Contenido

Página

1 Concepto de calidad de los datos

> *La calidad se mide por el costo de los desembolsos que implica el hecho de no actuar conforme con ella, es decir, el costo de hacer las cosas mal.*
>
> Philip Crosby
> Quality Is Free

La calidad de los datos no es un concepto nuevo. Por el contrario, este tema ha sido analizado desde hace mucho tiempo por especialistas en sistemas y tecnologías de información. Dichos especialistas han tratado de darle solución a este problema, sin embargo es mucho lo que hay por hacer aún en América Central en cuanto al tema.

Son muchas las definiciones que podemos encontrar en cuanto a calidad de datos. Sin embargo, la que encontramos más acertada es la de Ken Orr [Orr 1998], la cual dice que *"la calidad de los datos es la medida de conformidad entre la vista de datos presentada por un sistema de información y los mismos datos en el mundo real"*. Esto implica la necesidad de observar experimentalmente, medir y comparar.

Para Giri Kumar et al. [Kumar 1998], el término "calidad de datos" es mejor definido como "uso por conveniencia", lo que implica que el concepto de calidad de datos es relativo. Así, los datos que en determinado momento se considera tienen calidad apropiada para un uso específico, bien podrían no poseer la suficiente calidad para otro uso.

Las tendencias actuales hacia el uso de múltiples datos se manifiestan en la popularización de los sistemas de bases de datos relacionales, los depósitos de datos y la minería de datos, entre otros. Paralelamente, el poder computacional ya no está centralizado y llega a los departamentos y a los escritorios. Los datos y su uso crecen día a día. Esto ha traído consigo la necesidad de establecer mecanismos para mejorar la calidad de los datos. Además, el uso por conveniencia implica que se necesita mirar más allá de las preocupaciones tradicionales de la exactitud de los datos. Por ejemplo, los datos encontrados en los sistemas de contabilidad pueden ser exactos, pero son inadecuados para ciertos propósitos si no son lo suficiente oportunos. Frecuentemente encontramos que las bases de datos situadas en diferentes divisiones de la organización pueden generar datos correctos, pero son inadecuados si se desea combinar los datos con otros que tienen formatos incompatibles.

La mayoría de las organizaciones tienen que vivir con la realidad de la pobre calidad de los datos, debido principalmente a las pobres reglas de validación, al uso excesivo de campos o el mal empleo de algunos campos. Por ello, es muy importante que en las fases tempranas o iteraciones de un proyecto de sistemas se analice la calidad de los datos y los resultados se documenten.

El uso alternativo de datos impone requerimientos diferentes de calidad. Los sistemas transaccionales con frecuencia no motivan la satisfacción de la calidad de los datos para el desarrollo de nuevas aplicaciones. Para ilustrar cómo varios negocios utilizan los datos, lo cual impone requerimientos diferentes en estos, se considera la comparación de un sistema de procesamiento transaccional y un sistema de apoyo a la toma de decisiones (DSS). Ambos

trabajan con datos para diferentes propósitos:

- Un sistema transaccional utilizado para las actividades de entrega cotidiana de productos, requiere información básica actual del estado de los vehículos y operadores en cada localización, para satisfacer rápidamente las solicitudes que entran. Cuando se deteriora un vehículo o un operador ha cumplido su período de trabajo y requiere un descanso, estos eventos son inmediatamente capturados por el sistema para apoyar al personal de entrega.
- Una aplicación DSS utilizada para asignar bien la entrega (por ejemplo, vehículos y operadores) para las localizaciones y evaluar el mantenimiento del vehículo, trata de capturar los datos relativos al tiempo de averías, cambio de operador y otros ítemes relacionados. Los datos de interés incluyen el número de horas operadas, las fechas de reparación de partes, las fechas de la utilización de vehículos, etc. Estos detalles son utilizados por gerentes de recursos, para derivar de ellos las cifras agregadas de la supervisión y administración del inventario bien entregado.

Ambos conjuntos de usuarios se interesan por las averías de los vehículos, pero de manera diferente. Al personal de inventario le interesa conocer qué vehículos están con averías y cuáles están disponibles en un período de tiempo dado. Por otro lado, a los gerentes les interesa conocer porqué los vehículos se averían, cuánto tiempo se tarda en solucionar los problemas y en reemplazar las partes.

2 Categorías y dimensiones de la calidad de los datos

Investigar acerca del proceso de análisis y mejoramiento de la calidad de los datos puede revelar muchas variables asociadas a los datos, las cuales pueden tener efectos en su calidad. Richard Wang et al. [Wang 1998b] identificaron dieciséis (16) dimensiones de la calidad de los datos, las cuales son agrupadas en cuatro (4) categorías. Estas se pueden apreciar en la tabla No. 1. Es importante señalar que las medidas de la calidad de los datos no pueden hacerse sin tomar en cuenta las decisiones acerca de las variables implícitas o explicitas. Un reconocimiento cuidadoso de estos valores y efectos es esencial para entender y controlar la calidad de los datos.

Categoría de la Calidad de Datos	Dimensiones de la Calidad de Datos
Intrínseca	• Exactitud, objetividad • Reputación • Credibilidad
Accesibilidad	• Accesibilidad • Seguridad de acceso
Contextual	• Relevancia • Cantidad de datos • Integridad • Valor agregado • Oportunidad
Representativa	• Interpretabilidad • Fácil entendimiento • Representación concisa • Representación consistente.

Tabla No. 1: Categorías y dimensiones de la calidad de los datos

El primer paso para mejorar la calidad de los datos según Levant Orman [Orman 1994] es entender las dimensiones de la calidad de los datos y su importancia relativa.

2.1 Análisis de las dimensiones

Es importante entender el significado de las dimensiones de la calidad de los datos. A continuación describimos cada una de ellas para tener un panorama más claro:

- ***Exactitud***: se da cuando el ítem de dato capturado en un sistema de información refleja el estado del mundo real que se pretende representar. Es la medida o el grado de conformidad entre el valor del dato (o conjuntos de datos) y el origen, que se supone debe ser correcto.
- ***Credibilidad***: los usuarios de los datos no conocen la fuente a quien se le debe atribuir el problema de la calidad, ellos sólo conocen que existen conflictos. Esas preocupaciones aparecen como problemas de credibilidad.
- ***Reputación***: con el transcurso del tiempo, la información acerca de las causas de malas correspondencias se acumula a partir de las evaluaciones de la exactitud de diferentes fuentes, lo cual conduce a una pobre reputación para fuentes menos exactas.

