Evaluación de Calidad en una Aplicación de Data Warehousing: de la Definición de Metas a la Especificación de Métricas

Diego Sastre, Verónika Peralta, Raúl Ruggia

Instituto de Computación, Facultad de Ingeniería, Universidad de la República
Julio Herrera y Reissig 565, 5to piso
11300 Montevideo, Uruguay
diegos@fing.edu.uy, vperalta@fing.edu.uy, ruggia@fing.edu.uy

Resumen. La calidad de los datos tiene una importancia vital para la toma de decisiones. Muchos trabajos han propuesto factores y métricas de calidad, tanto desde un punto de vista teórico como pragmático. El objetivo de este artículo es presentar un enfoque práctico de análisis de la calidad, basado en el paradigma "*Goal-Question-Metric*". Para ello analizamos una aplicación real de Data Warehousing, relevando problemas de calidad, definiendo metas de calidad para solucionar dichos problemas, refinando y asociando las metas a factores de calidad y estableciendo las métricas que permitan cuantificarlos.

Palabras clave. Data Warehouse, Aplicaciones de Bases de Datos

1. Introducción

Los Sistemas de Data Warehousing (SDW) son sistemas de apoyo a la toma de decisiones. Estos sistemas integran información de múltiples sistemas operacionales (fuentes de datos) y la presentan al usuario de forma intuitiva y eficiente para el análisis. Son muy utilizados por mandos medios y directivos para analizar situaciones y tomar decisiones en las empresas, determinando causas y consecuencias de problemas y definiendo estrategias. En los últimos años los SDW se han transformado en una herramienta fundamental para definir el rumbo del negocio. En este contexto, la calidad de los datos brindados por un SDW tiene una importancia vital para las empresas. La toma de decisiones en base a datos desactualizados, incompletos o imprecisos puede ocasionar grandes pérdidas en una organización.

A pesar de la amplia gama de técnicas de diseño y mantenimiento de SDW propuestos en la última década (por ejemplo [11][1][17][28][16], muchas organizaciones cuentan con aplicaciones de Data Warehousing que fueron desarrollados de manera ad-hoc o que no fueron evolucionando a medida que cambiaban las necesidades de los usuarios o los sistemas fuente. Dichas carencias de diseño y mantenimiento tienen como consecuencia una disminución de la calidad de los datos, sobre todo desde la percepción de los usuarios.

Este trabajo plantea el análisis de un SDW desde la perspectiva de la calidad de los datos. Concretamente, se plantea determinar la calidad de los datos analizando diferentes tipos de errores que conllevan a la toma de decisiones equivocadas. Si bien muchos problemas de calidad son percibidos por los usuarios y/o los técnicos, no siempre es fácil determinar sus causas, su magnitud o su impacto en los negocios. La medición de la calidad permite encarar con objetividad y cuantificadamente los diferentes problemas, facilitando la aplicación de estrategias de mejora de la calidad.

Un enfoque muy utilizado en ingeniería de software para la definición de métricas de calidad es el paradigma GQM (*Goal-Question-Metric*) [3]. GQM propone identificar un conjunto de metas de calidad de alto nivel (*goals*), descomponerlas en una serie de preguntas orientadas a evaluar el nivel de satisfacción de cada meta (*questions*) y definir un conjunto de métricas (*metrics)* que permitan cuantificar las respuestas a las preguntas. Una característica particularmente interesante de este enfoque es que asocia las propiedades de calidad a metas del usuario (alto nivel), permitiendo encarar la detección de errores y acciones para su corrección en forma cercana al uso real.

Basados en este paradigma, el proyecto DWQ propone una metodología para desarrollar y mantener un SDW guiados por requerimientos de calidad de los usuarios [28]. La metodología consta de tres fases: (i) una fase de diseño que consiste en relevar metas de calidad, descomponerlas en preguntas y elaborar métricas para responderlas; (ii) una fase de evaluación que consiste en medir la calidad del SDW de acuerdo a las métricas elaboradas; y (iii) una fase de análisis y mejora en la cual, a la luz de las mediciones, se sugieren acciones de mejora. En el proyecto Quadris, se propone un refinamiento del modelo de calidad de DWQ, de manera de construir una colección de

