I+D Computación

Una publicación de la Academia de Posgrado de Ciencias Computacionales

Marzo de 2003, Vol. 2 No. 1

ISSN 1665-238X

Calidad de los Almacenes de Datos

Manuel Serrano, Ismael Caballero, Coral Calero, Mario Piattini
Grupo de Investigación Alarcos, E.S. Informática
Paseo de la Universidad, 4, 13071 Ciudad Real (España)
{Manuel.Serrano, Ismael.Caballero, Coral.Calero, Mario.Piattini}@uclm.es

Abstract.– The quality of information systems is a very important aspect in the real world, however ensuring quality is crucial when these information systems are used for strategic decision making. In this paper we present several aspects of quality in data warehouses, including data modeling as well as the data itself, and from the user's point of view.

Keywords: data warehouses, quality, information systems.

1. INTRODUCCIÓN

Actualmente las organizaciones pueden almacenar inmensas cantidades de datos obtenidas a un precio relativamente bajo, sin embargo estos datos pueden no proporcionar información [11], porque en algunos casos adolecen de problemas de polución. Esta polución puede llegar a tener graves consecuencias tanto desde el punto de vista técnico –como en la implementación de almacenes de datos [8]–, como organizacional –pérdida de clientes [28], grandes pérdidas financieras[1] [22] o insatisfacción de los trabajadores [10]– e incluso legal –basta recordar el artículo 4 del Título II de la actual LOPD [9]–. Para resolver este problema, las organizaciones están adoptando almacenes de datos, los cuales se definen como *"una colección de datos orientados a temas, integrados y no volátiles, que soportan la gestión de la toma de decisiones"* [16]. Los almacenes de datos se han convertido en la tendencia más importante de la informática, así, en [18] vaticinaban que en la actualidad el mercado de los almacenes de datos alcanzaría los 12 millones de dólares americanos.

Se han propuesto diferentes ciclos de vida y técnicas para el desarrollo de los almacenes de datos [7], [13]; [20]; [21]. Sin embargo, el desarrollo de un almacén de datos es una tarea difícil y llena de riesgos. Es esencial poder asegurar la calidad de la información que contiene el almacén de datos ya que éste se ha convertido en la principal herramienta para la toma de decisiones estratégicas [10].

La calidad de la información viene determinada por la calidad tanto del almacén de datos como por la calidad de la presentación de los datos. De hecho, es muy importante que los datos del almacén reflejen correctamente el mundo real, pero es también muy importante que los datos sean interpretados correctamente. En la calidad del almacén de datos, al igual que una base de datos operacional [25], se deben considerar tres aspectos: la calidad del SGBD (Sistema Gestor de Base de Datos) relacional o multidimensional que lo soporta, la calidad del modelo de datos (tanto conceptual, lógico como físico) y la calidad de los propios datos contenidos en el almacén.

[1] El *Data Warehousing Institute* apunta en febrero de 2002 pérdidas anuales en empresas americanas del orden de seiscientos millones de dólares debido a problemas de calidad de datos. *(Intelligent Enterprise 5(6) – Marzo 2002, pp 12.)*

Artículo recibido el 19 de febrero de 2003.

Como es lógico la calidad del modelo del almacén de datos tiene una gran influencia en la calidad de la información. Modelo que puede existir a nivel conceptual –como en las propuestas de [4], [12], [7] y [31], aunque no suele ser habitual–; a nivel lógico –para lo que se ha universalizado la utilización del "diseño en estrella", que permite buenos tiempos de respuesta y una comprensión fácil de los datos y los metadatos por parte de los usuarios y los desarrolladores [21]–, y a nivel físico –ya que el diseñador tiene que elegir las tablas físicas, los índices y las particiones de datos que mejor representan el almacén de datos lógico y que facilitan su funcionalidad [3], [18]–.

La calidad de los propios datos viene determinada principalmente por los procesos de extracción, filtrado, limpiado, sincronización, agregación y carga [1], [2], [21] que deben tener en cuenta ciertos requisitos de calidad de datos de los usuarios. Este concepto es tan amplio que para un mejor estudio de la calidad de los propios datos es necesario tener en cuenta las llamadas dimensiones de calidad [33]. Estas dimensiones representan los requisitos de calidad de datos de los usuarios. Muchos autores han tratado de definir una colección de dimensiones de calidad que sea estándar para cualquier problema de calidad de datos, pero esa solución es imposible por el alto grado de dependencia que existe con respecto a los requisitos de calidad de datos de los usuarios, que pueden cambiar de unas circunstancias a otras.

En la siguiente sección hablaremos acerca de la calidad de los modelos de datos y en el apartado 3 sobre la calidad de los propios datos. En la sección 4 se abordarán los temas de calidad de los almacenes de datos desde el punto de vista de los usuarios y en la última sección se comentarán algunas conclusiones.

2. CALIDAD DE LOS MODELOS DE DATOS

2.1. Diseño en Estrella

Un diseño multidimensional es un reflejo directo de la manera en que se ven los procesos de negocio. Estos diseños capturan las mediciones de importancia de un negocio y los parámetros por los que se identifican estas mediciones. Las mediciones se denominan *hechos* o *medidas*. Los parámetros por los que un hecho puede ser visto se denominan *dimensiones* [1].