- ***Objetividad***: sólo los que tienen conocimientos de los procesos de producción son conscientes de los problemas que aparecen acerca de la objetividad de los datos. Con el paso del tiempo, la información con naturaleza subjetiva durante la producción de los datos se acumula, dando como resultando datos de cuestionable credibilidad y reputación, que por ende dan poco valor agregado a los usuarios.
- ***Accesibilidad***: en cuanto a la accesibilidad, ésta tiene que ver con los niveles de acceso que se puedan mantener, dependiendo del personal que deba manejar los datos.
- ***Seguridad de acceso***: la seguridad de acceso tiene que ver con los mecanismos de control que se tengan para evitar el acceso inapropiado a datos confidenciales de la organización. Se deben mantener, además, los permisos pertinentes para las barreras de acceso.
- ***Relevancia***: muchos son los datos que normalmente existen en una organización, pero la relevancia trata de aquellos que son de gran importancia para la organización, es decir, se busca determinar cuáles son los datos que realmente se necesitan y pueden servir, por ejemplo para la toma de decisiones. Los datos con relevancia son los que agregan valor a las tareas de una manera oportuna.
- ***Valor agregado***: cuando una reputación de pobre calidad llega a ser de conocimiento común (datos sin importancia), los datos fuentes tienen poco valor agregado para la organización, dando como resultado un uso reducido de ellos.
- ***Integridad***: ocurre cuando se registran todos los valores ciertos para una variable y las variables explícitamente relacionadas con ella.
- ***Oportunidad***: es definida cuando el valor de un registro está actualizando en el momento que se requiere. La oportunidad tiene dos componentes: vigencia y la volatilidad. La vigencia refleja la edad de los datos, mientras que la volatilidad refleja la velocidad en que los datos llegan a ser obsoletos.
- ***Cantidad de datos***: muchas veces un patrón de grandes cantidades de datos puede conducir a problemas de integridad, los cuales se pueden interpretar a su vez como problemas de accesibilidad.
- ***Interpretabilidad***: los datos deben ser fáciles de interpretar, su descripción debe ser clara y fácil de comprender. Los datos no deben inducir a errores de interpretación.
- ***Fácil entendimiento***: independientemente de quién vaya a trabajar con los datos, éstos deben ser fáciles de entender. Por ejemplo, cuando se trata de interpretar diversos tipos de códigos, estos datos deben ser fáciles de entender para evitar confusiones.
- ***Representación concisa***: los datos deben ser mantenidos de tal forma que puedan ser representados en forma resumida.
- ***Representación consistente***: los datos deben ser mantenidos de tal forma que estén libres de variación o contradicción. La consistencia es la medida o el grado en que un conjunto de datos satisface un conjunto de restricciones.

A pesar de que la precisión y la utilizabilidad no existen en el conjunto de dimensiones descritas anteriormente, consideramos de gran importancia definirlas.

- ***Precisión***: es definida por Richard Wang [Wang 1998a] como un caso especial de carencia de ambigüedad. Él define ambigüedad como múltiples estados del mundo real rastreados en un solo estado del sistema de información. Por ejemplo, si hay un número insuficiente de dígitos en un sistema de información para representar los múltiples estados del mundo real, entonces el usuario no puede interpretar cómo un ítem de dato en el sistema de información representa

o corresponde al estado original del mundo real, que se desea representar. Allí un valor actual (estado del mundo real) no es conocido (por ejemplo, pronóstico de salarios para años futuros), múltiples estados del mundo real pueden ser usados para representar el valor actual. Este caso se puede clasificar como un caso de precisión sin ambigüedad. Algunas aplicaciones son más tolerantes que otras con respecto de la precisión.

- ***Utilizabilidad***: mide la utilidad de los datos para una aplicación particular. La utilizabilidad de los datos es relativa a las necesidades de procesamiento de una aplicación, ya que algunas aplicaciones necesitan procesar más datos que otras.

La falta de alguna de estas dimensiones en los datos puede generar problemas en la calidad de los datos y estos problemas pueden impactar en el desempeño de un sistema o de un usuario.

3 Impacto de la calidad de los datos

Resolver el problema de la falta de calidad de los datos no es una labor fácil. Sin embargo, son muchos los esfuerzos que hoy en día están orientados a esta tarea. Por ejemplo, el trabajo desarrollado por Edward Deming, Philip Crosby y J.M. Juran, llamado Administración Total de la Calidad, principalmente para procesos de manufactura, puede ser directamente aplicado a los datos. Esto se debe a que por ser productos de procesos de negocios, entonces a los datos les podemos aplicar esos mismos principios para mejorar su calidad. Esto puede hacerse para mejorar los procesos donde se crean y actualizan los datos [English 1996].

Larry English [English 1996], señala que para lograr un sistema de información exitoso es necesario implementar dos procesos en paralelo: la limpieza de datos y el mejoramiento en su operación de la calidad de los datos, además a las tareas usuales en el desarrollo. Este proceso de calidad de datos es llamado *Administración Total de la Calidad de Datos*. La limpieza de datos es requerida para eliminar el resultado de ejecutar procesos defectuosos por años, que llevan a producir gigabytes y terabytes de basura binaria. Por otro lado, English agrega que la mejor vía para mejorar la calidad de los datos es prevenir la entrada a la base de datos de datos que no son de calidad. Un beneficio que se puede adquirir al eliminar los datos que no son de calidad es la reducción de costos en la solución de problemas causados por este tipo de datos.

Hoy en día son muchos los problemas que acarrea el tener sistemas con pobre calidad de los datos. A continuación describimos algunos problemas que se pueden generar, las dimensiones que pueden ser afectadas, los efectos que pueden causar en la organización, las señales de advertencia, así como posibles soluciones, problemas con esa posible solución y la solución real que se le puede dar al problema. Los ejemplos están enfocados a los siguientes tipos de información: información de producción, información almacenada, información utilizada.

3.1 Fallas en la información de producción

De acuerdo con cómo la información se produce y cómo son los procesos de trabajo, se pueden dar diversas fallas. Por ejemplo, un pobre diseño y una pobre administración de un proceso de producción de la información pueden generar información de pobre calidad.

Múltiples orígenes de la misma información producen diferentes valores

Ejemplo # 1: en una empresa manufacturera se utilizan dos procedimientos separados para evaluar la calidad de un producto x. Los datos no se concilian.
Dimensiones afectadas: consistencia y credibilidad.
Efectos en la organización: financieros y problemas legales.
Señales de advertencia: diferentes sistemas son desarrollados para propósitos diferentes, los cuales requieren la misma información.
Posible solución: utilice un solo sistema. Proporcione un conjunto de información para los consumidores.
Problema con la posible solución: se pierde el propósito de sistemas que fueron desarrollados y no se utilizan.
Solución Real: desarrollar definiciones comunes y procedimientos consistentes.

Sistemas heterogéneos distribuidos inducen a definiciones, formatos y valores inconsistentes

Ejemplo # 2: diferentes sistemas en cada división o departamento, cada uno utilizando formatos diferentes.
Dimensiones afectadas: representación consistente, valor agregado y oportunidad.
Efectos en la organización: información inconsistente, la cual es difícil de accesar y agregar.
Señales de advertencia: múltiples sistemas a través de los departamentos.
Posible solución: los consumidores administran la extracción y agregación de información para cada sistema.
Problema con la posible solución: consumidores no entienden los datos y las estructuras de los archivos.
Solución Real: crear un depósito de datos (data warehouse)

3.2 Fallas en la información almacenada

Grandes cantidades de variada información por sistemas diferentes pueden ocasionar fallas con la información almacenada. Mantener más información almacenada no necesariamente es mejor.

Información no numérica es difícil de indizar

Ejemplo # 3: La información de imágenes médicas es relativamente fácil de almacenar, pero difícil de acceder.
Dimensiones afectadas: representación concisa, valor agregado y accesibilidad.
Efectos en la organización: costo de almacenamiento de información con beneficios potencialmente bajos.
Señales de advertencia: mucha información operacional está en forma de imágenes o texto. ***Posible solución***: almacene la información de texto e imágenes electrónicamente.
Problema con la posible solución: el almacenamiento electrónico puede ser costoso por el lado de la entrada con limitados beneficios por parte de la salida.
Solución Real: evaluar los beneficios del almacenamiento electrónico y comparar los costos de entrada y el almacenamiento de información. Establecer descriptores que faciliten las búsquedas.