métricas y métodos de medición de la calidad que sea reutilizable, tanto en el seno de una organización como a diferentes contextos de aplicación [8].
En este artículo presentamos un caso práctico de aplicación de la metodología desarrollada en [8]. Para ello analizamos una parte de un SDW real de una empresa, tomando como base los problemas más importantes y/o urgentes que encuentran los usuarios en la explotación de la información del SDW, y relacionándolos con propiedades de calidad de los datos involucrados. La propuesta tiene como objetivo principal determinar los factores de calidad que pueden tener más impacto en el uso real del sistema teniendo en cuenta los problemas planteados a los usuarios, estableciendo métricas apropiadas para la medición de dichos factores, y proponiendo mecanismos para hacer las mediciones y sugerir mejoras cuando sea posible.
Los factores de calidad a considerar se determinan a partir de un estudio de los datos manejados por el sistema, considerando tanto los problemas de calidad percibidos directamente o indirectamente por los usuarios así como algunos problemas potenciales que se desea verificar. Por ejemplo, se sospecha que se cometen errores en las definiciones de campañas de marketing dirigidas al lanzamiento de nuevos productos debido a información incompleta en el SDW (datos incompletos, datos faltantes, datos que llegan con retraso). Este trabajo posibilita un tratamiento objetivo y cuantificado de dichos problemas, proponiendo métricas apropiadas para los factores de calidad considerados.
La organización del artículo es la siguiente. La sección 2 describe trabajos relacionados. La sección 3 presenta nuestro enfoque de evaluación de la calidad. La sección 4 describe el caso de estudio presentando una descripción del SDW y sus fuentes de datos. En la sección 5 se presenta la problemática de calidad de datos a abordar y en la sección 6 se definen factores y métricas de calidad para cuantificar dichos problemas. En la sección 7 se establecen recomendaciones de rediseño del sistema y en la sección 8 se presentan las conclusiones del trabajo realizado y lineamientos de trabajo futuro.

2. Trabajo relacionado

Son muchos los trabajos que proponen mecanismos para definir, modelizar o evaluar la calidad de los datos. En general, hay consenso en que la calidad es un concepto multidimensional. En efecto, la calidad suele estudiarse vía múltiples dimensiones que caracterizan diferentes facetas de los datos, por ejemplo, *exactitud, accesibilidad, integridad, precisión, confiabilidad, completitud, consistencia, flexibilidad, trazabilidad, seguridad en el acceso, facilidad de manipulación, relevancia,* entre otros [24][29][19]. Algunos trabajos proponen clasificaciones de las dimensiones de calidad de acuerdo a criterios semánticos [29][21], orientados a procesos [20][30] u orientados a metas [13]. Otros trabajos proponen marcos formales para definer dimensiones de calidad y se enfocan en el manejo y almacenamiento de los metadatos apropiados [27][14][15][12][9][18].
No obstante, los diferentes autores difieren en las dimensiones de calidad que deberían estudiarse: Wang y Strong afirman que para mejorar la calidad de los datos, es necesario entender qué significa la calidad para los usuarios [29] y presentan un ranking de dimensiones de calidad que son más relevantes para éstos: *relevancia, exactitud, interpretabilidad, accesibilidad,* entre otros. Otros autores proponen listas de dimensiones principales para algunos tipos de sistemas o dominios de aplicación, por ejemplo, *completitud, unicidad, consistencia, frescura* y *exactitud* para Sistemas de Integración de Datos [2], *completitud, credibilidad, exactitud, consistencia e interpretabilidad* para sistemas de Data Warehousing [14], *exactitud, completitud, frescura y consistencia* para Sistemas Web [10]. Otros trabajos se focalizan en el estudio en profundidad de algunas dimensiones de calidad, por ejemplo *completitud* [22], *frescura* [23] o *exactitud* [7].
Cada dominio de aplicación tiene su propia visión de la calidad además de una serie de soluciones, generalmente ad-hoc, para enfrentar los problemas de calidad [4]. Esto motiva la necesidad de mecanismos para definir métricas de calidad teniendo en cuenta el dominio de aplicación y los requerimientos de los usuarios.
Basados en el paradigma GQM, el proyecto DWQ (Foundation son Data Warehouse Quality) propone definir métricas de calidad a partir de requerimientos de calidad de los usuarios [28]. Dicho proyecto estudia un conjunto de factores de calidad y sus relaciones con tipos de metas de calidad de usuarios potenciales, presentando un meta-modelo de calidad que tiene en cuenta dichas relaciones. Ese meta-modelo de calidad es refinado en el proyecto Quadris para adaptarlo a una gama más amplia de aplicaciones [2]. El meta-modelo de Quadris hace hincapié en la reutilización de métricas y métodos de calidad [8]. En particular, se propone la construcción de una librería de conceptos de calidad que puedan ser instanciados considerando las particularidades de cada meta de calidad. Este artículo pone en práctica dichos mecanismos mediante un caso de estudio.