Normalmente los modelos de datos multidimensionales se representan como esquemas en estrella, los cuales consisten en una tabla central y varias tablas de dimensión. Las medidas de interés es almacenan en la tabla de hechos (por ejemplo, ventas, o inventario). Para cada dimensión del modelo multidimensional existe una tabla dimensional (por ejemplo, producto, tiempo) que almacena la información acerca de estas dimensiones [18].

En la figura 1 se puede observar el ejemplo de un esquema multidimensional, en el que tenemos dos tablas de hechos (Production_Facts e Ingredient_Usage) y cinco tablas dimensionales (Production, Ingredient, Facility, Time y Production_Run).

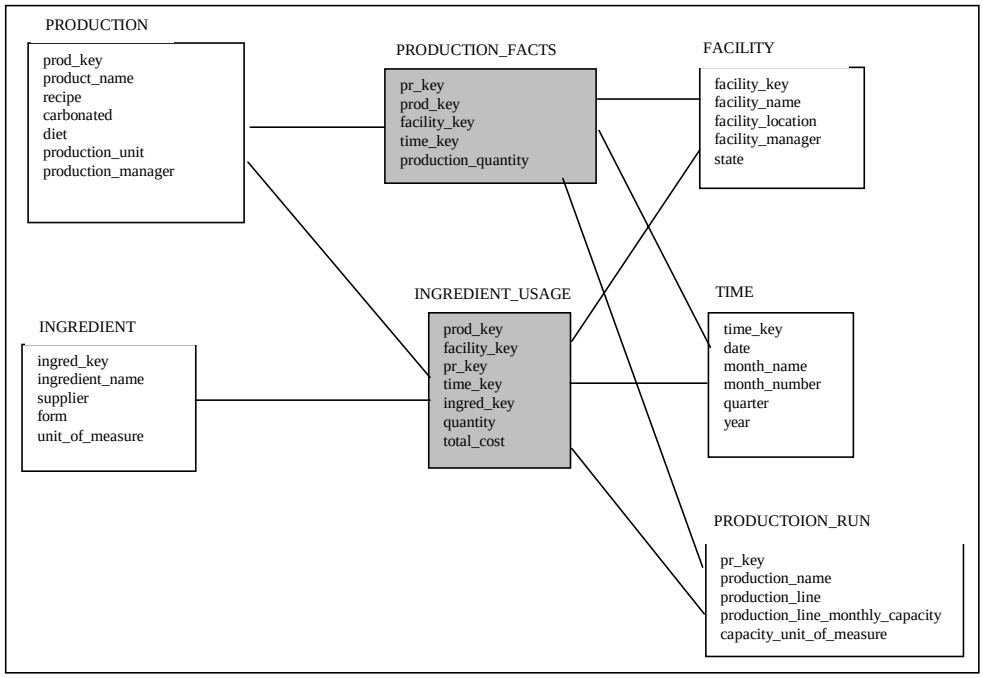

Figura 1. Ejemplo de un almacén de datos [1].

2.2. Métricas

Según la ISO 9126 [17] la calidad del producto está influenciada por la complejidad del mismo. El objetivo de nuestras métricas es poder medir la complejidad del esquema de estrella y así poder tener un indicativo objetivo de la calidad del sistema.

A la hora de definir métricas para modelos de datos de almacenes de datos podemos fijarnos en tres niveles distintos:

- Nivel de tabla.
- Nivel de estrella.
- Nivel de esquema.

Por ello, definimos métricas para estos tres niveles [6].

Métricas a nivel de tabla

En los últimos años se han desarrollado diferentes métricas para asegurar la calidad de las bases de datos relacionales [5]. Dos de estas métricas pueden ser útiles para los almacenes de datos:

Tabla 1. Métricas a nivel de tabla.

Métrica	Descripción
NA(T)	Número de atributos de una tabla
NFK(T)	Número de claves ajenas de una tabla

En la tabla 2 se pueden observar los valores de las métricas del nivel de tabla para el esquema de la figura 1.

Tabla 2. Valores de las métricas a nivel de tabla.

	NA	NFK
PRODUCTION	7	0
INGREDIENT	5	0
FACILITY	5	0
TIME	6	0
PRODUCTION-RUN	5	0
PRODUCTION-FACTS	5	4
INGREDIENT-USAGE	7	5

Métricas a nivel de estrella

A continuación se detallan, en la tabla 3, las métricas propuestas para el nivel de estrella, elemento principal de un almacén de datos.

Tabla 3. Métricas a nivel de estrella.

Métrica	Descripción
NDT(S)	Número de tablas dimensionales de una estrella
NT(S)	Número de tablas de la estrella
NADT(S)	Número de atributos de las tablas dimensionales de una estrella
NAFT(S)	Número de atributos de la tabla de hechos de la estrella
NA(S)	Número de atributos de la estrella.
NFK(S)	Número de claves ajenas de una estrella
RSA(S)	Ratio de atributos de la estrella. Número de atributos de las tablas dimensionales dividido por el número de atributos de las tabla de hechos
RFK(S)	Ratio de claves ajenas. Número de atributos de la tabla de hechos que son claves ajenas

En la tabla 4 se puede observar los valores de las métricas de nivel de estrella para el esquema de la figura 1:

Tabla 4. Valores para las métricas a nivel de estrella.