Grandes volúmenes de información dificultan el acceso a la información en un tiempo razonable

Ejemplo # 4: el análisis de las tendencias de múltiples años necesitan alrededor de 4 Gb. de información por cada año y necesitan analizar varios miles de registros entre millones.
Dimensiones afectadas: representación concisa, valor agregado, oportunidad y accesibilidad.
Efectos en la organización: se requiere de tiempo excesivo para extraer y sumarizar información.
Señales de advertencia: cantidades grandes de información operacional con necesidad de análisis estratégico o ejecutivo.
Posible solución: condense información utilizando códigos; cree subconjuntos de información extraída según sea necesario.

Problema con la posible solución: dificultades para los consumidores de información en la interpretación del código; no oportuno para análisis.
Solución Real: analizar las necesidades de información y desarrollar regularmente subconjuntos de esta información. Establecer bases de datos multidimensionales y procesamiento analítico en línea (OLAP)

3.3 Fallas de la información utilizada

El fácil acceso de la información puede estar en conflicto con los requerimientos de seguridad, privacidad y confidencialidad

Ejemplo # 5: la información de clientes bancarios debe ser segura y confidencial, sin embargo, analistas e investigadores pueden necesitar acceso a ésta.
Dimensiones afectadas: seguridad, accesibilidad y valor agregado.
Efectos en la organización: mecanismos para acceder la seguridad, de esta manera la información provee menos valor.
Señales de advertencia: la información vital no es accesible, con restricciones razonables.
Posible solución: solución local para violaciones de seguridad y demanda de accesibilidad conforme estas ocurran.
Problema con la posible solución: mezca de soluciones para una situación única, con incrementos de tiempo para resolver la accesibilidad.
Solución Real: desarrolle políticas consistentes y procedimientos para asegurar la información.

La falta de recursos informáticos limita el acceso

Ejemplo # 6: Líneas de comunicaciones no fiables llevan a información incompleta. La escasez de terminales reduce el valor de la información.
Dimensiones afectadas: accesibilidad y valor agregado.
Efectos en la organización: carencia de recurso computacional limita la información de calidad.
Señales de advertencia: los usuarios reclaman más recursos y su fiabilidad.
Posible solución: proveer más recurso computacional con base en la demanda de los usuarios o tener usuarios que presten su propio recurso computacional.
Problema con la posible solución: la localización de recurso computacional llega a ser un proceso político que carece de una base racional.
Solución Real: desarrolle políticas de actualización tecnológica, de tal manera que los usuarios conozcan que se esperan más recursos. Use clientes "delgados" para presentación en una arquitectura de múltiples capas y/o interfaces basadas en tecnología Web (ejecutadas por navegadores como Netscape o Internet Explorer).

3.4 Impacto de la calidad de los datos

En las organizaciones regularmente se presentan muchos problemas, pero principalmente para los altos ejecutivos, quienes frecuentemente están abrumados por situaciones como: baja satisfacción del cliente, altos costos, falta de elementos para la toma de decisiones, entre otros. El impacto que puede tener una pobre calidad de los datos es considerable si lo enfocamos desde los problemas antes mencionados.

Según Thomas Redman [Redman 1998], los problemas asociados a la calidad de los datos podemos categorizarlos como:

- Problemas asociados con la "vista" de datos (los modelos del mundo real capturados en los datos), tales como relevancia y la granularidad.
- Problemas asociados con el valor de los datos, tales como consistencia, exactitud, vigencia e integridad.
- Problemas asociados con la presentación de los datos, tales como la facilidad de interpretación y formatos apropiados.
- Falta de capacidad para generar información correcta: sin los datos apropiados, los usuarios no pueden conocer información segura, ya existente en las bases de datos.
- Pérdida de productividad: los usuarios no tienen la capacidad para generar información correcta.
- Desempeño impredecible: la falta de información en consultas estimadas, deteriora la capacidad del planificador para el funcionamiento de grandes consultas. Esto puede dar como resultado un desempeño impredecible.
- Dificultad para hacer actualizaciones y dar mantenimiento: no es fácil para los diseñadores de sistemas brindar un buen mantenimiento o actualización, si no se conoce cuál es la versión actual del sistema.
- Otros problemas como la privacidad, seguridad y la propiedad.

Estos problemas están relacionados estrechamente con las dimensiones que se mencionaron anteriormente. Sin embargo, la ciencia de la calidad de los datos todavía no ha avanzado hasta el punto donde existan medidas y métodos estándares para abordar estos problemas. Por otra parte, pocas son las empresas que rutinariamente miden la calidad de los datos. Sin embargo, muchos casos de estudio presentan medidas de exactitud. Medido al nivel de campos, el rango de errores varía ampliamente de 0.5% a 30%. Naturalmente, hay dificultades al comparar estos porcentajes de errores.

El impacto que puede presentar la calidad de los datos se puede esquematizar en tres niveles:

- **Nivel operacional**: la pobre calidad de los datos en el nivel operacional puede llevar directamente al descontento e insatisfacción del cliente, al aumento de los costos y al bajo rendimiento de los empleados, entre otros. Por ejemplo, el costo operacional se puede incrementar debido a que el tiempo de otros recursos es utilizado para detectar y corregir errores. En este sentido, los clientes tienen el derecho de suponer que los nombres y direcciones deben estar correctas, las órdenes de los productos o servicios se entregarán correctamente, que se cargarán en sus cuentas apropiadamente. Pero se dan simples errores, como recibir a los clientes en ubicaciones incorrectas, por ejemplo huéspedes que se envían a una habitación pequeña en lugar de una grande y además son forzados a esperar un tiempo

mientras se resuelven los errores que fueron cargados a su cuenta.

- **Nivel táctico**: cuando se tiene pobre calidad de los datos, ésta puede repercutir en forma significativa si se desea implementar un proyecto de sistemas o hacer una reingeniería, además puede provocar un incremento en la desconfianza organizacional. Primero, una pobre calidad de los datos compromete la toma de decisiones. Es comúnmente aceptado que las decisiones por tomar no son mejores que los datos en que éstas se basan. Segundo, una pobre calidad de los datos puede ocasionar dificultad al realizar una reingeniería. Una manera de ver los proyectos de reingeniería es que ellos apuntan a tener los datos correctos, en el sitio correcto, en el tiempo correcto, para servir mejor a los clientes. Finalmente, así como datos pobres en calidad pueden decrementar la satisfacción del trabajo de los empleados, también puede incrementar la desconfianza interna en la organización.
- **Nivel estratégico**: en este nivel el tener una pobre calidad de los datos puede repercutir en el proceso de toma de decisiones. Existe una menor evidencia directa del impacto de la pobre calidad de los datos sobre el nivel estratégico. Sin embargo, las consecuencias que provienen de los aspectos tácticos y operacionales se están tornando cada vez más claros. Primero, ya que la selección de una estrategia es en sí misma es un proceso de toma de decisión, se debería esperar que la estrategia seleccionada sea afectada adversamente. La carencia de datos relevantes, exactos, oportunos y completos acerca de los clientes, competidores, tecnologías y otras características importantes podrían ser un impedimento para desarrollar una estrategia sólida. Segundo, conforme se desarrollen estrategias y se implementen planes específicos, así éstos van siendo modificados a partir de los resultados que se obtienen. Si los resultados reportados son inexactos, tardíos o en alguna otra forma de pobre calidad, la ejecución de la estrategia es mucho más difícil.

