Conciencia Tecnológica
ISSN: 1405-5597
contec@mail.ita.mx
Instituto Tecnológico de Aguascalientes
México

García Merayo, Félix; Luna Ramírez, Enrique
El proceso Data Warehousing y los meta datos
Conciencia Tecnológica, núm. 15, diciembre, 2000, p. 0
Instituto Tecnológico de Aguascalientes
Aguascalientes, México

EL PROCESO DATA WAREHOUSING Y LOS META DATOS

Félix García Merayo[1]
fgmerayo@fi.upm.es
Facultad de Informática
Universidad Politécnica de Madrid
España

Enrique Luna Ramírez[2]
fp22067@zipi.fi.upm.es
Depto. de Sistemas y Computación
Instituto Tecnológico de Aguascalientes
México

RESUMEN

La tecnología Data Warehousing ha aparecido estos últimos años tras la convergencia entre las nuevas necesidades en el manejo de la información de las empresas y la capacidad que existe para integrar e implementar tecnologías aptas para responder a tales necesidades. Este trabajo intenta describir la lógica existente para construir uno de los componentes más importantes de un Data Warehouse: el repositorio de meta datos. Su importancia radica en el hecho de que todo el conocimiento sobre la creación de un Data Warehouse es almacenado en dicho repositorio.

Palabras clave: *Data Warehouse, Meta datos, Repositorio, Data Mart, OLAP.*

INTRODUCCIÓN

En la empresa actual se genera una gran cantidad de información que puede provenir de diversas fuentes. Esta información requiere de un tratamiento apropiado para que los responsables de tomar decisiones puedan sacar partido de ella. Para ello, resulta fundamental implementar una nueva informática de decisión para obtener una mejor comprensión del valor de las informaciones disponibles, definir indicadores de negocio pertinentes para facilitar la toma de decisiones y conservar la memoria de la empresa.

Para responder a estas necesidades, el nuevo papel de la informática es definir e integrar una arquitectura que sirva como base a las aplicaciones de soporte a la toma de decisiones. Esta arquitectura global es el Data Warehouse (DW). El DW ha aparecido estos últimos años (inicio de los 90) tras la convergencia entre las nuevas necesidades en el manejo de la información de las empresas y la capacidad de integrar e implementar tecnologías aptas para responder a ello [3].

Un DW no se compra, se construye. Este trabajo intenta describir la lógica existente en la construcción del repositorio de meta datos de un DW, y está estructurado en dos partes principales. La primera parte permite comprender la tecnología Data Warehousing y sus objetivos. La segunda parte aborda el tema de los meta datos, que como lo han señalado diversos autores [4,5,13], es uno de los aspectos más importantes en el ciclo de desarrollo de un DW. El trabajo termina con las conclusiones y referencias bibliográficas correspondientes.

CONCEPTOS BÁSICOS

En esta sección se definen los conceptos necesarios para poder tratar el tema del desarrollo de un DW, y particularmente, el tema de los meta datos, punto de interés de este trabajo.

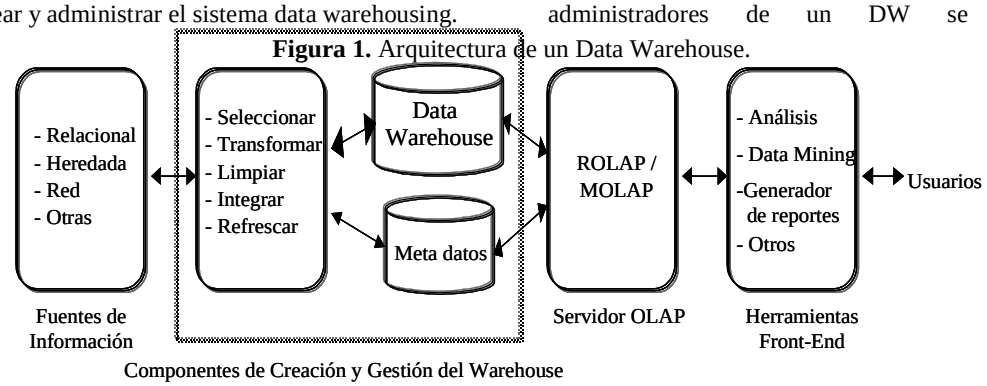
Iniciaremos por definir lo qué es un DW. Así entonces, la definición clásica de DW dada por Bill Inmon [6], reconocido como el padre del DW, es la siguiente: "Un Data Warehouse es una colección de datos orientados al tema, integrados, no volátiles e historiados, organizados para dar soporte al proceso de ayuda a la toma de decisiones". Otras definiciones se pueden ver en [2]… Esta colección de datos en un DW posee la arquitectura conceptual mostrada en la figura 1 [2,10].