	Production-Facts	Ingredient-Usage
NA	28	35
NFK	4	5
RSA	23/5	28/7
NDT	4	5
NT	5	6
NADT	23	28
NAFT	5	7
RFK	4/28	5/35

Métricas a nivel de esquema

Por último se presentan, en la tabla 5, las métricas al nivel del esquema de almacén de datos completo, el cual puede contener una o varias estrellas.

Tabla 5. Métricas a nivel de esquema.

Métrica	Descripción
NFT(Sc)	Número de tablas de hechos del esquema
NDT(Sc)	Número de tablas de dimensión del esquema
NSDT(Sc)	Número de tablas dimensionales compartidas por más de una estrella
NT(Sc)	Número de tablas del esquema
NAFT(Sc)	Número de atributos de las tablas de hechos del esquema
NADT(Sc)	Número de atributos de las tablas de dimensión del esquema
NASDT(Sc)	Número de atributos de las tablas de dimensión compartidas
NA(Sc)	Número de atributos del esquema
NFK(Sc)	Número de claves ajenas del esquema.
RSDT(Sc)	Ratio de de tablas dimensionales compartidas. Cantidad de tablas dimensionales que están relacionadas con más de una estrella
RT(Sc)	Ratio de tablas. Cantidad de tablas dimensionales por cada tabla de hechos
RScA(Sc)	Ratio de atributos del esquema. Número de atributos de las tablas dimensionales dividido por el número de atributos de las tablas de hechos
RFK(Sc)	Ratio de claves ajenas. Número de atributos que son claves ajenas
RSDTA(Sc)	Ratio de atributos de las tablas dimensionales compartidas. Número de atributos del esquema que son compartidos

En la tabla 6 se puede observar los valores que obtienen las métricas del nivel de esquema para el esquema de la figura 1.

Tabla 6. Valores de las métricas a nivel de esquema.

Métrica	Valor
NA	40
NFK	9
NDT	5
NT	7
NADT	28
NAFT	12
RFK	9/40
NFT	2
NSDT	4
NASDT	23
RSDT	4/5
RT	5/2
RScA	28/12
RSDTA	23/40

Como se puede observar, se han propuesto una gran cantidad de métricas. Actualmente estamos validando formal y empíricamente de manera que podamos elegir aquellas métricas que puedan ser útiles para medir la calidad de los almacenes de datos o podamos redefinir o refinar las métricas que tenemos.

Validación de las métricas

A fin de comprobar la validez de las métricas tanto teórica como prácticamente, se ha realizado la validación formal [6] y empírica de las métricas. En la tabla 7 se puede observar el resumen de la validación teórica viendo que todas las métricas se encuentran en la escala ordinal o alguna superior con lo que se puede concluir que todas ellas son formalmente válidas.

Tabla 6. Valores de las métricas a nivel de esquema.

Métrica	Escala
NA	Encima de la ordinal
NFK	Encima de la ordinal
NDT	Encima de la ordinal
NT	Ratio
NADT	Encima de la ordinal
NAFT	Encima de la ordinal
NFT	Ratio
NSDT	Encima de la ordinal
NASDT	Ratio
RSA	Absoluta
RFK	Absoluta
RSDT	Absoluta
RT	Absoluta
RSDTA	Absoluta

Una vez comprobada la validez teórica de las métricas, es necesario realizar el proceso de validación empírica para comprobar si las métricas son realmente útiles en la práctica, para ello se han realizado diversos experimentos [29], [30], en los que hemos obtenido que las métricas NFT (Número de tablas de hechos), NT (Número de tablas) y NFK (Número de claves ajenas) parecen ser buenos indicadores de la complejidad de los almacenes de datos. Actualmente seguimos trabajando en la validación empírica del conjunto de métricas propuestas para poder conseguir un conjunto válido y útil de métricas para almacenes de datos.

3. CALIDAD DE LOS PROPIOS DATOS

Dado que la calidad tiene componentes objetivas y subjetivas [27], es necesario catalogar los requisitos de calidad de datos de los usuarios según unas determinadas dimensiones de calidad. En la literatura consultada los autores intentan definir el concepto de calidad de datos y catalogar las dimensiones de calidad en función de unos determinados criterios, como pueden ser el ciclo de vida de los datos [28] o los tipos de investigación realizadas [14], [32], o simplemente la forma en la que se usan los datos [10]. Pero todos están de acuerdo en que la calidad de datos es un concepto multidimensional que comprende distintos aspectos según las necesidades de los consumidores de datos o de los diseñadores de sistemas, y que dichas necesidades deberían introducirse desde las fases

más tempranas del desarrollo a modo de requisitos de calidad de datos, con herramientas o interfaces como el propuesto en [26]. La tabla 8 muestra las dimensiones de calidad más importantes.

Tabla 8. Dimensiones de calidad más importantes [27] y preguntas relacionadas.