Ese impacto no sólo se dará en función de los niveles organizacionales, sino también a nivel de las tecnologías de manejo de datos. En el siguiente punto se describen el impacto de la calidad de los datos frente a estas tecnologías.

3.5 La calidad de los datos frente a las tecnologías de manejo de datos

3.5.1 Sistemas de bases de datos relacionales (SBDR)

En los diseños tradicionales de bases de datos los aspectos de calidad de datos no son explícitamente incorporados. El diseño conceptual típicamente se formula en términos de entidades y relaciones. A pesar de que no cambia la importancia y el valor de los datos respecto de la aplicación para la cual estos fueron diseñados, posteriormente estos datos son procesados junto con otros datos y al pasar el tiempo siguen siendo usados por usuarios no familiarizados con ellos, por consiguiente debería darse una atención más explicita a calidad de los datos.

En general, diferentes usuarios tienen diferentes requerimientos de calidad de datos y diferentes datos son de diferentes calidades. Sin embargo los aspectos fundamentales de la calidad deben mantenerse, independientemente del tipo de usuario que sea.

3.5.2 Depósitos de datos

Ahora que las grandes organizaciones han empezado a crear depósitos de datos (*data*

warehouses)[1] para el soporte a la toma de decisiones, los problemas de la calidad de los datos han resultado ser lastimosamente claros y evidentes. Estas organizaciones han descubierto que la calidad de los datos en bases de datos heredadas (legadas) es su problema más grande. Ken Orr [Orr 1998], señaló que un administrador de datos de una gran compañía reportó que un total el 60% de los datos transferidos hacia su depósito de datos fallaron al pasarlos por las reglas del negocio. Sin embargo, los operadores del sistema habían dicho que los datos estaban en vigencia, situación que puedo haberse previsto basado en la pobre utilidad de los datos.

El extremo superior del desarrollo de depósitos de datos representa un salto cuantitativo en términos del uso de datos por usuarios finales. La ventaja de crear depósitos de datos es que las personas utilizan los datos de una manera más estricta. La necesidad de una mejor calidad de los datos ha contribuido a que se preste más atención a la administración del estado de estos en bases de datos y depósitos de datos.

3.5.3 Minería de datos

Los negocios actuales están experimentando cambios a un ritmo cada vez más acelerado. Estos impactan en las decisiones y las estrategias de aquellas empresas que pretenden ser dominantes, aumentar su participación en el mercado e incluso mantener su competitividad. Hoy en día la cantidad de información recolectada y almacenada en las organizaciones presenta volúmenes muy elevados. Esta situación genera modificaciones en las estructuras, así como los procesos, con los que se pretende aumentar la competitividad. Para lograr estos cambios se hace necesario un nivel diferente de acceso para la información que fluye por toda la organización. El análisis de enormes cantidades de datos repercute en el incremento del tiempo de espera.

A pesar de que existen herramientas para análisis estadístico y para administración de archivos, esto ya no es suficiente, debido a la gran cantidad de información corporativa que existe en las organizaciones. Las técnicas "ad-hoc" son poco adecuadas en la revisión de grandes cantidades de datos. Por ello, se están implementado métodos y técnicas, como la minería de datos, para convertir la información corporativa en una ventaja competitiva en los negocios, desde el área financiera hasta el área de comunicaciones. La minería de datos o descubrimiento del conocimiento, consiste en extraer de los datos información implícita, no trivial, potencialmente útil, que no se conocía previamente para descubrir "relaciones insospechadas", por ejemplo, entre productos y clientes o patrones de compra de los clientes. La meta es descubrir "revelaciones estratégicas competitivas" para controlar la participación en el mercado y las utilidades.

[1]Repositorio único, completo y consistente de datos, obtenidos de diferentes fuentes – bases de datos, plataformas, archivos – y que están disponibles a los usuarios finales de forma que puedan ser comprendidos y usados en un contexto de empresa. Véase el informe *Depósitos de datos*, de Beatriz Jiménez y Rafael Ávalos, publicado por el Club en noviembre de 1998.

Las principales entradas para el sistema de minería de datos son los datos del depósito de datos. Para poder obtener los resultados esperados, estos datos pasan por un proceso de limpieza y extracción, los datos con fallas obvias deben ser eliminados, así como los registros con errores y datos insignificantes que están fuera de lugar; asimismo, se eliminan todas las inconsistencias y heterogeneidades. Por consiguiente, los datos que se obtienen pasan por un proceso de control de calidad. En este sentido, el concepto de calidad de los datos es ampliamente utilizado en la minería de datos.

3.5.4 Procesamiento analítico en línea

En los noventas, el reto de la tecnología de información ha sido el de procesar bases de datos cada vez más grandes, con un alto nivel de complejidad, sin sacrificar tiempos de respuesta. El procesamiento analítico en línea (OLAP) surge como un proceso para ser usado en el análisis de ventas y mercadotecnia, para elaborar reportes administrativos y consolidaciones, para presupuestación y planeación, para análisis de rentabilidad, reportes de calidad y otras aplicaciones que requieren datos de calidad. OLAP provee los reportes y gráficos sumarios que los ejecutivos requieren para tomar decisiones, así como la facilidad de elaborar cálculos complejos, enfoques a detalles operativos y consultas no programadas. Se alimenta principalmente de los sistemas transaccionales y debe considerar una eficiente administración de la base de datos y proveer un nivel adecuado de seguridad. Es importante señalar que todas las dimensiones de la calidad de los datos desempeñan un papel de gran importancia en las aplicaciones OLAP.

4 Aseguramiento de la calidad de los datos

El aseguramiento de la calidad de los datos no es una labor fácil. Muy por el contrario, conlleva una serie de elementos que deben ser tomados en cuenta al determinar cuál es el método o plan que se vaya a seguir. Debido principalmente a la importancia de asegurar la calidad de los datos, en este apartado se plantean las dificultades que se pueden encontrar al intentar asegurar la calidad, se resalta la importancia de contar con un programa de calidad de datos, se describe un proceso para estimar la calidad de éstos, y finalmente se presenta un proceso de calidad basado en usos.

4.1 Dificultades para asegurar la calidad de los datos

Uno puede argumentar que asegurar la calidad de los datos es mucho más difícil que la fabricación de un producto. La materia prima (los datos) puede ser de calidad incierta y el uso puede ser sólo parcialmente conocido. La efectividad de los posibles procedimientos de control de calidad son inciertos, si los datos experimentan una serie de procesos "ad hoc". Esto es posible técnicamente al combinar colecciones de datos que nunca se pensó anticipadamente en combinar.

Otro problema es la baja prioridad que a menudo se le asigna a la calidad de los datos. De alguna manera la garantía en la calidad de los datos es similar a la seguridad en los ambientes computacionales. Casi todo el mundo está de acuerdo en que la seguridad en tales ambientes es una actividad muy importante. Sin embargo, y a pesar de que asegurar la calidad de los datos es una actividad importante y valiosa, en la práctica son pocas las personas que listan esto como una alta prioridad.