La arquitectura de un DW incluye herramientas para extraer datos de diversas bases de datos operativas y fuentes externas; para limpiar, transformar e integrar estos datos; para cargar los datos dentro del DW; y para refrescar periódicamente el DW y así reflejar las actualizaciones en las fuentes y purgar los datos. Los datos en un DW son almacenados y gestionados por uno o más servidores OLAP (On-Line Analytical Processing, véase [2]), que pueden ser MOLAP (Multidimensional On-Line Analytical Processing) o ROLAP (Relational On-Line Analytical Processing),

[1] Doctor en Informática y Profesor Titular de la Universidad Politécnica de Madrid.
[2] Doctorando en Informática en la Universidad Politécnica de Madrid, España.
Este trabajo cuenta con el auspicio del Consejo Nacional de Ciencia y Tecnología de México (CONACYT).

y pueden presentar vistas multidimensionales de los datos hacia una gran variedad de herramientas front-end (herramientas de consulta, generadores de reportes, herramientas de análisis, y herramientas de data mining) y estas herramientas formatean los datos de acuerdo a los requerimientos del usuario. Finalmente, existe un repositorio para almacenar y gestionar los meta datos, y herramientas para monitorear y administrar el sistema data warehousing.

Figura 1. Arquitectura de un Data Warehouse.



EL ROL DE LOS META DATOS EN EL PROCESO DATA WAREHOUSING

El diseño del repositorio de los meta datos es uno de los aspectos más importantes para el éxito de un DW, aunque su valor en los proyectos de desarrollo de DWs es subestimado [12,13]. Su importancia radica en el hecho de que todo el conocimiento sobre la creación de un DW es almacenado en el repositorio de meta datos. En esta sección se discutirán los conceptos asociados al tema de los meta datos tales como su definición, clasificación, gestión, arquitectura, y representación, entre otros, necesarios para una mejor comprensión de este tema.

Definición de meta datos

En general, los meta datos son definidos como información sobre los datos [7,12], es decir, información sobre la estructura, contenido e interdependencias de los componentes del DW [11]. En un DW, los meta datos describen los tipos de datos en el DW, las definiciones física y lógica de los datos, consultas y reportes predefinidos, reglas de validación y orientadas al tema, definiciones de fuentes de datos, rutinas de transformación y de proceso, e información del usuario. Los meta datos se refieren a cualquier cosa que define un objeto del DW (una tabla, una columna, una consulta, un reporte, una regla orientada al tema, o un algoritmo de transformación). Los meta datos guían los procesos de extracción, de limpieza y de carga, además de que hacen que las herramientas de consulta y los generadores de reportes funcionen correctamente [4,5,7].

Clasificación de los meta datos

Usualmente, los meta datos son divididos en meta datos técnicos y meta datos semánticos u orientados al tema [8,11]. Así por ejemplo, los desarrolladores y administradores de un DW se interesan principalmente en los meta datos a un nivel de implementación técnica. Los desarrolladores de software usan los meta datos técnicos para conocer las definiciones física y lógica de los datos para poder diseñar y escribir aplicaciones, mientras que los administradores accesan a este tipo de meta datos para ejecutar sus tareas administrativas tales como la gestión de los objetos y usuarios del DW, afinamiento de la base de datos y almacenamiento de los datos. Por su parte, los usuarios finales tales como los analistas y gerentes, que no están familiarizados con los formatos de descripción del DW tales como los archivos SQL-DDL de la base de datos, están interesados en entender la semántica orientada al tema y por lo tanto necesitan representaciones semánticamente ricas de la estructura y contenidos del DW.

Gestión de los meta datos

A menudo, un repositorio de meta datos es usado para almacenar y gestionar todos los meta datos asociados a un DW. El repositorio permite compartir los meta datos entre las diversas herramientas y procesos utilizados para diseñar, establecer, usar, operar, y administrar un DW [2,5]. El beneficio de gestionar los meta datos técnicos de un DW es similar al beneficio que se obtiene de gestionar los meta datos en un ambiente de procesamiento de transacciones OLTP (On-Line Transaction Processing). Los meta datos técnicos integrados y consistentes crean un ambiente de desarrollo más eficiente para el staff técnico responsable de construir y mantener los sistemas de procesamiento de decisiones [4]. Un

beneficio adicional en el ambiente data warehousing es la habilidad de rastrear como cambian los meta datos a lo largo del tiempo. Por su parte, los beneficios obtenidos gracias a la gestión de los meta datos semánticos son exclusivos de un ambiente de procesamiento de decisiones y son la clave para explotar el valor de un DW una vez que ha sido puesto en operación.