Dimensión	Definición
Accesibilidad	Los datos o están disponibles o son fácil y rápidamente recuperables
Cantidad Apropiada de datos	El volumen de datos es adecuado para la tarea que se está realizando
Compleción	Los datos son completos y suficientes para la tarea que se está desarrollando
Comprensibilidad	Los datos son fácilmente comprensibles
Credibilidad	Los datos pueden ser considerados como creíbles y verdaderos
Disponibilidad Temporal	Los datos están lo suficientemente actualizados para la tarea que se está desarrollando
Facilidad de manipulación	Los datos son fácilmente aplicables y manipulables en diferentes tareas
Interpretabilidad	Los datos están representados en el idioma apropiado, con una simbología correcta y adecuada y con la definición apropiada.
Libres de error	Los datos son correctos y fiables
Objetividad	Los datos son imparciales, sin prejuicios y sin connotaciones.
Relevancia	Los datos son útiles y aplicables en la tarea que se está desarrollando
Representación Concisa	Los datos están representados de una forma compacta
Representación Consistente	Todos los datos se representan en el mismo formato, que además es el más adecuado para la tarea que se está desarrollando.
Reputación	Los datos están altamente relacionados en términos de sus fuentes o contenidos.
Seguridad	El acceso a los datos está restringido apropiadamente para garantizar su seguridad
Valor Añadido	Los datos son beneficiosos y ofrecen ventajas al usarlos.

3.1. Metodología para la Medición de la Calidad de los Datos

En este apartado se muestra una metodología para medir la calidad de los datos guardados en un almacén de datos. Basada en las ideas propuestas en [23] donde se propone almacenar información referente a la calidad de los datos en el mismo almacén de datos, dicha metodología propone una serie de pasos bien estructurados y definidos, que partiendo de los requisitos de calidad de datos de los usuarios, trata de identificar las dimensiones de calidad que mejor describen esos requisitos, para después obtener métricas a partir de ese conjunto de dimensiones; después se realiza el proceso de medición propiamente dicho, que consiste en generar un valor numérico como resultado de un juicio de un determinado valor del dato con respecto a la dimensión elegida; posteriormente los resultados se guardan en el mismo almacén de datos, para después analizar los resultados. La forma de guardar los datos depende fuertemente del modelo de datos elegido para el almacén de datos.

El objetivo fundamental de la metodología que a continuación se describe, es ofrecer al usuario un marco de trabajo para determinar la calidad de los datos de un almacén de datos atendiendo a la calidad de los datos propiamente dicho. Lo que se propone en este marco de trabajo es, tras analizar

los requisitos de calidad de datos para la aplicación, buscar las dimensiones más significativas según dichos requisitos, obtener valores para dichas dimensiones según los datos y analizar las medidas aplicando algún criterio de valoración.