3. Nuestro Enfoque de Evaluación de la Calidad

La calidad de productos y procesos suele analizarse de manera *top-down*, comenzando por analizar las metas de la organización desde una perspectiva de alto nivel y refinando sucesivamente esas metas hasta obtener valores cuantitativos de calidad. El paradigma GQM (Goal-Question-Metric; meta-pregunta-métrica) [3] propone tres niveles de abstracción: (i) a nivel conceptual, se definen metas de calidad de alto nivel para los productos y procesos, (ii) a nivel operacional, para cada meta se detalla un conjunto de preguntas que caracterizan la forma de evaluarla, y (iii) a nivel cuantitativo, a cada pregunta se le asocia un conjunto de métricas orientadas a responderla. La calidad de la información también puede analizarse siguiendo el paradigma GQM. El modelo de calidad desarrollado en el proyecto DWQ [28] reutiliza y extiende el modelo GQM. En el contexto del proyecto Quadris, dicho modelo fue refinado y adaptado a una clase más amplia de aplicaciones [3].
Nuestro enfoque de evaluación de la calidad consiste de 3 pasos principales, según el paradigma GQM: (1) Definición de metas de calidad que apunten a resolver los problemas de calidad de la organización. (2) Definición de preguntas de calidad orientadas a evaluar el nivel de satisfacción de cada meta. (3) Definición de métricas de calidad que permitan cuantificar las respuestas a las preguntas.
El primer paso consiste en el relevamiento de los problemas de calidad de la organización y la definición de metas que apunten a resolver esos problemas. Un ejemplo de meta es "reducir la cantidad de cartas que no llegan a los clientes". El segundo paso consiste en descomponer las metas en sub-metas más simples, de manera de poder definir mecanismos para evaluar si son satisfechas o no. Las preguntas surgen de descomponer suficientemente las metas hasta poder asociarlas a un único *factor de calidad*. Un factor de calidad representa un aspecto particular de la calidad, por ejemplo la correctitud sintáctica, la precisión o la actualidad de los datos. Siguiendo el ejemplo anterior, una pregunta posible es "¿Cuántas direcciones de clientes están desactualizadas?". Muchos factores de calidad han sido propuestos en la literatura; véase por ejemplo [29][25]. La tabla 1 presenta algunas definiciones de factores que son usados en este trabajo, tomadas de [23][22][26]. El tercer paso implica seleccionar las métricas apropiadas de un factor de calidad e implementar los métodos de medición apropiados. Una métrica es un instrumento usado en la medición de un factor de calidad, por ej. "la cantidad de días pasados desde la última actualización de los datos" es una métrica del factor actualidad. Un método es un programa o proceso que se utiliza para obtener medidas de calidad. Estas medidas permiten cuantificar las metas de calidad y así obtener una visión precisa del estado de los negocios.
En las siguientes secciones presentamos la utilización práctica de nuestro enfoque en una aplicación de Data Warehousing.

Tabla 1. Definición de algunos factores de calidad, agrupados según dimensiones de alto nivel

Dimensión	Factor	Definición
Exactitud	Correctitud Semántica	Indica qué tan bien se corresponden los datos con la realidad
	Correctitud Sintáctica	Indica qué tan libres de errores sintácticos y de discordancias de formato están los datos
Completitud	Cobertura	Indica la porción de la realidad (entidades) representada en el sistema
	Densidad	Indica la proporción de valores no nulos de los datos
Frescura	Actualidad	Indica qué tan actualizados son los datos con respecto a la realidad
Consistencia	Integridad de dominio	Indica el nivel de satisfacción de reglas de dominio
Unicidad	Unicidad	Indica el nivel de duplicación de los datos

4. Descripción del Caso de Estudio

En esta sección describimos una porción de un SDW de una empresa que brinda servicios profesionales en informática. Estos servicios se brindan a través de contratos que regulan el vínculo entre la empresa y sus clientes. Los contratos pueden corresponder a proyectos puntuales con duración limitada, o a servicios de asistencia, soporte y ejecución de tareas en forma periódica durante largos periodos de tiempo.
El sub-sistema de SDW a analizar, ilustrado en la Figura 1, se compone de tres fuentes de datos principales: (i) *CRM*, que dispone de información de los clientes y sus contratos, (ii) *Mesa de Ayuda* que dispone de la

información técnica operativa, y (iii) *Contabilidad y facturación*, que contiene la información operativa contable. Esta información se almacena en una base de datos integrada y a partir de ella se cargan diariamente dos bases de datos OLAP[1] (o "cubos"), los cuales se utilizan para el análisis y toma de decisiones sobre el negocio de la empresa.