Esfuerzos de estandarización

Existen varios esfuerzos paralelos en la industria que han conducido a algún desarrollo de estándares de meta datos, entre los que destacan dos de ellos. El primero llevado a cabo por la Meta Data Coalition, establecida en 1995, ha conducido a la Meta Data Interchange Specification, llamada MDIS [9]. Este enfoque modela información de esquemas de diferentes tipos de almacenes de datos tales como relacionales, multidimensionales, orientados a objetos, sistemas de bases de datos tipo red o jerárquicos, además de estructuras de archivos. MDIS no es exclusivo del data warehousing y está limitado a las relaciones entre los esquemas. No cubre la parte de los meta datos semánticos y ofrece poco soporte para propósitos de movimiento de datos [11]. Un segundo esfuerzo, hecho por Microsoft Corp., está basado en el estándar del Unified Modeling Language (UML) del Object Management Group. El Repositorio de Microsoft [1], llamado Open Information Model (OIM), proporciona un formato común para que las herramientas compartan la información que describe a los objetos, componentes y módulos a lo largo del ciclo de vida del desarrollo de una aplicación. Este repositorio ofrece una interfaz de meta datos y capacidades de exportación para el intercambio de meta datos entre repositorios. Microsoft también anunció que hará futuras extensiones al modelo OIM que soporten todas las formas de meta datos en el proceso data warehousing: reglas de extracción y transformación, mapeos de datos e información del modelo de datos.

Un enfoque reciente, basado en el Extensible Markup Language (XML), para intercambiar un rango amplio de meta datos es el formato XML Metadata Interchange, llamado XMI y desarrollado por el Object Management Group. Su objetivo es intercambiar datos de programación de los desarrolladores que trabajan con tecnología de objetos sobre Internet. Sin embargo, este formato no está orientado al data warehousing.

Arquitectura de los meta datos

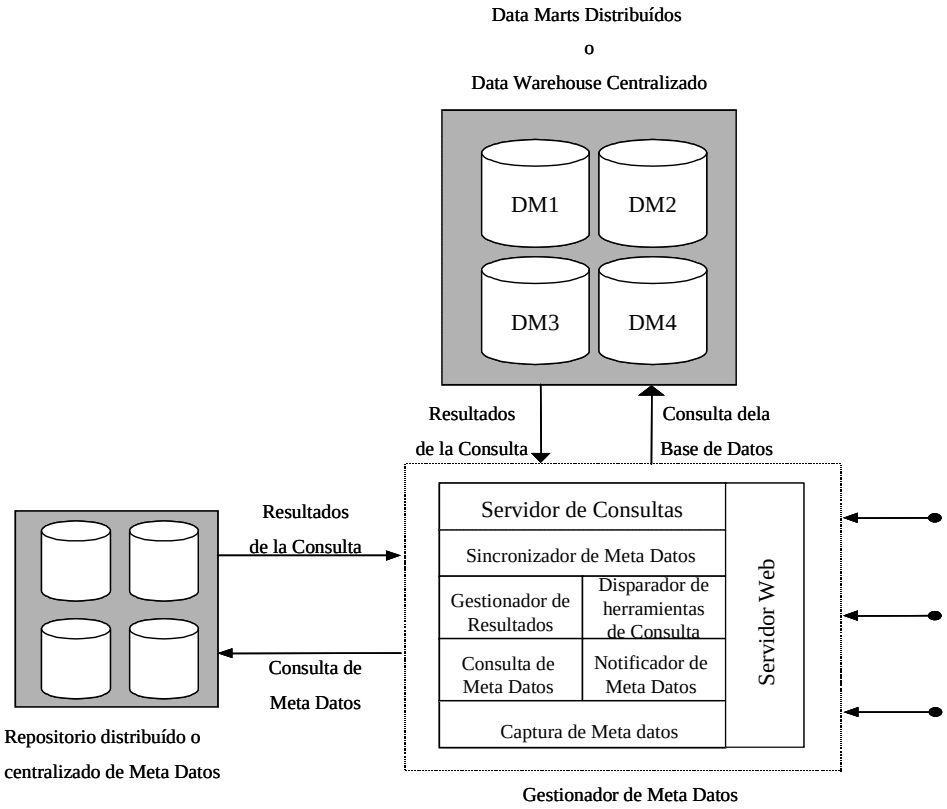
Ya sea que se esté construyendo un sólo Data Mart[3] o un DW complejo para una gran compañía, la arquitectura de los meta datos debería ser una parte integral del proceso de diseño. Desarrollar una arquitectura al inicio del proyecto ayuda a tener una visión a futuro, guiando al equipo del proyecto data warehousing a través de las diferentes fases. Los usuarios de un DW deberían contar no sólo con meta datos que sean precisos, sino también contextuales, ya que de otra manera se podría obtener información engañosa y ambigua, la cuál puede conducir a decisiones equivocadas. Así, la arquitectura de los meta datos en un proyecto data warehousing debería ser un punto obligatorio y bien planificado de toda la arquitectura del DW en su conjunto [8,13].