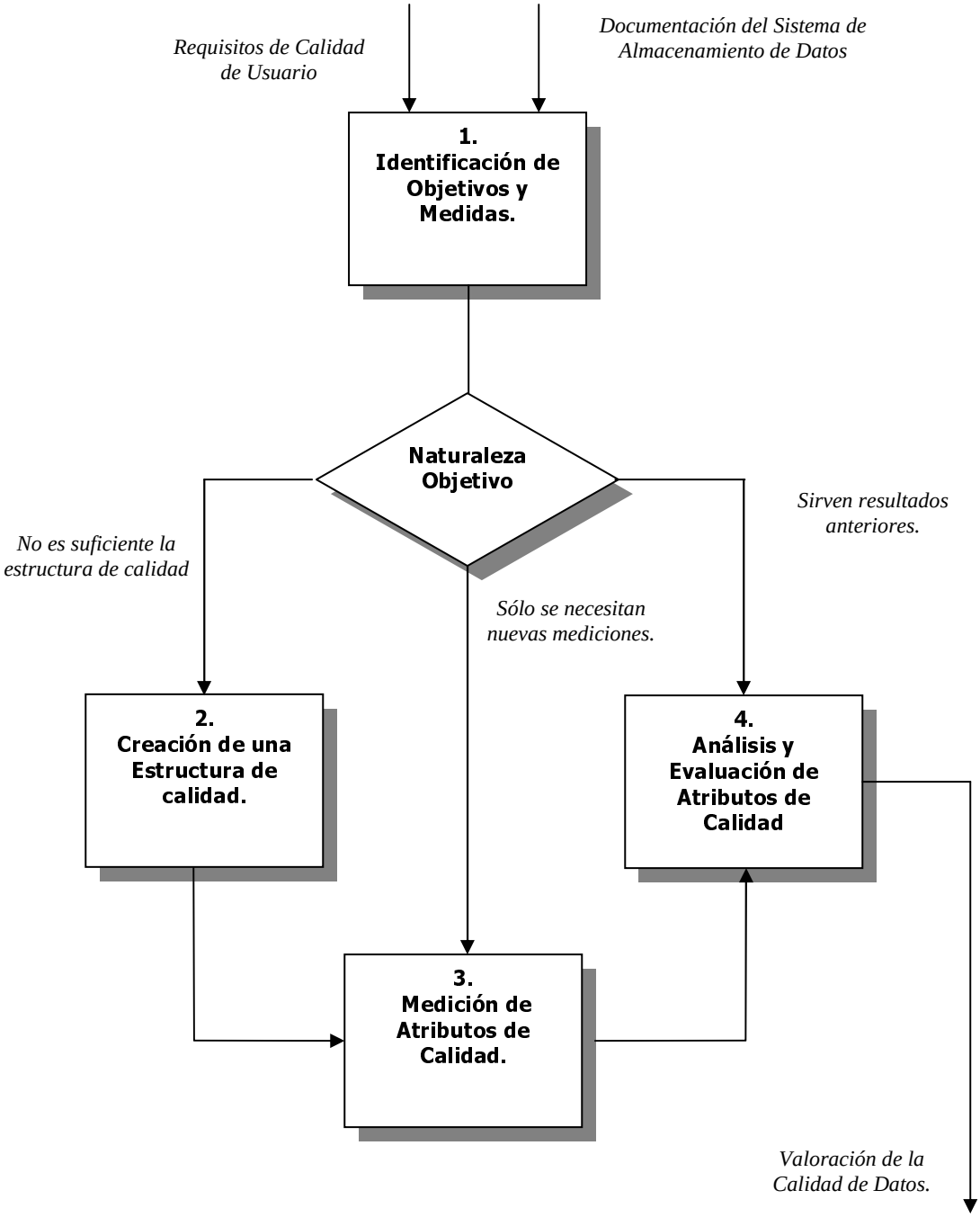

Figura 2. Metodología para la medición de la calidad de los datos.

La metodología se compone de un total de cuatro fases independientes bien diferenciadas. Cada una de estas fases está a su vez estar formada por unos pasos o actividades. Es recomendable seguir las fases de manera consecutiva, pero habrá ocasiones en las que sea necesario saltar alguna por no contemplarlas los objetivos de la medición. Las fases, representadas en la figura 2, son las siguientes:

1. **Fase 1: Identificación de los objetivos y de las medidas.** Es una fase de **análisis,** donde a partir de los requisitos de calidad de los usuarios se obtendrían una serie de productos de trabajo tras completar cada una de las siguientes actividades:

   1.1. Determinar el objetivo de la medición. Se trata de determinar las razones por las que se quiere medir el nivel de calidad de datos.

   1.2. Determinar los parámetros e indicadores de calidad. A partir de los requisitos de los usuarios se identifican las dimensiones y métricas de calidad de datos más significativos para acotar el problema de calidad de datos.

   1.3. Localizar los datos a valorar. Esta actividad se divide en las siguientes subactividades.

   1.4. Definición de criterios de calidad. Se trata de establecer criterios de valoración para juzgar la bondad de un dato y de definir criterios de evaluación para determinar la bondad del conjunto de los datos.

2. **Fase 2: Creación de una estructura de calidad.** Es la fase de **diseño,** donde el objetivo es dotar al almacén de datos de una estructura para guardar los valores que más tarde se recogerán para las medidas de calidad.

3. **Fase 3: Medición de los atributos de calidad.** Una vez que el almacén de datos disponga de una estructura para guardar las medidas de las dimensiones de calidad, esta fase consiste en recoger valores para dichas medidas en las dimensiones especificadas. Puede llegar a ser necesario que para algunas dimensiones de calidad se deba conocer el valor del dato real y compararlo con el del dato almacenado. En función de la cantidad de datos y del nivel de calidad exigido puede ser necesario medir los valores de todos los datos o seleccionar por muestreo solo una parte de esa totalidad. En cualquier caso estas mediciones se guardaran en el almacén de datos.

4. **Fase 4: Análisis y Evaluación de los valores de los atributos de calidad.** En esta fase, se someterá los valores individuales medidos en la fase anterior a los criterios de valoración para determinar el grado de bondad de un dato y según el número de datos con calidad y los criterios de evaluación establecidos se juzgarán si esos datos tienen o no el grado de calidad deseado. Si es así, se certifican los datos como válidos para la aplicación. En caso contrario se desechan como inválidos, procediendo posteriormente como mejor convenga: corrección de los datos existentes o captura de nuevos datos.

3.2. Un Modelo de Madurez de Calidad de Datos

La existencia de un modelo de madurez de calidad de datos se justifica por la necesidad crítica de una metodología que permita determinar el estado en el que las organizaciones manejan sus datos [19] y por el hecho de que ninguna de los modelos de valoración y evaluación de procesos, incluyendo CMM, CMMI, ISO 9001, BootStrap, o SPICE observan la calidad de datos entre sus objetivos de calidad.

El modelo que a continuación se presenta debería ser usado por un equipo especializado de aseguramiento de calidad de datos para:

- determinar el estado de madurez de una base/almacén de datos,

- proponer mejoras al tratamiento que el sistema de información hace de la calidad de los datos para alcanzar niveles de madurez más altos con el correspondiente mantenimiento.

Para determinar el nivel de madurez de calidad de datos de una base/almacén de datos, es posible plantearse su ciclo de vida para un mejor entendimiento del problema como un Proceso de Gestión de Datos en el que se va a obtener un determinado producto de datos [34]. Teniendo esto en cuenta, se desarrolla CALDEA como un marco de referencia para determinar el nivel de calidad de un determinado Proceso de Gestión de Datos.