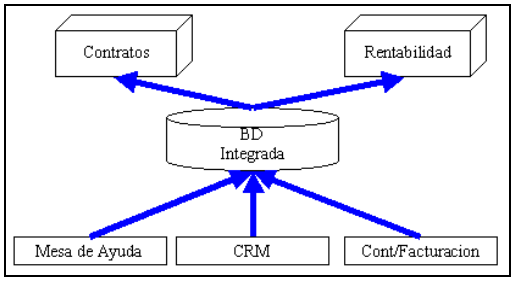

Fig. 1. Arquitectura del sub-sistema analizado

El cubo *Contratos* es utilizado para analizar la evolución de los contratos con los clientes. Interesa verificar qué tipos de servicio se ejecutan y sobre qué plataformas de software o hardware, de forma de clasificar los contratos y productos involucrados. Por otro lado es importante analizar qué sectores técnicos son responsables de atender más contratos y qué componentes del contrato le corresponde a cada uno, buscando un equilibrio entre dichos sectores. Desde un punto de vista comercial interesa analizar qué vendedores venden más contratos y de qué tipo, así como estudiar qué tiene contratado un cliente con el fin de profundizar la relación con el mismo. La figura 2a muestra el esquema conceptual del cubo *Contratos* (usando la notación del modelo CMDM [6]). Las dimensiones son: Cliente, Plataformas, Nro Contrato, Vendedor, Sector Técnico, Forma de Pago, Tipo de Servicio y Componentes. La medida es Monto mensual del contrato.
El cubo *Rentabilidad* se concentra en el rendimiento económico de los servicios ejecutados. Se debe poder responder consultas del tipo "listado ordenado de los clientes por rentabilidad", o "los 10 clientes menos rentables", o "los clientes de mayor consumo de tiempo". El esquema conceptual de este cubo se presenta en la figura 2b. Las dimensiones son: Cliente, Sector Técnico y Fecha. Las medidas son: Importe factura, Tiempo, Costo, Importe/Tiempo y Rentabilidad.

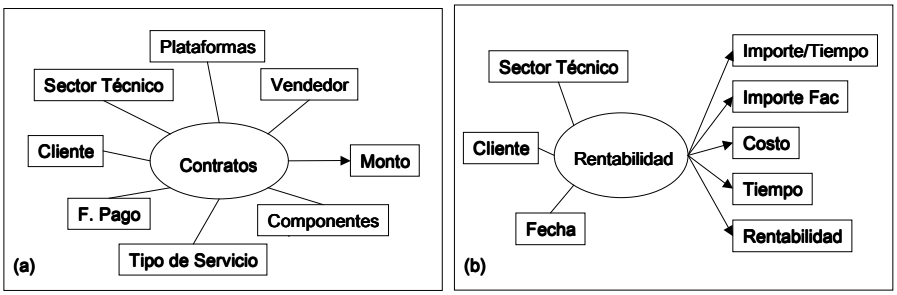

Fig. 2. Esquema conceptual de los cubos: (a) *Contratos*, (b) *Rentabilidad*

5. Problemática de Calidad

En esta sección describimos y clasificamos los principales problemas de calidad encontrados en el SDW. Algunos problemas son consecuencia de problemas de calidad de las bases de datos fuentes, mientras que otros son introducidos por un mal diseño del SDW. Para el relevamiento se tuvieron en cuenta los problemas detectados y reportados por los usuarios además de indicios y sospechas de potenciales problemas a verificar. Los problemas encontrados son los siguientes:

[1] OLAP: On-line Analytical Processing. Tecnología utilizada para almacenar y analizar información, la cual se representa mediante Modelos Multidimensionales.

- *Problema 1*: Hay clientes cargados dos o más veces en el sistema por errores de codificación, por ejemplo, por ingresar su nombre en dos formas diferentes. Este problema conlleva a errores en el análisis de rentabilidad del cliente, dado que no se le considera en forma completa, y es de alto impacto en el cubo *Rentabilidad*.
- *Problema 2*: Algunos datos se cargan a un cliente equivocado, por ejemplo, se le ingresa un contrato que no le corresponde o se le asocia tiempo técnico invertido en un servicio que no contrató. El primer ejemplo impacta directamente el análisis de contratos y el segundo genera una distorsión de la rentabilidad de los clientes. Este tipo de situaciones generalmente tienen origen cuando un cliente contrata servicios para terceros (por ej. sus propios clientes). Nos interesa ingresar el contrato y adjudicar actividades al cliente que contrató el servicio, que es en definitiva a quien se le factura.
- *Problema 3*: Algunos datos no fueron cargados aún en la base de datos integrada en el momento que se consulta el cubo. Por ejemplo, el registro de tiempo técnico invertido en servicios y la facturación de servicios suelen cargarse con retrasos. Estos casos distorsionan los análisis de rentabilidad por cliente, ya que el sistema no está en condiciones de dar información completa, y por lo tanto fiable, de su situación.
- *Problema 4*: Hay valores nulos debidos a errores en la integración de datos de las fuentes, lo que lleva a que algunas entidades no se visualicen en los cubos. Por ejemplo, hay casos en los que el monto de un contrato se carga como un valor nulo en la BD integrada, y dicho contrato no se carga en el cubo *Contratos*. Esta situación generalmente se corrige en la siguiente actualización de montos (mensual), pero mientras tanto el cubo de *Contratos* esta dando información incompleta o inválida, la cual no es detectable por el usuario.
- *Problema 5*: Hay errores de consistencia en ciertos datos, por ejemplo, para un servicio se ingresaron 90 horas de trabajo de un técnico en un día. Estos errores se producen por "bugs" en el sistema operacional o en el proceso de carga de los datos al cubo. Este tipo de situaciones generan errores de impacto en el cubo de *Rentabilidad.*
- *Problema 6*: No se mantienen fechas en el cubo *Contratos*, lo cual impide el manejo de información histórica. Muchas consultas clave para los usuarios no pueden realizarse por falta de la perspectiva histórica, por ejemplo, no se puede analizar la evolución de los contratos y sus montos en el tiempo.
- *Problema 7*: No se guardan versiones anteriores de los datos en la base de datos integrada, por ejemplo de los ajustes en los montos de los contratos, la evolución del dólar, la evolución de los costos hora de los técnicos, etc. Esto impide un estudio sobre la evolución económica del negocio y hace que las comparaciones de rentabilidad en diferentes períodos de tiempo muchas veces carezcan de valor real.