En el futuro, los meta datos adquirirán mucha más importancia dada la unión (complementación) que existe entre la tecnología Web y el Data Warehousing [3,13]. Esta unión resultará en un browser de meta datos como punto de acceso a la información orientada al tema. Los meta datos se convertirán en un componente crítico de la arquitectura de cualquier DW. La figura 2 muestra una representación lógica de una propuesta de arquitectura para los meta datos en un DW que considera la unión de la tecnología Web y el Data Warehousing [8,13].

El repositorio de meta datos puede estar centralizado o distribuido dependiendo de las necesidades y requerimientos organizacionales. Este se vería como un grupo de almacenes de datos, compuestos de objetos y datos relacionales. El gestionador de meta datos sería el componente esencial de la arquitectura de meta datos e idealmente consistiría en los siguientes componentes:

- **Captura de meta datos**: captura inicial de los meta datos desde diversas fuentes.
- **Sincronizador de meta datos**: procesos para mantener los meta datos actualizados.
- **Motor de búsqueda de meta datos**: sería el componente front-end para los usuarios que buscan y accesan a los meta datos.
- **Gestionador de resultados de meta datos**: procesaría los resultados de la búsqueda de meta datos y permitiría al usuario hacer una selección apropiada.
- **Notificador de meta datos**: notificaría a los suscriptores sobre cualquier cambio en el contenido de los meta datos dependiendo del perfil del usuario.

[3] Data Mart (DM): base de datos orientada al tema puesta a disposición de los usuarios en un contexto de decisión descentralizado [3].



- **Disparador de consultas de meta datos**: dispararía una herramienta de consulta apropiada para obtener datos desde un DW o cualquier otra fuente basada en la selección hecha por el usuario en el gestionador de resultados de meta datos.

Figura 2. Arquitectura de los meta datos en un Data Warehouse.

Recientemente, la Meta Data Coalition ha construido una versión alfa de un puente entre la especificación MDIS de la Meta Data Coalition y el repositorio basado en el modelo OIM de Microsoft. Este puente permitirá un intercambio de información en ambos sentidos entre los archivos MDIS y el repositorio de Microsoft.

Representación de los meta datos

En la actualidad, las herramientas existentes en el mercado para el ambiente data warehousing usualmente almacenan los meta datos en una base de datos relacional u orientada a objetos, la cuál es manejada como una caja negra, y sólo soportan un subconjunto aislado de meta datos. En particular, no existe ninguna solución comercial que integre los meta datos para el movimiento de los datos técnicos, análisis multidimensional y modelado de la semántica en un DW con una única herramienta. Por ejemplo, el Repository Information Model (RIM) de ROCHADE cubre un rango amplio de meta datos semánticos y multidimensionales además de meta datos técnicos. Sin embargo, la representación de las reglas de transformación sobre la capa técnica y entre los meta datos técnicos y semánticos no llega a ser clara [11].

En forma similar al mercado comercial, la mayoría de los enfoques de investigación están limitados a subconjuntos de meta datos. A continuación se discuten algunos de los enfoques desarrollados en torno a la representación de los meta datos en un DW.

Gorczynska et al. proponen un modelo para representar los meta datos en un DW y lo describen para datos multidimensionales. Este modelo está basado en el modelo entidad-relación y es independiente del enfoque usado para modelar los datos multidimensionales (ROLAP o MOLAP). Dado que el modelo está diseñado para catalogar sólo datos multidimensionales, este necesita ser extendido para poder representar la información técnica sobre los DWs, así como los datos operativos [5].