Los niveles que se definen en el modelo, para un Proceso de Gestión de Datos son los siguientes:

1. **Inicial**, donde técnicamente no existe ningún objetivo de calidad de datos. Las organizaciones no prestan especial atención a este aspecto, y si llegan a tenerla es mera casualidad. No existen objetivos de calidad de datos propiamente dichos, por lo que no se emprenden explícitamente acciones para asegurar la calidad en el Proceso de Gestión de Datos.

2. **Gestionado o de Definición,** donde se empieza a tomar conciencia de la necesidad de incorporar la calidad de datos al desarrollo de un producto o a la prestación de un servicio. Para ello se deben realizar una serie de esfuerzos a fin de delimitar y diseñar el Proceso entero de Gestión de Datos, identificando todos sus componentes y el modo en que se relacionan entre ellos. Se hace necesario agrupar y coordinar todas estas actividades en un proyecto para el Proceso de Gestión de Datos. Para ello se tienen en cuenta las siguientes áreas de proceso:

   - *Gestión de Proyecto de Proceso de Gestión de Datos.*
   - *Gestión de Requisitos de Datos.*
   - *Gestión de Dimensiones y métricas de Calidad de Datos.*
   - *Gestión de fuentes y sumideros de datos.*
   - *Gestión del proyecto de adquisición o desarrollo de una base de datos o de un almacén de datos.*

3. **Definido o de Integración**, un Proceso de Gestión de Datos está en este nivel cuando se ha alcanzado el nivel anterior y se realizan esfuerzos a fin de desarrollar y ejecutar políticas de calidad de datos. Esto implica tener estandarizados a nivel organizacional ciertos aspectos de calidad de datos. Este nivel se centra en capturar el conocimiento adquirido a través de la experiencia y hacerlo reutilizable añadiéndolo a la cultura de la organización a fin de evitar errores pasados mejorando la calidad. Esta filosofía hace necesaria la redacción de distintos estándares y documentos maestros de calidad de datos, que proporcionarán un mejor uso de las herramientas y de los métodos [15]. Para conseguir este objetivo se deben realizar las siguientes actividades:

   - *Gestión de un equipo de calidad de datos.*
   - *Validación y verificación de un producto de calidad de datos*
   - *Gestión del riesgo y del impacto de la baja calidad de datos.*
   - *Gestión de la estandarización de la calidad de datos.*
   - *Gestión de los procesos organizacionales.*

4. **Gestionado cuantitativamente o de Medición.** Un Proceso de Gestión de Datos se dice en este nivel cuando está integrado en una cultura organizacional de calidad de datos (ha alcanzado el nivel

tres) y se realizan mediciones relacionadas con el propio proceso o con sus componentes. El objetivo de este nivel es obtener la conformidad numérica de que el rendimiento del proceso de gestión de datos es consistente en un periodo de tiempo más o menos largo en unas circunstancias estables. La actividad que debe realizarse es la siguiente:

- *Gestión de Medidas del Proceso de Gestión de datos.*

5. **Optimizante o de Mejora Continua.** Un Proceso de gestión de datos se dice en el nivel optimizante o de mejora continua cuando las medidas definidas el nivel cuatro se utilizan para identificar causas de defectos o para mejorar su rendimiento. Se deben realizar las actividades siguientes:

- *Análisis causal para la prevención de defectos.*
- *Desarrollo organizacional e innovación.*

La tabla 9 presenta un resumen de los distintos niveles y los correspondientes objetivos de calidad para cada uno de ellos.

Tabla 9. Objetivos de calidad de datos para cada nivel de madurez.

Nivel	Enfoque	Áreas de Proceso
5. Optimizante o de mejora continua	Mejora continua de procesos	• *Análisis causal para la prevención de defectos.* • *Desarrollo organizacional e innovación.*
4. Gestionado cuantitativamente o de medición	Gestión cuantitativa	• *Gestión de medidas del proceso de gestión de datos.*
3. Definido o de integración	Estandarización de procesos	• *Gestión de un equipo de calidad de datos.* • *Validación y verificación de un producto de calidad de datos.* • *Gestión del riesgo y del impacto de la baja calidad de datos.* • *Gestión de la estandarización de la calidad de datos.* • *Gestión de los procesos organizacionales.*
2. Gestionado o de definición	Gestión básica de procesos	• *Gestión de proyecto de proceso de gestión de datos.* • *Gestión de requisitos de datos.* • *Gestión de dimensiones y métricas de calidad de datos.* • *Gestión de fuentes y sumideros de datos.* • *Gestión del proyecto de adquisición o desarrollo de una base de datos o de un almacén de datos.*
1. Inicial	No se observa ningún aspecto de calidad de datos.	

