entre datos de diversas fuentes,

3. La fuente original es la fuente donde se registra el dato por primera vez, por ejemplo, en el caso de datos de producción, los documentos donde se registran estos datos en el área productiva.

y la posterior consolidación de los datos en una única base (López y Pérez, 2002; Loshin, 2001, http:// www.dataflux.com/datamanagement).

También son grandes los esfuerzos que se dedican a garantizar un adecuado diseño de entrada de datos cuando se construye la base de datos del sistema. El uso de cuadros de edición, listas desplegables, controles numéricos, etc, ayudan a lograr una entrada de datos casi libre de errores. Ésta es una práctica especialmente importante pues se ha demostrado que cuando este esfuerzo no se hace, el proceso de entrada de datos engendra una razón de error de un 5% o más (López y Pérez, 2002).

Para mejorar definitivamente la calidad de los datos, la tecnología informática no es la única solución, ya que para encontrar las causas de los problemas e incluso para mejorarlos, se necesita enfocarse en cuestiones no sólo relacionadas con el uso de tecnología de punta (Redman, 2001). Pretender mejorar la calidad de los datos sólo a partir de buscar y corregir errores sería actuar sobre los efectos del problema, cuando realmente lo más importante es detectar las causas de los errores de calidad.

En esta lógica se basan los llamados sistemas de calidad de datos de segunda generación, los cuales tratan de identificar y eliminar las causas de familias completas de errores para evitar errores futuros (más que para detectarlos y corregirlos) y de implantar una infraestructura de gestión que permita desarrollar adecuadamente este objetivo (Redman, 2001). Un elemento importante considerado en estos sistemas es el diagnóstico. En ese sentido también se han desarrollado procedimientos o metodologías para el diagnóstico de la calidad de los datos que, entre otros aspectos, inciden en la búsqueda de las causas que provocan los problemas de calidad (Heredia y Vilalta, 2008; Vilalta, 2008).

Diversas investigaciones han mostrado que algunos de los problemas de calidad de datos son consecuencia de aspectos como cadenas de información diseñadas de forma deficiente, poca motivación de los trabajadores en cuanto a la calidad de los datos, inadecuada capacitación de los trabajadores vinculados a los procesos de producción de datos, condiciones ergonómicas inadecuadas en estos procesos, entre otras (http://web.mit.edu/tdqm; Heredia y Vilalta, 2008). Todos estos factores requieren de soluciones no vinculadas precisamente al uso de un software de calidad de datos.

En aquellas organizaciones que presentan problemas de calidad de datos, no se emprenden iniciativas para la mejora de la misma, motivado en muchas ocasiones por la dificultad de medir la ganancia esperada de éstas. No obstante, la ganancia en este caso siempre se puede materializar en el ahorro de costos (costos relacionados con los diversos aspectos comentados en este artículo) que pueda tener la empresa al poner en práctica la iniciativa de calidad de datos. En última instancia, si al poner en práctica la iniciativa, se espera mejorar el servicio al cliente y en definitiva las relaciones con éstos, pues estas serán razones suficientes para aplicar la mejora proyectada.

Conclusiones

El uso de datos con calidad se hace imperante, ya no sólo para desarrollar apropiadamente los procesos de toma de decisiones, sino para la correcta gestión de la empresa misma. Para

lograr una adecuada calidad de los datos, se debe partir de medir los niveles de calidad en este sentido, para lo cual se pueden calcular, desde tasas de error en las bases de datos hasta los costos que se generen por los errores en los datos. Los esfuerzos para mejorar la calidad de los datos que se deriven de la evaluación, deben poner en justo orden el uso de la tecnología informática y las técnicas de gestión disponibles.

Bibliografía

Abate, Marcey L & Diegert, Kathleen V. (1998). *A Hierarchical Approach to Improving Data Quality*. http:// www.dataquality.com.

Caro, A., Calero C. & Caballero, I. (2006). *A first Approach to a data Quality Model for Web Portals*. International Conference on Computational Sciences and its Applications (ICCSA).

Cong, G., Fan, W., Geerts, F., Jia, X. & Ma, S. (2007). *Improving data quality: consistency and accuracy*. Proceedings of the 33 International Conference on Very Large Data Base.

Gendron, M. & D'Onofrio, M. (2001). *Data Quality in the Healthcare Industry*. Data Quality, Vol. 7, No. 1.

Gil – Aluja, J. (2000). *Las decisiones y la incertidumbre*. Barcelona.

Heredia, J & Vilalta, J. (2008). *Procedimiento para el Diagnóstico de la Calidad de los Datos*. Sexto Taller de Calidad. Universidad de la Habana.

Javed, B. & Hussain, S. (2003). *Data quality – A problem and an approach*. Wipro Technologies.

Klein, Barbara. (1998). *Data Quality in the Practice of Consumer Product Management*. http://dataquality.com.

Levy, S. (2004). *Model Documents and forms for Organizing and Maintaining a Data Quality Program*. www .dataqualitymodeldocument.com

López, Beatriz & Pérez, Ramiro. (2002) *¿Tiene usted datos sucios?* Revista: GIGA. La Habana, Cuba.

Loshin, David. (2001). *Integration and the Data Quality Imperative: The Data Quality Monitor*. http:// www.datajunction.com.

Maynard, J. (1982). *Dictionary of Data Processing*. Londres, Inglaterra.

Naveh, E. & Halevy, A. (2000). *A hierarchical framework for a quality information system*. Total Quality Management, Vol. 11, No. 1, p 87-111.

Olson, J. (2002). Data Profiling: *The Data Quality Analyst's Best Tool*. DM Direct, December. DMReview.com.

Pipino, Leo; Lee, Yang & Wang, Richard (2002). *Data Quality Assessment*. http://web.mit.edu/tdqm/www/tdqmpub/PipinoLeeWangCACMApr02.pdf.

Ponjuán, Gloria. (2006). *Gestión de Información en las Organizaciones*. Editorial Félix Varela.

Rahm, E. & Hong, H. (2000). *Data cleaning: Problems and current Approache*. IEEE Techn. Bulletin on Data Engineering.

Redman, Thomas C. (2001). *Sistemas de calidad de datos de segunda generación*. Manual de Calidad de Juran. McGraw Hill.

Redman, Thomas C. (2004). *Data an unfolding quality disaster*. http://www.dmreview.com/portals/dataquality.

Strong, D.M., Lee, Y.W. & Wang R.Y. (1997). *10 Potholes in the Road to Information Quality*. IEEE Computer, Vol. 30, No. 8, pp. 38 – 46.

Strong, Diane & Lee, Yang. (2003). *Process knowledge and data quality outcomes*. http://web.mit.edu/tdqm/www/tdqmpub/LeeStrong.pdf.

Tayi, G. & Ballou, D. (1998). *Examining Data Quality*. Communications of the ACM. Vol. 41, No. 2.

Vilalta, J. (2008). *Procedimiento para el Diagnóstico de la Calidad de los Datos*. Una nueva versión. 14 Conferencia de Ingeniería y Arquitectura., Cujae, La Habana, Cuba.

Norma Cubana ISO 9000:2000: *Sistemas de gestión de la calidad*. Fundamentos y vocabulario.

Good Clinical Data Management Practices Committee. (2002). Good Clinical Data Management Practices Guide.

Data Management. http:// www.dataflux.com/datamanagement.

The MIT Total Data Quality Management Program. http://web.mit.edu/tdqm.