Nos planteamos como metas de calidad, la mejora de los problemas planteados anteriormente. La tabla 2 lista las metas planteadas. Las mismas se pueden clasificar en dos grandes grupos: las metas 1 a 5 se corresponden directamente con situaciones típicas de mejora de la calidad de los datos y las metas 6 y 7 se corresponden con mejoras del diseño del sistema, tanto a nivel de la base de datos integrada como a nivel del diseño conceptual y lógico de los cubos. Las primeras se profundizan en la sección 5, donde se definen preguntas y métricas con el objetivo de diagnosticar y cuantificar la incidencia de los problemas de calidad en el sistema; las segundas se consideran en la sección 6, donde se sugieren acciones para mejorar el diseño del sistema y se motiva la necesidad de trabajo futuro por parte de los dueños del sistema.

Tabla 2. Metas de calidad

1	Detectar y consolidar los clientes duplicados
2	Corregir la información de tiempo trabajado cargado en clientes equivocados
3	Disponer de toda la información de tiempos trabajados y facturación durante las consultas
4	Corregir los montos de contratos con valores nulos por su valor real
5	Eliminar los errores en el cargado de la horas trabajadas
6	Disponer de información histórica en el cubo *Contratos*
7	Tener la capacidad de mantener información histórica y versiones en ambos cubos

6. Definición de Factores y Métricas de Calidad de Datos

En esta sección descomponemos las metas presentadas en la sección anterior en un conjunto de preguntas que apuntan, cada una, a describir una faceta particular del problema y las asociamos a factores de calidad. Para cada

factor definimos un conjunto de métricas que permiten cuantificar los problemas de calidad y describimos como llevar a cabo las mediciones.

6.1. Definición de preguntas de calidad

En una primera instancia descomponemos cada meta de calidad en una serie de preguntas que capturan diferentes facetas de los problemas, y asociamos cada pregunta a un factor de calidad, como se muestra en la Tabla 3.
La meta 1 posee dos facetas: el nivel de duplicación de los datos y la cantidad de errores de codificación. La meta 2 se refiere a errores conceptuales en el cargado de los datos que hace que la información no se corresponda con la realidad. La meta 3 surge porque no están disponibles en el sistema la totalidad de los eventos (actividades realizadas). Complementariamente, se relaciona con el tiempo que transcurre entre que se produce el evento en la realidad (ejecución de la actividad) y que se carga en el sistema. La meta 4 surge de la falta de datos para algunas entidades (por ej. el monto de un contrato). La meta 5 refiere a valores que no son consistentes con el dominio de sus atributos respectivos. Por otro lado, esto puede deberse a errores de digitación.

Tabla 3. Asociación de las metas de calidad definidas a factores de calidad

Meta	Pregunta	Factor de calidad
1	¿Cuántas entidades están repetidas?	Unicidad
	¿Cuántos errores de digitalización y formateo tenemos?	Correctitud sintáctica
2	¿Cuantos clientes no tienen cargado el tiempo de trabajo realizado que les corresponde?	Correctitud semántica
3	¿Qué porcentaje de los eventos están representados?	Cobertura
	¿Cuánto tiempo se demora en cargar dichos eventos?	Actualidad
4	¿Cuántos valores nulos se introducen durante la carga?	Densidad
5	¿Cuántos valores están fuera de rango?	Integridad de dominio
	¿Cuántos errores de digitación tienen los datos?	Correctitud sintáctica

6.2. Definición de métricas

En esta sub-sección se establecen las mediciones a realizar en la base de datos integrada, con el objetivo de establecer los niveles de calidad de los factores definidos y poder dar seguimiento a su evolución. Para cada factor definimos las métricas a utilizar, describimos los métodos o programas de medición y presentamos posibles correcciones para mejorar sus niveles. Los detalles de implementación de los métodos quedan fuera del alcance de este artículo. La recolección periódica de valores para cada métrica que se defina y el mantenimiento estadístico de estos valores, van a dar una idea del comportamiento del sistema frente a cada factor de calidad y de los resultados que se producen luego de aplicar posibles acciones correctivas.