El problema de asegurar la calidad de los datos es estimulado por la multiplicidad de problemas potenciales con los datos. Hay muchas maneras en que los datos pueden ser erróneos. Por ejemplo, los datos sobre intercambio de las divisas extranjeras en los periódicos de hoy están al día; sin embargo, a pesar de ello estaban desactualizadas antes de que el periódico se imprimiera. Para un archivo en particular cada valor de los datos puede ser exacto y oportuno, pero algunos registros (filas) pueden haberse perdido en su totalidad. Nadie puede anticipar todas las circunstancias que podrían comprometer la integridad de los datos en la organización.

Muchas veces es difícil determinar qué tan serias son las deficiencias con los datos. Por ejemplo, algunas personas suelen utilizar diferentes códigos alfanuméricos en diferentes divisiones de la organización para representar el mismo dato. Esto probablemente no cree alguna dificultad, ya que si estos datos permanecen aislados en la división de origen y nunca son combinados por otras unidades organizacionales, entonces no hay necesidad de cambiarlos. Sin embargo, si los datos están disponibles para otras divisiones organizacionales, como puede ser el caso de un *data mart* (mercado de datos), entonces se necesita resolver las inconsistencias. Un problema similar es la determinación de la naturaleza de los datos deficientes. Esto es cierto especialmente en ambientes multiusarios, por ejemplo para usuarios que tengan requerimientos diferentes de calidad de los datos. Un primer paso para entender cómo los datos pueden tener deficiencias, es reconocer en realidad que los datos tienen múltiples atributos o dimensiones así como múltiples usos.

Por otro lado, es importante, pero a menudo difícil, determinar el nivel apropiado de la calidad de los datos. Aunque uno pueda desear que todos los datos de la organización sean correctos por todas partes, alcanzar esto podría hacer fracasar la organización. Determinar la necesidad de calidad es difícil, sobre todo cuando diferentes usuarios tienen diferentes necesidades. Uno podría ser tentado a establecer que el uso de los datos requiere la más alta calidad, lo cual debería ser determinado en todos los niveles de la organización. ¿Pero qué tal si el uso de los datos es mucho menor y sin importancia para la organización, mientras que el mayor uso de los datos no requiere de tal nivel de calidad? Es necesario balancear el conflicto de requerimientos para la calidad de los datos, es decir, analizar los costos y beneficios.

4.2 Desarrollo de un programa de calidad de datos

Un programa de calidad de datos debe iniciar tan pronto como se detecten fallas o errores que puedan obstaculizar el desempeño de las labores de la organización. El éxito del programa depende de la participación de los usuarios expertos, personal administrativo y del grupo de calidad.

El programa de calidad de datos debe incluir:

- Eliminar datos redundantes
- Validar datos de entrada
- Eliminar valores nulos
- Resolver conflictos de datos
- Establecer y aplicar estándares
- Asegurar una definición apropiada y el uso del valor de los datos
- Asegurar que los valores de los datos está dentro de los dominios que se definieron.

Para cada área problema, los datos problema son identificados en detalle. Una parte importante del proceso identifica exactamente dónde en el ciclo de vida se origina el problema. A menudo, el problema de la calidad de datos requiere de dos esfuerzos separados: un proyecto para corregir los datos que ya existen y un proyecto para corregir la causa detrás de los datos problema.

En cada área seleccionada puede existir un conjunto de problemas comunes, para los cuales se especifica un conjunto de métodos que pueden ser aplicados. Primero se definen los temas y los objetivos de cada área. Se debe realizar un análisis de costo/beneficio para determinar si el proyecto es factible. De ser factible se deben asignar los recursos para realizar el proyecto. En la etapa de análisis se deben identificar los temas que tienen prioridad, además se debe revisar la documentación (políticas, regulaciones, procesos, etc.). Asimismo, se deben desarrollar detalladamente parámetros de calidad. Finalmente, se debe seleccionar la solución más apropiada y definir cursos de acción. De ser posible, adquirir herramientas que puedan contribuir a mejorar y mejorar la calidad de los datos (ver apéndice).

En la figura siguiente se presenta un bosquejo de una estrategia para mejorar la calidad de los datos. La estrategia inicia con la definición del problema, seguido por los pasos de identificación de los datos problema, el análisis y la solución (mejoramiento) para cada problema.

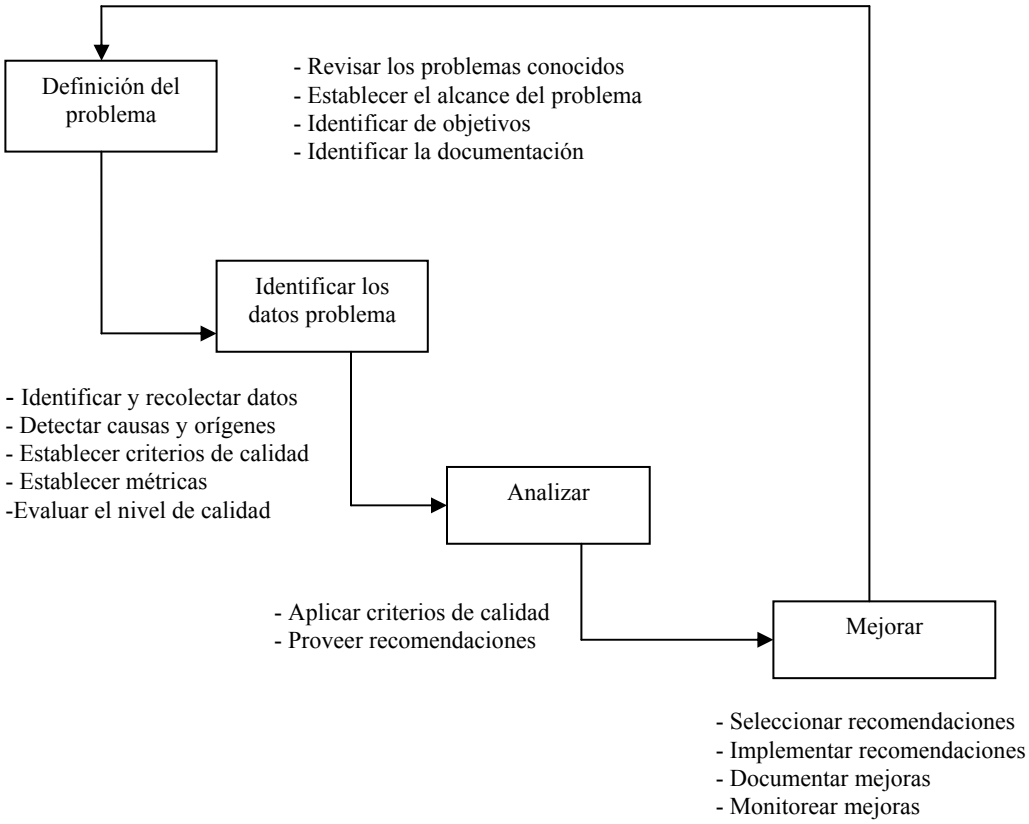

Figura No.1: Estrategia para la calidad de los datos

4.3 Aseguramiento de la calidad de los datos: un enfoque basado en riesgos

Por lo general, el riesgo de la calidad de los datos se identifica tempranamente en el desarrollo de sistemas, pero el enfoque también puede ser aplicado en las situaciones de post implementación. Duane Hufford [Hufford 1996a] presenta un enfoque basado en riesgos para asegurar que los datos en el sistema soporten la toma decisiones "basadas en la realidad". Hufford recomienda incorporar el aseguramiento de la calidad de los datos en el desarrollo del sistema, lo cual se logra a través de cuatro actividades claves:

- Definir las expectativas y métricas de los usuarios respecto de los datos. Las expectativas son definidas utilizando métricas de calidad de los datos, las cuales miden las características apropiadas de los datos (por ejemplo, consistencia, legibilidad, integridad, exactitud, oportunidad, etc.) para cada uso.
- Definir el riesgo en términos de qué puede causar la falta de calidad de los datos para las expectativas que se plantean y de la iniciativa de acciones o proyectos que puedan minimizar el riesgo. Los costos pueden ser expresados en términos de la pérdida de financiamiento, de producción, de ventajas o de lealtad de clientes. El riesgo es identificado durante todo el ciclo de vida de la administración de los datos por el sistema. Las posibilidades de disminución del riesgo varían y principalmente dependen de la fase del ciclo de vida en que se identifica el riesgo. El proceso de mejoramiento de la calidad de datos debe realizarse después de introducir y actualizar los datos fuentes (por ejemplo, en los sistemas operacionales).

- Evaluar el riesgo de la calidad de los datos. Implementar mecanismos apropiados de detección y reporte de datos defectuosos, que permitan ayudar a clarificar los problemas de la calidad y decidir cómo disminuir los riesgos.
- Incorporar validación, verificación y certificación como una práctica integral para detectar los problemas potenciales de calidad de datos. El objetivo es integrar el aseguramiento de la calidad de los datos en los procedimientos de adquisición, mantenimiento y archivo de datos. Esto puede dar mejores resultados que tratarlo como una actividad separada. Los resultados deben ser revisados y evaluados para determinar cómo las prácticas de administración de la calidad de los datos pueden mejorar.

4.4 Proceso para estimar la calidad de los datos

Actualmente, no se han establecido métodos estándar para darle solución a los problemas de la calidad de los datos. Sin embargo, Larry English [English 1998] plantea un conjunto de fases para conducir un proceso de estimación de la calidad de los datos, los cuales fundamentan su metodología TQDM (Total Quality Data Management). English enfatiza que TQDM es un proceso continuo de mejoramiento de los procesos de definición de los datos y los procesos de la organización que crean, actualizan, eliminan o presentan información, al integrar, dentro de la cultura organizacional, un conjunto de principios, creencias y métodos de calidad total.

A continuación presentamos las fases descritas por English:

1. Establecer objetivos. Usualmente los objetivos se enfocarán en áreas, procesos y variables de la organización que exhiben problemas en cuanto a calidad de datos. La calidad de datos comprende: calidad en cuanto a la definición, contenido y presentación de los datos.
2. Definición de las métricas de calidad de los datos. Por ejemplo: ¿qué componentes o partes de la antigua aplicación son lo suficientemente portátiles y reutilizables para tomarlos para el nuevo sistema? Es importante señalar que un programa de métricas debe ser gradual.
3. Establecer el valor de la información y sus costos. Usualmente esto involucra cuantificar cuánto estamos perdiendo por la mala calidad de los datos, cuánto estamos dejando de hacer por la mala calidad de los datos, o de qué manera nos beneficiamos si nuestra información es altamente confiable.
4. Analizar la calidad de la definición de los datos, de acuerdo con su claridad, completitud, precisión y nivel de importancia para los trabajadores del conocimiento. Revisar la completitud y precisión en las reglas de negocio. Actualmente, existen herramientas que apoyan el análisis de calidad de los datos (ver apéndice).
5. Analizar la calidad del contenido de los datos. Antes se hacen estudios de valoración. Para analizar la calidad del contenido de los datos se debe seguir los siguientes pasos:
   - Se identifica sobre qué archivos, bases de datos, procesos, se va a realizar el estudio de valoración.
   - Se definen las pruebas que se van a implementar sobre los datos existentes, considerando los objetivos propuestos del conjunto de datos y los procesos identificados.
   - Se elaboran los flujos detallados de información para los datos y procesos que se quieren analizar. En este paso se identifican las bases de datos, archivos, elementos de datos y relaciones entre archivos. Los catálogos de datos y/o archivos ofrecen una fuente valiosa de información para esta actividad.
   - Se extrae una muestra aleatoria de los datos que refleje con alto nivel de precisión la

calidad de la población.
- Se cuantifica la calidad de los datos para la muestra generada en el paso anterior, en términos de integridad, precisión, duplicación de llaves, validez en los rangos de los dominios de definición, correspondencia con las reglas de negocio, consistencia en la derivación.
- Finalmente, el nivel de calidad de los datos es reportado en términos de tipos de errores y calidad observada. Se analizan los valores obtenidos con respecto de niveles de tolerancia y se analizan las causas de estos errores.

6. Implementación de acciones para la limpieza de los datos. El apoyo de las unidades del negocio es clave.
7. Implementación de mejoras para la calidad de la información. Esto involucra mejorar los sistemas y procesos de información y las estructuras de los datos. El apoyo de la alta gerencia de la organización y de tecnologías de información es clave para obtener los resultados necesarios.
8. Continuar el proceso de mejoramiento de la calidad de la información.

4.5 Programas de calidad de datos basados en uso

Si la calidad de los datos es una función de su uso (como se explicó en el capítulo uno), hay una sola forma segura de obtener la calidad de los datos: mejorar su uso. Según Ken Orr [Orr 1998] los programas de calidad de datos basados en uso se construyen alrededor de resultados innovadores y de formas sistemáticas para asegurar que se da uso a los datos críticos. A continuación describimos cuatro programas de calidad basada en uso.

4.5.1 Auditoría de calidad de datos basada en uso

Para mejorar la calidad de los datos, es necesario poder determinar cuán buenos son los datos que están en la base de datos. Este tipo de programa involucra responder un conjunto de preguntas importantes, tales como:
- *¿En cuántos datos estamos interesados?*
- *¿Cuál es el diseño de los datos?*
- *¿Cuál es el modelo de datos?*
- *¿Cómo es el uso de los datos?*
- *¿Con qué propósito se utilizan los datos?*
- *¿Qué tan a menudo se utilizan los datos?*

En la mayoría de los casos, la auditoría de la calidad de los datos da mejores resultados cuando se utilizan muestras estadísticas. Raramente es buena idea verificar todos los datos de una base de datos real; es necesario crear una muestra suficientemente grande que nos permita obtener conclusiones significativas.

4.5.2 Rediseño de calidad de datos basado en uso

Para mejorar la calidad de los datos, es necesario mejorar la unión entre los varios usos de los datos en todo el sistema. Uno de los problemas es decidir por dónde empezar. Por lo general, entre más ambientes de tecnología de información existen, estos contienen cientos de archivos (tablas) y miles de elementos de datos y comúnmente todos estos datos no son igualmente

importantes. En muchos sistemas, hay pocos conjuntos de datos críticos que hacen la diferencia. A menudo, "el consumidor", "el producto", "la orden" ("el pedido") y/o la "estructura organizacional" son las categorías de datos más importantes para una empresa.