4. CALIDAD DE LOS ALMACENES DE DATOS DESDE EL PUNTO DE VISTA DE LOS USUARIOS

Las encuestas son probablemente el método de investigación más comúnmente usado [24]. En este apartado presentaremos los resultados de una encuesta que hemos realizado a las principales empresas del sector de los sistemas de información, con el objetivo de conocer el estado actual del desarrollo y mantenimiento de los almacenes de datos, de manera que podamos obtener una panorámica acerca de los usos, técnicas, metodologías y problemas. Estas encuestas pueden servirnos también para poder obtener una visión global de las necesidades y comportamientos de los usuarios.

Las encuestas fueron enviadas a empresas cuyos sectores comerciales eran bastante heterogéneos (empresas públicas, educación, alimentación, aeronáutica, ...) y con una gran disparidad en el número de empleados. Estas organizaciones tienen un departamento de informática propio de pequeño tamaño (generalmente no más de 10 especialistas) con una edad media en el tramo de 31 a 40 años.

Los especialistas en bases de datos han recibido formación generalista a nivel académico (Formación Profesional, Ingenieros Técnicos y Superiores Informáticos, Físicos, Químicos, Matemáticos) y específica para las herramientas con las que trabajan. Generalmente siguen alguna metodología, propia o estandarizada, para el desarrollo de proyectos de almacenes de datos, aunque no se usa ningún tipo de medición del producto ni del proceso.

Los proyectos de almacenes de datos suelen ser implementados usando herramientas comerciales estandarizadas y ampliamente utilizadas, para las que, normalmente y con los datos aportados respecto a su formación, se podría decir que tienen suficiente experiencia en ellas. Entre estas herramientas se pueden encontrar, como las más usadas, *Ingres* y *Oracle* ejecutándose sobre *Unix/Linux*, y con menor frecuencia sobre *Windows NT / 2000 Server*.

Respecto a la experiencia de las empresas consultadas en el desarrollo de proyectos con almacenes de datos el 50% de las organizaciones consultadas ha realizado de 1 a 5 proyectos y sólo el 25% desarrolló más de 15 proyectos de almacenes de datos.

En cuanto a los recursos, podemos apreciar que el tiempo de desarrollo de estos proyectos no suele superar 1 año y que la inversión media en estos proyectos se encuentra en torno a los 60.000 €. Por último, podemos observar que existen dos tipos de proyectos que realizan las empresas, aquellos en los que el tamaño del sistema se encuentra entorno a los 50Gb (probablemente datamarts) y aquellos en que el volumen de información supera los 500Gb.

5. CONCLUSIONES

Una de las obligaciones principales de los profesionales de las tecnologías de la información debe ser asegurar la calidad de la información, ya que esta es uno de los principales activos de las organizaciones. Las consideraciones acerca de la calidad han acompañado a los almacenes de datos desde sus inicios [18].

La calidad de la información viene determinada por la calidad tanto del almacén de datos como por la calidad de la presentación de los datos. De hecho, es muy importante que los datos del almacén reflejen correctamente el mundo real, pero es también muy importante que los datos sean interpretados correctamente. En la calidad del almacén de datos, al igual que una base de datos operacional [25], se

deben considerar tres aspectos: la calidad del SGBD (Sistema Gestor de Base de Datos) relacional o multidimensional que lo soporta, la calidad del modelo de datos (tanto conceptual, lógico como físico) y la calidad de los propios datos contenidos en el almacén.

En este artículo se ha presentado un conjunto de métricas para medir y controlar la calidad de los modelos multidimensionales, así como se han hecho varias consideraciones acerca de la calidad de los propios datos que se encuentran almacenados en el almacén de datos y que características deben tenerse en cuenta a la hora controlar la calidad de esta información. También hemos recogido algunos resultados de una encuesta que hemos realizado para conocer los aspectos importantes de la calidad de los almacenes de datos desde el punto de vista de los usuarios.

AGRADECIMIENTOS

Este trabajo es parte del proyecto CALDEA (TIC 2000-0024-P4-02) financiado por la Subdirección General de Proyectos de Investigación, Ministerio de Ciencia y Tecnología de España.

REFERENCIAS

[1] C. Adamson y M. Venerable, *Data Warehouse Design Solutions*, John Wiley and Sons, E.U.A., 1998.

[2] M. Bouzeghoub, F. Fabret y H. Galhardas, "Datawarehouse Refreshment", Capítulo 4, *Fundamentals of Data Warehouses*, Springer-Verlag, 2000.

[3] M. Bouzeghoub y Z. Kedad, "Quality in Data Warehousing", *Information and Database Quality*, eds. Piattini, Calero y Genero, Kluwer Academic Publisher, pp. 163-198, 2002.

[4] L. Cabbibo y R. Torlone, "A Logical Approach to Multidimensional Databases", *Sixth Internationational Conference on Extending Database Technology (EDBT'98)*, Valencia. España. *Lecture Notes in Computer Science 1377*, Springer-Verlag, pp 183-197, 1998

[5] C. Calero, M. Piattini y M. Genero, "Empirical Validation of Referential Integrity Metrics", *International and Software Technology*, vol. 43, pp. 949-957, 2001.

[6] C. Calero, M. Piattini, C. Pascual y M.A. Serrano, "Towards Data Warehouse Quality Metrics", *Proc. 3rd Workshop on Design and Management of Data Warehouses (DMDW'01)*, Interlaken, Suiza, Jun. 2001.