Meta 1 - Unicidad

Métrica: Cantidad de clientes repetidos que se ingresan al sistema.
Forma de medir: Para contar los clientes repetidos se dispone de un programa que recorre la base y busca por distintos mecanismos las entradas que pueden ser duplicadas, por ejemplo palabras que se repitan en varias tuplas (ej. nombre de empresa) o posibles abreviaturas de una palabra. Es el usuario del sistema quien determina qué casos son repetidos y cuáles no.
Posibles correcciones: Detectados los repetidos, la herramienta permite unificar dos clientes que estén duplicados, reconciliando los atributos de ambos.

Meta 1 - Correctitud Sintáctica

Métrica: Cantidad de clientes con datos mal ingresados al sistema.
Forma de medir: Vamos a obtener la métrica a partir de las quejas de clientes cuando reciben sus facturas con errores sintácticos en sus nombres. Se debe mantener un registro de estas situaciones para generar información estadística al respecto.

Meta 2 - Correctitud Semántica

Métrica: Cantidad de ingresos de actividades (tiempo de trabajo) en clientes incorrectos.
Forma de Medir: Analizaremos la asignación de tiempos de trabajo a clientes, comparando contra otros sistemas e identificando potenciales errores, los cuales se deberán presentar al usuario para validar y proponer las correcciones que sean necesarias. Usaremos dos estrategias para detectar potenciales errores. La primera consiste en consultar los datos de facturación e identificar aquellos casos en que haya actividades para clientes que no registran facturación o viceversa. La segunda estrategia consulta los contratos activos por cliente, identificando aquellos casos en que se detecten actividades de clientes sin contrato o viceversa.
Posibles correcciones: Esta búsqueda de inconsistencias se debe realizar en forma sistemática periódicamente, presentando los potenciales errores al usuario y corrigiendo los sistemas operacionales para evitar que se sigan generando errores.

Meta 3 - Cobertura

Métricas: (i) Porcentaje de horas trabajadas cargadas en el sistema respecto al total de horas trabajadas (en cada fecha). (ii) Porcentaje del monto facturado ingresado en el sistema respecto al monto total a facturar (en cada fecha).
Forma de Medir: Para medir la cobertura de horas trabajadas se compara la cantidad de horas cargadas en el sistema por fecha contra la cantidad real de horas trabajadas en esa fecha, la cual se estima a partir de dos fuentes alternativas: La primera es obtener el total de horas registradas en el sistema de marcas de entrada/salida. La segunda consiste en considerar las ausencias reportadas por fecha y restarlas del total presupuestado en plantilla. Estos dos mecanismos se complementan y aportan una fuente bastante certera del tiempo real trabajado en una fecha.
Para medir la cobertura de la facturación se compara el monto correspondiente a las facturas registradas en el sistema en una fecha dada, contra el total de facturación real generada para dicha fecha. Para obtener este último se debe recurrir a todos los documentos de ventas de servicios del sistema contable, que figuran como ejecutados en esa fecha en el sistema de mesa de ayuda. A esto se le debe sumar el monto de la facturación recurrente de los contratos que tienen esa fecha de facturación.
Posibles correcciones: Si medimos regularmente la cobertura (facturas/horas trabajadas) en una fecha fija referente, es esperable que la cobertura vaya aumentando progresivamente (al ritmo que se ingrese la información al sistema) hasta llegar a un nivel razonable (cercano a 1). A partir de estas mediciones se puede determinar el período de tiempo que es necesario esperar para contar con información suficientemente completa y por lo tanto estar en condiciones de tomar decisiones. Este dato ayudará a determinar las frecuencias de refresque de los cubos; es decir, la ventana de tiempo se dejará disponible a los usuarios en cada carga. Si este retraso penaliza al negocio, se deberán mejorar los procesos operativos para acelerar el ingreso de información al sistema.

Meta 3 - Actualidad

Métricas: Tiempo promedio y máximo transcurrido entre que se genera información en la realidad y ésta queda disponible en el sistema.
Forma de medir: Se calcula la diferencia de tiempo entre la fecha en que ocurre un evento (tarea o facturación) y la fecha en que el evento se carga en la BD integrada. Luego calculamos tiempos máximos y tiempos promedio de dichas medidas. También se pueden calcular máximos y promedios parciales para cada tipo de datos, lo cual puede aportar información adicional para la mejora del sistema, estableciendo donde se producen las mayores demoras.

Meta 4 - Densidad

Métrica: Porcentaje de contratos con monto no nulo respecto al total de contratos.
Forma de medirlo: Haciendo consultas simples en el sistema integrado (joins y totalizaciones) se puede obtener el total de contratos y el total de contratos con monto nulo. Esta medida se puede verificar al menos una vez al mes, que es el período en que varían los montos de contratos.
Posibles correcciones: Se debe tratar de corregir el problema de base en el proceso que carga los montos actualizados de los contratos en la base integrada.

Meta 5 - Integridad de Dominio

Métrica: Porcentaje de registros de tiempo trabajado que están fuera de rango respecto al total de registros.