El primer paso en los programas de rediseño de la calidad de los datos es identificar las áreas críticas de datos. Esto involucra una cuidadosa revisión de cómo se utilizan las piezas críticas de los datos. Normalmente esto se manifiesta en dos áreas: los procesos de negocio básicos (por ejemplo, satisfacer una orden de entrega) y el apoyo a las decisiones. El diseño basado en usos enfoca la forma en cómo exactamente se utilizarán los datos y en tratar de identificar maneras para asegurar que los datos sean utilizados más activamente. En muchos casos, estas medidas creativas persuaden a las personas que tienen más conocimiento de los datos para que tomen responsabilidad sobre éstos.

4.5.3 Aprendizaje de calidad de datos basado en uso

Uno de los mayores problemas con la calidad de los datos es hacer entender a usuarios y gerentes los principios de la calidad de los datos. En cualquier programa de calidad de los datos, se debe dedicar una cantidad significativa de tiempo a la educación y la capacitación. Es muy difícil convencer a los usuarios y administradores, quienes han contribuido al levantamiento de requerimientos de los datos, sin pensar acerca de su uso, de que mucho de lo que ellos recolectaron nunca será correcto. Afortunadamente, las personas llegan a hacer conciencia en un tiempo relativamente corto, cuando ven los efectos de un programa de calidad de datos.

4.5.4 Medición continua de calidad de datos basada en uso

La calidad de los datos requiere una constante medición para asegurar que las prácticas basadas en uso son seguidas durante todos los procesos. Por lo general, los problemas de la calidad son problemas de los sistemas, no problemas de los trabajadores. Sin embargo, los errores individuales también contribuyen a la pobre calidad de los datos. Las mediciones y los programas de calidad deben ir de la mano. Todas las preguntas que son formuladas en la auditoría de la calidad de los datos necesitan ser repetidas regularmente para el rediseño del sistema.

Una nota respecto de las mediciones: no se convenza de mediciones internas si las verificaciones externas no son adecuadas. Toda medida interior puede asegurar que los datos son consistentes internamente. Por ejemplo, ninguna organización grande puede contar con registros de inventario confiables, si no realiza periódicos inventarios físicos; un archivo puede mostrar que existen 23 computadoras, pero eso no significa que en los estantes de un almacén X existan realmente 23 computadoras. Si queremos que los datos que residen en nuestras bases de datos estén de acuerdo con el mundo real, debemos verificar periódicamente que esas computadoras existen realmente donde nuestro sistema dice que existen y debemos además tomar acciones para reconciliar cualquier tipo de diferencias. Los datos que son realmente vitales deben de ser físicamente auditados.

5 Recomendaciones

Existen varios métodos para mejorar la calidad de los datos. Estos métodos son agrupados en cuatro categorías interrelacionadas, las cuales se mencionan a continuación:

- Rediseño del negocio: el rediseño del negocio intenta obtener un conjunto de procesos y operaciones simple y moderno que minimicen la oportunidad de que ocurran errores en los datos.
- Motivación para mejorar la calidad de los datos: la motivación para mejorar la calidad de los datos intenta brindar estímulos y beneficios a los empleados, para fomentar una cuidadosa atención que contribuya a mejorar la calidad de los datos y que sean manejados por los miembros más indicados de la organización.
- Uso de nuevas tecnologías: las nuevas tecnologías para captura de datos, pueden mejorar significativamente la calidad a través de técnicas, tales como la automatización de entradas y la comunicación directa entre computadoras.
- Uso de tecnologías de interpretación de datos: las tecnologías de interpretación de datos asisten a los usuarios en entender el significado de los datos para que estos no se utilicen incorrectamente.

5.1 El factor humano y la calidad

Hacer las cosas bien depende de las personas. Es un hecho y una realidad constatable que el éxito de las empresas depende de las personas que las forman. No solamente desde el punto de vista empresarial, sino desde el punto de vista personal, es de trascendental importancia conseguir que los empleados, trabajadores y seres humanos, se sepan y se sientan realizados, seguros de sí mismos e integrados en su entorno y en su propia individualidad. Este aspecto indudablemente nos lleva a crear un ambiente propicio para el desempeño apropiado de las labores. Además de que se crea un entorno propicio para desarrollar una cultura de calidad de los datos.

5.2 Desarrollo de una cultura de calidad de datos

Juan Uría [Uría 1997] plantea que las organizaciones que han abordado exitosamente procesos de mejoramiento de la calidad están mejor preparadas para cambios radicales, tales como innovación del negocio habilitada por la tecnología y su desplazamiento hacia nuevas formas de organización.

Las características de un marco efectivo de implementación de un programa de calidad son:
- Holístico: contempla la gente, los procesos y la tecnología.
- Considera la cultura y el clima de tecnología de información existente.
- Hace conciencia respecto de la complejidad inherente de cualquier programa de mejoramiento de la calidad.

El desarrollo de una cultura de calidad de datos demanda una adecuada gerencia del cambio. Para gerenciar el cambio, se deben considerar los siguientes elementos: contexto del negocio, infraestructura e implementación. Por otro lado, los parámetros de comparación permiten que las organizaciones se impongan a sí mismas nuevas metas de desempeño realistas y rigurosas; además, el proceso contribuye a que la gente se convenza de la credibilidad de tales metas. Esto tiende a superar el síndrome de "eso no fue inventado aquí" y la justificación de "nosotros somos diferentes", que se esgrime para mantener el nivel presente. Para cada elemento se presentan a continuación los factores más relevantes.

5.2.1 Contexto del negocio

- Cuáles son las estrategias de TI.
- Alineación de las estrategias de TI con las estrategias del negocio.
- Qué estrategia de TI impulsa la iniciativa de la calidad de datos.
- Cuáles son los costos o implicaciones de no implementar un proceso de mejoramiento de calidad de datos.

5.2.2 Infraestructura

- Procesos de TI, diseño organizacional y estructura.
- Infraestructura tecnológica para soportar la calidad de los datos.
- Gerencia y medidas de desempeño.
- Métricas organizacionales como expectativas para un proceso exitoso de mejoramiento de los datos.

5.2.3 Implementación

- Liderazgo e influencia. Motivar la dinámica organizacional hacia el logro.
- Comunicación. Diálogo persuasivo, retroalimentación continua y verificación acerca de los resultados que se van logrando progresivamente.
- Desarrollo de capacidades. Para lograr el éxito en un proceso de mejoramiento de la calidad de los datos, los elementos considerados en cada dimensión deben ser integrados en un conjunto de acciones, estrategias y esfuerzos gerenciales, dentro de un proceso continuo de retroalimentación y análisis de resultados que gradualmente se van obteniendo para: entender en dónde estamos, comprender el contexto del negocio, conocer nuestras capacidades y estimar cuán capaces somos para el cambio.

Apéndice: Recursos

Herramientas

En la siguiente tabla se presenta información sobre los diferentes productos que actualmente existen en el mercado para la calidad de los datos. En esta tabla se presentará información referente al producto, vendedor, tipo de producto, plataformas en las que operan y una breve descripción de los mismos. Para un mejor entendimiento hemos seleccionado los productos por tipo y los hemos descrito con las siguientes siglas:

A = Auditoría, M = Metadatos, R = Descubrimiento de reglas, PS (S) = Proveedor de Servicio (solamente), L= Limpieza, LG = Limpieza General, P = Prevención de errores

Vendedor del Producto	Nombre del Producto	Tipo	Plataformas en las que opera	Descripción
Ardent Software	DataStage	LG	Win95/NT	Máquina de transformación de datos Cliente/Servidor: • Extrae datos de una variedad de fuentes. • Transforma bibliotecas de partes integrales • Incluye correspondencia de patrones • Conversión de formatos y conversiones aritméticas
Carleton Corp.	Enterprise Integrator Passport	LG	Unix, Win95/NT, C++, HP 9000, OS/2	Enterprise Integrator: • Utiliza una "máquina de transformación" comprensiva para transformar datos que involucran múltiples fuentes y bases de datos. Passport: • Herramienta dirigida por metadatos, la cual permite extraer, buscar, equiparar, unir y verificar información de una o varias bases de datos fuentes. • Define y aplica reglas técnicas y de negocios.