Müller et al. proponen un modelo, basado en UML, para la representación uniforme de las interdependencias entre los meta datos técnicos y los meta datos semánticos, así como para su integración. Sin embargo, este enfoque no considera la redefinición de meta datos técnicos en el repositorio y su propagación hacia las herramientas afectadas (arquitectura bidireccional), ni el soporte de consultas a nivel semántico y su translación automática hacia los programas de consulta al nivel del DW [11].

CONCLUSIONES

- En la actualidad, las empresas generan una gran cantidad de información que es necesario saber utilizar para sacarle el mayor provecho en pro de la toma de decisiones. En este sentido, la tecnología Data Warehousing proporciona un soporte importante para lograr este propósito.
- Los meta datos son una componente fundamental de un DW, ya que son estos los que se encargan de guiar los procesos de extracción, limpieza, transformación y carga de los datos dentro del DW. En particular, la arquitectura de los meta datos debería ser una parte integral, y bien planificada, de toda la arquitectura de un DW en su conjunto, ya que de esto dependerá la buena gestión de los meta datos, y por lo tanto, el buen funcionamiento del DW.
- El repositorio de los meta datos de un DW es utilizado para almacenar y gestionar todos los meta datos asociados al DW, tanto técnicos como semánticos. El beneficio de gestionar los meta datos técnicos es similar al beneficio obtenido por los desarrolladores y administradores de sistemas de información en un ambiente de procesamiento de transacciones OLTP, mientras que los beneficios de gestionar los meta datos semánticos, exclusivos de un ambiente de procesamiento de decisiones (ambiente OLAP), son obtenidos por los usuarios finales que, el no estar familiarizados con los formatos de descripción del DW, logran entender la semántica orientada al tema gracias a las representaciones de la estructura y contenidos del DW.
- Otro aspecto importante de los meta datos de un DW es precisamente su representación. Una representación explícita de los meta datos da soporte a los usuarios orientados al tema en las tareas de navegación, consultas adhoc y data mining. Existen algunas propuestas en torno a este tema, unas más completas que otras, pero todas ellas adolecen de alguna carencia de acuerdo a lo discutido en la sección correspondiente a la representación de los meta datos. En este sentido, existen diversos aspectos abiertos a la investigación.

REFERENCIAS BIBLIOGRÁFICAS

[1] Bernstein, P.A. et al.: "Microsoft Repository Version 2 and The Open Information Model", Information Systems, Vol. 24, No. 2, 1999, pp. 71-98.

[2] Chaudhuri, S. et al.: "An Overview of Data Warehousing and OLAP Technology", SIGMOD Record, Vol. 26, No. 1, March 1997, pp. 65-74.

[3] Franco, J.M.: "Le Data Warehouse, Le Data Mining", Editions Eyrolles, Paris, 1997.

[4] Gardner, S.R.: "Data Warehouses and Metadata: The Importance of Metadata Management", Data Mining, Data Warehousing, and Client/Server Databases. Proceedings of the 8th International Database Workshop. Springer-Verlag Singapore, 1997, pp. 61-71.

[5] Gorczynska, R. et al.: "Modeling Meta Data for Multidimensional Data", Journal of Data Warehousing, Vol. 3, No. 4, Winter 1998, pp. 32-42.

[6] Inmon, W.H.: "Building the Data Warehouse", QED Technical Publishing Group, 1992.

[7] Kimball, R.: "Meta Meta Data Data", DBMS, Vol. 11, No. 3, March 1998, pp. 18-20.

[8] Marco, D.: "Managing Meta Data", DM Review Magazine, March 1998.

[9] Meta Data Coalition: "Metadata Interchange Specification", Vers. 1.1, Aug. 1997, http://www.MDCinfo.com/standards/toc.html (visitado en Julio de 2000).

[10] Mohania, M. et al.: "Advances and Research Directions in Data Warehousing Technology", Australian Journal of Information Systems, Vol. 7, No. 1, 1999, pp. 41-59.

[11] Müller, R. et al.: "An Integrative and Uniform Model for Metadata Management in Data Warehousing Environments", Proceedings of the International Workshop on Design and Management of Data Warehouses (DMDW'99), Heidelberg, Germany, 1999.

[12] Ramasubbu, R.: "The Power of Meta Data", DM Review Magazine, April 1999.

[13] Sachdeva, S.: "Meta Data Arquitecture for Data Warehousing", DM Review Magazine, April 1998.