Forma de medirlo: Se debe recorrer la tabla que contiene los tiempos trabajados y verificar si los valores entran en el rango de valores permitidos.
Posibles correcciones: Se deberían mejorar los controles de integridad en el sistema de mesa de ayuda, donde se cargan inicialmente estos datos.

Meta 5 - Correctitud Sintáctica

Métrica: Porcentaje de registros con errores de digitación.
Forma de medirlo: Toda actividad realizada por un técnico se registra en papel (boletas de trabajo), incluyendo el tiempo dedicado. Luego un digitador o el propio técnico las ingresa al sistema. Se propone realizar un chequeo de las actividades sobre un muestreo de boletas de trabajo, pudiendo determinar una tasa de errores de digitación en ese muestreo y luego extrapolarlo al resto de la información.
Posibles correcciones: En función de los valores obtenidos se pueden tomar medidas correctivas. Primeramente, se puede concienciar a los digitadores y técnicos que cargan la información. O en caso de obtener valores muy altos, se pueden considerar mecanismos de cargado automático (tipo OCR) o la asignación de un segundo digitador que valide o corrija los valores ingresados.

7. Mejora del diseño del SDW para mejorar la calidad de los datos

Tal como se observó anteriormente, existen problemas de calidad de datos que son provocados por un diseño incorrecto o inadecuado del SDW. En esta sección comentamos brevemente los problemas 6 y 7 relevados en la sección 4 y esbozamos una posible solución.
Del problema 6 se desprende la necesidad de almacenar la fecha de los contratos en el cubo *Contratos*, para poder analizar la perspectiva temporal de los mismos. El problema 7 también revela la necesidad de mantener datos históricos, con marcas de tiempo asociadas avarios tipos de entidades.
Estos problemas no son solucionables directamente con la arquitectura actual del sistema. Para lograr una solución de fondo es necesario historizar la base de datos integrada. Una forma de realizarlo es incluyendo una base de datos con información histórica, que se cargue periódicamente a partir de la BD integrada y que provea a los cubos con información temporizada y versiones de ciertas entidades. Otras decisiones de diseño tendrán también impacto en la calidad de los datos.

8. Conclusiones

Este artículo presenta un análisis sobre la calidad de los datos en un Sistema de Data Warehousing concreto, cuantificando la calidad de los datos, y realizando recomendaciones de diseño y de gestión de datos para mejorar la calidad de los mismos. El análisis realizado sigue un encare de tipo "*Goal-Question-Metric*", en el cual se plantea la evaluación y mejora en la calidad de los datos a partir de problemas que encuentran los usuarios al utilizar el SDW, especificándose su relación con la mejora de la calidad de los datos a través de factores y métricas de calidad.
La aplicación de este encare metodológico en un caso concreto ha permitido evaluar su aplicabilidad práctica, lo cual nos ha resultado extremadamente interesante en el marco del desarrollo de métodos para la gestión de calidad de datos.
El estudio realizado se basa en la calidad de los datos fuentes integrados, sin llegar a propagar los valores de calidad a las estructuras OLAP (cubos). Esto último permitiría contar con valores evaluados a un nivel más cercano al usuario, y por lo tanto más fáciles de asociar con los problemas que se le presentan así como con los objetivos de mejora. Planteamos este tema como trabajo futuro.

9. Referencias

[1] Abelló, A., Samos, J., Saltor, F.: YAM²: A Multidimensional Conceptual Model Extending UML. Information Systems, 31 (6), September, 2006.