Century Analysis	CAI Integration Toolset: TDM Interface Engine	L	Win/NT	• Sincroniza y transforma datos a través de las aplicaciones. • Captura y filtra transacciones y las explota en una o más transacciones de salida que pueden ser procesadas por las aplicaciones.
Compedia	SA NameCop, EnComp	A, M	Win95/NT, DOS	NameCop: • Chequea el nombre de los diagramas ER, entidades, atributos en las arquitecturas de sistemas, conforme se renombran. EnComp: • Compara el contenido de dos repositorios para dar las diferencias.
DataFlux Corp.	Inspector, IntelliMerge, Datalogue	LG, A, R	Win 95/NT	Inspector: • Examina archivos de datos de varios tipos de registros duplicados. IntelliMerge: • Automatiza el proceso de eliminar registros identificados como duplicados Datalogue: • Asiste con la determinación de mejores estandarizaciones, eliminación de redundancia y corrección de errores.
Evoke Software	Migration Architect	R	Sun solaris 2.4+	• Analiza datos heredados de fuentes heterogéneas que han estado formateados para análisis, descubrimiento de reglas. • Deriva una tercera forma normal del modelo relacional para el análisis de datos.

Integral Solutions	Clementine	R, LG	Unix, Sun Solaris, HP, Silicon Graphics, RS 6000, Data General, NCR, Tandem, Win NT, ODBC Compliant	• Permite la extracción y descubrimiento de reglas utilizando redes neuronales y reglas de inducción. • Identifica relaciones entre los datos y las reglas generadas.
Platinum Technology	InfoRefiner, InfoPump, InfoTransport, InfoHub, DecisionBase	LG	DB2 en Mainframe, Unix, OS/2, Win3/95/NT	InfoRefiner: • Mapea y extrae datos heredados de múltiples fuentes mainframe. • Refina datos e integra éstos. InfoTransport: • Mueve datos a través de diversas plataformas. • Lleva a cabo conversiones de tipos de datos en el servidor y los mueve a las bases de datos. InfoPump: • Provee replicación bi-direccional de datos entre bases de datos diferentes. • Replica datos en el servidor para el acceso de usuarios finales.
Prism Solutions	Prism Warehouse Executive; Prism Quality Manager (Anteriormente QDB Analyze)	LG, A	Win3.x/95/NT, DOS 5.0; Prism Quality Manager Accesa ODBC	Prism Quality Manager: • Detecta y mide en un amplio rango los problemas de la calidad de los datos, incluyendo valores inválidos, datos incompletos, carencia de integridad referencial y duplicación de registros. Prism Warehouse: • Genera programas para extraer datos operacionales y datos externos de bases de datos fuentes. • Integra los datos de varias fuentes.
SAS Institute Inc.	SAS Warehouse	LG	Win 95/NT	• Permite crear metadatos para

	Administrator: Transformation Engine			definir reglas de negocios utilizando lenguaje SAS. • Transforma y limpia los datos.
Trillium Software	Trillium Software System	LG, R, PS, P	IBM Mainframe, Unix, DEC, HP, AS/400, Tandem	• Integra múltiples bases de datos heterogéneas en una base de datos única.
Unitech Systems	ACR/Plus	A	MVS, Unix, AS/400, PC	• Monitorea los datos durante todo el ciclo de vida, incluyendo la detección de errores en los datos heredados
Vality Technology	Integrity Data Re-engineering System	LG, R	MVS, Unix	• Conjunto de herramientas computacionales para investigar, estandarizar, enriquecer, condicionar, transformar e integrar datos a partir de archivos heredados o fuentes externas.
WizSoft	WizRule, WizWhy	R, A	Win 3.1/95	• Utiliza tecnología de minería de datos basada en un algoritmo propietario para analizar datos de uno o más archivos para descubrir patrones de reglas.

Bibliografía

[Ballou 1995] Ballou, D. et al. "Modeling data and process quality in multi-input, multi-output information systems." Management Science, Vol 31, Nº 2, págs. 150-162, 1995.

[Bertino 1998] Bertino, E. "Data Security." Data & Knowledge Engineering, Vol. 25, Nº (1-2), págs.199-216. Marzo, 1998.

[English 1996] English, L. "Help for Data Quality", http://www.startneting.com/articles, 7 octubre, 1996

[English 1998] English, L. "Plain English on Data Quality: DQ Point 7: Institute Leadership for Data Quality", DM Review Magazine, junio 1998

[Firth 1996] Firth, Ch. "Data quality in practice: experience from the frontline", Citibank Singapore, octubre, 1996

[Haebich 1998] Haebich, W. "Data Quality in the Real World", Database Programming and Design, febrero 1998

[Hearold 1994] Hearold, S. et al. "Beyond Security: A Data Quality Perspective on Defensive Information Warfare" Conferencia de Data Quality Management, 1994

[Hufford 1996a] Hufford, D. "Data Warehouse Quality: Part I", DM Review Magazine http://www.dmreview com/, 1996

[Hufford 1996b] Hufford, D. "Data Warehouse Quality: Part II", DM Review Magazine http://www.dmreview com/, 1996

[Kaplan 1998] Kaplan D. et al. "Assessing Data Quality in Accounting Information Systems", Communications of the ACM, vol 41, Nº 2, febrero 1998

[Kumar 1998] Kumar, G. et al. "Examining Data Quality", Communications of the ACM, vol 41, Nº 2, febrero 1998

[Orman 1994] Orman, L. et al. "Systems Approaches to Improving Data Quality". Conferencia de Data Quality Management, 1994

[Orr 1998] Orr, K. "Data Quality and Systems Theory", Communications of the ACM, vol 41, Nº 2, febrero 1998

[Redman 1998] Redman, T. "The Impact of Poor Data Quality on the Typical Enterprise", Communications of the ACM, vol 41, Nº 2, febrero 1998

[Storey 1994] Storey, V. et al. "Modeling Quality Requirements in Conceptual Database Design". Conferencia de Data Quality Management, 1994

[Strong 1994] Strong. D. et al. "Beyond Accuracy: How Organizations are Redefining Data Quality". Conferencia de Data Quality Management, 1994

[Strong 1997] Strong, D., et al. "10 Potholes in the Road to Information Quality", Computer, vol 30, Nº 8, agosto 1997

[Strong 1998] Strong, D. et al. "Data Quality in Context", Communications of the ACM, vol 40, Nº 5, mayo 1997

[Uría 1997] Uría J. "Resumen Conferencia de Data Quality Management", Banco Banesco, Venezuela, octubre 1997

[Wang 1998a] Wang, R. "A Product Perspective on Total Data Quality Management", Communications of the ACM, vol 41, Nº 2, febrero 1998

[Wang 1998b] Wang, R. et al. "Manage your Information as a Product". Sloan Management Review, verano 1998