[7] J. Cavero, E. Marcos y M. Piattini, "Metodología para el Diseño de Almacenes de Datos: Etapa de Modelado Conceptual", *4º Encontro para a Qualidade nas Tecnologias de Informação e Comunicações (QUATIC 2001)*, Lisboa, Portugal, 2001.

[8] J. Celko, "Don´t Warehouse Dirty Data", *Datamation*, vol. 15, pp. 42-52, Oct. 1995.

[9] E. Del Peso, *Ley de Protección de Datos: La Nueva LORTAD*, Madrid, España, Díaz de Santos, 2000.

[10] L.P. English, *Improving Data Warehouse and Business Information Quality: Methods for Reducing Costs andIincreasing Profits*, Willey & Sons, 1999.

[11] S.R. Gardner, "Building the Data Warehouse", *Communications of the ACM*, vol. 41, No. 9, pp. 52-60, Sep. 1998.

[12] M. Golfarelli, D. Maio y S. Rizzi, "Conceptual Design of Data Warehouses from E/R Schemes", *31st Hawaii International Conference on System Sciences*, 1998.

[13] T. Hammergren, *Data Warehousing Building the Corporate Knowledge Base*, International Thomson Computer Press, Milford, 1996.

[14] K.T. Huang, Y. Lee y R. Wang, *Quality Information and Knowledge*, Prentice Hall, Upper Saddle River, 1999.
[15] W. Humphrey, *Managing the Software Process*, Addison–Wesley, Reading Mass., 1989.
[16] W.H. Inmon, *Building the Data Warehouse*, segunda edición, John Wiley and Sons, E.U.A., 1997.
[17] ISO, *Software Product Evaluation – Quality Characteristics and Guidelines for their Use*, ISO/IEC Standard 9126, Ginebra, Suiza, 2001.
[18] M. Jarke, M. Lenzerini, Y. Vassiliou y P. Vassiliadis, *Fundamentals of Data Warehouses*, Springer-Verlag, 2000.
[19] B. Kahn, D. Strong y R. Wang, "Information Quality Benchmarks: Product and Service Performance", *Communications of the ACM*, vol. 45, No. 4, Abr. 2002.
[20] S. Kelly, *Data Warehousing in Action*, John Wiley & Sons, 1997.
[21] R. Kimball, L. Reeves, M. Ross y W. Thornthwaite, *The Data Warehouse Lifecycle Toolkit*, John Wiley and Sons, E.U.A., 1998.
[22] D. Loshin, *Enterprises Knowledge Management: The Data Quality Approach*. Morgan Kauffman, San Francisco, E.U.A., 2001.
[23] L. Orman, V. Storey y R. Wang, "Systems Approaches to Improving Data Quality", Total Data Quality Management Research Program, TDQM-94-05, http://web.mit.edu/tdqm/www/papers/94/94-05.html, Ago. 1994.
[24] S.A. Pfleeger y B.A. Kitchenham, "Principles of Survey Research", *Software Engineering Notes*, vol. 26, No. 6, pp. 16-18, 2001.
[25] M. Piattini, M. Genero, C. Calero, C., M. Polo y F. Ruiz, "Database Quality", *Advanced Database Technology and Design*, eds. O. Diaz, y M. Piattini, Artech House, Londres, Inglaterra, 2000.
[26] M. Piattini e I. Caballero, "A Data Quality Interface for a Software Development Methodology", *Proc. of Fifth IASTED International Conference*, Anaheim, E.U.A., pp. 94 a 99, ISBN- 0-88986-305-9, Ago. 2001.
[27] L. Pipino, Y. Lee y R. Wang, "Data Quality Assessment", *Communications of the ACM*, vol. 45, No. 4, Abr. 2002.
[28] T.C. Redman, *Data Quality for the Information Age*, Artech House Publishers, Boston, E.U.A., 1996.
[29] M. Serrano, C. Calero y M. Piattini, "Validating Metrics for Datawarehouses", *IEE Proceedings SOFTWARE*, ISSN: 1462-5970, vol. 149, No. 5, pp. 161-166, 2002.
[30] M. Serrano, C. Calero y M. Piattini, (2003). "Experimental Validation of Multidimensional Data Models Metrics", *Proc. of the Hawaii International Conference on System Sciences (HICSS'36)*, Ene. 2003.
[31] J. Trujillo, M. Palomar, J. Gómez e I. Song, "Designing Data Warehouses with OO Conceptual Models", *Computer*, pp. 66-75, Dic. 2001.
[32] Y. Wand y R. Wang, "Anchoring Data Quality Dimensions in Ontological Foundations", *Communications of the ACM*, vol. 39, No. 11, pp. 86-95, 1994.
[33] R. Wang, M.P. Reddy y H. Kon, "Toward Quality Data: An Atribute-based Approach", *Journal of Decision Support Systems*, vol. 13, pp. 349-372, 1992.
[34] R. Wang, H. Kon y S. Madnick, "Data Quality Requirements Analisys and Modeling", *Proc. Ninth International Conference of Data Engineering*, Viena, Austria, pp. 670-677, 1993.