[2] Akoka, J., L. Berti-Equille, O. Boucelma, M. Bouzeghoub, I. Comyn-Wattiau, M. Cosquer, V. Goasdoué-Thion, Z. Kedad, S. Nugier, V. Peralta and S. Sisaid-Cherfi (2007). A Framework for Quality Evaluation in Data Integration Systems. *9th International Conference on Enterprise Information Systems (ICEIS'2007)*, Funchal, Portugal.
[3] Basili, V., Caldiera, G., Rombach, H.D.: The Goal Question Metric Approach. Encyclopedia of Software Engineering, 528-532, John Wiley & Sons, Inc. (1994)
[4] Berti-Equille, L.: *Un état de l'art sur la qualité des données*". Ingénierie des systèmes d'information (ISI), Hermès, Vol. 9(5-6) :117-143, 2004
[5] Bouzeghoub, M.; Peralta, V.: "*A Framework for Analysis of Data Freshness*" In Proc. of the 1st Int. Workshop on Information Quality in Information Systems (IQIS'2004), Paris, France, 2004.
[6] Carpani, F.: CMDM: A conceptual multidimensional model for Data Warehouse. Master thesis, InCo - Pedeciba, Universidad de la República, Montevideo, Uruguay (2000)
[7] Etcheverry, L.; Tercia, S.; Marotta, A.; Peralta, V.: "Medición de la Exactitud de Datos Fuente: Un caso de Estudio". Technical Report, InCo, Universidad de la República, Uruguay, 2006.
[8] Etcheverry, L., Peralta, V., Bouzeghoub, M.: Qbox-Foundation: a Metadata Platform for Quality Measurement. In Proc. of the 4th Data and Knowledge Quality Workshop (DKQ'2008), Sophia-Antipolis, France (2008)
[9] Gertz, M.; Tamer Ozsu, M.; Saake, G.; Sattler, K.: "*Managing Data Quality and Integrity in Federated Databases*". In Proc. of the 2nd Working Conf. on Integrity and Internal Control in Information Systems (IICIS'98), Warrenton, USA, 1998.
[10] Gertz, M.; Tamer Ozsu, M.; Saake, G.; Sattler, K.: "Report on the Dagstuhl Seminar: Data Quality on the Web". SIGMOD Record Vol. 33(1), March 2004.
[11] Golfarelli, M. Rizzi, S.:"Methodological Framework for Data Warehouse Design.", DOLAP'98, USA, 1998.
[12] Helfert, M.; Herrmann, C.: "*Proactive Data Quality Management for Data Warehouse Systems*". In Proc. of the Int. Workshop on Design and Management of Data Warehouses, Toronto, Canada, 2002.
[13] Jarke, M.; Vassiliou, Y.: "*Data Warehouse Quality: A Review of the DWQ Project*". In Proc. 2nd Conference on Information Quality (IQ'1997), Cambridge, USA, 1997.
[14] Jarke, M.; Jeusfeld, M.A.; Quix, C.; Vassiliadis, P.: "*Architecture and Quality in Data Warehouses: An Extended Repository Approach*". Information Systems, vol. 24(3), 1999.
[15] Jeusfeld, M. A.; Quix, C.; Jarke, M.: "*Design and Analysis of Quality Information for Data Warehouses*". In Proc. of 17th Int. Conf. on Conceptual Modeling (ER), Singapore, November 16-19, 1998.
[16] Mazón, J.N., Trujillo, J., Serrano, M., Piattini, M.: Applying MDA to the development of data warehouses. In Proc. of ACM 8th International Workshop on Data Warehousing and OLAP (DOLAP'2005), Germany, 2005.
[17] Moody, D. Kortnik, M.: "From Enterprise Models to Dimensionals Models: A Methodology for Data Warehouse and Data Mart Design". DMDW'00, Sweden, 2000.
[18] Missier, P.; Scannapieco, M.; Batini, C.: "*Cooperative Architectures: Introducing Data Quality*". Technical Report 14-2001, Dipartimento di Informatica e Sistemistica, Universita' di Roma "La Sapienza", Roma, Italy, 2001.
[19] Motro, A.; Rakov, I.: "Estimating the quality of databases". In Proc of the 3rd Int. Conf on Flexible Query Answering Systems (FQAS'98), Roskilde, Denmark, 1998.
[20] Naumann, F.; Leser, U.; Freytag, J.C.: "Quality-driven Integration of Heterogeneous Information Systems". In Proc. of the 25th Int. Conf. on Very Large Databases (VLDB'99), Scotland, 1999.
[21] Naumann, F.; Rolker, C.: "Assessment Methods for Information Quality Criteria". In Proc. of the MIT Conf. on Information Quality (IQ'00), Cambridge, USA, 2000.
[22] Naumann, F., Freytag, J.C., Leser, U.: Completeness of Information Sources. In Proc. of the Workshop on Data Quality in Cooperative Information Systems (DQCIS'03), Siena, ITALY (2003)
[23] Peralta, V.: Data Quality Evaluation in Data Integration Systems. PhD Thesis, Université de Versailles, France & Universidad de la República, Uruguay (2006)
[24] Pipino, L.L.; Lee, Y.W.; Wang, R.: "*Data Quality Assessment*". Communications of the ACM, vol. 45, No. 4ve, April 2002.
[25] Redman, T.: "Data Quality for the Information Age". Artech House (1996)
[26] Scannapieco, M., Missier, P., Batini, C.: Data Quality at a Glance. Datenbank-Spektrum**,** German data base technology journal, vol. 14, pp. 6-14 (2005)
[27] Strong, D.; Lee, Y.; Wang, R.: "*Data Quality in Context*". Communications of the ACM, Vol. 40(5), May 1997.
[28] Vassiliadis, P., Bouzeghoub, M., Quix, C.: Towards Quality-oriented Data Warehouse Usage and Evolution. Information Systems, 25(2): 89-115 (2000)
[29] Wang, R., Strong, D.: Beyond accuracy: What data quality means to data consumers. Journal on Management of Information Systems, Vol. 12 (4), pp. 5-34 (1996)
[30] Weikum, G.: "*Towards guaranteed quality and dependability of information systems*". In Proc. of the Conf. Datenbanksysteme in B*uro, Technik und Wissenschaft, Freiburg, Germany, 1999.